

# **Automatic speech recognition: state-of-the-art and performance testing**

**Research thesis submitted for the degree of  
Master of Science HES-SO in Information Sciences**

**by**

**Aparna SCHWANDER**

**Cécile MOLLET**

Under the direction of:

**Prof. Christian MUMENTHALER, PhD.**

**Geneva, January 12, 2024**

**Information Sciences**

**Haute École de Gestion de Genève (HEG-GE)**

## Declaration

This research was carried out as a part of the master's degree in information sciences at the Geneva School of Business Administration.

The students certify that the work submitted is the result of their personal reflection and has been written independently without using sources other than those cited in the bibliography.

The students accept, where applicable, the confidentiality clause.

The use of the conclusions and recommendations formulated in this work, without prejudging their value, does not engage the responsibility of either the students or the director.

Geneva, January 12, 2024

Aparna Schwander  
Cécile Mollet

## Acknowledgments

We would like to warmly thank our advisor and director Prof. Christian Mumenthaler for his help and support throughout this research. We thank Grégoire Urvoy as well for his support.

## Summary

The performance of automatic speech recognition systems (ASR) has dramatically improved over the last decade. A multitude of commercial and open-source models are available to a researcher who wishes to choose one for his or her study. Commercial vendors tend to test their models on standard benchmark corpora which do not reflect real world scenarios. We test three state-of-the-art ASR systems (Amazon Transcribe, Google Speech-To-Text and Whisper from OpenAI) on a corpus of YouTube climate change videos. We compare their performances using the standard word error rate metric and conduct fine grained analysis of the transcripts produced by the systems. We find that amongst the three tested systems Amazon Transcribe performs the best on the climate change corpus. The best performing model will be subsequently used to transcribe the answers to self-registered questionnaires that examines barriers to climate change.

**Keywords :** Automatic Speech recognition systems ; Word error rate; climate change corpus

# Table of contents

Declaration .....	i
Acknowledgments .....	ii
Summary.....	iii
List of Tables.....	v
List of figures.....	vi
1. Introduction .....	1
2. Literature review.....	4
2.1 Process of speech recognition .....	5
2.1.1 Conversion of sound waves into spectrograms .....	5
2.1.2 Finding patterns in spectrograms using a DNN .....	6
2.2 ASR evaluation: Metrics .....	6
2.2.1 Evaluating word accuracy: WER .....	6
2.2.2 Evaluating groups of words: BLEU and METEOR .....	7
2.2.3 Evaluating semantics: BERTscore .....	7
2.2.4 Rationale for choice of metric for this study .....	7
2.3 Different benchmark corpora used in ASR evaluation.....	7
2.4 ASR performance on standard benchmark corpora .....	8
2.5 ASR performance: in the wild.....	8
2.6 Better corpora for benchmarking ASR.....	9
2.7 Inherent bias in ASR systems .....	10
3. Research question .....	11
4. Methods .....	12
4.1 Climate change YouTube Corpus creation and transcript generation.....	12
4.2 Brief description of ASR models that were used in this study .....	15
4.3 Global WER Calculation.....	15
4.3.1 Nuanced WER Calculation .....	15
4.3.2 Fine-grained ASR transcript analysis .....	16
5. Results .....	17
5.1 General overview of ASR Performance .....	17
5.2 ASR Performance Across Different Presentation Styles .....	18
5.3 Impact of noise on ASR performance .....	20
5.4 Difference between human and ASR transcripts.....	21
6. Discussion.....	24
7. Conclusions .....	25
Bibliography .....	26
Appendix 1 : Metadata of the YouTube Corpus .....	1
Appendix 2 : Definition of metadata .....	1
Appendix 3 : Examples of rare words and spelling convention differences .....	3

