Harnessing YouTube for Evaluating General-Purpose Speech Recognition Machine Learning Models

Tomasz Wojnar, Jarosław Hryszko, Adam Roman

JAGIELLONIAN UNIVERSITY IN KRAKÓW

Agenda

- Why YouTube as data source for speech evaluation?
- Mi-Go toolkit - automating the process
- Used model architectures & our evaluation dataset
- What is the Word Error Rate?
- Results
- Reasons for high WER
- Comparison to other datasets
- Conclusions

Why YouTube as data source for speech evaluation?

The world's audiovisual encyclopedia:

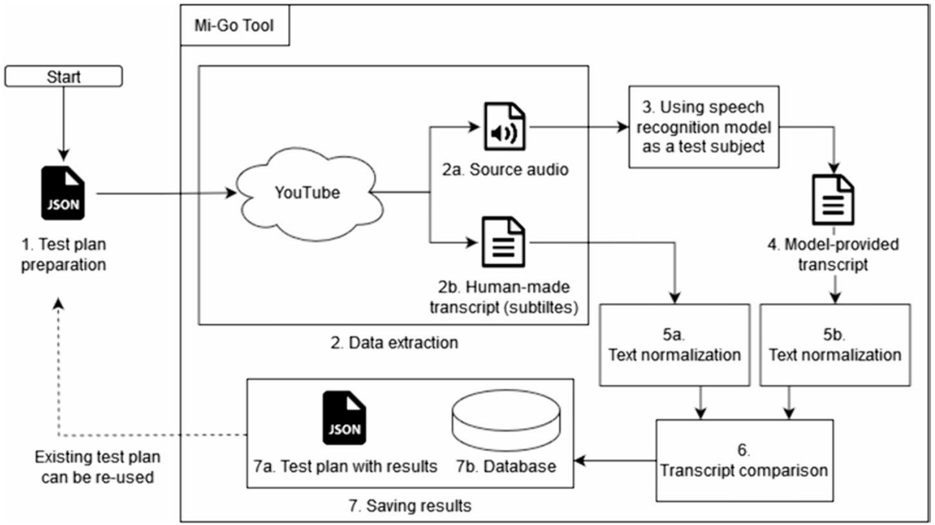
- Content from every corner of the world.
- Covers almost all languages, dialects, and accents.
- Background noises: outdoor events, cafes, concerts.
- Different vocal tones, emotions, and speaking speeds.
- **Large number of human-made transcriptions.**

Opportunity for ethical use:

- Evaluation only, the model does not learn from the data.
- “Fair Use” policy, ability to filter only Creative Common licensed videos.

Mi-Go toolkit - automating the process

- The test plan is created automatically using the YouTube Data API.
- Audio and subtitles are downloaded during testing.
- Both texts are normalised (lower case, removal of punctuation, etc.).
- The results are saved in SQLite for easy access and analysis.

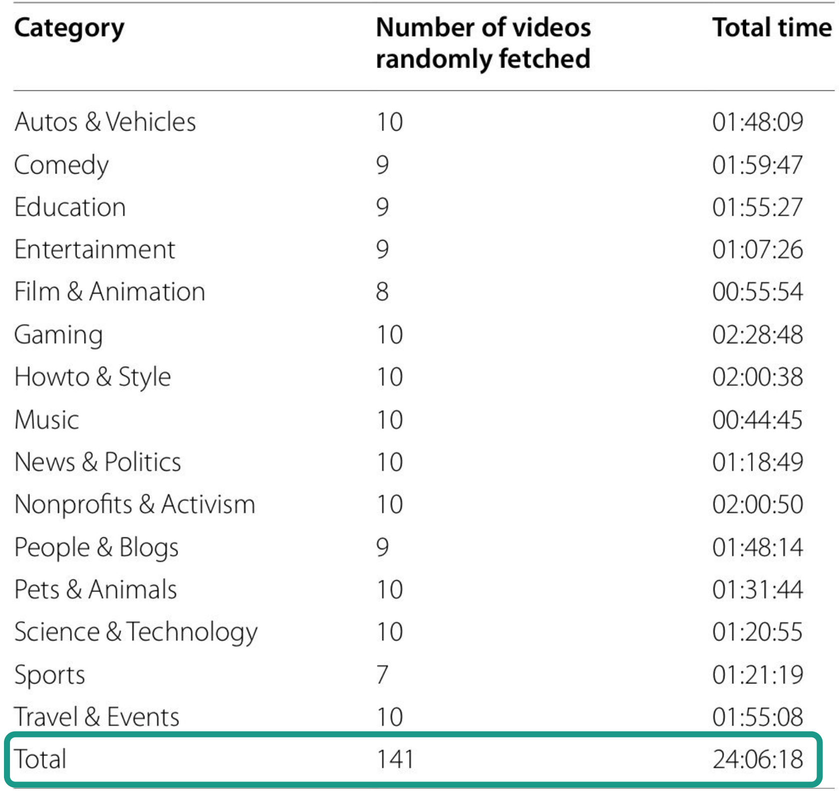


https://github.com/Kowalski1024/Mi-Go

Used model architectures & our evaluation dataset

Used model architectures:

- Whisper (OpenAI)
- Wav2Vec 2.0 (Meta)
- Conformer-Transducer X-Large (NVIDIA NeMo)
- Conformer (ESPnet2)



Category	Number of videos randomly fetched	Total time
Autos & Vehicles	10	01:48:09
Comedy	9	01:59:47
Education	9	01:55:27
Entertainment	9	01:07:26
Film & Animation	8	00:55:54
Gaming	10	02:28:48
Howto & Style	10	02:00:38
Music	10	00:44:45
News & Politics	10	01:18:49
Nonprofits & Activism	10	02:00:50
People & Blogs	9	01:48:14
Pets & Animals	10	01:31:44
Science & Technology	10	01:20:55
Sports	7	01:21:19
Travel & Events	10	01:55:08
Total	141	24:06:18

What is the Word Error Rate?

- **Common metric of the performance of a speech recognition.**
- Computed as:

$$WER = \frac{Substitutions + Deletions + Insertions}{Reference\_Words}$$

- Example

  Ground truth: *I am going to the park*

  Model output: *Now I am going to bark*

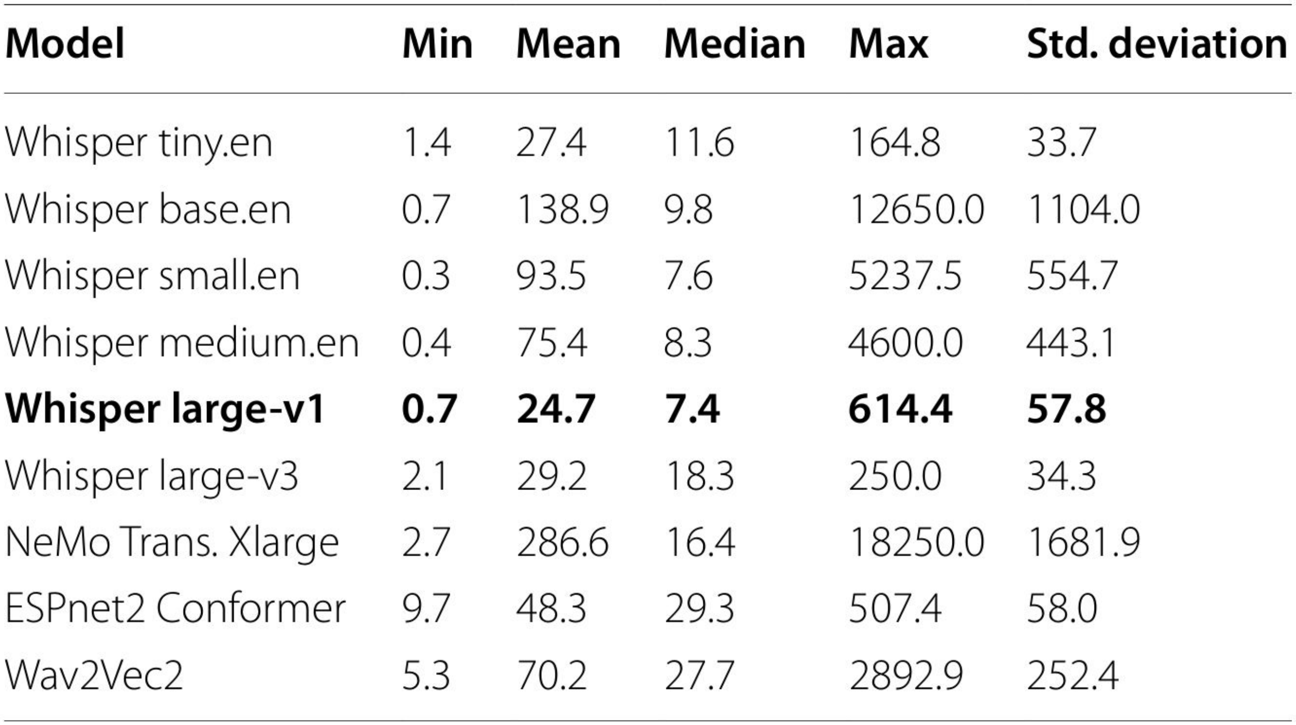
  Differences: *Now I am going to ~~the~~ bark*

Substitutions = 1, Deletions = 1, Insertions = 1, Reference Words = 6

$$WER = \frac{1+1+1}{6} = 0.5 = 50\%$$

Results

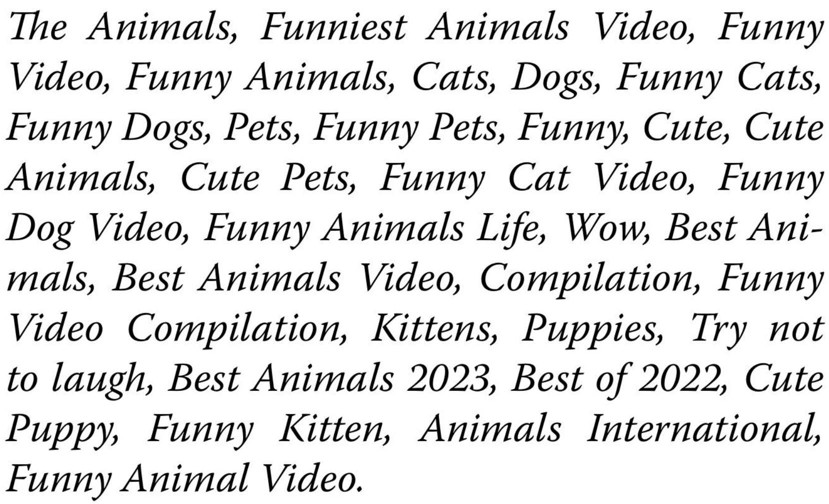
Metric: WER [%]



