EURASIP Journal on Audio, Speech, and Music Processing

SOFTWARE

Open Access

Mi-Go: tool which uses YouTube as data source for evaluating general-purpose speech recognition machine learning models

Tomasz Wojnar[1], Jarosław Hryszko[1*] and Adam Roman[1]

Abstract

This article introduces Mi-Go, a tool aimed at evaluating the performance and adaptability of general-purpose speech recognition machine learning models across diverse real-world scenarios. The tool leverages YouTube as a rich and continuously updated data source, accounting for multiple languages, accents, dialects, speaking styles, and audio quality levels. To demonstrate the effectiveness of the tool, an experiment was conducted, by using Mi-Go to evaluate state-of-the-art automatic speech recognition machine learning models. The evaluation involved a total of 141 randomly selected YouTube videos. The results underscore the utility of YouTube as a valuable data source for evaluation of speech recognition models, ensuring their robustness, accuracy, and adaptability to diverse languages and acoustic conditions. Additionally, by contrasting the machine-generated transcriptions against human-made subtitles, the Mi-Go tool can help pinpoint potential misuse of YouTube subtitles, like search engine optimization.

1 Introduction

Speech recognition has become a critical component in numerous applications, ranging from virtual assistants and transcription services to voice-controlled devices and accessibility tools. The increasing reliance on speech recognition machine learning models necessitates robust and comprehensive evaluation methodologies to ensure their performance, reliability, and adaptability across diverse scenarios.

Existing speech recognition models evaluations often rely on curated datasets, such as LibriSpeech [25], CommonVoice [4], and TIMIT [32]. While these datasets provide a controlled environment for evaluation, they may not capture the full spectrum of real-world scenarios, potentially limiting the model's generalizability. Additionally, these datasets may not be updated frequently, resulting in potential stagnation in performance evaluation.

In this article, we introduce Mi-Go (the name will be explained further), a tool designed to evaluate the prediction performance of general-purpose speech recognition machine learning models. Mi-Go harnesses the power of YouTube as a data source, providing access to a virtually unlimited repository of diverse audio-visual content. YouTube offers a rich and continuously updated collection of spoken language data, encompassing various languages, accents, dialects, speaking styles, and audio quality levels. This makes it an ideal source of data which can be used to evaluate the adaptability and performance of speech recognition models in real-world situations.

In recent years, there has been a growing interest in harnessing the vast amount of data available on platforms such as YouTube for machine learning tasks. Various approaches have been proposed to collect and process data from YouTube, including YouTube-8M [1], AudioSet [11], and GigaSpeech [6]. However, these methods primarily focus on video and audio

*Correspondence:
Jarosław Hryszko
jaroslaw.hryszko@uj.edu.pl
[1] Jagiellonian University, Faculty of Mathematics and Computer Science, Division of Software Engineering, Łojasiewicza 6, Krakow 30-348, Poland

classification tasks rather than the evaluation of speech recognition models.

The landscape of speech recognition technology has witnessed a paradigm shift, driven by rapid advancements in deep learning and artificial intelligence. Groundbreaking architectures, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and, more recently, transformer-based models, have revolutionized this domain, offering unprecedented accuracy in transcribing human speech. These models, trained on vast datasets, have demonstrated remarkable proficiency in navigating the complexities of language, including accents, dialects, and noise interference. The emergence of these models not only underscores the accelerated pace of development in this field but also leads one to believe that in the near future seamless human-computer interaction will become the norm. It should be noted that while these advancements present exciting prospects, they also raise compelling questions concerning data privacy, algorithmic bias, and the digital divide.

In our study, we address this need by proposing——and then empirically investigating the prediction performance of speech recognition model——evaluation tool which utilizes YouTube as a data source, providing access to an extensive and diverse collection of audio samples for evaluation purposes. This approach ensures that the performance assessment remains up-to-date and relevant, capturing the nuances of real-world speech more accurately than curated datasets. To the best of our knowledge, there is little or even no research on using YouTube and video subtitles provided by the YouTube users for speech recognition evaluation. Considering all the above, our goal is to answer the following research question:

(RQ) Will evaluation of the selected speech recognition machine learning model using YouTube as a data source, as made possible by Mi-Go, produce similar results (measured using the same metric) as the evaluation conducted by the model creators?

Mi-Go automates the process of data extraction, annotation, and evaluation from YouTube, ensuring an up-to-date and representative sample for evaluation purposes. By leveraging algorithms for data filtering and annotation, Mi-Go facilitates a thorough and unbiased evaluation of the speech recognition models. Moreover, Mi-Go is designed to be easily adaptable, allowing for seamless integration with variety of different speech recognition solutions, making it a versatile and valuable tool in the speech recognition research community.

The primary motivation behind the development of the Mi-Go tool stems from the recognition of several limitations in existing approaches to evaluate speech recognition models. As speech recognition technology continues to play a critical role in various applications, including voice assistants, transcription services, and accessibility tools, ensuring the robustness and accuracy of these models is crucial.

Other speech recognition model evaluation methods often rely on static, curated datasets which, while useful for establishing a controlled environment, may not fully represent the diversity and complexity of real-world speech scenarios. This can lead to overfitting and limit the model's generalizability, ultimately affecting its performance in real-world applications.

Additionally, as the field of speech recognition rapidly advances, existing evaluation methods may struggle to keep pace with new developments and challenges, potentially hindering the progress of these models. By utilizing YouTube as a data source, Mi-Go aims to overcome these limitations and offers a more comprehensive and dynamic evaluation environment.

Another motivation for the development of Mi-Go is the need for a flexible and adaptable tool capable of accommodating variety of speech recognition models. This adaptability allows researchers and developers to compare and contrast the performance of various models, facilitating the continuous improvement and refinement of speech recognition systems.

By addressing these limitations and providing a dynamic, diverse, and adaptable evaluation tool, Mi-Go aspires to contribute significantly to the field of speech recognition research, driving innovation and fostering the development of highly accurate and robust models for various applications.

In a summary, the Mi-Go tool is a contribution to the scientific and speech recognition community for the following reasons:

- *Rich and diverse test data source.* Mi-Go leverages YouTube, a platform with vast and continuously updated content, to provide a rich source of diverse audio-visual content. This includes various languages, accents, dialects, speaking styles, and audio quality levels. Such diversity is ideal for evaluating the adaptability and performance of speech recognition models in real-world situations, ensuring robustness, accuracy, and adaptability to diverse languages and acoustic conditions.
- *Dynamic evaluation environment.* By using YouTube as a data source, Mi-Go addresses limitations of previous approaches that often relied on static and potentially outdated datasets. It offers a more com-

prehensive and dynamic evaluation environment that reflects current real-world scenarios. This adaptability allows for the comparison of various models and facilitates the continuous improvement and refinement of speech recognition systems.

– *Practical and theoretical contributions*. The experimental results obtained through Mi-Go highlight the utility of YouTube as a valuable data source for the evaluation of speech recognition models. This not only underscores the platform's potential in enhancing model robustness and adaptability but also contributes to the academic discourse by providing a novel methodology for speech recognition research. Additionally, Mi-Go's approach to contrasting machine-generated transcriptions against human-made subtitles offers insights into potential misuse of subtitles, such as for search engine optimization purposes, thereby adding a layer of practical utility in detecting transcription anomalies.

2 YouTube as a data source for speech recognition model evaluation

With over 2 billion monthly active users and a diverse array of content uploaded every day, YouTube offers a rich resource for researchers and developers working on speech recognition technology. By tapping into this wealth of multilingual and multi-genre content, it is possible to evaluate and refine speech recognition models across various languages, dialects, and acoustic environments.

A vast digital archive. YouTube stands as a colossal repository of digital content, presenting an unparalleled resource for research across various disciplines. As the world's largest video sharing platform, it hosts an estimated billions of videos, a number that continues to grow exponentially with about 500 hours of new content uploaded every minute. Exact number of hosted videos is not known, but is estimated for not less than 2.5 billion of videos [3]. The number of YouTube "Shorts" videos only, identified through the usage of the hashtag #shorts, reaches approximately 828 million in February 2024[1].

Diversity of content. YouTube's vast library of user-generated content covers an extensive range of topics, languages, and styles. This diversity enables the evaluation of speech recognition models in real-world scenarios, such as noisy environments, various accents, and even low-quality audio recordings. By evaluating models on such a diverse dataset, researchers can identify potential weaknesses and areas for improvement, ultimately resulting in more robust and accurate speech recognition systems.

Multilingual corpus. One of the key advantages of using YouTube for speech recognition model evaluation is the platform's multilingual nature. Videos on the site are available in numerous languages, allowing for the assessment of models' performance across different linguistic settings. This multilingual corpus is invaluable for developing models that can handle a variety of languages, accents, and dialects, thereby expanding their utility and applicability.

