

(Infra)structural considerations for high quality ASR for a variety of research domains

25 June 2024

Henk van den Heuvel, Christoph Draxler, Arjan van Hessen

Seminar SURF / Stichting Open Spraaktechnologie:
Exploring infrastructure for Dutch speech recognition



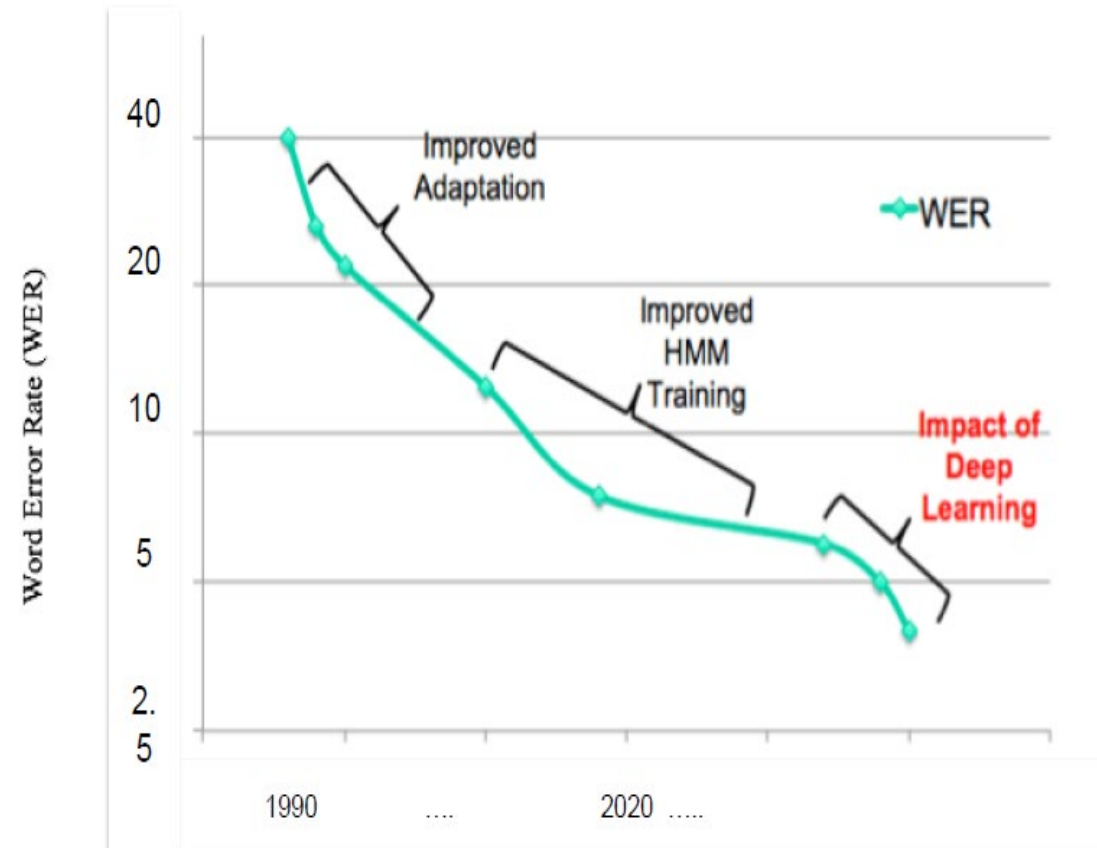
OUTLINE

1. ASR paradigm shift
2. Webservice at Radboud University
3. Webservices at BAS
4. Do it yourself with Whisper
5. Commercial solutions
6. Remaining challenges



1. ASR IN (ORAL) HISTORY

- ASR traditionally was knowledge based built on three components: Acoustic Models, Pronunciation Lexicon, Language Model
- Since 2020: Data-driven Deep Learning, Transformers
- Large Speech models are **not** Large Language Models
 - LSM: Speech input, text output
 - LSMs: Whisper, MMS, Chirp
- LSMs and LLMs can collaborate in applications, e.g. for oral history



TRADITIONAL ASR OUTPUT

mmm <unk> <unk> <unk> linkerhand dan zit je hier de gewrichten van hans en is 't eigenlijk
hoef je niet uh ruimte loopt zijn om 't te zien hè want die is kapot dus dat is deze hè van
hun rechterhand <unk> links is ook beschadigd hè want die ziet ook dat hier eigenlijk geen
met tussen zit uhm en dat die ook een beetje hij staat ook een beetje scheef <unk> mooie
recht op mekaar dat klopt dat is dat dat u uh aan de buitenkant mmm <unk> uh vertellen hoe
d'r maar 'ns uh begonnen dat dus uh dat gebeurt uh eerst leek 't op colitis ulcerosa <unk>
<unk> boek bloed geconstateerd twintigste-eeuwse definitief eigenlijk vastgesteld <unk>
<unk> uh <unk> uh <unk> <unk> <unk> <unk> waardoor de dieven invloed met hoofd en drummer
klachten bleven wel uh bestaan voornamelijk in mijn uh <unk> mijn ervaringen met het
voorschrijven van <unk> zijn heel positief uhm 't belangrijkste punt daarin is dat ze goed
werken dat wil zeggen dat mensen uh zelf minder pijn ervaren en dat wij als dokter ook
minder ontstekingen aan die vent richten zien 't voordeel uh van het gebruik van <unk> ten
opzichte van de conventionele middelen is dat ze uhm uh voor een goed patiënten die niet
reageert op conventionele middelen dus beter werken uhm daarnaast is het zo dat uh de
patiënten ook aangeven dat ze vaak sneller werken is dat mensen sneller van hun klachten af
zijn uh <unk> die ik nu krijgt 't is uh <unk> één keer een half jaar en uh twee maal daar is
honderd milligram <unk> en nou gebruikte dieren turkse maar aan een aantal jaren sinds
tweeduizend zes daarvoor heeft u ook een hele hoop medicijnen verbruikt waaronder ook uh uh
de <unk> uh die vond u dat hoeft je in de hand dat hier ook wel goed we op een gegeven moment
denk ik dat 't toch niet helemaal uh voldoende meer hielp <unk> die een bril gegaan dat ging
ook wel wel wel goed maar 'k heb toch 't idee dat dit 't beste loopt dus je bent echt
gewisseld van de medicijnen omdat het onvoldoende werkte op tv en had u

