SURF HPML Group 25/06/2024

Speech technology

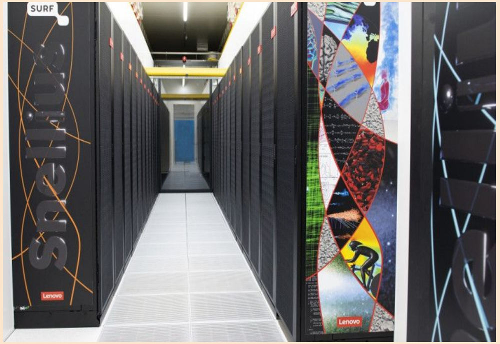
Trends, limitations & future

Vivian van Oijen

SURF

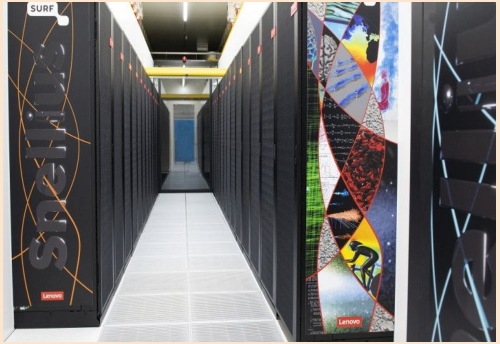
Hello!

Hello!



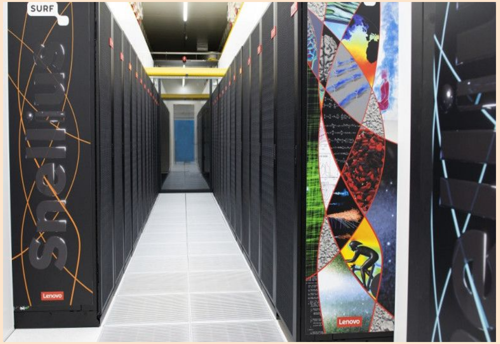
Machine learning
@SURF

Hello!