Availability of human-generated transcripts. Many YouTube videos come with human-generated subtitles, either provided by content creators or contributed by users through the platform's community contributions feature. These transcripts serve as valuable ground-truth data for evaluating speech recognition models, as they offer a reliable source of comparison for the models' output. By comparing model-generated transcriptions with human-generated ones, researchers can assess the accuracy and performance of their models, identifying areas where improvements are needed.

Potential for continuous model improvement. The ever-growing volume of content on YouTube presents an opportunity for continuous improvement and adaptation of speech recognition models. As new videos are uploaded, models can be re-evaluated and fine-tuned to ensure they remain up-to-date and effective in an ever-changing linguistic landscape. This continuous feedback loop helps researchers identify trends, challenges, and emerging language patterns, which can be incorporated into model updates.

YouTube is an invaluable platform for speech recognition model evaluation due to its diverse, multilingual content and the availability of human-generated transcripts. By leveraging this vast resource, researchers and developers can evaluate and refine their models, ensuring they are robust, accurate, and adaptable to a variety of languages and acoustic conditions.

3 Related work

Studies leveraging YouTube in the area of automatic speech recognition have made significant strides across various facets of the field. These investigations utilize YouTube's extensive library of videos to create datasets, improve speech recognition systems, and explore new approaches to automatic speech recognition, showcasing the platform's value in advancing speech recognition technology research. Key insights from these works include:

[1] Number of #shorts marked videos can be checked in the top left-hand corner of the page: https://www.youtube.com/hashtag/shorts

- *Datasets for automatic speech recognition models creation*. Researchers have developed methodologies for creating databases for audio/visual speech recognition using YouTube videos, such as the comprehensive Spanish dataset by Córdova Esparza et al [7]. In their work, researchers presented a novel approach for creating an audio/visual speech recognition database, particularly addressing the scarcity of datasets in languages other than English, with a focus on Spanish. By selecting hundreds of YouTube videos, the researchers were able to extract facial features and align voice with text with millisecond accuracy, creating a dataset of over 100,000 samples. That methodology not only facilitated the development of automatic speech recognition systems in underrepresented languages but also provided a blueprint for creating datasets in any language by selecting appropriate YouTube content. Takamichi et al. [29] contributed to the diversification of automatic speech recognition research resources through the JTubeSpeech corpus, which consists of Japanese speech collected from YouTube. This corpus was designed for both speech recognition and speaker verification tasks, addressing the need for comprehensive datasets in Japanese for training and evaluating automatic speech recognition systems. The corpus's creation from YouTube videos ensured a variety of speech contexts and speaker demographics, enhancing the robustness of automatic speech recognition models trained on it. Lakomkin et al. [20] developed the KT-speech-crawler, an automated tool for constructing speech recognition datasets from YouTube videos. This tool leveraged automatic captioning provided by YouTube to generate datasets, significantly reducing the manual effort required in dataset creation and enabling researchers to easily compile large-scale datasets tailored to specific speech recognition research needs. Latest work in the field——creation of *Yodas*, a YouTube-derrived Dataset, by Li et al. [22], showcases the ongoing efforts to harness YouTube content as diverse and comprehensive training data resource for developing new, robust speech recognition models. By compiling a diverse set of audio and speech samples from YouTube, *Yodas* aims to provide a versatile dataset that supports a wide range of automatic speech recognition tasks, including dialect and accent recognition, speech-to-text conversion, and speaker verification.
- *Improvement of automatic speech recognition systems*. Liao et al. [23], from Google, explored usage of new large scale deep neural network acoustic modeling for using in YouTube video transcription. By leveraging the massive amount of unlabeled audiovisual content on YouTube, the researchers were able to enhance the modeling process, by using video transcripts uploaded by YouTube users and thus demonstrating the potential of semi-supervised learning approaches in improving automatic speech recognition systems' performance, especially in noisy and challenging acoustic environments. Their findings then were used in actual YouTube automatic speech transcription improvements.
- *Audio-visual speech recognition*. In their work, Serdyuk et al. [28] delved into the enhancement of automatic speech recognition by incorporating video content from YouTube, a novel approach that significantly improved speech recognition accuracy. That study leveraged a large corpus of YouTube videos to train models, focusing on how the visual modality, particularly the movement of the speaker's mouth, could augment audio features for speech recognition tasks. By replacing traditional 3D convolutional neural networks with a video transformer to extract visual features, Serdyuk and his team demonstrated a substantial improvement in word error rates on both a labeled subset of YouTube videos and the LRS3-TED public corpus (described in [2]). Their methodology highlighted the potential of utilizing video content alongside audio data to advance the capabilities of automatic speech recognition systems. This research not only showcased the importance of YouTube as a rich data source for speech recognition technologies but also opened new pathways for enhancing speech recognition accuracy by integrating audio-visual data, paving the way for more sophisticated and efficient automatic speech recognition systems.
- *Bias and inclusivity in automatic speech recognition*. Koenecke et al. [18] uncovered significant racial disparities in the performance of commercial automatic speech recognition systems, including those developed by major tech companies. By analyzing speech from white and African American speakers, the study revealed a higher word error rate for African American speakers, highlighting a critical area for improvement in making automatic speech recognition technologies more inclusive and equitable. Tatman and Kasten [30] investigated the effects of talker dialect, gender, and race on the accuracy of Bing Speech and YouTube automatic captions. Their findings emphasized the impact of sociolinguistic factors on automatic speech recognition accuracy, urging the development of more sophisticated models that could better accommodate the diversity of human speech.

– *Utilizing YouTube as automatic speech recognition tool.* Kim et al. [17] embarked on an insightful exploration into the capabilities of automatic speech recognition tools by utilizing YouTube's automatic transcription service as a benchmark for automatic speech recognition accuracy. In their study, they meticulously compared manual transcriptions with those generated automatically by YouTube, alongside other leading speech recognition platforms such as Google Cloud, IBM Watson, Microsoft Azure, and Trint. Their analysis provided a comprehensive evaluation of the relative performance of these services, with a particular focus on YouTube's efficacy in providing accurate transcriptions. This approach not only highlighted YouTube's potential as an accessible and effective tool for automatic speech recognition but also contributed to the broader discourse on the reliability and accuracy of free, platform-based speech recognition services. Through their comparative study, Kim et al. shed light on the strengths and limitations of YouTube's transcription capabilities, offering valuable insights for researchers, developers, and users seeking to leverage automatic speech recognition technology in various contexts.

These studies illustrate the extensive use of YouTube as a rich data source for automatic speech recognition research, ranging from training dataset creation to addressing biases and inclusivity in speech technologies. However, to the best of our knowledge, there is no work describing the direct use of YouTube to evaluate the functional performance of the existing machine learning models used for automatic speech recognition.

4 Mi-Go Tool

Mi-Go was written in Python programming language. Its source code is available for download under Apache-2.0 license at the following address: https://github.com/Kowalski1024/Mi-Go

In the following, we will describe the tool by focusing on the subsequent operations of the tool – from launching it to saving the evaluation results of the selected speech recognition model.

4.1 Test Plan preparation

To start working with the tool, we need a file in JSON format, called a Test Plan. This is illustrated as number 1 in Fig. 1. In a special circumstances, Test Plan file can be manually written, but it is more efficient to generate it, using an additional script named the *Test Plan Generator*. This script queries YouTube's API to compile a random list of videos, basing on the command line parameters specifying the category of the videos, language, duration,

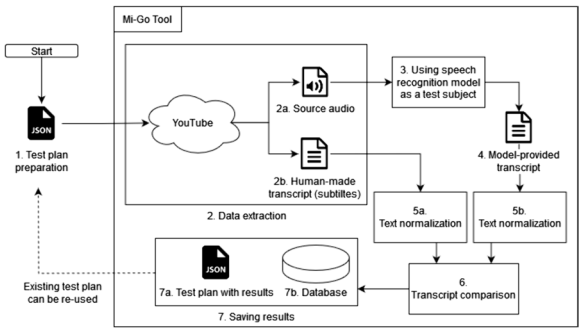

Fig. 1 Mi-Go and speech recognition model evaluation phases (described in text)

and desired quantity of list items (details can be found in Appendix 1). It is essential that the YouTube clearly indicates, that video has human-made subtitles, and only such videos are considered. To query the API, Test Plan Generator uses external Python library called *youtube-transcript-api*[2]. After querying the API, the Test Plan file contains all the necessary metadata about the videos being used in further evaluation and it also stores information about the selected parameters and token for YouTube Data API, which can be used in next test iterations, if needed.

