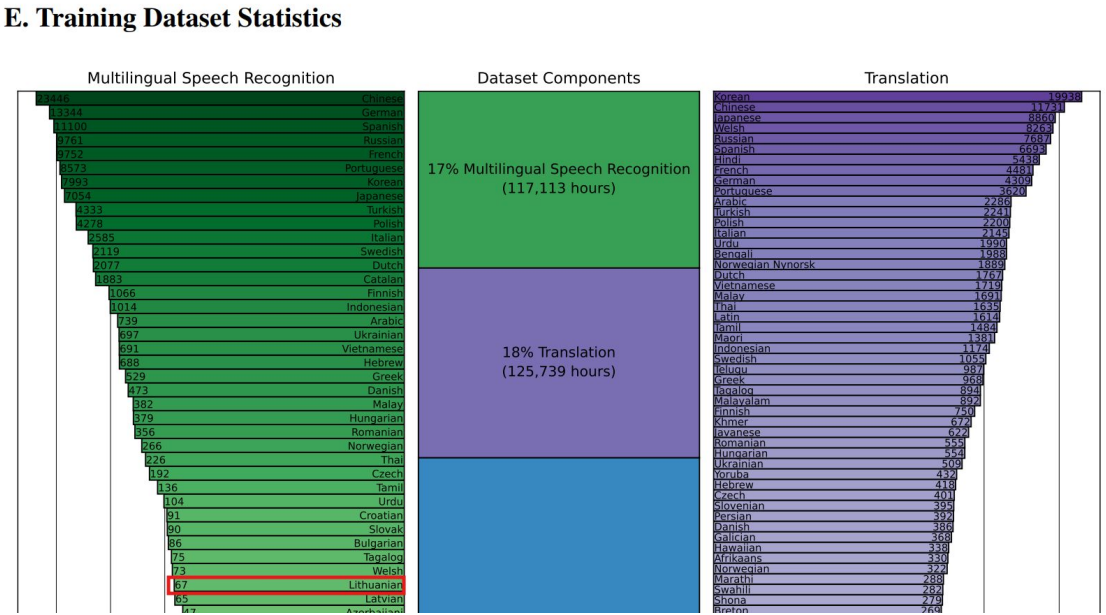
ASR Speech to Text LoRa fine-tuning with low-resource language using Whisper model

Povilas Kvedaras
AI Engineer at Novian PRO
VU MIF lecturer
AI Methods Lab researcher

Github repo with Google Colab script:
https://github.com/PovilasKvedaras/STT_workshop

Whisper model architecture (1)

- Robust Speech Recognition via Large-Scale Weak Supervision was introduced in 2022 DEC http://arxiv.org/pdf/2212.04356
- Model was trained with 680K hours of multilingual data for automatic speech transcription and translation. Lithuanian has only 67 hours as low-resource language



Whisper model architecture (2)

- Training data has linear relation to WER
- With 10K hours transcription model can reach up to 5 percent WER
- My analysis has shown that 5-8 percent WER can be reached even with 1K hours of transcription data if data quality is very high (low error rate, high number of different speakers, reduced audio silences, correctly formatted text, model aligned labels)

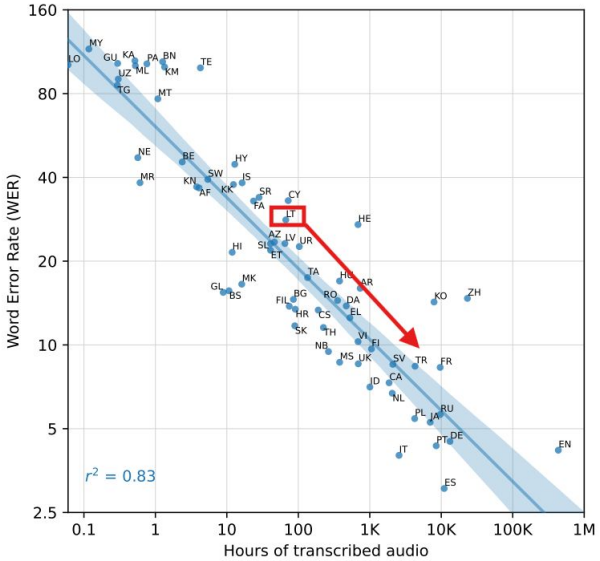
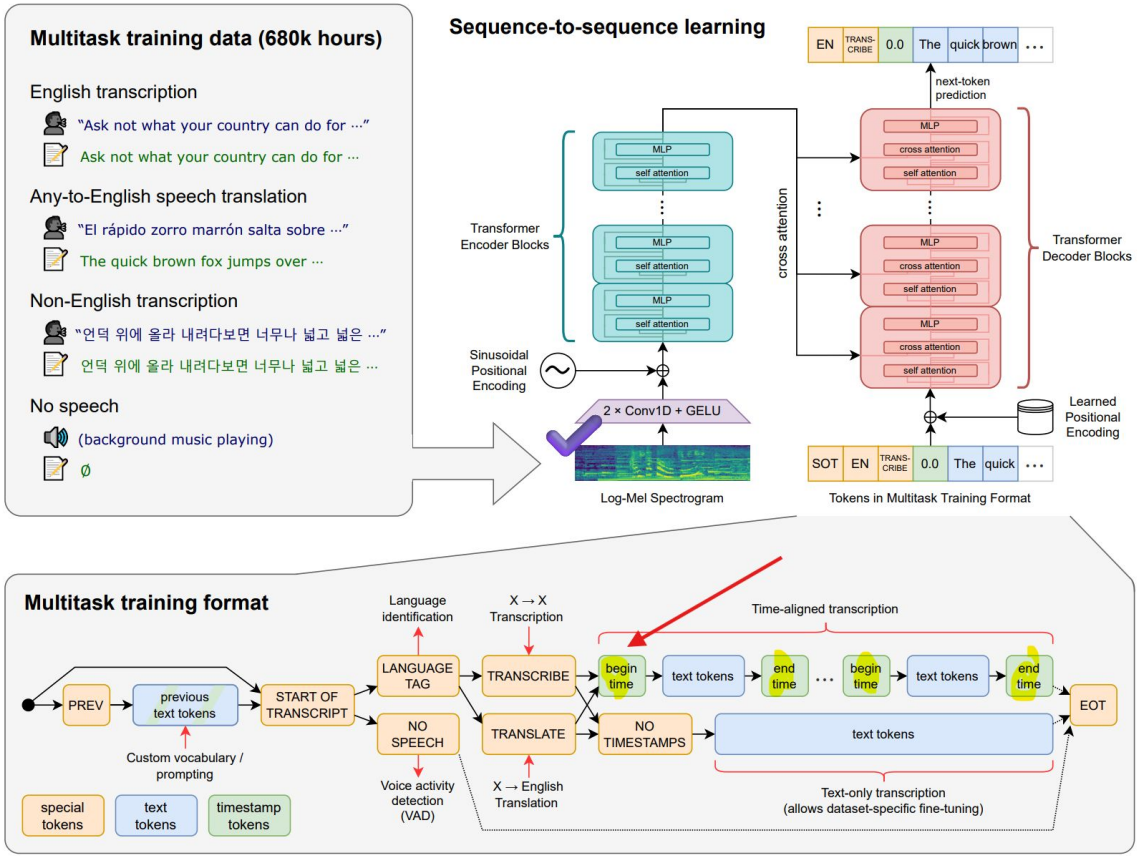


