

What did you say?

Automatic Speech Recognition in a World of Whisper

Lisa Becker

Machine Learning Engineer at ML6
Natural Language Processing

https://www.linkedin.com/in/becker-lisa/
or search “Lisa Becker ML6”

ML6 is a leading tech company specialized in AI services.

100+ machine learning & AI experts.
300+ use cases.

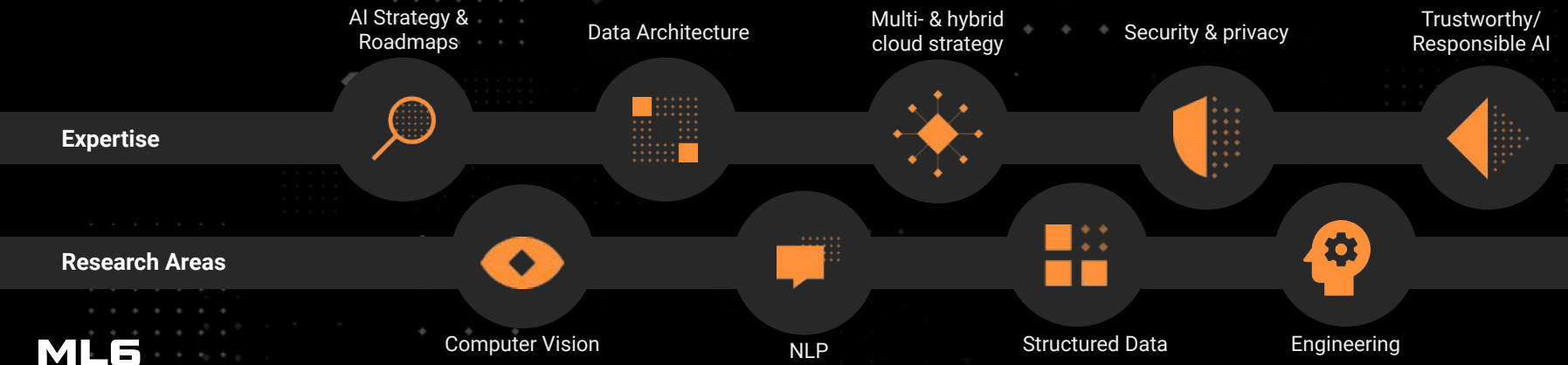
European presence at
150+ clients across industries.

International award winning
AI company

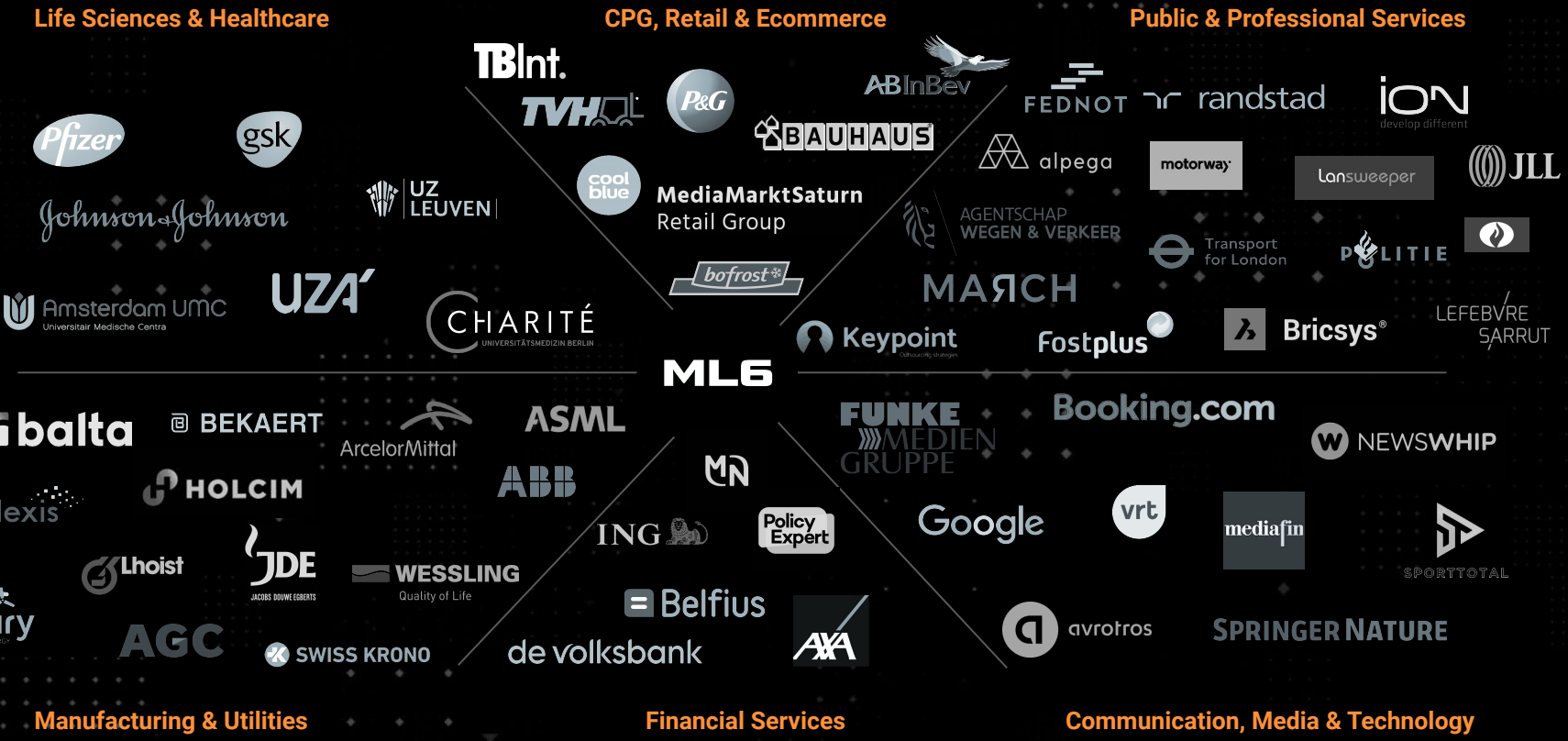
Cloud agnostic to match your
unique needs and preferences.