4.2 Data extraction and transcription

In the next step, marked with number 2 in Fig. 1, Mi-Go reads the Test Plan and, basing on that plan, downloads from YouTube the audio track of each video from the plan and the subtitles for that video. Thus, for each video, we have a pair consisting of an audio file (2a) in and human-generated subtitles (marked as 2b).

In the next step (number 3 in Fig. 1), a speech recognition model is employed to convert the downloaded audio into a textual transcript. It is done by the *TranscriptTest* component that executes the speech recognition machine learning model against audio data collected from YouTube. That component can be adjusted for specified speech recognition model by extending that component with model-specific code. It allows to use different models from popular "Hugging Face" machine learning models repository[3] as well as models dedicated for such toolkits like ESPnet or NeMo.

To eliminate inconsequential textual differences, both the subtitles downloaded from YouTube (number 2b in Fig. 1) and those generated by the speech recognition

[2] Available from https://pypi.org/project/youtube-transcript-api/, access: 2024.03.13

[3] Refer to https://huggingface.co/docs/hub/repositories, access: 2024.03.14

model (4) undergo a normalization process (5a and 5b) using an OpenAI's normalization function[4].

4.3 Evaluation and metrics

Speech recognition model evaluation involves comparing the human-made subtitles downloaded from YouTube and those generated by the model (number 6 in Fig. 1). For that evaluation, Mi-Go tool uses a open-source JiWER library[5] to calculate Word Error Rate (WER) measure [27]. WER is a common metric used to assess the performance of speech recognition systems, automatic translation systems, and other tasks involving transcription or translation. It is calculated by determining the minimum number of operations needed to transform the system output into the correct output. These operations include (see Eq. 1): word insertions I, word deletions D, and word substitutions S. To compute the WER, the total number of these operations is divided by the total number of words in the correct output N (in our case: total number of words in subtitles attached to a particular YouTube video), yielding a ratio that represents the rate of errors per word. The lower the WER, the better the performance of the system, as it means fewer errors were made.

$$\mathrm{WER} = \frac{S + D + I}{N} \cdot 100\% \tag{1}$$

The concept of WER has been part of the field of automatic speech recognition and computational linguistics for many years. It is based on the Levenshtein distance or edit distance, a string metric for measuring the difference between two sequences, introduced by Vladimir Levenshtein in 1965 [21]. The exact individual or group that first applied this concept specifically as Word Error Rate in speech recognition or translation systems is not clearly documented. It likely emerged from the academic and industry communities working on speech and language processing technologies. WER has since become a standard measure in these fields. In some cases, WER is expressed as a percentage (by multiplying the original formula by 100%), especially when easy understanding of the measure is a main concern.

The comparison results are stored both in the SQLite database (7b in Fig. 1) and directly in the previously used Test Plan file (7a). Such a Test Plan file, with its evaluation results recorded, can be reused for subsequent evaluation iterations, for instance, to augment results not previously gained, or to retest the same videos, specified within it. Such dual storage approach (database and Test Plan file) facilitates simple access, filtering, and analysis of the evaluation results.

5 Experimental setup

Here, we describe an experimental setup that leverages the Mi-Go tool to use YouTube videos, across all categories, as a data to evaluate speech recognition models by comparing their output with human-made transcripts. The purpose of the experiment is to confirm whether the following setup (Mi-Go and YouTube as evaluation data source) will allow to evaluate the speech recognition models and obtain evaluation results similar to those obtained by the model creators.

5.1 Machine learning models used in the experiment

5.1.1 OpenAI's Whisper

OpenAI, a company most notably recognized for its contribution to the field of artificial intelligence through the development of advanced large language models like GPT-3 and GPT-4, also developed state-of-the-art, general-purpose speech recognition models, which demonstrate exceptional performance in various applications, called Whisper [27].

Due to proven outstanding performance of that model family, as well as the fact that it has been made available under a open-source MIT Licence, we decided that, in our experiment, we will mainly focus on evaluation of the Whisper models. At this point, we should explain that the name "Mi-Go" comes from a novella by H.P. Lovecraft called "The Whisperer in Darkness"; thus, in our opinion, it would make a good name for the tool initially created to evaluate the Whisper models.

The model is based on a Transformer sequence-to-sequence architecture and is trained on a range of speech processing tasks, including multilingual speech recognition, speech translation, spoken language identification, and voice activity detection. These tasks are collectively represented as a sequence of tokens to be predicted by the decoder, enabling a single model to supplant multiple stages of a conventional speech processing pipeline. The multitask training approach employs a series of unique tokens that act as task specifiers or classification targets [27].

Whisper model is available in five different sizes. Four of them (tiny, base, small, medium) having additional English-only versions, which——according to the creators—— perform better when used in English-only applications [16]. Thus, in our research, we decided to use English-only model versions. The "large" model was improved twice; thus, in our experiment, we used two versions of "large" model——initial version, marked as "Whisper large-v1" and latest version, marked as

[4] Refer to https://github.com/openai/whisper/blob/main/whisper/normalizers/english.py, access: 2024.03.14

[5] Available from https://github.com/jitsi/jiwer, access: 2024.03.14

Table 1 Comparison of Whisper models [16]

Name of the model	Number of parameters	Layers	Relative speed	Required VRAM
tiny.en	39,000,000	4	~ 32x	~ 1 GB
base.en	74,000,000	6	~ 16x	~ 1 GB
small.en	244,000,000	12	~ 6x	~ 2 GB
medium.en	769,000,000	24	~ 2x	~ 5 GB
large (all versions)	1,550,000,000	32	~ 1x	~ 10 GB

"Whisper large-v3." Each model offers a balance between speed and accuracy. The names of the used models, their approximate memory requirements and relative speeds are provided in Table 1.

5.1.2 NVIDIA's Conformer-Transducer X-Large

To prove that Mi-Go can be used for evaluation of different speech recognition models, apart from OpenAI's Whisper, in our experiment, we also included models provided by other companies, like one developed by NVIDIA, built upon the Conformer-Transducer architecture, which blends the strengths of transformer and convolutional neural network architectures [13]. The "X-Large" variant of this model signifies its substantial size and capacity, enabling it to process and understand complex audio inputs with higher accuracy compared to its predecessors. It is distributed on Creative Commons BY 4.0 license [24].

When comparing the Conformer-Transducer X-Large model to OpenAI's Whisper model, there are several key points of differentiation. The Whisper model, as we stated before, is based on a different architectural approach, primarily leveraging transformer neural networks. While both models aim to provide high accuracy in speech-to-text conversion, the NVIDIA model's use of the Conformer-Transducer architecture may offer advantages in handling real-time or streaming audio applications. Additionally, the specific design choices in the NVIDIA model might result in better performance in certain scenarios, such as dealing with background noise or low-quality audio inputs [8].

Conformer-Transducer X-Large model is primarily used by NVIDIA in their open-source NeMo toolkit——designed to simplify the process of building, training, and fine-tuning complex neural network models, particularly for speech and natural language processing tasks [19]. To indicate this fact, as well as to use shorter name, in the following text, we will refer to the model as "NeMo Transducer Xlarge."

5.1.3 ESPnet2 model

Similarly to NeMo, ESPnet2 (End-to-End Speech Processing Toolkit, version 2) is an open-source (using Apache 2.0 license) software toolkit designed for speech processing tasks, including automatic speech recognition, text-to-speech, and language modeling. Key features of ESPnet2 include its support for state-of-the-art machine learning models, its flexibility in handling different types of neural network architectures, and its comprehensive set of tools for training, evaluating, and fine-tuning models. ESPnet2 is widely used in the academic and research community for experimenting with novel ideas in speech processing and for developing systems that are more efficient and accurate in real-world applications [15].

Among different speech recognition models available for ESPnet2 toolkit, we chose one of the models trained by Shinji Watanabe, shortly called in this work as "ESPnet2 Conformer[6]" to use it as a reference point for Whisper models evaluation in our experiment. Selection of this particular model was motivated by fact that it was used with success in official ESPnet2 demonstration material [31].

5.1.4 Facebook's wav2vec2-base-960h

Facebook's Wav2Vec 2.0 is an advanced neural network-based framework for speech recognition developed by Facebook AI researchers. It employs a self-supervised learning approach where the model is initially trained on a 53,000 of hours of unlabeled audio [5]. This pre-training allows the model to learn representations of speech from the raw audio itself. Once pre-trained, Wav2Vec 2.0-derived models can be fine-tuned with a smaller amount of labeled data to achieve high performance in transcribing speech. Model selected to be used in our experiment——"wav2vec2-base-960h"——was fine-tuned on 960 h of LibriSpeech [25] dataset on 16 kHz sampled speech audio.