Figure 3. **Correlation of pre-training supervision amount with downstream speech recognition performance.** The amount of pre-training speech recognition data for a given language is very predictive of zero-shot performance on that language in Fleurs.

Word Error Rate (WER): Measures how often a speech-to-text system makes mistakes (substitutions, insertions, deletions) compared to a correct transcript, calculated as (S+I+D)/N.

Whisper model architecture (3)



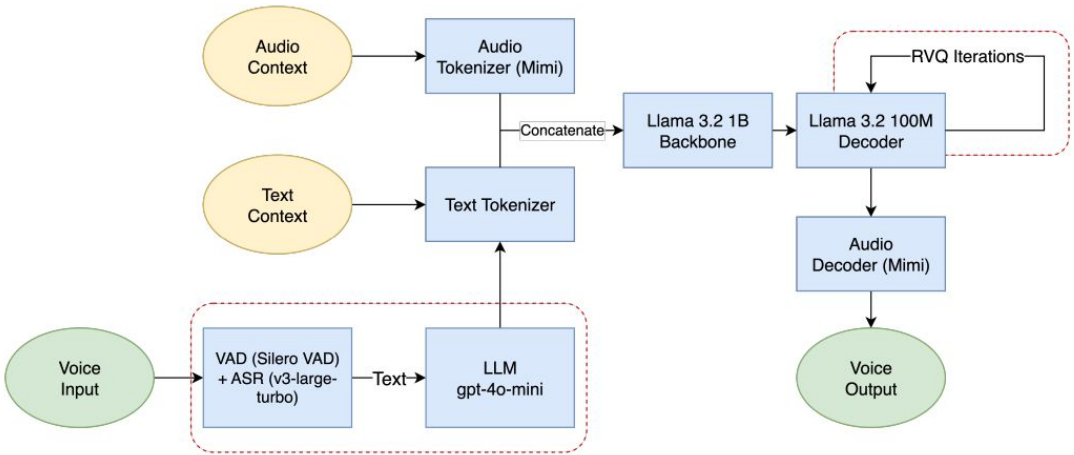
Correct labels formatting is IMPORTANT if we want to minimize catastrophic forgetting.

[SOT] [LANG] [TRANSCRIBE]
<|0.00|> Sentence One <|4.02|>
<|4.02|> Sentence Two <|12.60|>
[EOT]

Special tokens:
SOT - task “start of transcription”
LANG - language (better to provide)
<> - begin and end time tokens!

Why Whisper is used in current days?

- Newer models haven't outperformed - Whisper turbo v3 model shows good performance to resource ratio (only 0.8b parameters) and can run even on CPU
- Tokenizer is trained on big compute with 100+ Languages
- Good support and big community - At least three different well supported inference engines (Faster Whisper, Transformers, Unsloth), ease of fine-tuning using different frameworks, quantization reduces latency by 19% and model size by 45%, while preserving WER https://arxiv.org/pdf/2503.09905
- transcription accuracy
- Various use cases:
  - Transcription services on mass scale (1 second GPU = 120 seconds of transcription)
  - Speech2Speech - https://arxiv.org/pdf/2509.20971 L4 GPU latency 0.5 seconds with 200 USD deployment costs and up to three concurrent speakers



Demo