17% of our workforce is fully
dedicated to ML research.

Trusted open-source
contributions.



ML6 is a leading tech company specialized in AI services.



Agenda.

- What is Automatic Speech Recognition (ASR)?
- ASR at ML6
- Why is Whisper so good?
- Other cool things around Whisper

What is ASR?
An overview.

Automatic Speech Recognition ≈ Speech-to-Text

- Recognition & translation of spoken language into text
- Downstream NLP tasks
- Reverse: Speech synthesis

1952 — Audrey

Single speaker digit recognition system.

1970s — Hidden Markov Models

Dominant for a long time.
Implemented in tools like Kaldi.

2007 — LSTM + Connectionist Temporal Classification

Increases Google's speech recognition accuracy by 49%.

2016 — Listen, Attend and Spell

Sequence to Sequence attention based models (transformers).

2020 — wav2vec 2.0 Meta

Transformer with semi-supervised pre-training approach.

2022 — Whisper OpenAI

Transformer with weakly supervised pre-training approach.

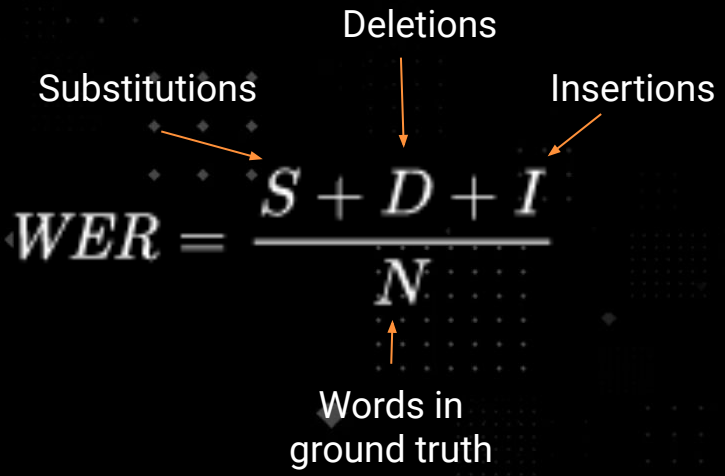
ASR at ML6.
Phase 1: DPV & Memo

- Transcription of customer calls
- dpv: German
- Memo: Flemish
- wav2vec 2.0 finetuned on custom labelled data

memo

dpv Deutscher Pressevertrieb

ASR at ML6.
Phase 1: DPV & Memo



Ground Truth	"Word error rate (WER) is a common metric of the performance of a speech recognition or machine translation system."
Hypothesis	Substitution Deletion "Ward error rate (WER) is a _ metric of the performance of a speech recognition system or machine translation system." Insertion

WER: Word Error Rate

CER: Character Error Rate

1 + 1 + 1 / 19 = 15% WER

→ **goal is to minimize WER**

ASR at ML6.
Phase 1: DPV & Memo

	dpv Phase 1	memo+ Phase 1
Model	wav2vec 2.0	wav2vec 2.0
Finetuned	~10 hrs	~20 hrs
WER (word)	32%	32%
CER (letter)	20%	19%

ASR at ML6.
Phase 2: Memo

- 2022/2023
- Whisper (large v2)
- No finetuning
  - Generating insights from call transcriptions

memo+

	dpv Phase 1	memo+ Phase 1	memo+ Phase 2
Model	wav2vec 2.0	wav2vec 2.0	**Whisper large**
Finetuned	~10 hrs	~20 hrs	❌
WER (word)	32%	32%	**27%**
CER (letter)	20%	19%	**17%**
Size		14GB	**10GB**
Multitask	❌	❌	✅

