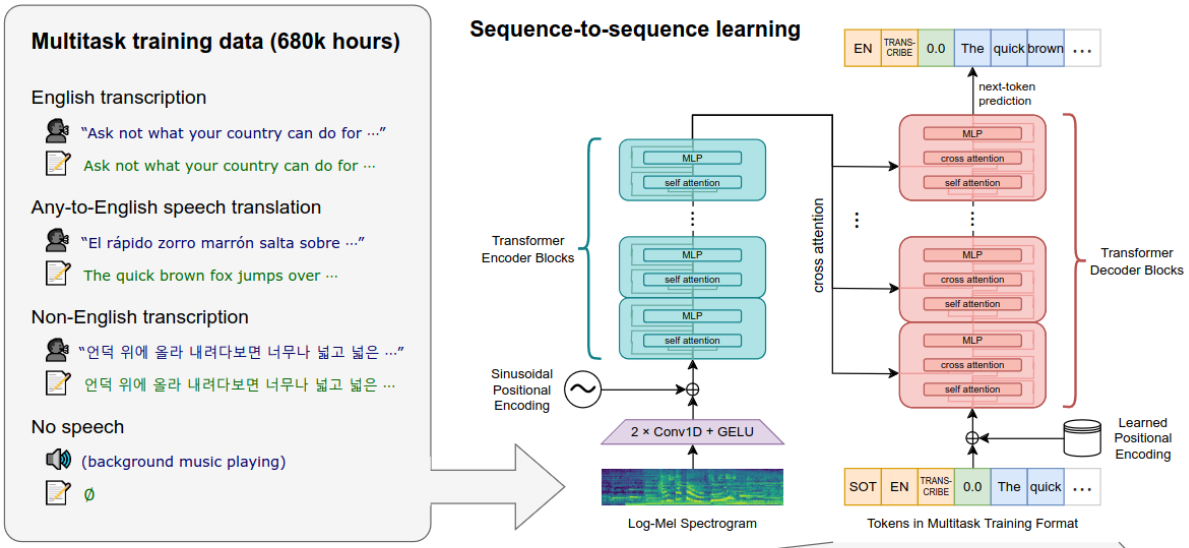
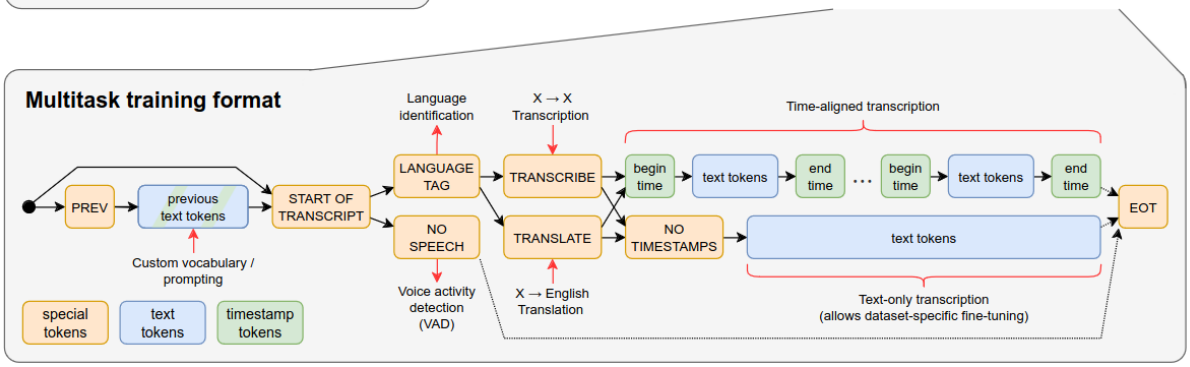
Folker + Whisper Notes

Goal: Speech to cGAT

Fine-tune Whisper on **German language** and **cGAT annotations**

- Why Whisper?:
  - Robust through pre-training
  - Collection of models
- Why **not** Whisper?
  - Audio-conditional language model:
    - Context
    - Corrections: “zwo” → “zwei”
  - Spectrogram “compression” of raw signal → information lost*