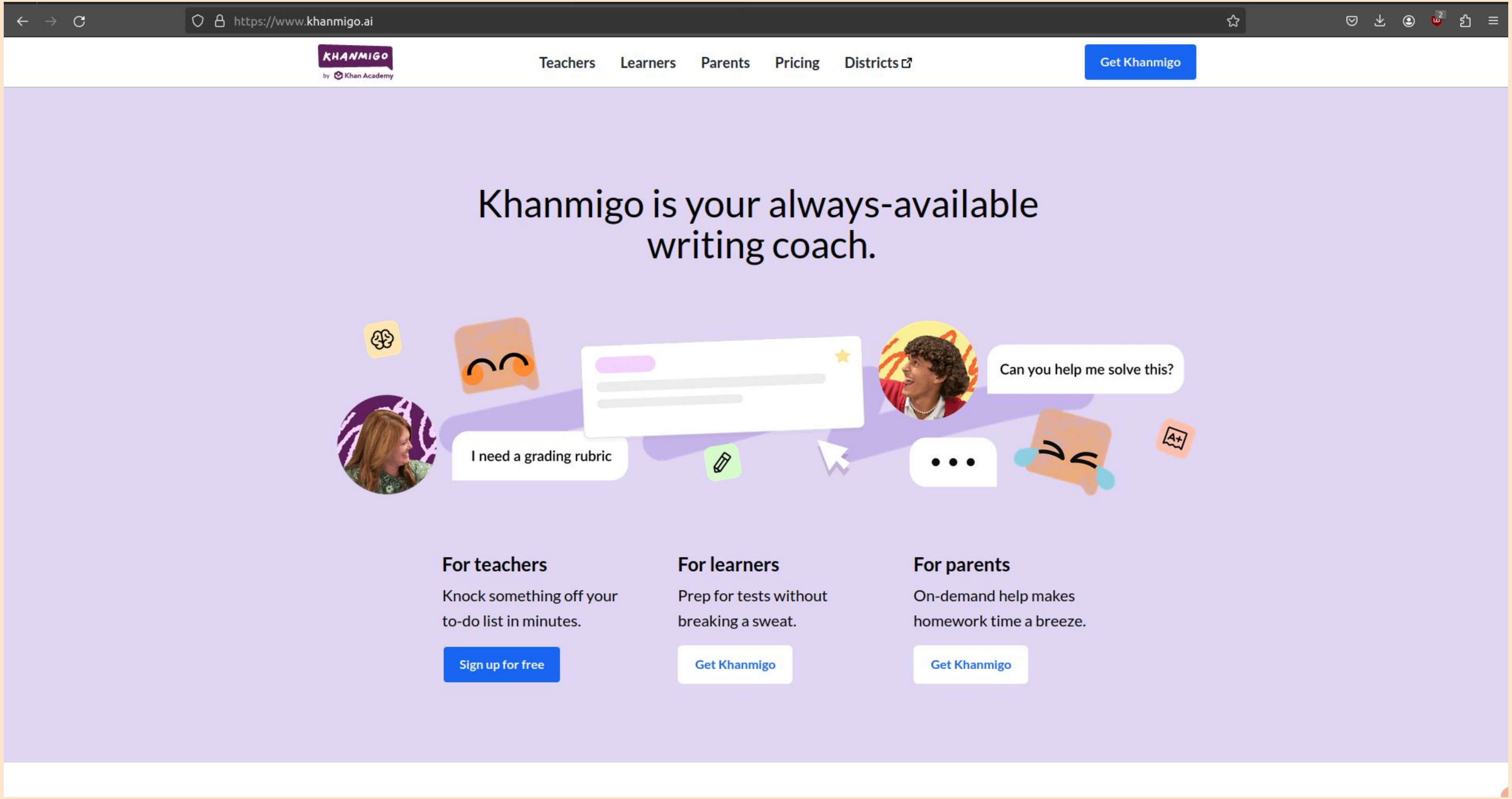
Machine learning
@SURF

Hello!



Machine learning
@SURF

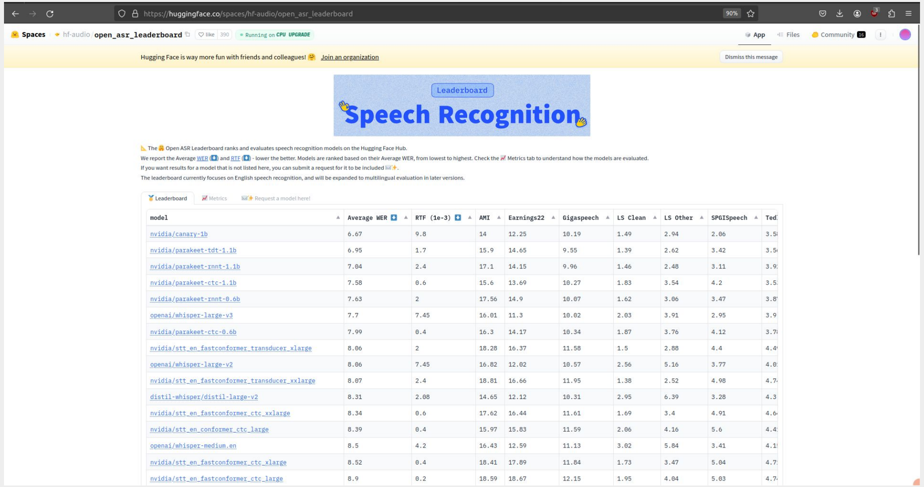
Speech technology


https://www.khanmigo.ai
KHANMIGO
by Khan Academy
Teachers
Learners
Parents
Pricing
Districts
Get Khanmigo
Khanmigo is your always-available writing coach.
I need a grading rubric
Can you help me solve this?
For teachers
Knock something off your to-do list in minutes.
Sign up for free
For learners
Prep for tests without breaking a sweat.
Get Khanmigo
For parents
On-demand help makes homework time a breeze.
Get Khanmigo

State of the Art

State of the Art

Hugging Face Leaderboard



https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

Spaces hf-audio/open_asr_leaderboard like 390 Running on CPU UPGRADE

App Files Community 16

Hugging Face is way more fun with friends and colleagues! Join an organization

Dismiss this message

Leaderboard

Speech Recognition

The Open ASR Leaderboard ranks and evaluates speech recognition models on the Hugging Face Hub.