Why is Whisper so good?
Multitask

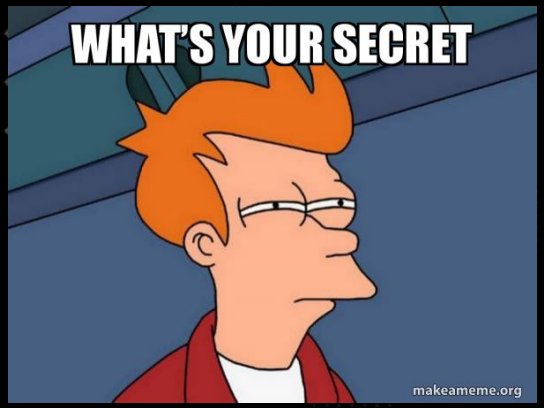
```python
import whisper

model = whisper.load_model("base")
result = model.transcribe("audio.mp3")
print(result["text"]
```

> of wat informatie? Ik zou graag willen... Ik was een keer op dit appartement en nu is het ongeveer een paar weken en dit appartement is nog steeds beschikbaar dus ik heb de vraag, als het nog steeds beschikbaar is, kan ik deze huis afwachten?

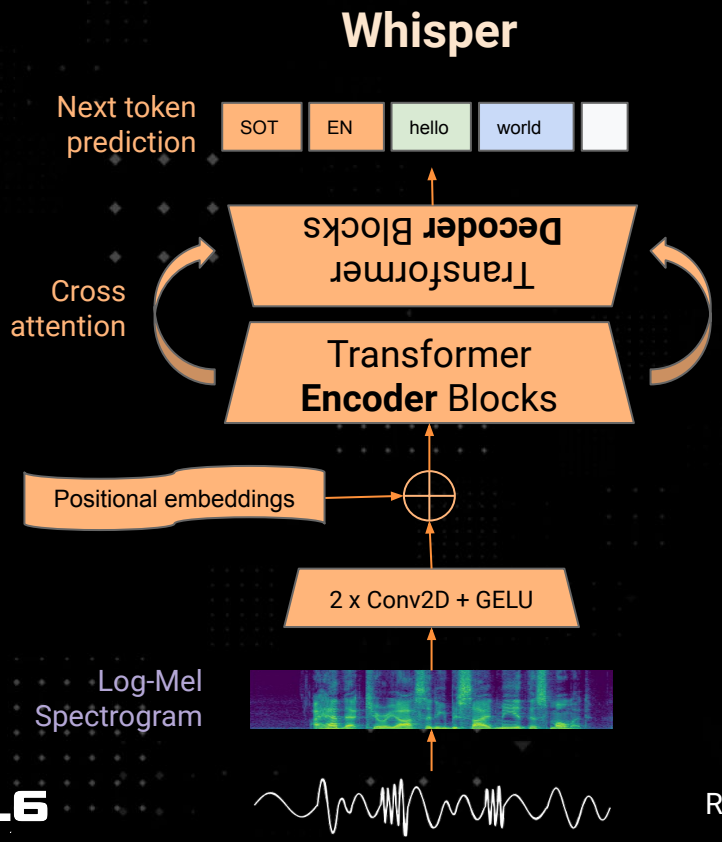
- Whisper large, 2022

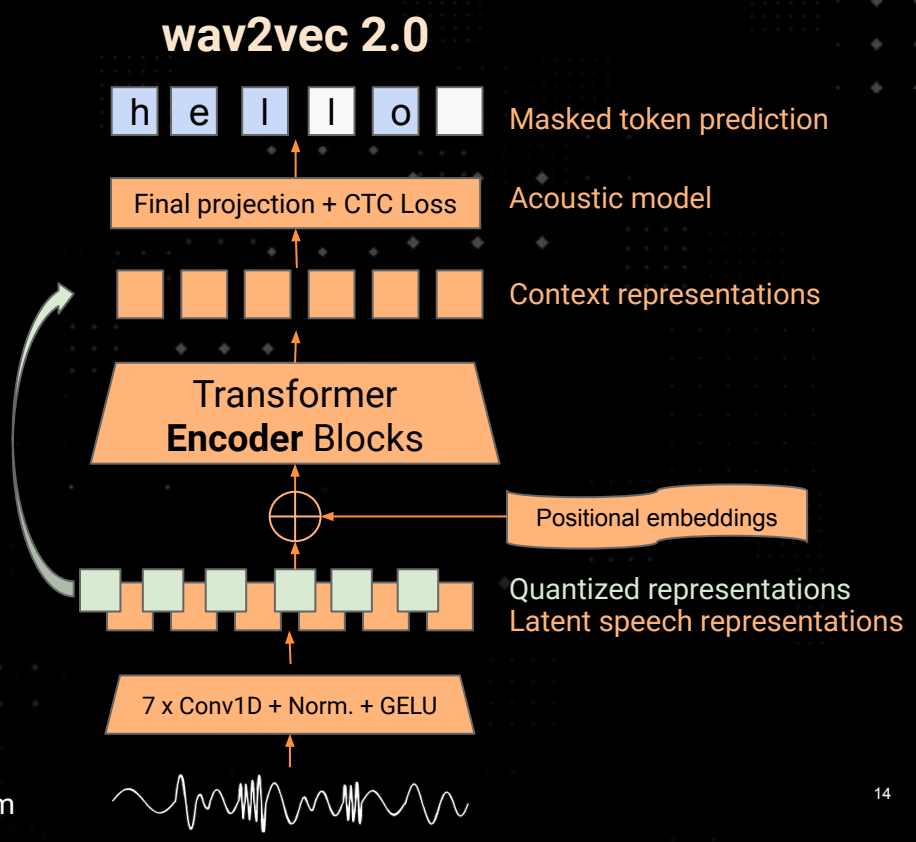
Why is Whisper so good?
What's your secret?



Dataset	wav2vec 2.0 Large WER (word)	Whisper Large v2 WER (word)
LibriSpeech	2.7	2.7
Artie	24.5	6.7
Fleurs (EN)	14.6	4.6
Common Voice	29.9	9.5
Tedlium	10.5	4.0
ChiME6	65.8	25.6
WSJ	7.7	3.1
…		
Average	**29.5**	**12.9**

Why is Whisper so good?
Because of the architecture?