## List of Tables

Table 1 : YouTube Search Criteria .....	12
Table 2 : Details of the YouTube Climate Change corpus .....	13
Table 3 : Details of post processing of automatic transcripts .....	14
Table 4 : Features that were used for sliced WER calculations and analysis.....	15
Table 5 : Bootstrapped CI values of the mean WER values of each of the tools examined in this study.....	18
Table 6 : Results from a Wilcoxon Signed Ranked Paired test between pairs of ASR tools.. .....	18
Table 7 : Results from the Wilcoxon Signed Rank Test for the three ASR models in quiet and noisy ambient conditions.....	21
Table 8 : Differential treatment of disfluencies in transcription .....	22
Table 9 : Comparison of reductions in transcription .....	22

## List of figures

Figure 1: Applications of ASR Technology in everyday life .....	4
Figure 2: Full spectrogram of the "hello" sound clip. ....	5
Figure 3: A RNN's word prediction process for "hello." .....	6
Figure 4: Pipeline of ASR performance evaluation on the YouTube Climate Change Corpus .....	13
Figure 5: Distribution of WER values obtained for Google (Google Speech-To-Text), Amazon (Amazon Transcribe), OpenAI (Whisper from OpenAI).....	17
Figure 6 : Bubble chart shows the WER values of different tools for each category of video.. ....	19
Figure 7 : Performance of Amazon Transcribe, Google Speech-To-Text and Whisper from OpenAI in quiet and noisy environments.. ....	20

# 1. Introduction

Humans learnt to speak before they learnt to write. As of today, the spoken word remains the most easy and effective means of communication among human beings. It seems only natural that we would want to communicate with machines using our voice and that we would want them to respond using a human like voice. This has been made possible with the advent of automatic speech recognition (ASR) technologies.

In the past decade, ASR research has rapidly advanced, paving the way for voice interaction with machines. This progress is attributed to three key factors. First, improvements in computer technology, including enhanced computational power through general-purpose graphical processing units (GPUs) and multicore processors, enable the training of complex deep learning models with vast amounts of data. Second, the availability of big data allows deep learning models to be trained on extensive datasets, resulting in reduced errors compared to traditional ASR models. Lastly, the inconvenience of communicating with mobile devices through written instructions, as seen in cell phones, in-vehicle infotainment systems, and wearable devices like smartwatches, have created a substantial market potential for voice recognition technology (Yu, Deng 2015).

The above cited conditions led to the development of ASR models whose current state-of-the-art Word Error Rate (WER) is now in the lower single digits. WER is a widely adopted metric for assessing ASR performance where a lower WER indicates a superior ASR system. For human transcription WER typically is around 5% (Bazillon, Estève, Luzzati 2008). In 2016, Microsoft researchers revealed that their ASR model had reached a WER of 5.9%, essentially matching human transcription capabilities (Allison 2016; Xiong et al. 2018).

A researcher who needs to choose an ASR model for his or her research needs is faced with a plethora of choices today. Speechmatics (Speechmatics 2023), AssemblyAI (AssemblyAI 2023), Whisper from OpenAI (Whisper 2023), Microsoft Azure (Microsoft 2023), Google Speech-To-Text (Google 2023), Amazon Transcribe (Amazon 2023), Rev.ai (Rev 2023), Wit.ai (Wit 2023) etc. are some of the ASR systems that exist in the market today. A host of well performing open source models exist as well (Ferraro et al. 2023). Both commercial vendors as well as scientific studies have evaluated ASR performances and published WER rates.

But choosing an ASR model for a particular context solely based on published WER rates is a fallacy. For one, ASR performance varies heavily depending on the type of corpus used to test it (Szymański et al. 2020). For example, Microsoft's purported "near human performance" ASR model was assessed on "LibriSpeech," a widely adopted standard benchmark corpus featuring "read speech" recorded in controlled studio environments. These testing conditions are far off from the reality of where ASR systems are typically deployed.