We report the Average WER and RTF - lower the better. Models are ranked based on their Average WER, from lowest to highest. Check the Metrics tab to understand how the models are evaluated.

If you want results for a model that is not listed here, you can submit a request for it to be included.

The leaderboard currently focuses on English speech recognition, and will be expanded to multilingual evaluation in later versions.

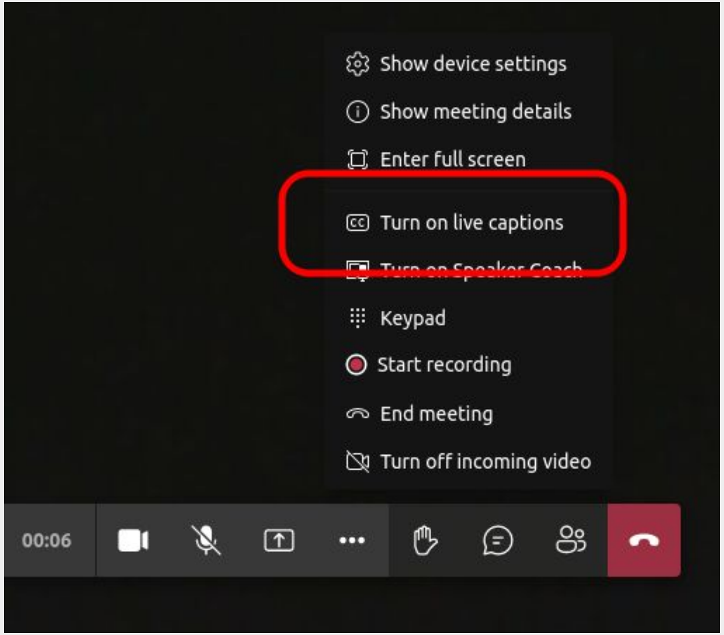
Leaderboard Metrics Request a model here!

model	Average WER	RTF (1e-3)	AMI	Earnings22	Gigaspeech	LS Clean	LS Other	SPGISpeech	Ted
nvidia/canary-1b	6.67	9.8	14	12.25	10.19	1.49	2.94	2.06	3.5
nvidia/parakeet-tdt-1.1b	6.95	1.7	15.9	14.65	9.55	1.39	2.62	3.42	3.5
nvidia/parakeet-rnnt-1.1b	7.04	2.4	17.1	14.15	9.96	1.46	2.48	3.11	3.9
nvidia/parakeet-ctc-1.1b	7.58	0.6	15.6	13.69	10.27	1.83	3.54	4.2	3.5
nvidia/parakeet-rnnt-0.6b	7.63	2	17.56	14.9	10.07	1.62	3.06	3.47	3.8
openai/whisper-large-v3	7.7	7.45	16.01	11.3	10.02	2.03	3.91	2.95	3.9
nvidia/parakeet-ctc-0.6b	7.99	0.4	16.3	14.17	10.34	1.87	3.76	4.12	3.7
nvidia/stt_en_fastconformer_transducer_xlarge	8.06	2	18.28	16.37	11.58	1.5	2.88	4.4	4.4
openai/whisper-large-v2	8.06	7.45	16.82	12.02	10.57	2.56	5.16	3.77	4.0
nvidia/stt_en_fastconformer_transducer_xxlarge	8.07	2.4	18.81	16.66	11.95	1.38	2.52	4.98	4.7
distil-whisper/distil-large-v2	8.31	2.08	14.65	12.12	10.31	2.95	6.39	3.28	4.3
nvidia/stt_en_fastconformer_ctc_xxlarge	8.34	0.6	17.62	16.44	11.61	1.69	3.4	4.91	4.6
nvidia/stt_en_conformer_ctc_large	8.39	0.4	15.97	15.83	11.59	2.06	4.16	5.6	4.4
openai/whisper-medium.en	8.5	4.2	16.43	12.59	11.13	3.02	5.84	3.41	4.1
nvidia/stt_en_fastconformer_ctc_xlarge	8.52	0.4	18.41	17.89	11.84	1.73	3.47	5.04	4.7
nvidia/stt_en_fastconformer_ctc_large	8.9	0.2	18.59	18.67	12.15	1.95	4.04	5.03	4.7