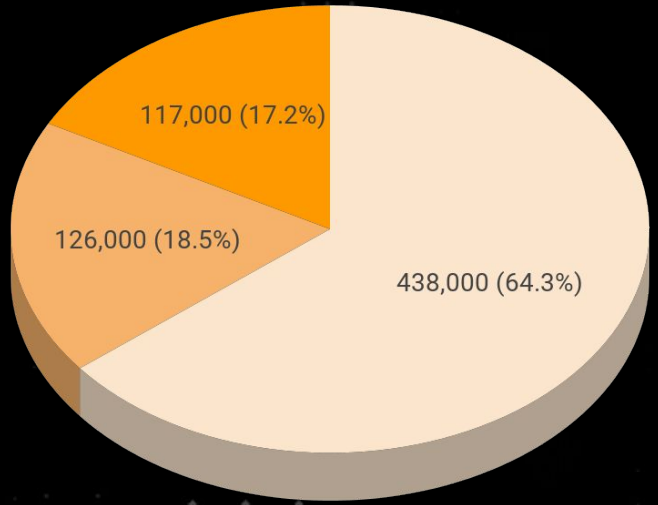
Why is Whisper so good?
Let's compare

Model	wav2vec 2.0 Meta	Whisper OpenAI
Released	2020	2022
# languages	53	99
Language detection	❌	✅
Language translation	❌	✅
Timestamps	(✅)	✅
Diarization	❌	❌
Casing	❌	✅
Punctuation	❌	✅
On	✅	✅
Training data	56k hours	680k hours

Why is Whisper so good?
Because of the training data!

Paper: Robust Speech Recognition via Large-Scale Weak Supervision

"If an acronym or basis for the name is desired,
***WSPSR** standing for Web-scale Supervised*
Pretraining for Speech Recognition can be used."



- Hours English audio + English transcript
- Hours non-English audio + English transcript
- Hours non-English audio + corresponding transcript

Why is Whisper so good?
Because of the training data!

Initial training & manual inspection

Identify and remove low-quality samples.