5.2 Data collection and preparation

To begin the experiment, we instruct the *TestPlan Generator*, a component of the Mi-Go tool, via command line interface, to randomly fetch 7–10 videos per category

[6] Mentioned model under its real name is available on https://zenodo.org/records/4585558, access: 2023.12.17

Table 2 YouTube videos categories considered in the experiment

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

Table 3 Word Error Rate [%] value statistics for all evaluated model versions

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

listed in Table 2. What is important, we decided to use such number of videos basing only on available computing resources; number of videos used for evaluation is not restricted and can be freely set by other Mi-Go users.

These videos are *randomly* selected, but basing on factors such as popularity, relevance, and the presence of human-generated subtitles, ensuring a diverse and high-quality dataset. The YouTube Data API is used to acquire the videos, while the *youtube-transcript-api* library retrieves their corresponding transcripts. Already fetched, the same set of videos is used to evaluate selected automatic speech recognition models (presented in Section 5.1). Full list of 141 videos used in experiment is provided in Appendix 4.

6 Results

To answer the research question, we used the proposed Mi-Go tool, to utilize 141 YouTube videos, representing all categories listed in Table 2, to evaluate selected automatic speech recognition models (as presented in Section 5.1) and collect Word Error Rate (WER) metrics as a result.

Statistics for collected Word Error Rate values for all evaluated models are presented in Table 3 and illustrated in Fig. 2. Detailed statistics of the WER value for each model by category are presented in Appendix 3. Results for different datasets compared to our YouTube-based results are gathered in Appendix 2.

Whisper model characteristics, published by its authors [27], concern only "large-v1" model——thus, in Table 3, we presented WER statistics for that model with bold font.

As we can see, the median for "large-v1" model evaluation results is WER = 7.4%. The worst median of results for Whisper "large-v1" model presented by its creators was 19.6% (see Table 4 in Appendix 2). That result was achieved by using CORAAL speech recording dataset

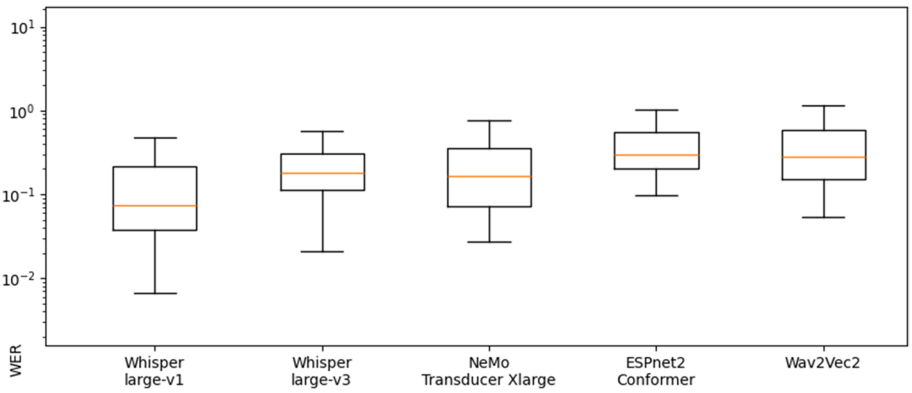

Fig. 2 Box plot of experiment results. Note the logarithmic scale

popularized by Gunter et al. [14]. Other datasets used to validate models by its creators were as follows: recordings of earning calls by Del Rio et al. [9], sets of recordings of online blogs and podcasts, and dataset containing recordings of The Late Show (sic!). Whisper "large-v1" model evaluation results from [27] compared to our results are presented in Table 4 in Appendix 2. By making this comparison, we can conclude that Whisper model evaluation, described in this work, produce similar results as the tests conducted by the Whisper model creators, using different data. Similarly, our results for ESPnet2 Conformer and wav2rec models are similar to those of other authors, achieved using different datasets (Tables 5 and 7 in Appendix 2). Low WER median of YouTube-based result of the Conformer-Transducer model compared to the results of other authors (Table 6 in Appendix 2) can be explained by the occurrence of the highest WER value for this model (18250%) due to fact that the model refused to transcribe the music video "All I Want For Christmas Is You" by Mariah Carey (other models did fine) possibly because of a model's failure[7].

New version of Whisper model——"large-v3"——resulted with worse WER median than "large-v1" version. However, at the same time, "large-v3" resulted with much lower——when compared to "large-v1"——maximum WER value and standard deviation. Thus, we can interpret that result as indication of higher stability of "large-v3" outcomes when compared to older Whisper model version.

One can find large WER values for selected results, significantly different from the median. However, by reviewing the YouTube videos that were used for the tests that ended with high WER values, we can conclude that the reason for this is not due to a malfunction of the Mi-Go tool or speech recognition model. Instead, the high WER values are due to the actual discrepancy between the human-made subtitles attached to the video and those generated by the model. We found that such discrepancies occur due to several reasons:

1. *Transcription errors*. Humans, despite their proficiency, are not infallible and may make mistakes when transcribing speech to text. This could involve mishearing words or phrases, particularly in a noisy environment, during rapid speech, or when dealing with dialectal variations or accents. On the other hand, automatic speech recognition models can "hallucinate" under certain conditions, causing high WER values. For example, in our experiment, for one video containing little speech[8], Whisper "large-v1" model returned such transcription:

   I'm not a dog. I'm a cat. (...)
2. *Interpretation differences*. Subtitling is not always a direct one-to-one transcription process. The transcriber's understanding and interpretation of the speech can influence the outcome. Homonyms, idiomatic expressions, cultural references, or ambiguous statements can all be interpreted differently depending on the transcriber's knowledge and perspective.
3. *Contextual adaptations*. Subtitle makers often make deliberate changes to the text for various reasons. They may simplify or clarify speech to make it more accessible to the audience, especially if the speech is complex or jargon-filled. They may also modify the text to match reading speed constraints, given that text must be readable within the time it is displayed. Cultural adaptations may also be made to make the content more comprehensible to a specific audience (as a form of video's localization).
4. *Descriptive transcriptions*. Some transcriptions go beyond the spoken content and provide descriptions of the visual elements in the video. These are often intended for visually impaired or blind viewers to provide them a more comprehensive understanding of the video content. Such case occurred with video resulted with second highest WER value in our experiment (WER = 12650%). While that video only consist of animal sounds, actual subtitles are as follows (original spelling)[9]:

   Cats Cats are very cute animals Animals that are close and affectionate with people Cat breed is a species with relatively high fertility, giving birth to 2-3 litters of kittens a year New born kittens only weighs about 100g and fits easily in the palm of your hand Horses are smart, wise animals Mother horses as young as 3 years old can start breeding (...)
5. *Search engine optimization (SEO)*. Some subtitles may be created or modified with the goal of improving the video's visibility in search engine results. The inclusion of relevant keywords and phrases can make

[7] Other explanation is that the model does not simply feel so-called "christmas spirit"

[8] https://www.youtube.com/watch?v=lCegfmeugdQ, access: 2024.03.14

[9] https://www.youtube.com/watch?v=4Co4mDeCIJ4, access: 2024.03.14

the video more likely to appear in search results related to those terms, hence enhancing the video's discoverability. Here is an example of such subtitles from one of the fetched videos[10]:

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.

By comparing model-made transcription to the existing human-made subtitles, discrepancies can be identified. Factors such as background noise, speaker accents, or low-quality audio can impact the model's performance. Hence, although speech recognition models can help identify potential inaccuracies in subtitles, a degree of human oversight and validation is typically necessary to confirm and rectify these inaccuracies. From a different perspective, automated setup which utilizes Mi-Go and selected speech recognition model, can significantly help in detection of video subtitles misuse.

7 Conclusions and future work

In this paper, we have introduced Mi-Go, a lightweight and flexible tool for evaluating general-purpose speech recognition models and using YouTube's vast and diverse content. Traditional evaluation methods, which employ curated datasets, may not capture the broad array of real-world scenarios, hence potentially limiting a model's generalizability. Mi-Go, by leveraging YouTube's dynamic content, offers an enriched platform for evaluating such models. An experiment was conducted, using randomly fetched 141 YouTube videos, demonstrating the usefulness of the Mi-Go tool in evaluation of model prediction performance and identification of discrepancies between model-generated transcriptions and human-made subtitles. The results underscore the necessity for human oversight in rectifying inaccuracies and the potential of the Mi-Go tool for enhancing speech recognition models' robustness and adaptability.