Annotation → Task

”Easy” Tasks

- **Disfluencies**: Mhm, Ähm, Hm, öh, …

- Textual
- Token: `<|Verzögerungssignale|>`(äh) , `<|Rezeptionssignale|>` (mh)

- **Pause**: No speech token + timestamp token: `<|0.0|><|nospeech|><|8.3|>`
  - Mikropausen* `<|micropause|>`
- **Fremdwörter**, **Fachwörter**, **Zahlen, …**: Provided by Multilanguage Large Pre-training
- **Nonverbale Handlungen**: lachen, kichern, husten, räuspern, schnauben, …
  - `<|startofnonverbal|>` Prediction `<|endofnonverbal|>`
  - separate tokens: `<|lachen|>`, ...
- Segment-based **timestamp** prediction

Difficult Tasks

- **Tönhöhenbewegung:** hoch steigend, steigend, gleichbleibend, fallend, tief fallend
  - ProsoTool: https://github.com/szekrenyesi/prosotool: Audio+Diarization **=** Intonation

> https://aclanthology.org/W16-4016.pdf

- Pipeline Praat

- **Fokusakzent**
- (**Dehnung:** 3 types: 0.2s-0.5s, 0.5s-0.8s, 0.8s-1.0s)
- **Ein- und Ausatmen***
- **Wortübergreifende Prozesse**: hat_n, gibt_s, so_ne, …
- **Speaker Diarization:** speaker swap token?
- **Unsicherheit**

Modelling

Size	Layers	Width	Heads	Parameters	English-only	Multilingual
tiny	4	384	6	39 M	√	√
base	6	512	8	74 M	√	√
small	12	768	12	244 M	√	√
medium	24	1024	16	769 M	√	√
large	32	1280	20	1550 M	x	√

MULTILINGUAL LIBRISPEECH

Model	Dutch	English	French	German	Italian	Polish	Portuguese	Spanish
Whisper tiny	39.4	15.7	36.8	24.9	41.7	34.2	31.3	19.2
Whisper base	28.4	11.7	26.6	17.7	31.1	22.8	21.9	12.8
Whisper small	17.2	8.3	16.2	10.5	21.4	11.2	13.0	7.8
Whisper medium	11.7	6.8	8.9	7.4	16.0	6.5	9.0	5.3
Whisper large	10.2	6.3	8.9	6.6	14.3	6.6	9.2	5.4
Whisper large-v2	9.3	6.2	7.3	5.5	13.8	5.0	6.8	4.2

Table 10. WER (%) on MLS

- Folker files → Tokens
- max. 30 sec segments
- 16,000 Hz, 80 channel log-mel-spectrogram, 25ms window, 10ms stride
  - 30sec audio → 80 x 3000 spectrogram
- Freeze first layers?
- Artificially extend dataset

▼ Examples

cGAT:

weil da da sieh siehtst du trotzdem sehr viele szenarien (.) aus denen du lernen kannst **mhm**

das sind zwar ausgedachte szenarien alles

(1.12)

Whisper:

Ja, weil da siehst du trotzdem sehr viele Szenarien, aus denen du lernen kannst.
Das sind zwar ausgedachte Szenarien alles, aber...

Whisper + Prefix:

Ja, weil da siehst du trotzdem sehr viele Szenarien, aus denen du lernen kannst.
Mhm.
Das sind zwar ausgedachte Szenarien alles, aber...