https://huggingface.co/spaces/hf-audio/open_asr_leaderboard

State of the Art

Study case 1: Teams

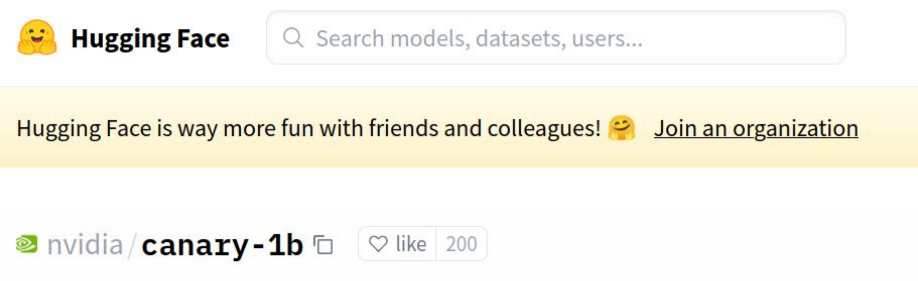


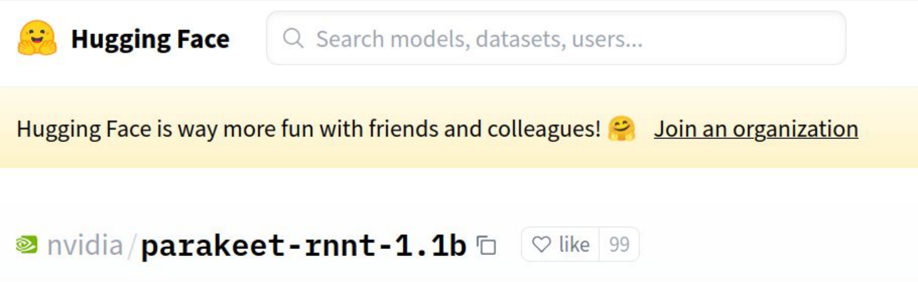
State of the Art

Study case 2: Canary & Parakeet

State of the Art

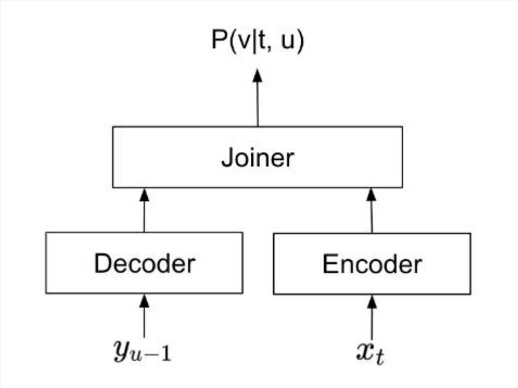
Study case 2: Canary & Parakeet





State of the Art

Study case 2: Canary & Parakeet



How to use Parakeet-TDT

To run speech recognition with Parakeet-TDT, you'll need to install NVIDIA NeMo. It can be installed as a pip package, as shown below. Cython and PyTorch (2.0 and above) should be installed before trying to install NeMo.

```
pip install nemo_toolkit['asr']
```

Once NeMo is installed, you can use Parakeet-TDT to recognize your audio files as follows:

```
import nemo.collections.asr as nemo_asr
asr_model = nemo_asr.models.ASRModel.from_pretrained(model_name="nvidia/parakeet-tdt-1.1b")
transcript = asr_model.transcribe(["some_audio_file.wav"])
```