MODERN ASR WHISPER OUTPUT WITH PUNCTUATION AND SPEAKER DIARIZATION

Speaker 1: Kijk, hier ziet u hem zien met uw hand. Dus dit is uw rechterhand, dit is uw linkerhand. Dan ziet u hier de gevrichte van de hand en u ziet dat u eigenlijk niet een rheumatoloog moet zijn om het te zien, want die is kapot. Dus dat is deze van uw rechterhand. Ja, die is kapot. Links is ook beschadigd, want u ziet ook dat hier eigenlijk geen kraakbeerhandje meer tussen zit. Hij staat ook een beetje scheef, ze staan niet mooi recht op elkaar. Dus dat klopt, dat is dat, dat u aan de buitenkant aan uw hand ziet.

Speaker 0: Ik kan vertellen hoe de reum is begonnen. Op mijn twaalfde is dat gebeurd. Eerst leek het op colitis ulcerosa, maar er waren van alle rheumafactoren in het bloed geconcentreerd. Op mijn 21e is definitief eigenlijk vastgesteld dat het om weg tref ging. Dit was nadat ik eigenlijk mijn colitis ulcerosa, dat is verholpen door mijn dikke darm te verwijderen, waardoor die geen invloed meer had. De rheumaklachten bleven wel bestaan, voornamelijk in mijn wervelkolom.

Speaker 1: Mijn ervaringen met het voorschrijf van TNF-alpha-remmers zijn heel positief. Het belangrijkste punt daarin is dat ze goed werken. Dat wil zeggen dat mensen zelf minder pijn ervaren en dat wij als dokter ook minder ontstekingen aan de gewrichten zien. Het voordeel van het gebruik van biologicals ten opzichte van de conventionele middelen is dat ze voor een groep patiënten die niet reageert op conventionele middelen, dus beter werken. Daarnaast is het zo dat patiënten ook aanreven dat ze vaak sneller werken. Dus dat mensen sneller van hun klachten af zijn.

WHISPER OUTPUT WITH TIMESTAMP AND SPEAKER DIARIZATION

1

00:00:06,780 --> 00:00:09,989

Speaker 1: Kijk, hier ziet u hem zien met uw hand.

2

00:00:10,009 --> 00:00:13,031

Speaker 1: Dus dit is uw rechterhand, dit is uw linkerhand.

3

00:00:13,071 --> 00:00:20,815

Speaker 1: Dan ziet u hier de gevrichte van de hand en u ziet dat u eigenlijk niet een rheumatoloog moet zijn om het te zien, want die is kapot.

4

00:00:20,855 --> 00:00:25,338

Speaker 1: Dus dat is deze van uw rechterhand.

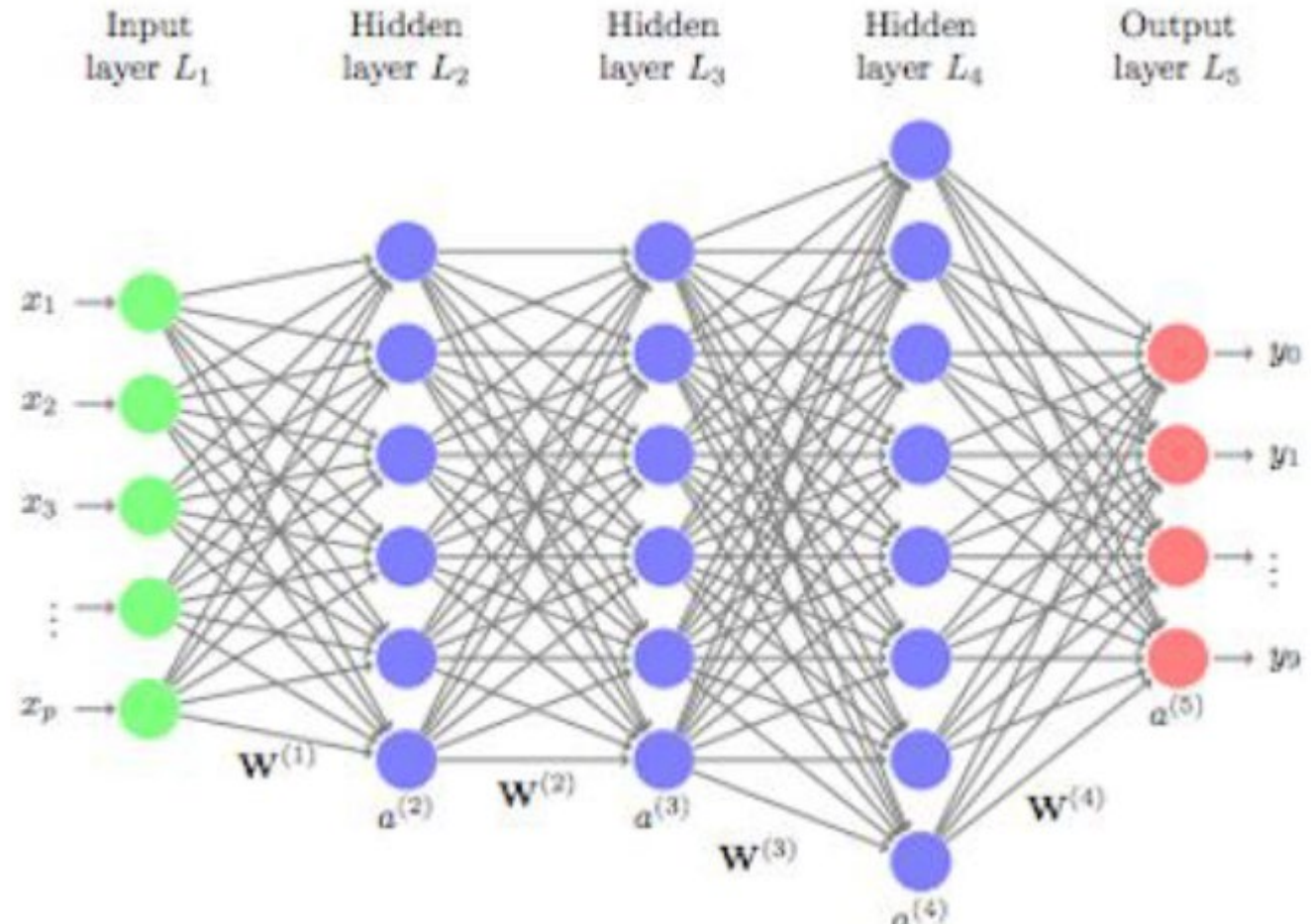
SRT (SubRip Subtitle file) output:
suitable for subtitling videos, and time aligned research