Moreover, factors like speech signal quality, environment in which the audio is recorded (noisy versus adapted environments), speaker variability (gender, age, accents etc), spoken language variability (dialects), type of speech recorded (monologue, conversational speech) play a significant role in determining ASR performance (Errattahi, El Hannani, Ouahmane 2018; O'Shaughnessy 2008). Consequently, even the most advanced speech recognition systems, though showcasing remarkable performance on benchmark corpora that may not faithfully represent these factors, encounter difficulties in adapting to the complexities of diverse real-life scenarios.



While numerous scientific studies assessing ASR performance have been published in academic journals, they exhibit variations in the choice of ASR tools, testing corpora, and metrics for performance evaluation. This diversity makes direct comparisons challenging. Furthermore, even when researchers use the same corpus for evaluation, they often employ different subsets, introducing additional complexity (Këpuska 2017; Gaida et al. 2014; Trabelsi et al. 2022; Kim et al. 2019; Siegert et al. 2020). Commercial ASR system vendors feature WER comparisons on their websites to showcase their tool's performance against competitors (Rev 2023) without giving much details on the size or the nature of the corpus that it was tested on.

Also, commercial cloud-based systems are continually updated at the backend, leading to constant improvements in their performance (Xiong et al. 2018). ASR systems are often trained on the benchmark corpora itself, leading to a potential bias that could influence the obtained results and limit the generalizability of the WER results reported. Hence, it is important to evaluate the performance of different ASR systems to determine their suitability for a particular deployment scenario or application.

The current research project evaluates the performance of three popular ASR tools namely Google Speech-to-Text, Whisper from OpenAI and Amazon Transcribe. As mentioned above testing ASR tools on standard benchmark corpus is not sufficient to ascertain performance in real use case scenarios. Hence a corpus containing “real world” speech data that is oriented around climate change from YouTube was put together specifically to test the selected ASR systems.

Accordingly, this research presents **three main contributions** to the existing ASR performance evaluation literature:

- 1) We present a YouTube Corpus on climate change that represents a “real world” corpus containing diverse speech styles (webinar, conversations, and interviews), diverse recording environments (studio registered speech, speech registered at homes, speech registered on the street) and diverse subtopics on climate change.
- 2) We test the performance of three ASR models Amazon Transcribe, Whisper from OpenAI and Google Speech-To-Text on the above corpus. We not only publish overall WER values of the 3 systems but also sliced WER values nuanced for speech styles and type of recording environment.
- 3) To our knowledge this is the first time ASR evaluation has been carried out in the climate change domain.

The best performing tool will then subsequently be used in a large-scale study that will investigate barriers against action for climate change. Responses to computer assisted self-interviews on this subject will be collected and transcribed before performing natural language processing to extract insights pertaining to barriers against climate change.

Computer aided self-interviewing is a cost-effective method to collect large scale audio responses to questions pertaining to topics of interest (Brown, Vanable, Eriksen 2008). They provide an opportunity for social science researchers to explore and analyze various phenomena, such as cultural practices, social attitudes, experiences, and beliefs. The voice data must often be transcribed before performing downstream processing and analysis (Pentland et al. 2022). Manual transcription of speech is a time and resource heavy process.

It takes approximately 4 hours to transcribe 1 hour of speech (Bazillon, Estève, Luzzati 2008). This becomes a problem especially when analyzing data at a large scale, and the current study represents the first step towards overcoming this barrier.

Accordingly, we have the following research objectives for the current study: 1. Collect a corpus of YouTube videos on climate change. 2. Perform transcription of the collected videos. 3. Identify performance metrics that can be used to evaluate performance using a survey of related scientific literature. 4. Analyze the performance of Google Speech-to-Text, Amazon Transcribe and Whisper from OpenAI using the above corpus.

## 2. Literature review

Automatic speech recognition can be **defined** (Aldarmaki et al. 2022) as follows:

*“Automatic Speech Recognition (ASR) is the process of automatically identifying patterns in a speech waveform. Patterns that could be detected from speech include the speaker’s identity, language, emotion and the textual transcription of the spoken utterance. The latter is what is typically sought in ASR “*

The textual transcripts resulting from ASR can be used in **three kinds of applications** (Palomäki 2014) illustrated in Figure 1 below.