Automated filtering methods

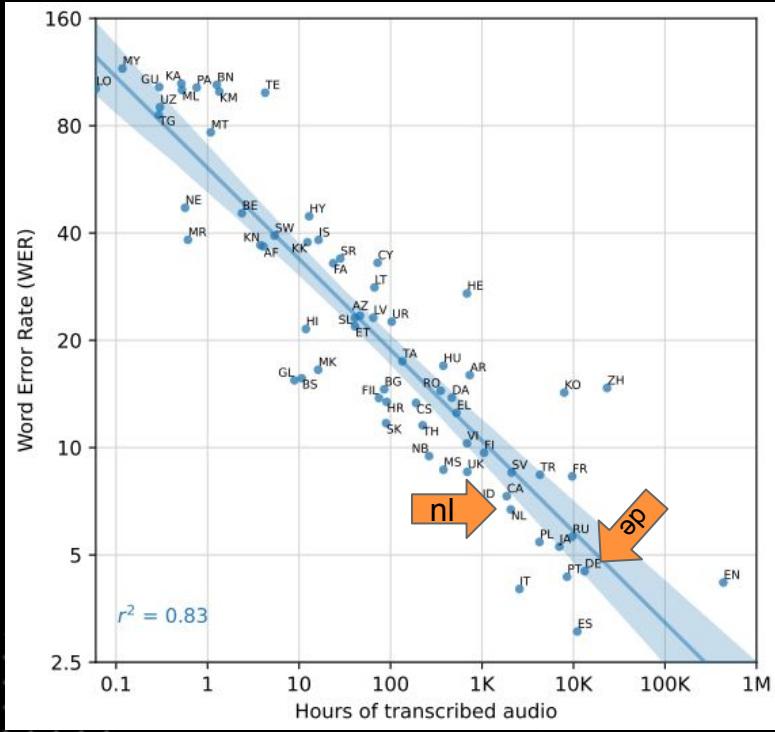
- Exclude if all upper-/ lowercase
- Exclude if no punctuation
- Check if language matches metadata
- Fuzzy de-duping
- …

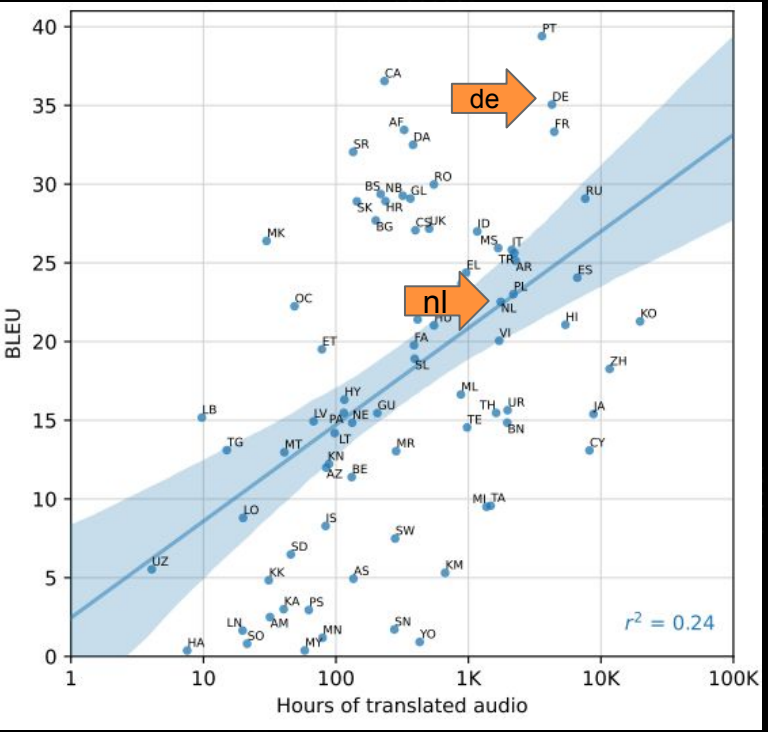
Scraping

“Audio that is paired with transcripts on the Internet… from many environments, recording setups, speakers, and languages.”

Other cool things about Whisper

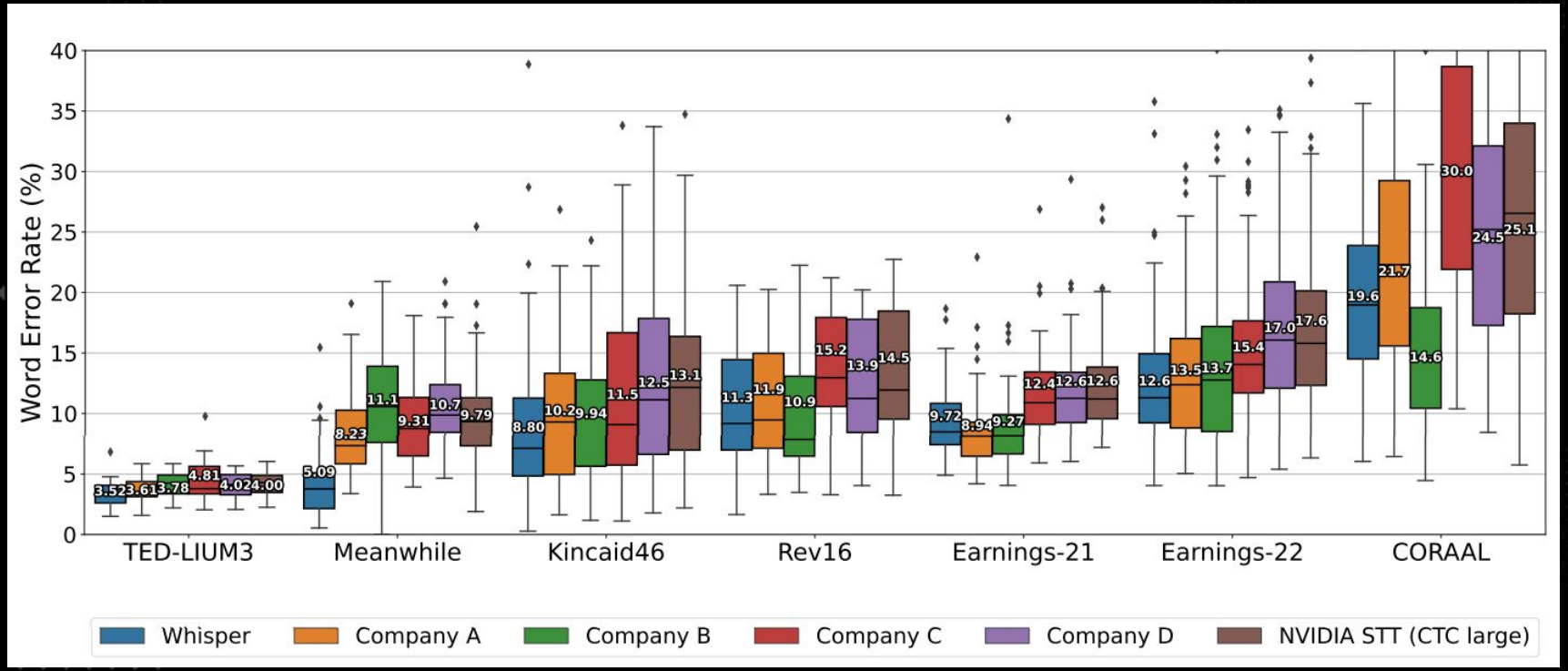
More data == better?