NLP & MORE

- Text **Summarisation**
- Topic modelling
- Sentiment analysis
- Text simplification
- Text generation
- Translation
- Literature review
(wider coverage)

LLM finetuning of LSM output:

- ◆ Conversational prompting
- ◆ Prompt design
- ◆ Prompt engineering



2. WEBSERVICES AT RADBOUD UNIVERSITY: HTTPS://WEBSERVICES2.CLS.RU.NL/ASRSERVICE

The screenshot shows a web browser window with the URL `webservices2.cls.ru.nl/asrservice`. The page title is "Automatic Speech Recognition Service". The navigation bar includes "1. Projects", "2. Staging", "3. Runtime", "4. Results", and "REST API Specification". The main content area features a "System Information" box with the following details: Version: 0.3, Author(s): Maarten van Gompel, and Contact: proycon@anaproy.nl. Below this is a "Data processing notice" paragraph. At the bottom, there is a light blue box with the text "You will be asked to authenticate when you continue to use the rest of this service, by clicking the button below." and a large blue "Continue" button. Below the button, it says "Or alternatively [continue using fallback authentication\(HTTP Basic Authentication\)](#)".

webservices2.cls.ru.nl/asrservice

Automatic Speech Recognition Service 1. Projects 2. Staging 3. Runtime 4. Results REST API Specification

An Automatic Speech Recognition Service for a variety of languages, powered by WhisperX

System Information

Version: 0.3
Author(s): Maarten van Gompel
Contact: proycon@anaproy.nl

Data processing notice: All data you upload to this service and data obtained using this service will remain yours and is accessible only by you and our technical staff. Your data will not be shared with third parties and not be used for any purpose other than this service's operation. You can remove your projects at any time and are encouraged to do so, which will remove your data from our servers permanently. We can not guarantee any long-term storage of your data so you are recommended to download the results and store it yourself immediately; projects on the server will be automatically deleted after 30 days. Despite our security precautions, we do discourage use of this service for highly confidential material as there is no encryption on the storage. Last, we also collect some statistics on the frequency of use of this service, when shared this will always be anonymised.

You will be asked to authenticate when you continue to use the rest of this service, by clicking the button below.

Continue

Or alternatively [continue using fallback authentication\(HTTP Basic Authentication\)](#)

- Login (CLARIN)
- -Create project
- Upload audio file(s)

Automatic Speech Recognition Service - version 0.3
by [Maarten van Gompel](#)

Powered by CLAM v3.2.10 - Computational Linguistics Application Mediator
by Maarten van Gompel - <https://proycon.github.io/clam>
Centre for Language and Speech Technology, Radboud University Nijmegen
& KNAW Humanities Cluster

Input

Input files

Show entries

Search:

Input File	Template	Format	Actions
KNMP2013najr.wav	Wav audio file	Wave Audio File	

Showing 1 to 1 of 1 entries

First Previous **1** Next Last

[Upload a file from disk](#)

Parameters

Global

Language

The language to recognize

Dutch / Nederlands

Model

The ASR model to use

large-v2

GPU

Use GPU (improves performance but may not always be available)



Diarization

Diarization

Enable speaker diarization?



Minimum speakers

Minimum number of speakers (this helps diarization)

2

Maximum speakers

Minimum number of speakers (this helps diarization)

3

Start

Activeer Windows
Ga naar Instellingen om W

Done.
Processing files
Starting...