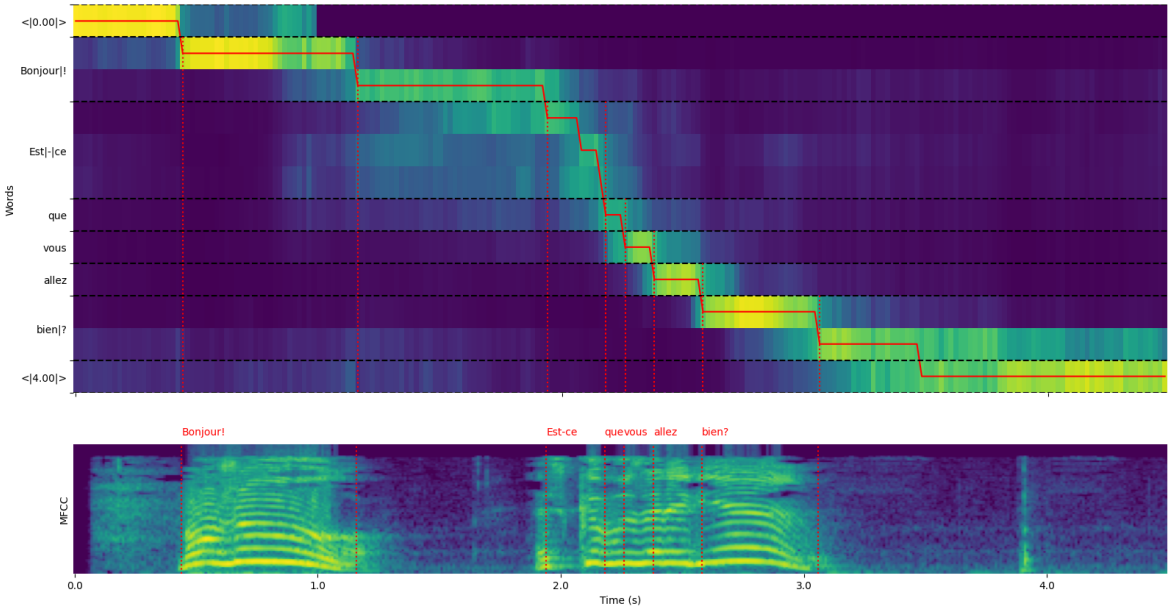
Prefix: `"Ähm, lass mich mal überlegen, mhm ... Okay."`

cGAT: Du hast_n (.) wie Helmut Kohl (.) so_n Aquarium

Whisper: Du hast wie Helmut Kohl ein Aquarium.

▼ Word-level timestamps

- WhisperX: add wav2vec model Paper, https://github.com/m-bain/whisperX
- https://github.com/linto-ai/whisper-timestamped: Dynamic Time Warping applied to cross-attention weights



Resources

- Fixed input dimension

- Reduce memory: https://github.com/huggingface/community-events/tree/main/whisper-fine-tuning-event#tips-and-tricks

GPU	Model	Train Batch Size	Eval Batch Size
V100 (16 GB)	small	16	8
V100 (16 GB)	medium	2	1
A100 (40GB)	small	64	32
A100 (40GB)	medium	32	16

- DeepSpeed

GPU	Model	Train Batch Size	Eval Batch Size	Speed
V100 (16GB)	small	32	16	1.3s/it
V100 (16GB)	medium	16	8	2.0s/it
V100 (16GB)	large	8	4	3.8s/it
A100 (40 GB)	small	64	32	2.3s/it
A100 (40 GB)	medium	64	32	5.8s/it
A100 (40 GB)	large	32	16	5.9s/it

Parameter-Efficient Fine-Tuning

- https://github.com/huggingface/peft
- https://github.com/huggingface/peft/blob/main/examples/int8_training/peft_bnb_whisper_large_v2_training.ipynb

Useful links

- https://huggingface.co/blog/fine-tune-whisper
- https://github.com/huggingface/community-events/tree/main/whisper-fine-tuning-event
- Adding a (special) token: https://github.com/openai/whisper/discussions/658
- Fine-tuning Whisper with timestamp tokens: https://github.com/openai/whisper/discussions/620
- How to make Whisper recognize more words: https://github.com/openai/whisper/discussions/288