State of the Art

Study case 3: Whisper

Paper

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford *[1] Jong Wook Kim *[1] Tao Xu [1] Greg Brockman [1] Christine McLeavey [1] Ilya Sutskever [1]

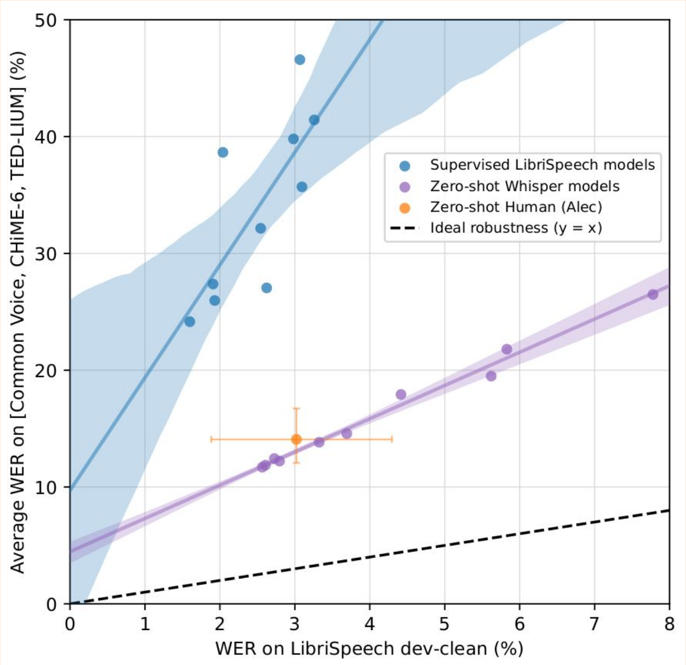
Abstract

We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results but in a zero-shot transfer setting without the need for any fine-tuning. When compared to humans, the models approach their accuracy and robustness. We are releasing models and inference code to serve as a foundation for further work on robust speech processing.

methods are exceedingly adept at finding patterns within a training dataset which boost performance on held-out data from the same dataset. However, some of these patterns are brittle and spurious and don't generalize to other datasets and distributions. In a particularly disturbing example, Radford et al. (2021) documented a 9.2% increase in object classification accuracy when fine-tuning a computer vision model on the ImageNet dataset (Russakovsky et al., 2015) without observing any improvement in average accuracy when classifying the same objects on seven other natural image datasets. A model that achieves "superhuman" performance when trained on a dataset can still make many basic errors when evaluated on another, possibly precisely because it is exploiting those dataset-specific quirks that humans are oblivious to (Geirhos et al., 2020).