Cancel and delete project

Discard output and restart

Show input files

Output files

(Download all as archive: [zip](#) | [tar.gz](#) | [tar.bz2](#))

Show entries

Search:

Output File	Template	Format	Viewers
error.log	Log file with (standard) error output	Plain Text Format	Download More...
KNMP2013najr.ctm	Transcription with full word segmentation/alignment	Conversation Time Marked File	Download More...
KNMP2013najr.json	Transcription with full word segmentation/alignment and speaker attribution	JSON Format (generic, not further specified)	Download More...
KNMP2013najr.srt	Timed transcriptions with speaker attribution (srt)	SubRip Text	Download More...
KNMP2013najr.tsv	Timed transcriptions with speaker attribution (TSV)	Tab Separated Values	Download More...
KNMP2013najr.txt	Plain text transcriptions without time stamps and speaker attribution	Plain Text Format	Download More...
KNMP2013najr.vtt	Timed transcriptions with speaker attribution	WebVTT	Download More...

Activeer Windows
Ga naar Instellingen om

3. WEBSERVICES AT BAS: TRANSCRIPTION PORTAL

The screenshot shows the TranscriptionPortal v1.0.7 interface. A 'Queue' dialog box is open, displaying the following content:

Queue [OK]

The following files are going to be processed one after the other. Please check if all options are set as you wish. While you are selecting the language you can click on the logos of the service providers for further information like data storage policy and terms and conditions.

When you click "OK" you agree with the terms and conditions of the selected (third-party) services and the files are marked for further processing.

Language: German (DE) [Watson]

Known Issues for Watson ASR: in very rare cases the service returns a generic REST transport error (92) instead of a result; this is reproducible, i.e. the same input signal always causes this error, but the reason is unknown.

Valid?	File	Language	Upload	Speech Recognition	Manual Transcription	Word alignment	Phonetic detail
				<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
✓	test_3.wav	deu-DE	✓	✓	✓	✓	✓
✓	test_1.wav	deu-DE	✓	✓	✓	✓	✓
✓	test_2.wav	deu-DE	✓	✓	✓	✓	✓

[Cancel] [OK]

3. WEBSERVICES AT BAS: TRANSCRIPTION PORTAL

The screenshot displays the TranscriptionPortal v1.0.7 interface. On the left, a 'Queue' modal is open, showing a list of files: test_3.wav, test_1.wav, and test_2.wav, all with a 'Valid?' status of '✓'. The language is set to 'German (DE) [Watson]'. A 'Known Issues for Watson ASR' warning is also visible. The main interface shows a 'Processing...' status with a progress bar and a '3. STOP PROCESSING' button. The progress bar includes steps: File, Upload, Speech Recognition, Manual Transcription, Word alignment, Phonetic detail, and Export. The 'Speech Recognition' and 'Manual Transcription' steps are completed for all files, while 'Word alignment' and 'Phonetic detail' are in progress.

File	Upload	Speech Recognition	Manual Transcription	Word alignment	Phonetic detail	Export
test_3.wav	✓	✓	✓	⏳	⏳	⏳
test_1.wav	✓	✓	✓	⏳	⏳	⏳
test_2.wav	✓	✓	✓	⏳	⏳	⏳

3. WEBSERVICES AT BAS: TRANSCRIPTION PORTAL

The screenshot displays the TranscriptionPortal web interface. The browser address bar shows the URL `clarin.phonetik.uni-muenchen.de/apps/TranscriptionPortal/`. The page title is "TranscriptionPortal v1.0.7". The main interface includes a navigation bar with "1. ADD FILES" and "2. VERIFY" buttons. A file list on the left shows three audio files: `test_3.wav`, `test_1.wav`, and `test_2.wav`. A modal window titled "Preview: test_2.txt" is open, displaying the following text:

teilzunehmen ist es nicht für sie die Gelegenheit ihn zu kritisieren bitte unterlassen sie das weil auch der die Bonus ganz darin die Bundesregierung sämtliche Altgatte ein verantwortlich für Millionenfachen Rechtsbruch im Rahmen der von Ihnen verursachten Masseneinwanderungen damit auch mitverantwortlich für die herausreduzierenden Morde tötungsdelikte Vergewaltigungen und viele andere Verbrechen und Vergehen massenhafte Verstöße gegen Artikel sechzehn AH Grund gesetz verantwortlich für die gesellschaftlichen Verwerfungen und kosten in vieler medialen Höhe meinen Abonnementen an denen noch Generationen von Deutschen leiden werden und über die die das Parlament nie entschieden hat Verantwortlich sind sie alle für politische Verfolgungen alles wirklich alles vernünftigen alles Deutschen im besten Sinne und für die politische Instrumentalisierung des Inlandsgeheimdienstes gegen die größte Oppositionsfraktion Sie sind verantwortlich

At the bottom of the preview window, there are two buttons: "Cancel" and "DOWNLOAD".

3. WEBSERVICES AT BAS: TRANSCRIPTION PORTAL

TranscriptionPortal v1.0.7

+ 1. ADD FILES 2. VERIFY

File

- test_3.wav
- test_1.wav
- test_2.wav

Preview: test_2.txt

teilzunehmen ist es nicht für sie die Gelegenheit ihn zu Bundesregierung sämtliche Altgatte ein verantwortlich Masseneinwanderungen damit auch mitverantwortlich andere Verbrechen und Vergehen massenhafte Verstöße gesellschaftlichen Verwertungen und kosten in vieler n Deutschen leiden werden und über die die das Parlam alles wirklich alles vernünftigen alles Deutschen im bes Inlandsgeheimdienstes gegen die größte Oppositionsf

Download results by line

This creates a zip-archive of all the selected results and optionally conversions to other formats. If you do not select any additional conversions only the original results are added to the zip-archive.

Please notice, that only the latest results are inserted to the package.