**Pure Voice-to-Text Transcription:** This application involves converting spoken words into written text. Use cases range from dictation devices, particularly prevalent in medical transcription, to captioning services like those provided by YouTube. The incorporation of text captions into videos benefits individuals with hearing impairments.

**Voice-Activated Digital Assistants:** The second application of ASR is in the domain of voice activated digital assistants. Here, ASR not only transcribes but also utilizes the transcript to execute actions. Examples include IBM’s Watson, Samsung’s Bixby, Apple’s Siri, and Amazon’s Alexa, enabling users to interact with devices like cars, phones, and television sets using voice commands.

**Language Interpretation:** The third application of ASR is in the field of language interpretation. In this context, ASR technology is used to convert spoken language from one language into written text and then to translate it into another language, all in real time. Examples of this application include multilingual meetings in international conferences.

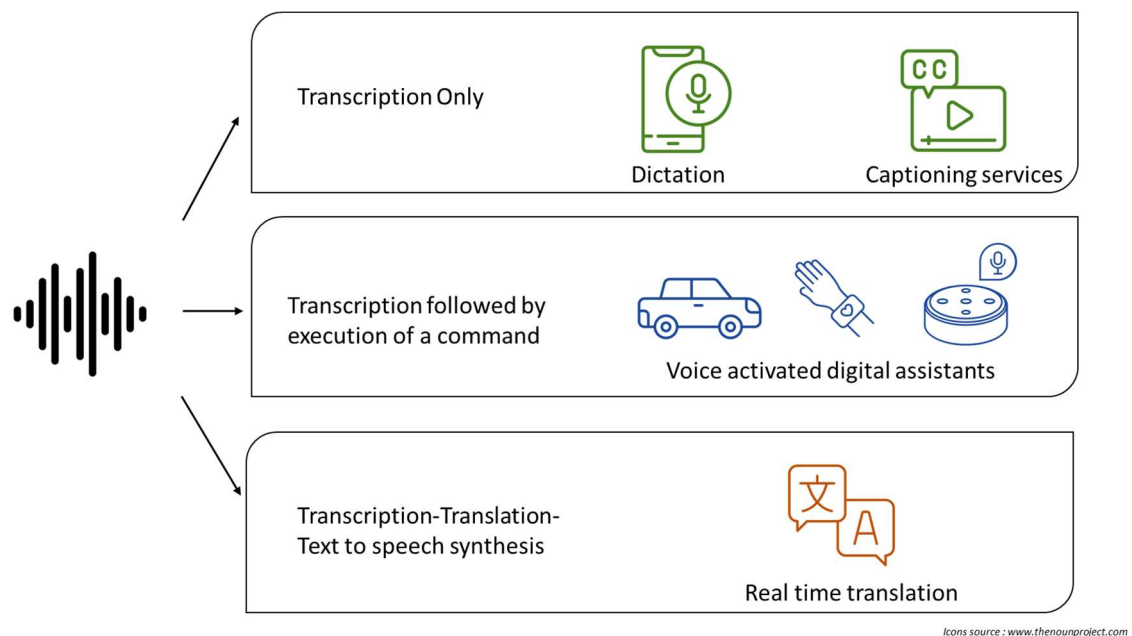


Figure 1: Applications of ASR Technology in everyday life

## 2.1 Process of speech recognition

Google Speech-To-Text (Chiu et al. 2018), Whisper from Open AI (Radford et al., 2022) and Amazon Transcribe (Guo et al. 2020) are ASR models which are based on Deep Neural Networks (DNN). DNNs recognize patterns in speech data by analyzing the digital representation of sound (i.e., spectrograms). They learn features such as frequencies and amplitudes, and employ this knowledge to discern language patterns.