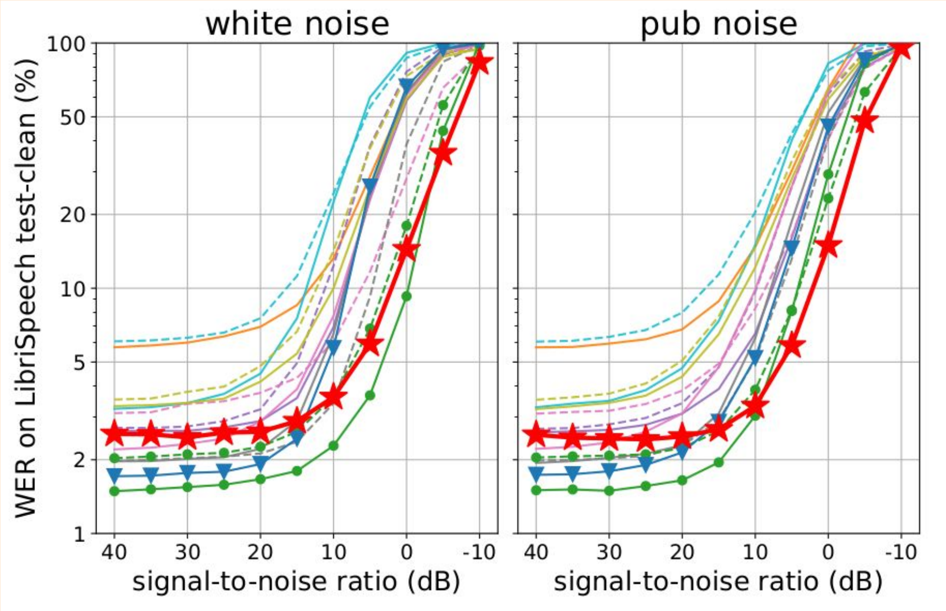
https://arxiv.org/abs/2212.04356

Paper

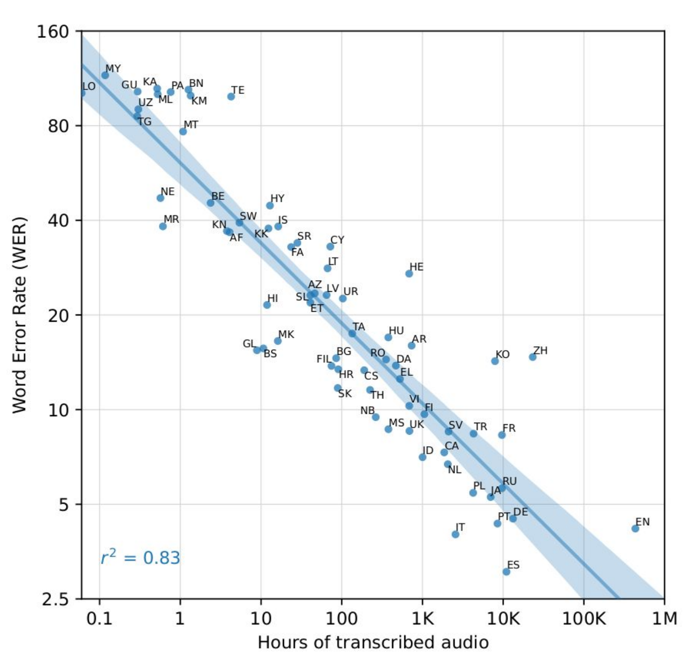


Supervised LibriSpeech models
Zero-shot Whisper models
Zero-shot Human (Alec)
Ideal robustness (y = x)
Average WER on [Common Voice, CHiME-6, TED-LIUM] (%)
WER on LibriSpeech dev-clean (%)
0
10
20
30
40
50
0
1
2
3
4
5
6
7
8

Paper


white noise
pub noise
WER on LibriSpeech test-clean (%)
100
50
20
10
5
2
1
40
30
20
10
0
-10
signal-to-noise ratio (dB)
signal-to-noise ratio (dB)

Paper



160
80
40
20
10
5
2.5
Word Error Rate (WER)
0.1
1
10
100
1K
10K
100K
1M
Hours of transcribed audio
$r^2 = 0.83$
MY
LO
GU
KA
PA
BN
ML
KM
TE
UZ
TG
MT
NE
BE
HY
MR
SW
IS
KN
AF
KK
SR
FA
CY
LT
HE
AZ
LV
UR
SL
ET
HI
TA
HU
MK
AR
GL
BS
BG
RO
DA
FIL
KO
ZH
HR
CS
EL
SK
TH
VI
NB
FI
MS
UK
SV
TR
FR
ID
CA
NL
RU
PL
JA
PT
DE
EN
IT
ES

Paper

Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford [*1] Jong Wook Kim [*1] Tao Xu [1] Greg Brockman [1] Christine McLeavey [1] Ilya Sutskever [1]