Add conversions (optional):

<input type="checkbox"/> CTM (.ctm)	<input type="checkbox"/> BAS Partitur Format (.par)
<input type="checkbox"/> AnnotJSON (_annot.json)	<input type="checkbox"/> TextGrid (.TextGrid)
<input type="checkbox"/> Text (.Table)	<input type="checkbox"/> Plain text (.txt)
<input type="checkbox"/> SRT Subtitles (.srt)	<input type="checkbox"/> WebVTT Subtitles (.vtt)

Get package

Close

3. WEBSERVICES AT BAS: OCTRA EDITOR

The screenshot displays the OCTRA v2.0.0 (local) web interface, a Dictaphone Editor. The browser tabs include 'OetraBackend', 'BAS | web service interface', 'BAS CLARIN Repository', and 'OCTRA - Orthographic Transcription'. The address bar shows the URL: `clarin.phonetik.uni-muenchen.de/apps/octra/octra-2/intern/transcr`. The interface features a navigation bar with 'OCTRA v2.0.0 (local)', 'Dictaphone Editor', 'Linear Editor', and '2D-Editor'. A 'beta' badge is visible. Below the navigation bar are links for 'Shortcuts [ALT + 8]', 'Guidelines [ALT + 9]', 'Overview [ALT + 0]', and 'Help'. The main area shows a timeline of audio tracks with corresponding text transcriptions. The audio is represented by green waveforms, and the text is shown in a light blue background. The timeline is marked with time intervals from 00:00 to 03:00. The text transcriptions include phrases such as 'Wir si...orden.', 'Ich bin in die Schule gegangen.', 'Auch in Neukönigsmark.', 'In keiner Hauptschule nicht, weil da war ich in Langenbach.', 'Und die acht Jahre bin ich auch nicht ganz gegangen, weil sie mich in der ...', 'tschaft geb...', 'Jetzt bin ich auch bei den Brüdern heimgegangen.', 'Und h...ilet?', 'Und was ... gemacht?', 'Alles.', 'Frucht gesät, gegac...', '...ker', 'Zuerst mit den Kühen, dann mit dem Raus und dann mit dem Traktor.', 'Da ...n.', 'Ja.', 'Alles, alles durchgemacht.', 'Mit der Hand gemäht.', 'Das machen he...hr so viele.', 'Naja, das können wir auch nicht viel.', 'Aber eins gefreut mich, beim Pinterbeck, der hat dann so einen Schnitt daheim gemacht.', 'Alle Ja...', 'Und da haben sie halt so groß gerät, die Männer.', 'Und hier hat es ... der Herr Sauer,', 'Sagt er, nein, in mir lässt er sich nicht gleich was sagen.', 'Und ich so, ich auch nicht.', 'Jetzt hat er mich so gesc...', 'eil der ist ja noch nicht so g...', 'Dann hat er gesagt, naja, mit soll man nicht haben.', 'Aber da mü...te kommen.', 'Dann hat er gesagt, was willst du da tun.', 'Und dann war die Bibsi da bei Sammersdorf.', 'Im Terrasse...ese gehabt.', 'Und ich habe gemäht.', 'Und auf ein... da stehen.', 'Ich habe aber den Buckel zum Weirrauch gezogen.', 'Und ich habe gesagt, na ja, kommt er auch nicht, der Herr Sauer.', 'Und er hat gesagt, ja Herr, ich muss mich bei dir...', 'tschuldigen, du kannst ja wir...', 'Hat er g...ewinnen?', 'Jetzt habe ich dann gesagt, na ja, ich tue dir das Eingst auch dem Gelingen.', 'Jnd j...', '02:40', '03:00'. The interface also includes a 'Quit' button and an 'Export Data' button.

3. WEBSERVICES AT BAS: OCTRA EDITOR

OCTRA v2.0.0 (local) — Dictaphone Editor Linear Editor 2D-Editor
beta
Shortcuts [ALT + 8] Guidelines [ALT + 9]

00:00 Wir si...orden. Ich bin in die Schule gegangen. Auch in Neukönigsmark. In keiner Ha
00:20 tschaft geb... Jetzt bin ich auch bei den Brüdern heimgegangen. Und h...ilet? Und
00:40 ...ker Zuerst mit den Kühen, dann mit dem Raus und dann mit dem Traktor. Da ...n. Ja. Alles, alles durchgemacht.
01:00 Naja, das können wir auch nicht viel. Aber eins gefreut mich
01:20 Und da haben sie halt so groß gerät, die Männer. Und hier hat es ... der Herr Sauer, Sagt e
01:40 eil der ist ja noch nicht so g... Dann hat er gesagt, naja, mit soll man nicht haben. Aber da mü...te kommen. Dann hat er gesagt, wa
02:00 Und ich habe gemalt. Und auf ein... da stehen. Ich habe aber den Buckel zum Weirrauch gezogen.
02:20 tschuldigen, du kannst ja wir... Hat er g...ewinnen? Jetzt habe ich dann gesagt, na ja, ich tue dir das Eingst auch dem Gelingen. Jnd j
02:40 03:00