While the Mi-Go tool demonstrates promising results in evaluating speech recognition models, several avenues for future work can further enhance its capabilities:

1. Expanding the tool to accommodate other data sources (like non-English YouTube videos or video hosting services other than YouTube), providing an even more diverse and representative set of audio samples for evaluation
2. Incorporating advanced techniques for data preprocessing and augmentation, which can help in simulating various real-world challenges, such as background noise and audio distortions
3. Developing a graphical user interface and API, making it easier for researchers and developers to integrate and utilize the Mi-Go tool in their projects
4. Extending the tool to support other tasks, such as speaker identification evaluation and language identification evaluation, in addition to automatic speech recognition evaluation

An important area for further work is the tool's lack to handle audio characteristics such as noise, the number of speakers, accents, and the distance of the speaker. This limitation stems from the tool's foundational approach, which uses a straightforward comparison between human-made YouTube subtitles and those generated by a speech recognition model. This approach inherently focuses on textual alignment without delving into the nuances of audio quality or speaker attributes.

To address the mentioned audio characteristics handling, an advanced feature could be integrated into the Mi-Go tool, employing audio analysis techniques to evaluate and adjust for different audio characteristics before the transcription process. This enhancement could involve the implementation of pre-processing algorithms capable of detecting and compensating for noise levels, identifying speaker count and accents, and adjusting for recording distance. Such improvements would not aim to refine the accuracy of the speech recognition, as it is not the tool's purpose, but enrich Mi-Go's speech recognition model evaluation results by adding possible root causes (such as high levels of noise or far-field speech) of potential poor model's performance.

In the pursuit of excellence within the realm of rapid speech-to-text models development, currently, the Mi-Go tool is undergoing a rigorous and comprehensive testing process, embodying the highest standards of software quality assurance [10]. This meticulous testing is crucial not only to ensure the tool's reliability and accuracy in evaluating speech-to-text models but also to guarantee an optimal user experience, free from technical glitches and usability hurdles. By subjecting Mi-Go to such thorough scrutiny, we aim to provide users with a seamless and efficient tool, facilitating effective and user-friendly interactions in the nuanced field of speech-to-text system evaluation.

[10] https://www.youtube.com/watch?v=Jk83I-z6C98, access: 2024.03.14

We hope that Mi-Go tool will find wide application in both speech recognition machine learning model evaluation and detection of anomalies in existing video transcriptions.

Appendix 1: Testplan generator command-line parameters

Usage

`python testplan_generator.py <NumberOfVideos> [options]`

Required arguments

`NumberOfVideos`: The number of randomly fetched videos planned to use in model's evaluation. This argument is required.

Optional arguments

`-o, -outputDirectory <directory>`: Destination directory for the testplan files. Defaults to `./testplans/`.

`-l, -relevanceLanguage <ISO 639-1 language code>`: Preferred language for the video's content. Defaults to `en`.

`-c, -videoCategoryId <video category ID>`: Use videos from a specific YouTube category, characterized by the YouTube API's category ID.

`-t, -topicId <topic ID>`: Use videos about a specific topic, characterized by the YouTube API's topic ID.

`-r, -regionCode <region code>`: Use videos targeted to a specific region. Defaults to `US`.

`-d, -videoDuration <duration>`: Video duration filter. Possible values are `any`, `long`, `medium`, and `short`. Defaults to `medium`.

`-lc, -videoLicense <license>`: Video license filter. Possible values are `any`, `creativeCommon`, and `youtube`. Defaults to `creativeCommon`.

`-q, -queryTerm <term>`: Query term for filtering the videos.

Examples

Generate testplan for using 100 random videos:

```
python testplan_generator.py 100
```

Generate testplan for using 50 videos, output to the specified directory, and filter by English language:

```
python testplan_generator.py 50 -o /path/to/directory -l en
```

Note: Replace `/path/to/directory` with the actual directory path where you want the testplan to be saved.

Appendix 2: Comparison of WER values for different datasets to our results, broken down by model types

Table 4 Comparison of WER values for Whisper large-v1 model presented in [27] and our results (highlighted)

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

Table 5 Comparison of WER values for wav2rec 2.0 Large model presented in [27] and our results (highlighted)

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

Table 6 Comparison of WER values for Conformer-Transducer model presented in [12] and our results (highlighted)

Dataset used	WER [%]
LibriSpeech Clean	3.1
LibriSpeech Other	6.6
GigaSpeech Middle Test	15.7
GigaSpeech Middle Dev	16.1
YouTube	**16.4**

Table 7 Comparison of WER values for ESPnet2 Conformer model presented in [26] and our results (highlighted)

Dataset used	WER [%]
LibriSpeech 960h Clean	1.85
LibriSpeech 960h Other	3.95
LibriSpeech 100h Clean	6.6
GigaSpeech Middle Test	10.8
GigaSpeech Middle Dev	10.9
LibriSpeech 100h Other	17.2
YouTube	**29.3**
AphasiaBank Control	35.3
Fisher-Callhome devtest	38.3
Fisher-Callhome evltest	38.8
AphasiaBank Patients	40.3

Appendix 3: Detailed results for particular models, broken down by YouTube categories

Table 8 WER [%] statistics for Whisper large-v3 model

Min	Max	Mean	Std. deviation	Median	Category
4.8	24.3	12.8	5.2	12.4	Autos & Vehicles
6.4	20.3	13.0	4.1	12.4	Comedy
4.5	94.1	29.1	34.1	11.8	Education
4.3	80.7	32.3	28.4	14.4	Entertainment
44.8	250.0	110.0	76.6	71.8	Film & Animation
2.1	22.3	9.9	6.1	7.9	Gaming
5.4	19.7	13.1	4.5	12.9	Howto & Style
6.7	49.3	28.8	10.7	28.5	Music
23.7	67.4	35.7	13.5	28.8	News & Politics
4.6	22.7	11.8	5.6	11.3	Nonprofits & Activism
13.9	56.8	27.9	12.2	24.2	People & Blogs
12.9	99.0	55.6	36.8	55.9	Pets & Animals
5.7	33.8	15.3	7.7	15.7	Science & Technology
20.9	81.8	41.4	17.9	38.2	Sports
8.7	40.4	20.5	9.7	21.4	Travel & Events

Table 9 WER [%] statistics for Whisper large-v1 model

Min	Max	Mean	Std. deviation	Median	Category
0.7	13.8	4.9	3.7	4.3	Autos & Vehicles
5.1	16.5	8.1	3.3	7.2	Comedy
2.7	98.1	24.6	37.1	3.9	Education
3.6	94.5	27.1	31.4	8.9	Entertainment
56.5	161.2	83.8	33.6	65.9	Film & Animation
1.6	15.0	4.7	3.9	2.9	Gaming
0.8	16.3	7.2	5.0	5.3	Howto & Style
0.7	614.4	80.8	178.2	20.9	Music
12.9	65.0	28.4	15.5	22.2	News & Politics
0.7	8.2	2.9	2.1	2.0	Nonprofits & Activism
1.5	47.6	11.1	13.6	6.0	People & Blogs
6.7	106.2	52.1	44.0	47.4	Pets & Animals
0.8	17.4	5.6	5.3	3.4	Science & Technology
16.8	87.8	35.3	22.1	31.0	Sports
3.7	10.3	5.6	2.3	4.5	Travel & Events

Table 10 WER [%] statistics for Whisper medium.en model

Min	Max	Mean	Std. deviation	Median	Category
1.5	14.4	4.6	3.6	3.6	Autos & Vehicles
4.7	18.0	8.6	3.8	7.5	Comedy
1.9	219.5	38.7	70.2	3.8	Education
4.5	108.3	34.8	36.0	20.5	Entertainment
34.4	4600.0	957.9	1612.0	90.1	Film & Animation
1.2	14.4	4.5	3.7	2.7	Gaming
0.9	16.4	7.0	5.4	6.1	Howto & Style
12.0	118.9	34.5	29.6	25.6	Music
10.5	64.2	28.8	15.2	23.7	News & Politics
0.6	8.3	3.1	2.2	2.7	Nonprofits & Activism
1.9	48.2	12.9	13.3	8.8	People & Blogs
5.9	411.5	94.5	120.3	54.0	Pets & Animals
0.4	17.2	6.0	5.4	4.3	Science & Technology
18.8	82.4	33.8	20.2	26.4	Sports
2.6	14.1	5.2	3.1	4.1	Travel & Events

Table 11 WER [%] statistics for Whisper small.en model

Min	Max	Mean	Std. deviation	Median	Category
1.0	12.5	4.8	3.3	3.9	Autos & Vehicles
5.9	22.4	9.1	5.2	6.7	Comedy
2.1	93.6	23.7	36.5	3.3	Education
3.9	107.0	34.1	37.1	14.0	Entertainment
54.4	3975.0	693.7	1284.5	91.2	Film & Animation
2.0	17.5	5.1	4.5	3.0	Gaming
1.1	18.6	7.2	5.8	5.4	Howto & Style
2.2	58.0	30.7	18.2	26.9	Music
11.2	63.8	28.5	15.2	22.9	News & Politics
1.1	8.2	3.1	2.0	2.6	Nonprofits & Activism