### 2.1.1 Conversion of sound waves into spectrograms

Sound energy travels in the forms of waves. The height of the wave is called its amplitude and the number of waves produced per second is its frequency (measured in Hertz). DNN's are incapable of processing analog data necessitating the conversion of analog waveforms into a digital format through the process known as "sampling". During sampling, the amplitude of a sound wave is meticulously read and recorded thousands of times per second. For effective speech recognition, a sampling rate of 16,000 Hz, capturing 16,000 samples per second, is sufficient for obtaining high-quality signals (*Nyquist frequency* 2023). Sampling yields a sequence of numerical values, for example, an 8,000 Hz sampling rate provides 8,000 amplitude values for every second of the sound wave. Following sampling, Fourier transformation extracts component frequencies from amplitude values, generating a spectrogram—a graphical representation depicting how frequency and amplitude vary over time. This is akin to taking a unique fingerprint of the sound wave, which encapsulates its most essential characteristics (*Fast Fourier transform* 2023). Figure 2, shows the spectrogram of an adult male saying hello. On the y-axis, we find the frequencies, while the x-axis records the passage of time. The intensity and type of color are utilized to signify amplitude. In this example of the male voice recording, the spectrogram reveals a prevalence of lower frequencies in comparison to higher ones.

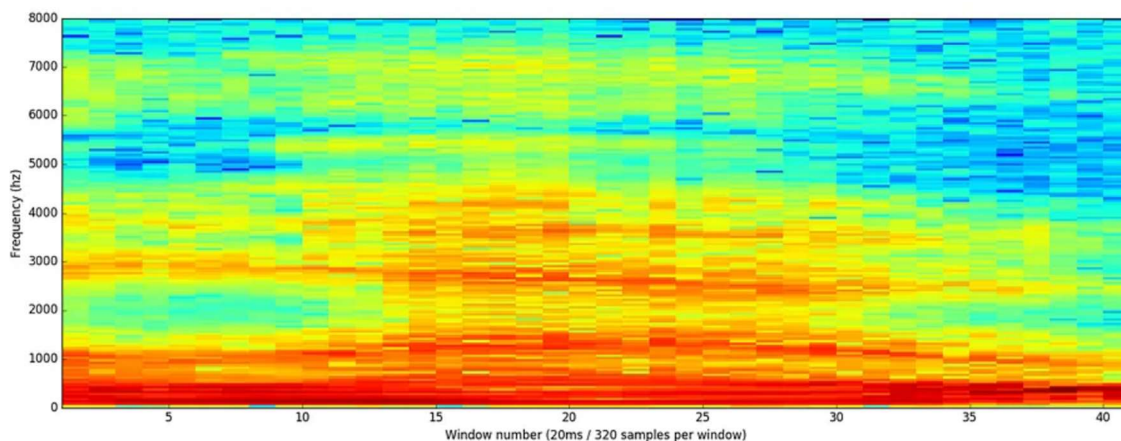


Figure 2: Full spectrogram of the "hello" sound clip. On the y-axis, we find the frequencies, while the x-axis records the passage of time. The intensity and type of color are utilized to signify amplitude. In this example of the male voice recording, the spectrogram reveals a prevalence of lower frequencies in comparison to higher ones (Geitgey 2020).

The next step involves additional processing of the Fourier-transformed signals. Frequencies outside the audible range of 20-20,000 Hz, encompassing the human hearing spectrum are typically filtered out (Stanford Online 2015). Subsequently, the spectrograms derived from brief segments of speech data, such as 20 milliseconds, are input into a DNN, which processes this data in a manner analogous to how it handles image data.

### 2.1.2 Finding patterns in spectrograms using a DNN

Different kinds of DNNs can be used at this stage and complex architectures can be deployed to predict speech into text. It is beyond the scope of this thesis to explain them in detail. A Recurrent Neural Net (RNN) is used as an example to illustrate the process of identifying speech patterns from a spectrogram. DNNs work at the level of graphemes (letters) when compared to conventional or traditional models of speech recognition like Hidden Markov Models, which recognize speech the level of phonemes (Georgescu et al. 2021).