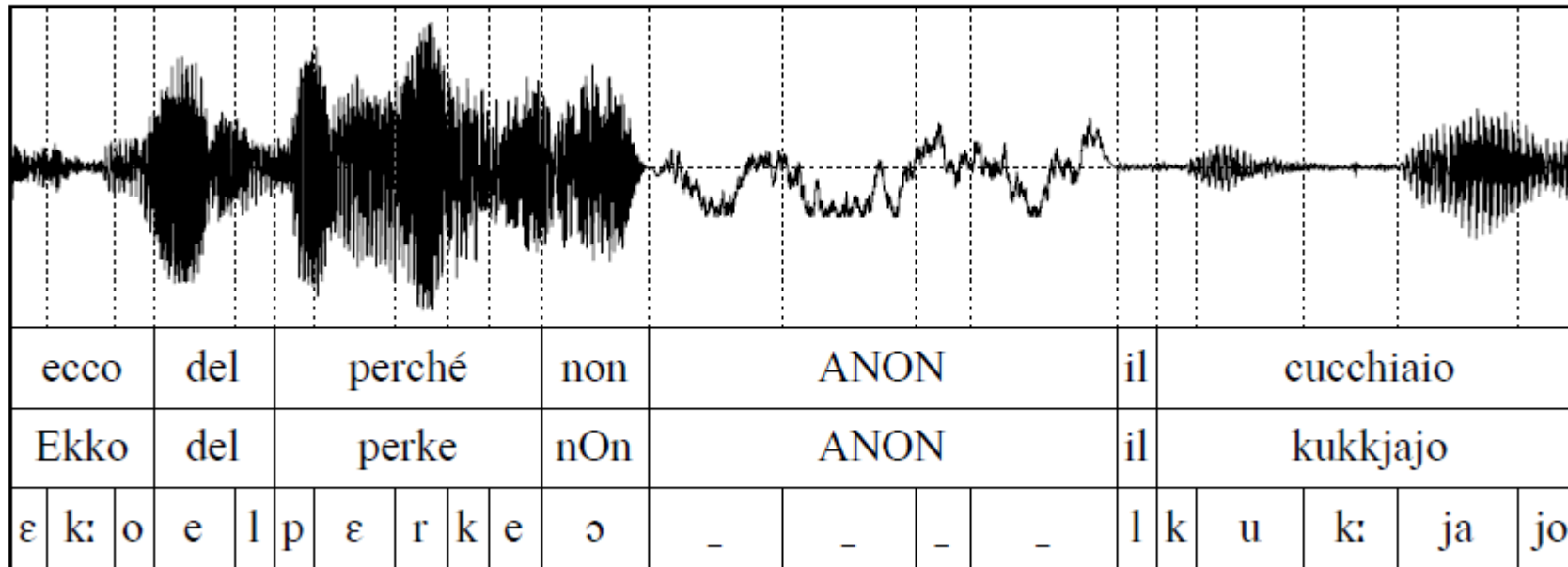
Quit Export Data

Octra 2D editor:

- Runs in a browser
- Displays large parts of longer recordings without visual clutter focuses on orthographic transcription
- May access local and third-party ASR providers and MAUS segmentation
- Exports transcripts to a large number of file formats.

3. WEBSERVICES AT BAS: ANONYMIZER

web service Anonymizer: removes words from a stop list from the transcript (on all levels) and replaces the corresponding fragment in the audio with a noise



3. WEBSERVICES AT BAS: OTHER SERVICES

- ChannelSeparator separates the individual channels of a stereo recording and retains in each channel only the voice of the dominant speaker
- G2P (Grapheme to phoneme) converts an orthographic text to its phonemic representation. The service allows a customized specification of pronunciation rules for vernacular language, dialects, and common coarticulation phenomena (e. g. 'haben wir' (we have) → /hamva/ or /hama/ in German). These rules improve the performance of automatic word alignment
- WebMAUS (**M**unich **a**utomatic **s**egmentation) aligns an orthographic transcript in one of the available languages and regional variants) with the audio signal

4. DO IT YOURSELF WITH WHISPER: GENERAL

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification. Whisper can recognise more than 99 different languages.

To download Whisper, go to their Github-page:

<https://github.com/openai/whisper/blob/main/README.md>

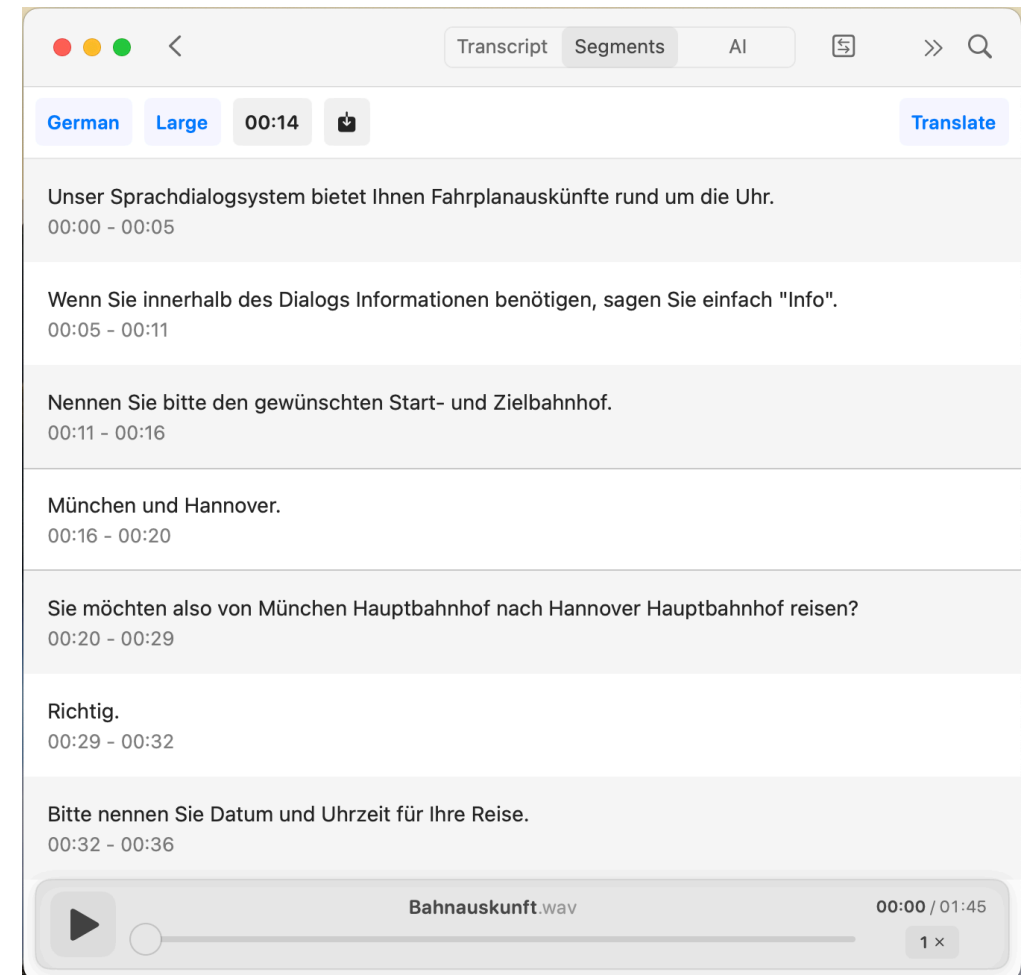
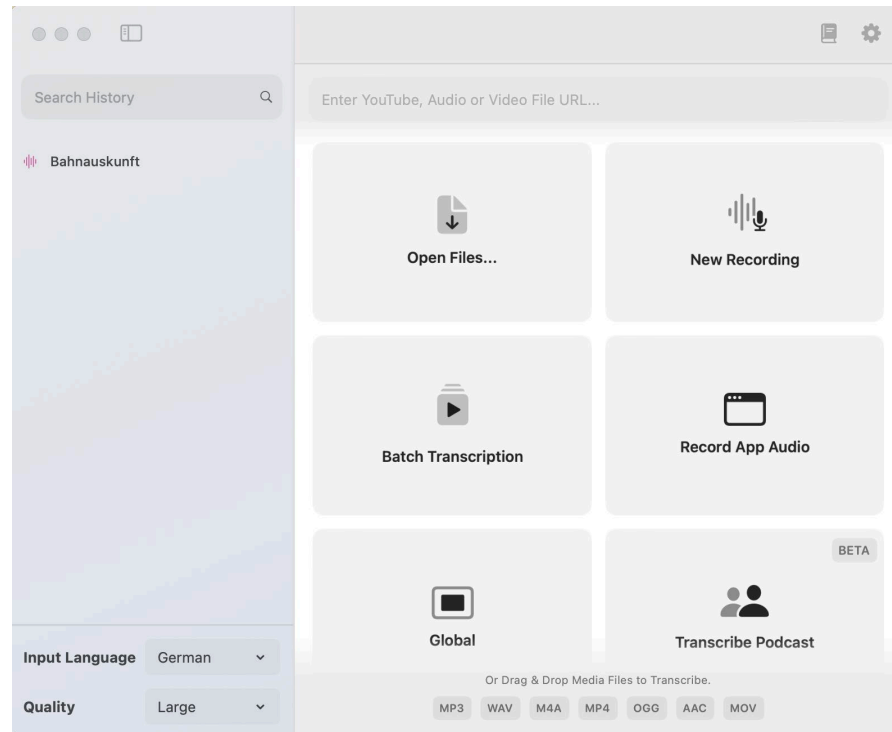
Because it is Open Source, other people can modify and add their own “additions” like Diarization & VAD (WhisperX) and can decrease the working time (Faster-Whisper).

Moreover, Georgi Gerganov developed a C++ version that can be used in stand-alone software packages like **MacWhisper** (Apple), and **aTrain** (Windows).

A more detailed description of Whisper and its derivatives, can be found at <https://speechandtech.eu/news>

4. DO IT YOURSELF WITH WHISPER: MACWHISPER

MacWhisper is a simple but complete ASR package that enables the recognition and correction of AV-files.