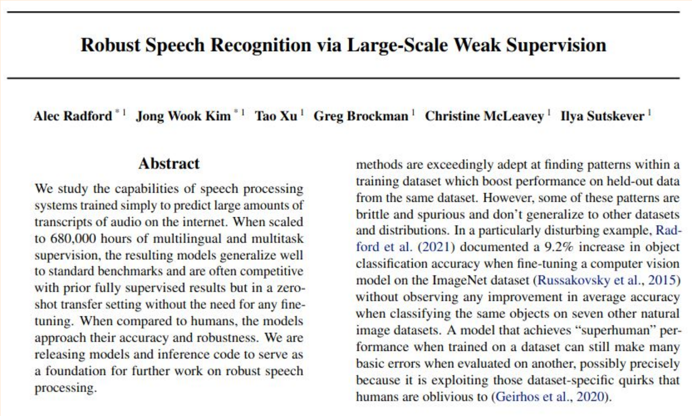
Abstract

We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results but in a zero-shot transfer setting without the need for any fine-tuning. When compared to humans, the models approach their accuracy and robustness. We are releasing models and inference code to serve as a foundation for further work on robust speech processing.

methods are exceedingly adept at finding patterns within a training dataset which boost performance on held-out data from the same dataset. However, some of these patterns are brittle and spurious and don't generalize to other datasets and distributions. In a particularly disturbing example, Radford et al. (2021) documented a 9.2% increase in object classification accuracy when fine-tuning a computer vision model on the ImageNet dataset (Russakovsky et al., 2015) without observing any improvement in average accuracy when classifying the same objects on seven other natural image datasets. A model that achieves "superhuman" performance when trained on a dataset can still make many basic errors when evaluated on another, possibly precisely because it is exploiting those dataset-specific quirks that humans are oblivious to (Geirhos et al., 2020).

https://arxiv.org/abs/2212.04356

Paper



Robust Speech Recognition via Large-Scale Weak Supervision

Alec Radford [*1] Jong Wook Kim [*1] Tao Xu [1] Greg Brockman [1] Christine McLeavey [1] Ilya Sutskever [1]

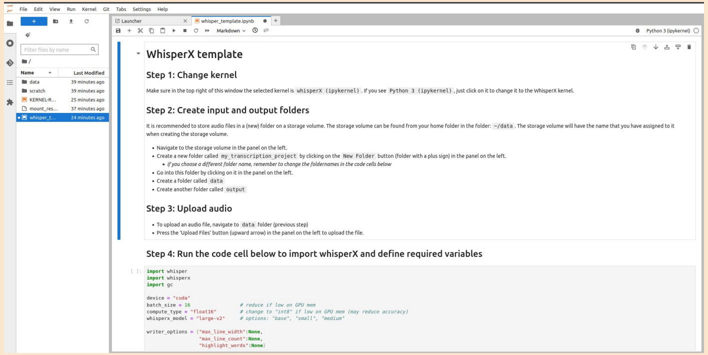
Abstract

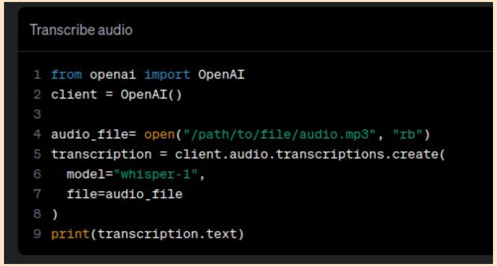
We study the capabilities of speech processing systems trained simply to predict large amounts of transcripts of audio on the internet. When scaled to 680,000 hours of multilingual and multitask supervision, the resulting models generalize well to standard benchmarks and are often competitive with prior fully supervised results but in a zero-shot transfer setting without the need for any fine-tuning. When compared to humans, the models approach their accuracy and robustness. We are releasing models and inference code to serve as a foundation for further work on robust speech processing.

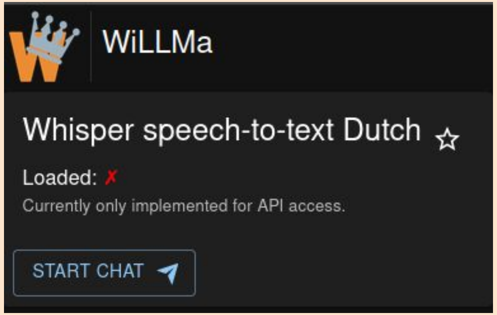
methods are exceedingly adept at finding patterns within a training dataset which boost performance on held-out data from the same dataset. However, some of these patterns are brittle and spurious and don't generalize to other datasets and distributions. In a particularly disturbing example, Radford et al. (2021) documented a 9.2% increase in object classification accuracy when fine-tuning a computer vision model on the ImageNet dataset (Russakovsky et al., 2015) without observing any improvement in average accuracy when classifying the same objects on seven other natural image datasets. A model that achieves "superhuman" performance when trained on a dataset can still make many basic errors when evaluated on another, possibly precisely because it is exploiting those dataset-specific quirks that humans are oblivious to (Geirhos et al., 2020).