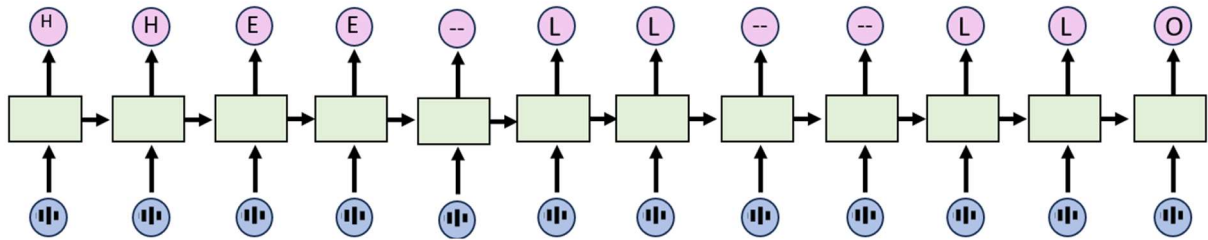


Figure 3: A RNN's word prediction process for "hello." Speech spectrograms are divided into 20ms segments each entering a neural net layer (represented by the light green rectangles).

In Figure 3, we observe a RNN's word prediction process for "hello." Speech spectrograms are divided into 20ms segments (represented by the blue bubbles) each entering a neural net layer (represented by the light green rectangles). The prediction (pink bubbles) made by the first layer is then transferred to the second layer, which considers the input from the previous layer before making predictions. For instance, if the current prediction reads "HEL," it is likely to be followed by "LO" (forming "HELLO"), "L" (resulting in "HELL"), or "LIX" (leading to "HELIX"), but not "TX" since "HELTX" is not a valid English word (Olah). Words that are output by the DNN are compared with large databases of words and only the most probable ones are retained. For example, if for the above example, "HELLO", "HULLO" and "ALLO" were predicted, only "HELLO" will be most likely retained.

## 2.2 ASR evaluation: Metrics

The accuracy of ASR systems has been assessed using various metrics in the scientific literature. WER, BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit ORdering), and BERTscore (Bidirectional Encoder Representations from Transformers score) will be briefly explained in this section.

### 2.2.1 Evaluating word accuracy: WER

The go-to metric for assessing ASR performance is the Word Error Rate (WER), which will also be a key metric in this study. WER can be computed using the formula:

$$W := \frac{\sum_i e_i}{\sum_i n_i}$$

Where  $n$  is the number of words in a sentence  $i$  and  $e$  is the Levenshtein distance or edit distance between the reference transcript and the transcript produced by the ASR. In other words, it is the minimum number of inserts, deletions, and substitutions necessary to transform one transcript to another.

### **2.2.2 Evaluating groups of words: BLEU and METEOR**

BLEU and METEOR are metrics that were originally designed for evaluating the quality of machine-generated translations (Papineni et al. 2001; Banerjee, Lavie 2005) but can also be used to evaluate the quality of ASR transcripts (Sampaio et al. 2021; Magalhães et al. 2022). BLEU is a score based on the precision of n-grams (contiguous sequences of words) in the machine-generated text compared to the reference text. Higher BLEU scores indicate better ASR performance, with a perfect score of 1.0 representing an exact match to the reference text. When compared to the WER which works at the level of the word, BLEU captures the importance of groups of words. METEOR evaluates machine-generated text quality by considering various linguistic features, including unigram precision, recall, stemming, synonymy, and word order. It provides a more comprehensive evaluation compared to BLEU.