Other cool things about Whisper

Comparable/better to commercial models, similarly costly

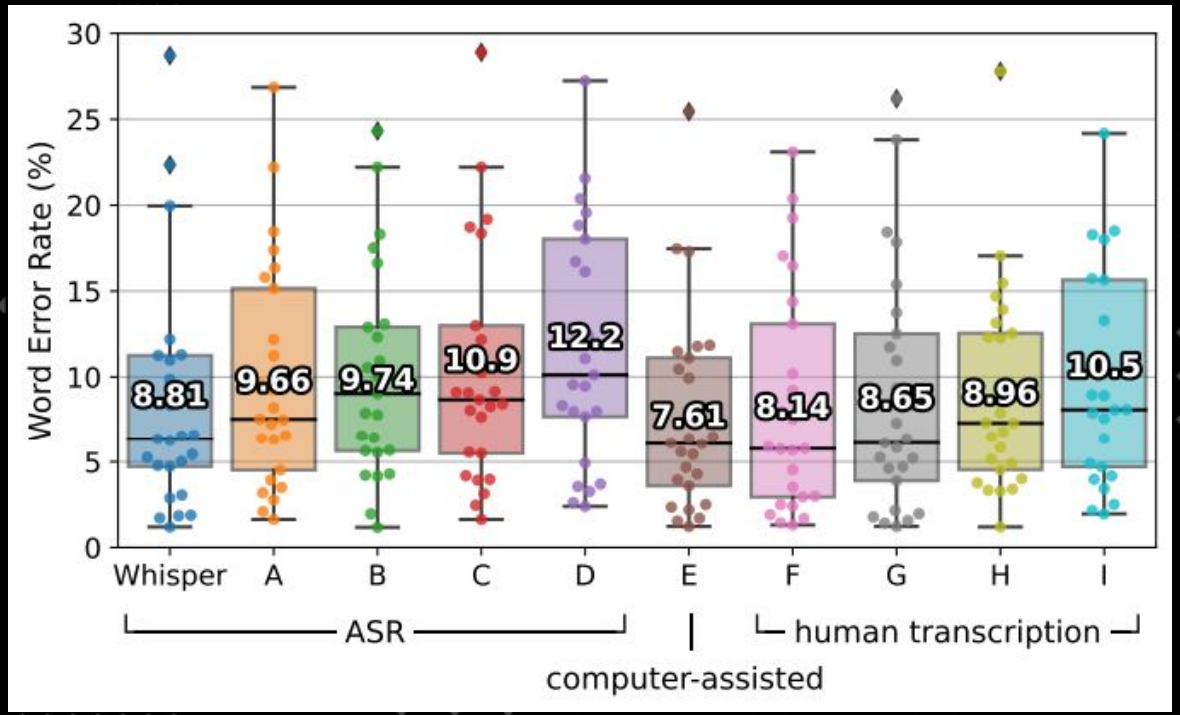


Other cool things about Whisper

“We are family” 🎶

Size	Parameters	Required VRAM	Relative speed
tiny	39 M	~1 GB	~32x
base	74 M	~1 GB	~16x
small	244 M	~2 GB	~6x
medium	769 M	~5 GB	~2x
large	1550 M	~10 GB	1x

