Guideline for Automatic Speech Recognition, Correction and Time Aligned Interview Transcripts

31 October 2021 S. Stigter, updated for Whisper by L. Wolfkamp 15 June 2023.

1. Automatically Transcribe

Whisper is an automatic speech recognition system developed by OpenAI. There are a several ways of using Whisper on your own computer. For Mac OS, MacWhisper is easiest. For Windows, use Subtitle Edit or Python.

An extensive overview can be seen here:

- https://speechandtech.eu/tech-tools/automatic-speech-recognition
- https://speechandtech.eu/tech-tools/asr-tools

1.1. Mac OS

1.1.1. MacWhisper

1. Download MacWhisper from https://goodsnooze.gumroad.com/l/macwhisper for (a free or paid) version and follow the installation instructions
2. If asked for an update refuse, as this brings you to the Appstore, which requires you to pay 4 euros for the free version. It overwrites your first install, making it unavailable!
3. If you opt for a paid version you can use the bigger models (large and medium) to get a better quality recognition. In most cases, medium is a perfect model to use.
4. You can either use MacWhisper to correct the ASR results if you only need subtitles, or use 3 Whisper Corrector to improve the ASR-subtitles (.srt), and get a full transcript including speaker turns (.txt) and an easy search-and-find file (.html).

1.2. Windows

1.2.1. Subtitle Edit

1. Download latest version of Subtitle Edit from https://github.com/SubtitleEdit/subtitleedit/releases
2. The first time you use the program for speech recognition, you need to change some settings, **only once**.
   a. Click on Options – Settings.
   b. In the window that opens, click on Video player, then on Download mpv lib, then click OK
   c. In the main menu, click on Video – Audio to text (Whisper)
   d. The program will ask permission to download some software.
   e. Click Yes to download **FFmpeg**, then click OK
   f. Click Yes to download **whisper.ccp**, then click OK
   g. In the window that opens, click on the three dots.
   h. Select the model(s) you want to use for speech recognition. Larger model means more accurate transcription, but slower processing. Models ending on .en are more accurate but only for full English text.
   i. Once you've downloaded the model(s), Subtitle Edit is ready to use for speech recognition.
3. To automatically transcribe an audio or video file with Subtitle Edit, use the following steps:
   a. Click on Video – Open video file, and select your audio or video file (if you cannot see your file, make sure to click on All files in the lower right corner)
   b. Click on Video – Audio to text (Whisper)
   c. Select the language and the desired model (make sure you have downloaded the model before)
   d. Make sure the boxes Auto adjust timings and Use post-processing are ticked.
   e. Click on Generate. The program will start automatic recognition. You can follow the progress in the green bar.
   f. When automatic recognition is completed, save the .srt file by clicking File – Save as… Save the file in the same folder as your audio or video file.
4. To correct the ASR results, you can either use Subtitle Edit if you only need subtitles (.srt), or use 3 Whisper Corrector to generate subtitles (.srt), a full transcript including speaker turns (.txt) and an easy search-and-find file (.html).

1.2.2.Python

1. To use Whisper in Python, you need to install Python, PyTorch, ffmpeg and finally, Whisper. This can all be done by hand, or in one go with following steps (only on Windows):
   a. Open the command line interface, for example Powershell (installed by default).
   b. Copy-paste the following command into the command line: iex (irm whisper.tc.ht) Press Enter.
   c. Wait until everything is installed.
2. To automatically transcribe an audio or video file, use the following steps:
   a. Open the command line interface, as described above.
   b. Navigate to the folder that contains your audio or video file by typing cd, followed by the folder location, and press Enter.
      For example: cd C:\Documents\Projects\Interviews\Struycken_20230615_Amsterdam
   c. Type the following command into the command line and fill in the filename, model and language: whisper [filename] --model [model name] --language [language] Press Enter.
      Example: whisper Struycken_20230615_Amsterdam.m4v --model medium --language nl
      Or: whisper Monahan_20221205_DenHaag.m4v --model small --language en
   d. Wait until recognition is complete. You can follow the progress in the terminal.
3. The results will automatically be saved in the same folder as your audio or video file. You will get three files: .txt, .srt, and .vtt. Use the .srt file to correct the results in 3 Whisper Corrector.

2. Safekeeping and Pre-archiving

1. The original audio or video recording and ASR results should now be stored in the same folder on your computer. This is a good moment to back up your material.
2. Back up your material on *UvA Research Drive* via a file drop link (ask datasteward-fgw@uva.nl). If no access, use a cloud alternative so long as you have no access to a safe repository.
3. Make one folder per project / interview and interviewee.
4. Important: leave material on the Research Drive until all material is archived at DANS-KNAW and the material has received a digital object identifier – DOI, a stable URL.

3. Whisper Corrector

1. Download latest version of Whisper Corrector from speechandtech.eu and follow the Whisper Corrector Manual. Note: the manual is not entirely up-to-date. Instead of a .ctm you load the .srt file that resulted from step 1. Automatically Transcribe into the corrector.
2. Only correct what is automatically transcribed and **do not add additional** information. Clarifications and annotations can be done separate from this process in the final transcript, for instance. Please note:
   a. Avoid freestanding interpunction marks, e.g. with spaces around it.
   b. Avoid square brackets in the running text […], for they indicate the speaker turns only.
   c. Each speaker turn should start on a new line. Also check the default speaker at the top of each turn. Right mouse click allows to change / insert a speaker.
3. Once you are done correcting, export the results in Expert mode (Settings – Show expert mode), click export for FA and save the new .txt file in the same folder as your audio or video file (overwriting the non-corrected .txt file).
4. So long as there is no spell-checker in the ASR-Corrector, check the .txt file manually for typos. Eliminate all double spaces in the file, floating interpunction marks (find-replace) or use the FACleaner.
5. Continue with step 4 Forced Alignment for an audio-text aligned transcription to facilitate search runs, subtitling and a written transcript.
6. For a quick transcript only, click on Export Word Transcription and continue with step 5 Transcript.

4. Forced Alignment

To have your spoken text aligned with written words makes your material suitable for digital research (ai), (translated) subtitling, search runs, and annotation while listening to the interview. A forced alignment tool using Whisper is currently in development. For the time being, continue with step 5 Transcript.

5. Transcript

1. Open the '.word.txt' in your folder and copy-paste into Word (or your preferred text editor).
2. Select all text and slide the left margin to the right and fix at 3,5 cm for proper speaker alignment.
3. Add metadata to provide context for the transcript. See the metadata templates on the Resources tab on www.uva.nl/ici

6. Consent

1. Before deposition, make sure you have consent from all parties involved in the recording: the interviewee and interviewer(s), videographer, everyone else visible or audible.
2. Templates are provided by the Interviews in Conservation Initiative www.uva.nl/ici under Resources
3. Always aim for open access to support FAIR principles in research. Alternatives are provided, but make sure to get consent for open access, even if you decide to make the material available with restricted access.

7. Deposition

1. Ask for a file drop link on Research Drive with UvA's data steward, fgw-datasteward@uva.nl to drop the prepared materials to be archived, including signed forms of consent.
2. Go to DANS Data Station Social Sciences and Humanities and log in.
3. Once logged in, click on Add Data – New Dataset. Here you can upload your files and the corresponding metadata.
4. For examples, see thematic collection *Interviews in Conservation Research* and browse other oral history collections.