Min	Max	Mean	Std. deviation	Median	Category
2.5	47.3	15.0	16.4	8.4	People & Blogs
5.7	5237.5	575.3	1555.0	54.3	Pets & Animals
0.3	17.1	6.0	5.5	3.5	Science & Technology
19.3	86.7	34.0	22.1	24.8	Sports
3.8	7.7	5.2	1.2	5.0	Travel & Events

Table 12 WER [%] statistics for Whisper base.en model

Min	Max	Mean	Std. deviation	Median	Category
1.8	14.8	5.9	3.9	4.9	Autos & Vehicles
6.0	31.3	12.0	7.4	9.3	Comedy
1.9	93.8	24.1	36.5	3.9	Education
4.4	97.1	34.8	35.3	13.9	Entertainment
40.7	145.8	81.4	33.0	84.5	Film & Animation
2.8	28.7	8.6	7.9	5.5	Gaming
1.3	19.3	7.8	5.8	5.5	Howto & Style
10.4	81.6	40.5	21.7	35.5	Music
13.0	61.5	28.6	14.4	22.9	News & Politics
0.7	8.0	3.4	2.2	2.4	Nonprofits & Activism
2.5	47.9	14.7	13.8	10.3	People & Blogs
7.9	12650.0	1680.8	3823.3	55.5	Pets & Animals
0.7	18.6	7.1	5.7	5.0	Science & Technology
23.0	87.6	39.7	21.7	28.7	Sports
4.6	9.3	6.3	1.4	6.0	Travel & Events

Table 13 WER [%] statistics for Whisper tiny.en model

Min	Max	Mean	Std. deviation	Median	Category
2.8	20.7	8.1	5.2	7.3	Autos & Vehicles
7.2	33.4	14.7	7.5	11.2	Comedy
2.9	94.0	24.7	36.3	4.0	Education
5.4	93.1	39.2	35.4	21.7	Entertainment
70.5	132.2	93.7	21.4	91.1	Film & Animation
3.1	33.6	8.6	8.9	4.4	Gaming
2.7	21.4	9.2	5.9	7.1	Howto & Style
18.5	85.8	50.4	19.8	55.8	Music
14.6	61.4	29.6	14.0	24.1	News & Politics
1.4	9.2	4.1	2.4	3.4	Nonprofits & Activism
3.6	49.9	17.4	14.4	9.6	People & Blogs
8.9	164.8	66.4	59.5	55.4	Pets & Animals
1.4	20.1	9.0	6.1	6.1	Science & Technology
27.6	92.9	45.3	20.5	37.9	Sports
5.0	16.2	8.6	3.4	7.1	Travel & Events

Table 14 WER [%] statistics for NeMo Transducer Xlarge model

Min	Max	Mean	Std. deviation	Median	Category
6.4	28.5	13.3	6.4	12.8	Autos & Vehicles
12.1	65.2	26.0	15.1	22.0	Comedy
3.4	95.2	25.9	36.0	6.0	Education
5.8	1587.5	303.0	543.8	32.7	Entertainment
55.3	2000.0	548.6	746.2	91.3	Film & Animation
5.0	67.4	14.3	18.1	7.4	Gaming
6.8	22.2	12.2	4.6	11.2	Howto & Style
68.4	18250.0	2481.8	5627.2	236.9	Music
17.1	61.7	31.1	13.7	24.9	News & Politics
2.7	16.4	5.8	3.8	4.6	Nonprofits & Activism
3.4	47.2	15.2	13.1	10.2	People & Blogs
11.6	6275.0	986.8	1998.0	23.9	Pets & Animals
4.6	22.6	13.6	6.7	12.7	Science & Technology
27.0	133.8	60.1	32.6	53.1	Sports
4.6	17.7	9.0	4.7	6.9	Travel & Events

Table 15 WER [%] statistics for ESPnet2 Conformer model

Min	Max	Mean	Std. deviation	Median	Category
14.0	38.1	24.4	7.9	24.5	Autos & Vehicles
27.0	71.5	40.9	12.5	39.4	Comedy
14.9	95.9	35.0	31.7	19.0	Education
19.2	95.2	49.1	26.4	42.9	Entertainment
64.0	127.5	92.1	17.6	92.7	Film & Animation
16.5	59.1	26.5	12.4	22.2	Gaming
17.9	57.0	28.3	10.8	24.8	Howto & Style
75.5	507.4	188.5	125.3	143.2	Music
28.9	67.8	40.6	12.0	37.3	News & Politics
9.7	23.4	15.6	4.5	14.6	Nonprofits & Activism
15.4	60.7	34.3	16.4	27.6	People & Blogs
14.5	29.9	20.7	5.5	18.1	Science & Technology
47.6	88.5	61.4	12.8	58.8	Sports
15.7	41.4	27.6	8.2	25.7	Travel & Events

Table 16 WER [%] statistics for Wav2Vec2 model

Min	Max	Mean	Std. deviation	Median	Category
10.6	38.2	22.0	9.0	21.7	Autos & Vehicles
25.1	67.5	37.3	12.4	32.8	Comedy
9.8	96.1	30.7	33.9	15.0	Education
12.6	102.4	46.9	30.8	39.2	Entertainment
59.4	635.7	160.5	180.3	99.5	Film & Animation
10.7	66.3	20.9	16.0	15.1	Gaming
9.9	27.5	19.2	5.1	19.7	Howto & Style
90.5	2892.9	462.0	828.3	120.4	Music
26.5	65.7	38.9	12.6	35.6	News & Politics
5.3	23.1	12.0	4.6	11.1	Nonprofits & Activism
8.7	53.8	26.5	15.7	20.3	People & Blogs
15.5	99.2	59.7	38.4	63.9	Pets & Animals
11.7	44.5	27.5	12.1	22.9	Science & Technology
52.2	152.1	74.5	32.4	63.4	Sports
8.7	38.6	20.6	11.4	15.3	Travel & Events

Appendix 4: List of YouTube videos randomly selected by Mi-Go tool for speech recognition model evaluation experiment

Category: Autos & Vehicles

1. https://www.youtube.com/watch?v=EM4odIQZVgw
2. https://www.youtube.com/watch?v=oUpDEsEle68
3. https://www.youtube.com/watch?v=ANZDDO9TKc4
4. https://www.youtube.com/watch?v=II7SZUBr8ig
5. https://www.youtube.com/watch?v=4d88gPxvmFI
6. https://www.youtube.com/watch?v=mYmNM8-XRP0
7. https://www.youtube.com/watch?v=diY4pmAnb1g
8. https://www.youtube.com/watch?v=ESc1GpDxieM
9. https://www.youtube.com/watch?v=_W2MLhH6O8o
10. https://www.youtube.com/watch?v=azwrKNmDkLE

Category: Comedy

1. https://www.youtube.com/watch?v=mfjnDLbCroQ
2. https://www.youtube.com/watch?v=dGPEBuTSmQg
3. https://www.youtube.com/watch?v=4TIIdrOfbls
4. https://www.youtube.com/watch?v=Bq7O57JOFAM
5. https://www.youtube.com/watch?v=WROByxR_ZLg
6. https://www.youtube.com/watch?v=iGqWc5EHeDc
7. https://www.youtube.com/watch?v=eP91xAGs0WE
8. https://www.youtube.com/watch?v=7wR3dnLWF6c
9. https://www.youtube.com/watch?v=yinYc-bwAw0

Category: Education

1. https://www.youtube.com/watch?v=wX78iKhInsc
2. https://www.youtube.com/watch?v=rhgwIhB58PA
3. https://www.youtube.com/watch?v=S294zRodS_4
4. https://www.youtube.com/watch?v=GEmuEWjHr5c
5. https://www.youtube.com/watch?v=fXsOlAYvgh0
6. https://www.youtube.com/watch?v=cPnbdAFrSLM
7. https://www.youtube.com/watch?v=TKQqKZ8EMes
8. https://www.youtube.com/watch?v=y3fm6wNzK70
9. https://www.youtube.com/watch?v=r5sw-6lJmTA

Category: Entertainment

1. https://www.youtube.com/watch?v=Y7JQfHGrjqc
2. https://www.youtube.com/watch?v=QeDumNeq5-w
3. https://www.youtube.com/watch?v=2wc5VpRc450
4. https://www.youtube.com/watch?v=LCygFRx2_DE
5. https://www.youtube.com/watch?v=YfXdrUfKOn8
6. https://www.youtube.com/watch?v=R5aiIWf5YGk
7. https://www.youtube.com/watch?v=mNIXRXikYDc
8. https://www.youtube.com/watch?v=CL0nHs73YO0
9. https://www.youtube.com/watch?v=IlavFAjBdWo