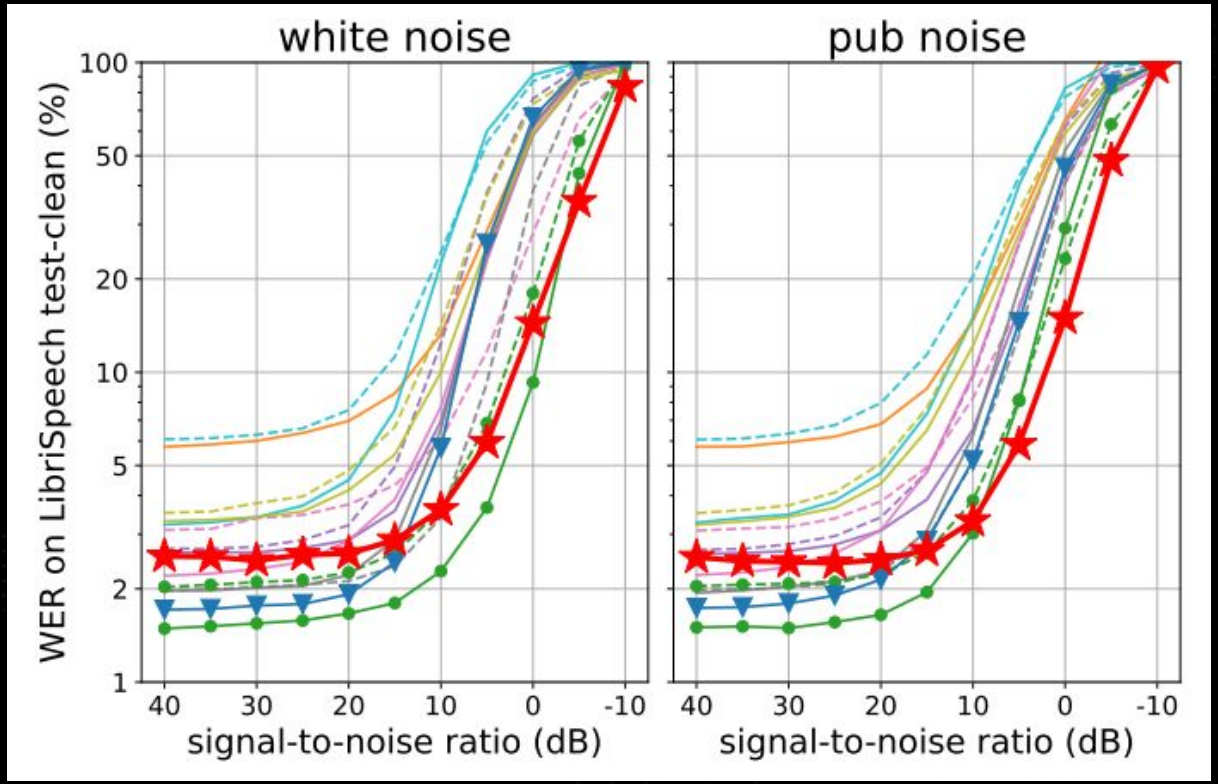
Other cool things about Whisper
Comparable to human transcription



A - D: Commercial models
F - I: Commercial human transcription services

Other cool things about Whisper

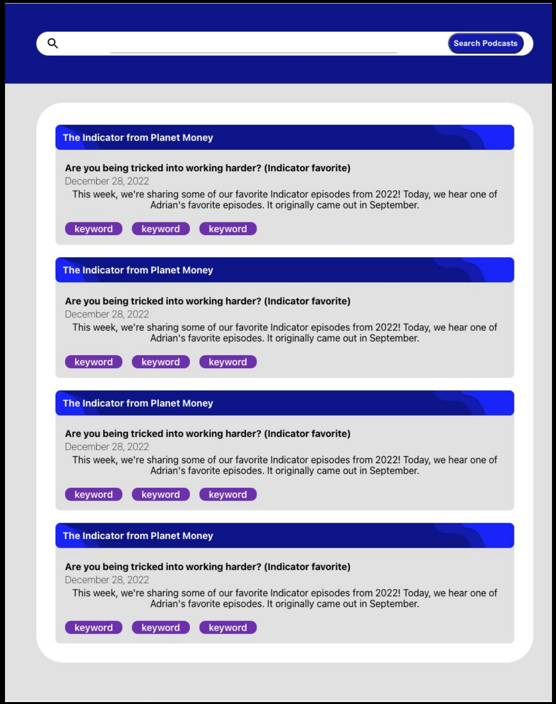
More noise == more robustness



Other cool things done with Whisper
Internal Project: "Make podcasts searchable"

Idea:
Tool for searching podcasts (+ summary & keywords)

Checkpod
data acquisition → transcription with Whisper → haystack by deepset
summarization, … →



Takeaways:

- Only large model yields acceptable multilingual results
- Processing time on GPU ≈ 3% of audio input duration (1.5 min for 1 hr in EN compared to 100% for NL)

Other cool things done with Whisper
Finetuning Event

- 2-week sprint in December 2022
- **Hugging Face**: Training scripts, notebooks, events…
- **Lambda**: A100 GPUs
- Very active, supportive community on Discord

- Evaluated on Common Voice 11
- 🔗Blog post, leaderboard, winners

🤗 Whisper Event: Final Leaderboard

This is the leaderboard for Common Voice 11 Dutch (nl).