Model	Min	Mean	Median	Max	Std. deviation
Whisper tiny.en	1.4	27.4	11.6	164.8	33.7
Whisper base.en	0.7	138.9	9.8	12650.0	1104.0
Whisper small.en	0.3	93.5	7.6	5237.5	554.7
Whisper medium.en	0.4	75.4	8.3	4600.0	443.1
Whisper large-v1	**0.7**	**24.7**	**7.4**	**614.4**	**57.8**
Whisper large-v3	2.1	29.2	18.3	250.0	34.3
NeMo Trans. Xlarge	2.7	286.6	16.4	18250.0	1681.9
ESPnet2 Conformer	9.7	48.3	29.3	507.4	58.0
Wav2Vec2	5.3	70.2	27.7	2892.9	252.4

Subtitles added for Search Engine Optimization?

- Search engine optimization - subtitles may be created or modified with the **goal of improving the video's visibility** in search engine results.

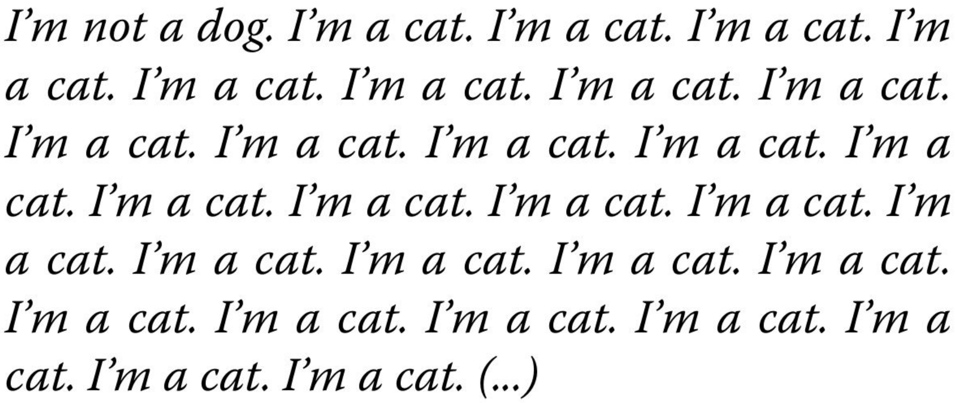


The Animals, Funniest Animals Video, Funny Video, Funny Animals, Cats, Dogs, Funny Cats, Funny Dogs, Pets, Funny Pets, Funny, Cute, Cute Animals, Cute Pets, Funny Cat Video, Funny Dog Video, Funny Animals Life, Wow, Best Animals, Best Animals Video, Compilation, Funny Video Compilation, Kittens, Puppies, Try not to laugh, Best Animals 2023, Best of 2022, Cute Puppy, Funny Kitten, Animals International, Funny Animal Video.

Subtitles from YouTube Video ID: **Jk83I-z6C98**

Sometimes model just hallucinate...

- Transcription errors - mistakes when transcribing speech to text or model **hallucinations**.



I'm not a dog. I'm a cat. (...)

~ Whisper large-v1

Model output from YouTube Video ID: **lCegfmeugdQ**

Comparison to other datasets

Compared to other datasets, **YouTube** does not fare badly and, despite its flaws, **can be considered as a data source** for evaluating speech-to-text models.

Dataset used	WER [%]
TED-LIUM3	3.5
Meanwhile	5.1
YouTube	**7.4**
Kincaid46	8.8
Earnings-21	9.7
Rev16	11.3
Earnings-22	12.6
CORAAL	19.6

Whisper large-v1

Dataset used	WER [%]
LibriSpeech Clean	2.7
LibriSpeech Other	6.2
WSJ	7.7
Tedlium	10.5
Fleurs En	14.6
VoxPopuli En	17.9
Artie	24.5
YouTube	**27.7**
Switchboard	28.3
Common Voice	29.9
CallHome	34.8
CORAAL	35.6
AMI IHM	37.0
CHiME6	65.8
AMI SDM1	67.6

Wav2Vec 2.0

Conclusions

- **YouTube is a very good place to look for Out of Distribution data.**
- The proposed Mi-Go toolkit helps to fully automate the process of evaluating models on the YouTube dataset.
- Among the models tested, Whisper large-v1 was the best.
- Relying on subtitles added to videos by users has its drawbacks.
- Despite its flaws, YouTube can be considered as a dataset for evaluating speech-to-text models.

Thank you for your attention!

in wojnar-tomasz

Link to the paper