https://arxiv.org/abs/2212.04356

"Demo"







SURF

The other way around: text to speech!

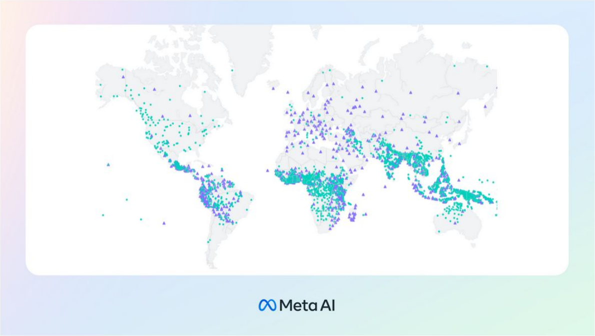
Suno

Bark

tacotron2

Introducing speech-to-text, text-to-speech, and more for 1,100+ languages

May 22, 2023 · 7 minute read



The other way around: text to speech!

```
[viviano@gcn59 Bark]$ python -m bark --text "Hello, my name is Suno and I like pizza [laughs]" --output_filename "bark_output3.wav"
/gpfs/home2/viviano/SpeechTech/Bark/venv/lib/python3.11/site-packages/torch/nn/utils/weight_norm.py:28: UserWarning: torch.nn.utils.weight_norm
 is deprecated in favor of torch.nn.utils.parametrizations.weight_norm.
  warnings.warn("torch.nn.utils.weight_norm is deprecated in favor of torch.nn.utils.parametrizations.weight_norm.")
100%|██████████████████████████████████████████| 316/316 [00:03<00:00, 82.49it/s]
100%|████████████████████████████████████████████| 16/16 [00:08<00:00,  1.90it/s]
Done! Output audio file is saved at: './bark_output3.wav'
```

Deepgram

Nova
Unmatched performance and value
Our next-gen model surpasses all competitors in speed, accuracy, and cost. Compared to the nearest competitor, Nova is 22% more accurate, more than 20 times faster, and over 3x cheaper.

Whisper
Improvements you can't miss
Our fully managed Whisper APIs are faster, more reliable, and cheaper than OpenAI's. Includes built-in diarization, word-level timestamps, and an 80x higher file size limit.

Custom
Boost performance using your data
Custom trained speech models give accuracy a noticeable boost, especially on unique customer jargon. High throughput models are also available to meet enterprise scalability requirements.

SPEECHMATICS

Great AI transcription: The bedrock of value

Accurate transcription. Valuable on its own. And the foundation for so much more.

Speak to sales

IIElevenLabs

Generative Voice AI

Amberscript

Transform your audio and video to text and subtitles

Our cutting-edge generative AI, paired with top-tier language professionals, collaboratively deliver highly accurate solutions tailored to your business needs.

Request a quote | Try it free

Some commercial alternatives…

SURF

Limitations

Limitations

Services

Open models

Limitations

Services

- Little transparency
- Requires sharing your data

Open models

Limitations

Services

- Little transparency
- Requires sharing your data

Open models

- Sometimes less quality
- Less user friendly

Limitations

View on GitHub

Dutch Open Speech Recognition Benchmark

Results of Dutch ASR models, collected by the community

https://opensource-spraakherkenning-nl.github.io/ASR_NL_results/

Future

Future

How do we overcome current limitations

1. **Quality**

   We need data

2. **Infrastructure**

   Trusted party that provides API and UI

Thank you!

Vivian van Oijen