Please click on the model's name to be redirected to its model card.

Want to beat the leaderboard? Don't see your model here? Ensure...

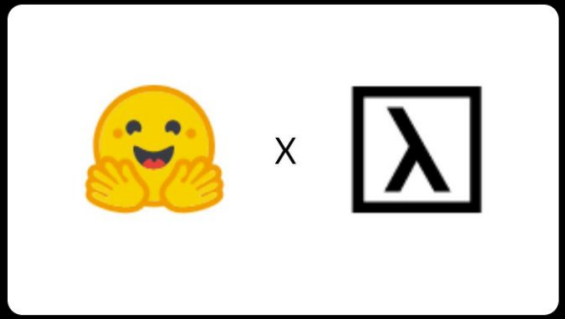
model_id	wer	cer
GeoffVdr/output1	5.90	-
GeoffVdr/whisper-medium-nlcv11	7.51	-
pplantinga/whisper-small-nl	11.51	-
thomashaas/parla-nl-v1	11.62	-
sanchit-gandhi/whisper-small-nl-1k-...	11.79	-
BerB2000/whisper-small-nl-last	307.71	-

Lambda
@LambdaAPI

Results from the Lambda x 🤗@huggingface sprint to fine-tune @OpenAI's Whisper model:

- 20K+ A100 hours provided for community use via Lambda Cloud credits
- 100+ languages covered
- 650 models created

Beyond impressed by the community. Giveaway announcement coming Friday!



8:55 pm · 4 Jan 2023 · **31.3K** Views

Other cool things done with Whisper

Lex Fridman podcast transcriptions: “Lexicap”

🔗 **Lex Fridman podcast:**

- Since 2018
- Originally AI, nowadays wide ranging topics

🔗 **Lexicap:**

- Andrej Karpathy
- Database of all episodes transcribed
- Whisper large

Lexicap: Lex Fridman Podcast Whisper captions

These are transcripts for Lex Fridman episodes. First we get all the episodes in the playlist (ty youtubesearchpython), see their docs.
Then we download the audio for all of them (ty yt-dlp): `yt-dlp -x --audio-format mp3 -o {mp3_file} -- {youtube_video_id}`
Then we transcribe them (ty OpenAI Whisper): `whisper --language en --model large -o {out_dir} -- {mp3_file}`
Download the raw captions data as a zip file here.
Send comments to @karpathy on Twitter.

Episodes:

1 Max Tegmark: Life 3.0 | Lex Fridman Podcast #1
2 Christof Koch: Consciousness | Lex Fridman Podcast #2
3 Steven Pinker: AI in the Age of Reason | Lex Fridman Podcast #3
4 Yoshua Bengio: Deep Learning | Lex Fridman Podcast #4
5 Vladimir Vapnik: Statistical Learning | Lex Fridman Podcast #5
6 Guido van Rossum: Python | Lex Fridman Podcast #6
7 Jeff Atwood: Stack Overflow and Coding Horror | Lex Fridman Podcast #7
8 Eric Schmidt: Google | Lex Fridman Podcast #8
9 Stuart Russell: Long-Term Future of Artificial Intelligence | Lex Fridman Podcast #9
10 Pieter Abbeel: Deep Reinforcement Learning | Lex Fridman Podcast #10

Recap
ASR in a World of Whisper

Given enough high quality data, we can achieve human performance

- Whisper & wav2vec 2.0:
  - Easy to use
  - Easy to finetune
  - wav2vec 2.0 throughput increases with average file length
  - Whisper throughput decreases with average file length
- Whisper:
  - Multitask
  - Better accuracy out-of-the-box
  - Best for English
  - Less GPU memory
- wav2vec 2.0
  - Significantly faster (15x - 40x more throughput)
  - Lower GPU utilization

https://www.linkedin.com/in/becker-lisa/
or search “Lisa Becker ML6”

Resources

- Whisper paper
- wav2vec 2.0 paper
- Whisper & wav2vec 2.0 benchmarking
- Whisper finetuning blog post, leaderboard, winners
- Lex Fridman podcast & transcripts
- Connectionist Temporal Classification blog post