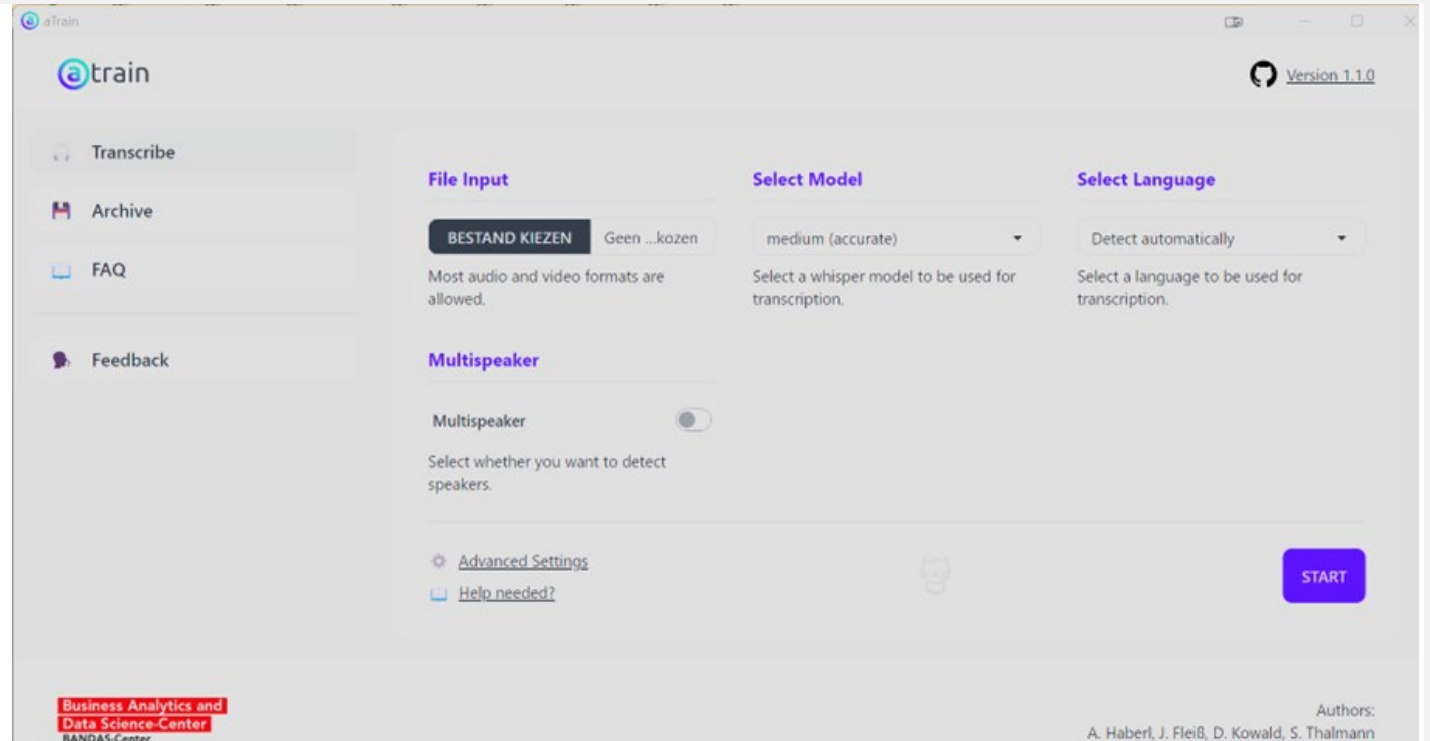


4. DO IT YOURSELF WITH WHISPER: ATRAIN

ICT people from the Universität of Graz developed aTrain: an ASR for Windows and Linux.

Whereas MacWhisper only does what Whisper can do, aTrain is more comprehensive and incorporates several new developments used today.

For example, Diarization can be done, and the programme is very fast by using Fast-Whisper.



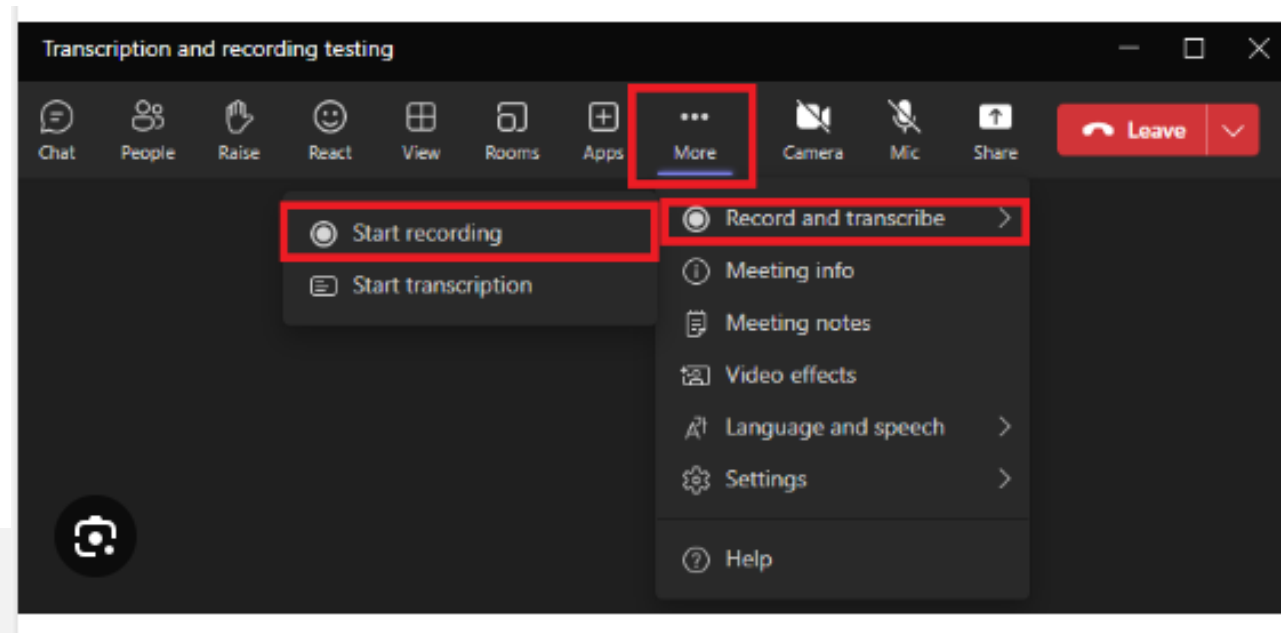
aTrain can be downloaded (>10GB) in the Microsoft store:
<https://apps.microsoft.com/detail/9n15q44szns2>

5. COMMERCIAL SOLUTIONS

Companies offering transcription services, e.g.

- Amberscript, Sonix, Scriptoman, Happy Scribe, Autoscriber
- Transcription solutions in Zoom and MS Teams for online interviews / conversations / focus groups

Special arrangements / licenses needed for sensitive recordings



6. REMAINING CHALLENGES

What remains under the ASR radar:

- Filled pauses and back channels
- Repetitions with word truncations
- Self-repairs
- Pronunciation peculiarities

However, these phenomena are very relevant for specific research domains



6. REMAINING CHALLENGES

Problematic conditions:

- Background noise
- Speaker overlap
- Dialect speech
- Children & older people
- Speakers with language impairments / speech disorders