Category: Film & Animation

1. https://www.youtube.com/watch?v=kNw8V_Fkw28
2. https://www.youtube.com/watch?v=2RALmFInHGg
3. https://www.youtube.com/watch?v=7GjJef2QkQU
4. https://www.youtube.com/watch?v=AZS5cgybKcI
5. https://www.youtube.com/watch?v=BCCwCSdXRSE
6. https://www.youtube.com/watch?v=ztpcMUH44jk
7. https://www.youtube.com/watch?v=gZyjJtBIlow
8. https://www.youtube.com/watch?v=MCKPIVszXUc

Category: Gaming

1. https://www.youtube.com/watch?v=kbNjpCeYuvE
2. https://www.youtube.com/watch?v=3JZeI9SF0lI
3. https://www.youtube.com/watch?v=4G_7obY14X0

4. https://www.youtube.com/watch?v=9GaROGghe3E
5. https://www.youtube.com/watch?v=30YEc779Imc
6. https://www.youtube.com/watch?v=lG2dXobAXLI
7. https://www.youtube.com/watch?v=IUSWXcuz-Vno
8. https://www.youtube.com/watch?v=gdrKuYwsq8s
9. https://www.youtube.com/watch?v=gvjVP56r0BA
10. https://www.youtube.com/watch?v=zViFnhVHPUI

Category: Howto & Style

1. https://www.youtube.com/watch?v=kUE2fPLOUxo
2. https://www.youtube.com/watch?v=SLfH9yOGs3o
3. https://www.youtube.com/watch?v=DHzJMa_pqPY
4. https://www.youtube.com/watch?v=-eqcnPq2xdE
5. https://www.youtube.com/watch?v=vOo88OyATpI
6. https://www.youtube.com/watch?v=rZhnLoHg0Sg
7. https://www.youtube.com/watch?v=meSiRSFSQNY
8. https://www.youtube.com/watch?v=WEGmOnpOvRM
9. https://www.youtube.com/watch?v=b5G-rWS8Xmk
10. https://www.youtube.com/watch?v=NmzyzsmQIxA

Category: Music

1. https://www.youtube.com/watch?v=XXYlFuWEuKI
2. https://www.youtube.com/watch?v=b1kbLwvqugk
3. https://www.youtube.com/watch?v=QcIy9NiN-bmo
4. https://www.youtube.com/watch?v=gl1aHhXnN1k
5. https://www.youtube.com/watch?v=aJOTlE1K90k
6. https://www.youtube.com/watch?v=LHCob76kigA
7. https://www.youtube.com/watch?v=fNFzfwLM72c
8. https://www.youtube.com/watch?v=uWRlisQu4fo
9. https://www.youtube.com/watch?v=aAkMkVFwAoo
10. https://www.youtube.com/watch?v=YVkUvmDQ3HY

Category: News & Politics

1. https://www.youtube.com/watch?v=PYooyPcRNVc
2. https://www.youtube.com/watch?v=23cJM6UEdTQ
3. https://www.youtube.com/watch?v=OWIrhn6KyNA
4. https://www.youtube.com/watch?v=ovbGQ1B4rhY
5. https://www.youtube.com/watch?v=L8uiUc5ivGs
6. https://www.youtube.com/watch?v=S9e0gPyAJbo
7. https://www.youtube.com/watch?v=9297wk_HG8M
8. https://www.youtube.com/watch?v=7dBkVC40tdU
9. https://www.youtube.com/watch?v=YQDdBR2ByqI
10. https://www.youtube.com/watch?v=5jEv98bHD6M

Category: Nonprofits & Activism

1. https://www.youtube.com/watch?v=qXHuQfZTH20
2. https://www.youtube.com/watch?v=mrPjz30rAVQ
3. https://www.youtube.com/watch?v=bfAzi6D5FpM
4. https://www.youtube.com/watch?v=UzdF2zpex8o
5. https://www.youtube.com/watch?v=3m6OGbLTQgY
6. https://www.youtube.com/watch?v=KEoxUw-gwec
7. https://www.youtube.com/watch?v=CxCsk-rvfTQ
8. https://www.youtube.com/watch?v=iX9fizsJfuU
9. https://www.youtube.com/watch?v=Yt38f7A_Rwo
10. https://www.youtube.com/watch?v=Z-6IfEoETyU

Category: People & Blogs

1. https://www.youtube.com/watch?v=7H3D-6nj_dY
2. https://www.youtube.com/watch?v=3pJdft6QIUA
3. https://www.youtube.com/watch?v=7AeMhVN-TFA
4. https://www.youtube.com/watch?v=WgPZt7WGZJk
5. https://www.youtube.com/watch?v=3zTR4ayDG38
6. https://www.youtube.com/watch?v=XHw0bDa16xA
7. https://www.youtube.com/watch?v=-wFsYY71wyk
8. https://www.youtube.com/watch?v=lj5GXZaE7qs
9. https://www.youtube.com/watch?v=u0uXzzW6bJ0

Category: Pets & Animals

1. https://www.youtube.com/watch?v=4Co4mDeCIJ4
2. https://www.youtube.com/watch?v=wRpvm3B5Ocg
3. https://www.youtube.com/watch?v=OI4Y-efFkzU
4. https://www.youtube.com/watch?v=Jk83I-z6C98
5. https://www.youtube.com/watch?v=lCegfmeugdQ
6. https://www.youtube.com/watch?v=Dl9Sa4H5TM0
7. https://www.youtube.com/watch?v=j0SF0A6aDOU
8. https://www.youtube.com/watch?v=mKoF48g89s4
9. https://www.youtube.com/watch?v=xl-GCjSsgho
10. https://www.youtube.com/watch?v=wlzMqe2ZqXo

Category: Science & Technology

1. https://www.youtube.com/watch?v=SEI0LtUmpn4
2. https://www.youtube.com/watch?v=Tf3QDABo4MA
3. https://www.youtube.com/watch?v=OyQ3B1U8_XY
4. https://www.youtube.com/watch?v=5pVjCJDAyhk
5. https://www.youtube.com/watch?v=z-2N3WoikqA
6. https://www.youtube.com/watch?v=_3TkeK2uK94
7. https://www.youtube.com/watch?v=t7RaVnEGkc0
8. https://www.youtube.com/watch?v=5s5uVZSdH7s
9. https://www.youtube.com/watch?v=rPJcY_UwlXc
10. https://www.youtube.com/watch?v=uxzbrkSxqqo

Sports

1. https://www.youtube.com/watch?v=dwV04XuiWq4
2. https://www.youtube.com/watch?v=bIDKhZ_4jLQ
3. https://www.youtube.com/watch?v=heIKaaamvdc
4. https://www.youtube.com/watch?v=luR70V5gdS0
5. https://www.youtube.com/watch?v=No8-mBek3rs
6. https://www.youtube.com/watch?v=-RmUADCWI4A
7. https://www.youtube.com/watch?v=hOtv5V9II8o

Category: Travel & Events

1. https://www.youtube.com/watch?v=7vqfjBZ9864
2. https://www.youtube.com/watch?v=MQROYY0dY9A
3. https://www.youtube.com/watch?v=DNNMS7l6A-g
4. https://www.youtube.com/watch?v=dNU1lJiDaSY
5. https://www.youtube.com/watch?v=9p_GPYW0nO0
6. https://www.youtube.com/watch?v=RB1MN0QoXH0
7. https://www.youtube.com/watch?v=yQCBAaJg1LE
8. https://www.youtube.com/watch?v=9wbNabuP6aM
9. https://www.youtube.com/watch?v=Wt4XODPm4hA
10. https://www.youtube.com/watch?v=kFMHx6XwBk0

Acknowledgements
We extend our heartfelt gratitude to YouTube for the *Fair Use* policy allowing to use their platform and videos for research purposes. This study would not have been possible without the rich and diverse content available on YouTube, which has been instrumental in evaluating and demonstrating the adaptability and performance of speech recognition models in various real-world scenarios.
We would also like to express our appreciation to the creators of the Whisper speech recognition model for their outstanding contribution to the field of automatic speech recognition. Their innovative work has provided an excellent benchmark for assessing the effectiveness of our Mi-Go tool and has made a significant impact in advancing the capabilities of speech recognition technologies.
The resources provided by both YouTube and the Whisper have been invaluable, enabling us to conduct this research with a great scope and depth. Thank you for advancing the frontiers of audio, speech, and music processing.

Authors' contributions
All authors, Tomasz Wojnar, Jarosław Hryszko and Adam Roman contributed to this research work. The specific roles and contributions are elaborated as follows: *Conceptualization and design*: all authors participated in formulating the research questions, designing the experiments, and setting the methodology. *Data selection and analysis*: Tomasz Wojnar, Jarosław Hryszko, and Adam Roman equally shared the responsibilities of data selection and analysis. Each author independently verified the analyses carried out by the others to ensure accuracy and reliability. *Tool development*: the development of the "Mi-Go" tool was primarily carried out by Tomasz Wojnar, who took the lead in the creation of various modules and functionalities. *Writing and revision*: the majority of the manuscript drafting and substantial editing was led by Jarosław Hryszko. While each section of the paper was collectively discussed and revised by all authors, Jarosław Hryszko took on the primary role of composing and refining the text. *Review and validation*: every author took part in the validation of the experimental results. They also reviewed and approved the final version of the manuscript prior to submission. *Project management*: all authors were involved in the administration and logistics of the research project. All authors have read and approved the final version of this manuscript. By explicitly detailing the contributions of each author, we aim to provide a transparent account of the roles played in this research.

Authors' information
Tomasz Wojnar is currently a computer science student with a keen interest in the real-life applications of machine learning. His academic focus lies in understanding how machine learning models can be optimized and deployed to solve everyday challenges. As a young researcher, Tomasz brings a fresh perspective to the team, particularly in the realm of speech recognition and its practical applications.
Jarosław Hryszko holds a Ph.D. in Computer Science, specializing in the use of machine learning for software quality assurance. With a strong background in both machine learning and software development practices, Dr. Hryszko provides a nuanced understanding of how quality assurance can be enhanced through machine learning technologies. His experience in the field adds considerable depth to the team's expertise.

Adam Roman serves as an assistant professor and is the head of the Software Engineering Division of Faculty of Mathematics and Computer Science, Jagiellonian University, Poland. His primary research interests are centered on software testing, including AI testing. Professor Roman has contributed significantly to both academia and industry through his comprehensive studies on various aspects of software engineering and testing methodologies. His leadership and extensive experience provide the team with strategic direction, academic rigor and absurd sense of humor.
Each author brings a unique set of skills and expertise to this research project, collectively forming a multidisciplinary team capable of tackling complex problems in the field of audio, speech, and music processing.

Funding
Not applicable.

Availability of data and materials
The data we use is available on the YouTube platform on a *Fair Use* policy (more information on Fair Use on YouTube one can find on https://support.google.com/youtube/answer/9783148?hl=en (access: 2023.09.09)). Specific video URLs are listed in Appendix 4.
The source code of the Mi-Go tool is available under Apache 2.0 open source licence at https://github.com/Kowalski1024/Mi-Go.

Declarations

Competing interests
Not applicable.[[

Received: 9 September 2023 Accepted: 10 April 2024
Published online: 01 May 2024

References

1. S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, S. Vijayanarasimhan, Youtube-8m: A large-scale video classification benchmark. (2016). arXiv preprint arXiv:1609.08675
2. T. Afouras, J.S. Chung, A. Zisserman, Lrs3-ted: A large-scale dataset for visual speech recognition. (2018). arXiv preprint arXiv:1809.00496
3. S. Allen. How many videos are on YouTube? 33+ interesting stats. (2023). https://www.nichepursuits.com/how-many-videos-are-on-youtube/. Accessed 17 Dec 2023
4. R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F.M. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus. Proceedings of the Twelfth Language Resources and Evaluation Conference. (European Language Resources Association, Marseille, 2020), p. 4218–4222
5. A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
6. G. Chen, S. Chai, G. Wang, J. Du, W.Q. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, et al., Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. Proceedings of the Interspeech 2021. (International Speech Communication Association (ISCA), Brno, 2021), p. 3670–3674
7. D.M. Córdova-Esparza, J. Terven, A. Romero, A.M. Herrera-Navarro. Audio-Visual Database for Spanish-Based Speech Recognition Systems, in *Advances in Soft Computing: 18th Mexican International Conference on Artificial Intelligence,* Xalapa*,* 2019,452–460
8. M. Cui, J. Kang, J. Deng, X. Yin, Y. Xie, X. Chen, X. Liu, Towards effective and compact contextual representation for conformer transducer speech recognition systems. Proceedings of the Interspeech 2023. (International Speech Communication Association (ISCA), Dublin, 2023), p. 2223–2227
9. M. Del Rio, N. Delworth, R. Westerman, M. Huang, N. Bhandari, J. Palakapilly, Q. McNamara, J. Dong, P. Zelasko, M. Jetté, Earnings-21: A practical benchmark for asr in the wild. Proceedings of the Interspeech 2021. (International Speech Communication Association (ISCA), Brno, 2021), p. 3465–3469
10. M. Drąg, J. Hryszko, Testing of Mi-Go application - Technical report (2023). https://frege.ii.uj.edu.pl/dragmigo2023.pdf. Accessed 27 July 2023
11. J.F. Gemmeke, D.P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, M. Ritter. Audio set: An ontology and human-labeled dataset for audio events, in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP),* New Orleans, 2017, p. 776–780
12. X. Gong, Y. Wu, J. Li, S. Liu, R. Zhao, X. Chen, Y. Qian, Longfnt: Long-form speech recognition with factorized neural transducer, in *ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Ialissos, 2023, p. 1–5
13. A. Gulati, J. Qin, C.C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al., Conformer: Convolution-augmented transformer for speech recognition. (2020). arXiv preprint arXiv:2005.08100
14. K. Gunter, C. Vaughn, T. Kendall, Contextualizing/s/retraction: Sibilant variation and change in Washington DC African American Language. Lang. Var. Chang. **33**(3), 331–357 (2021)
15. T. Hayashi, R. Yamamoto, T. Yoshimura, P. Wu, J. Shi, T. Saeki, Y. Ju, Y. Yasuda, S. Takamichi, S. Watanabe, Espnet2-tts: Extending the edge of tts research. (2021). arXiv preprint arXiv:2110.07840
16. J.W. Kim. Whisper GitHub Project Readme. (2023). https://github.com/openai/whisper#readme. Accessed 27 July 2023
17. J.Y. Kim, C. Liu, R.A. Calvo, K. McCabe, S.C. Taylor, B.W. Schuller, K. Wu, A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech. (2019). arXiv preprint arXiv:1904.12403
18. A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J.R. Rickford, D. Jurafsky, S. Goel, Racial disparities in automated speech recognition. Proc. Natl. Acad. Sci. **117**(14), 7684–7689 (2020)
19. O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, et al., Nemo: A toolkit for building ai applications using neural modules. (2019). arXiv preprint arXiv:1909.09577
20. E. Lakomkin, S. Magg, C. Weber, S. Wermter, Kt-speech-crawler: Automatic dataset construction for speech recognition from YouTube videos. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Brussels, 2018, p. 90–95
21. V. Levenshtein, Binary codes capable of correcting spurious insertions and deletions of ones. Russ. Probl. Peredachi Informatsii **1**, 12–25 (1965)
22. X. Li, S. Takamichi, T. Saeki, W. Chen, S. Shiota, S. Watanabe, Yodas: YouTube-oriented dataset for audio and speech, in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),* Taipei, 2023, p. 1–8
23. H. Liao, E. McDermott, A. Senior. Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription, in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding,* Olomouc, 2013, p. 368–373
24. NVIDIA. Conformer-Transducer X-Large description (2023). https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge. Accessed 17 Dec 2023
25. V. Panayotov, G. Chen, D. Povey, S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books, in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP),* Brisbane, 2015, p. 5206–5210
26. Y. Peng, K. Kim, F. Wu, B. Yan, S. Arora, W. Chen, J. Tang, S. Shon, P. Sridhar, S. Watanabe, A comparative study on e-branchformer vs conformer in speech recognition, translation, and understanding tasks. Proceedings of the Interspeech 2023. (International Speech Communication Association (ISCA), Dublin, 2023), p. 2208–2212
27. A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision. (2022). arXiv preprint arXiv:2212.04356
28. D. Serdyuk, O. Braga, O. Siohan, Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video. (2022). arXiv preprint arXiv:2201.10439
29. S. Takamichi, L. Kürzinger, T. Saeki, S. Shiota, S. Watanabe, JTubeSpeech: corpus of Japanese speech collected from YouTube for speech recognition and speaker verification. (2021). arXiv preprint arXiv:2112.09323
30. Tatman, R., Kasten, C, Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. Proceedings of the Interspeech 2017. (International Speech Communication Association (ISCA), Stockholm, 2017), p. 934–938
31. S. Watanabe, ESPnet2-ASR realtime demonstration (2023). https://espnet.github.io/espnet/notebook/espnet2_asr_realtime_demo.html. Accessed 17 Dec 2023
32. V. Zue, S. Seneff, J. Glass, Speech database development at MIT: TIMIT and beyond. Speech Commun. **9**(4), 351–356 (1990)

Publisher's Note