### **2.2.3 Evaluating semantics: BERTscore**

While the metrics listed above help access the quality of the transcripts at the level of words (WER) or n-grams (BLEU and METEOR) none of them can account for the quality of the semantics of the transcription. BERTscore relies on contextual embeddings generated by models like BERT to measure similarity between reference and machine generated text (Zhang et al. 2020). It calculates word embeddings for each word in the reference and the machine generated transcript and gives a measure of semantic similarity of the transcripts. Clinical BERTscore, is a BERTscore adapted specifically for the medical domain. It uses the word embeddings of a BERT specifically trained on a corpus of medical sentences and gives a weighted preference to medical words. (Shor et al. 2023)

### **2.2.4 Rationale for choice of metric in this study**

We have chosen to use WER as the metric for quantifying ASR performance. This choice is motivated by several reasons. Firstly, being the most popular metric of choice reporting it facilitates comparison with other studies. Metrics like METEOR or BLEU are being mostly used in the automatic translation literature and do not add any more value than WER for this study. The BERTscore however could have been a valuable metric to understand semantic difficulties encountered by ASR systems. Unfortunately employing this metric necessitates specialist knowledge and resources which are both out of scope for the current study.

## **2.3 Different benchmark corpora used in ASR evaluation**

A benchmark corpus is a collection of audio data that is used to “test” the performance of ASR systems. To date, several of these corpora are publicly available. Corpora can contain speech data belonging to one or more of the different kinds of speaking styles. Monologues (spontaneous), conversations (spontaneous), interviews (spontaneous or semi-spontaneous), read speech (sentences read from a written script) are some of the different kinds of speech that can be represented by corpora. Different recording environments that could have been used to record speech data include studio (clean speech), telephone, noisy environments like dinner table conversations etc. Rich corpora might also include speakers from all genders, speaking different dialects of a language, and other speaker demographics like different ages, native and non-native speakers of a language (Szymański et al. 2020). Two of the benchmark corpora that are popularly used in ASR literature will be explained briefly.

TIMIT (Texas Instruments/Massachusetts Institute of Technology) corpus contains recordings of 6300 utterances of read speech from eight major American English dialect regions. 70% speakers are male and 30% female (Garofolo et al. 1993)

SWITCHBOARD - It was created by the Linguistic Data Consortium and funded by the U.S. government's Defense Advanced Research Projects Agency. It contains telephonic conversations between 500 paid actors in all the major American dialects. Actors were engaged to speak on predefined topics. About 500 hours of conversations each 3-5 minutes in length were recorded (Godfrey, Holliman, McDaniel 1992).

## **2.4 ASR performance on standard benchmark corpora**

ASR performance on benchmark corpora has evolved significantly over the years. In the 1990s, a subset of the SWITCHBOARD corpus yielded an ASR performance of around 78%, with only 22% of the words recognized correctly (Gillick, Cox 1989). Fast forward 15 years, and state-of-the-art ASR systems had made substantial progress, achieving a WER between 20% and 30% on a subset of the SWITCHBOARD corpus (Hain et al. 2005). In 2015, IBM researchers achieved a 6.6% WER on the SWITCHBOARD Hub5 2000 evaluation benchmark subset (Saon et al. 2015). Just two years ago, IBM reported an even lower WER of 5% on another subset of the SWITCHBOARD corpora (Tüske, Saon, Kingsbury 2021).

## **2.5 ASR performance: in the wild**

ASR performance on real world corpora however tell a different story. The following studies were conducted either on interview audio data or conversational audio data collected in settings that were closer to real life situations.

One study compared several online ASR systems, including Google Cloud, IBM Watson, Microsoft Azure, Trint, and YouTube, and examined the nonverbal responses to unintelligible speech (Kim et al. 2019). The authors collected speech data from interactions between medical students (trainee doctors) and mock patients (people acting as patients) using an online videoconferencing platform, resulting in 24 videos from 12 consultations being analyzed. They found that manual transcriptions had the best performance, with a WER of 17.4%, followed by YouTube with a WER of 28%.

The efficiency of a LF-MMI trained acoustic model on the transcription of oral German history interviews (interviews of contemporary witnesses to historical events) was investigated in a recent study (Gref et al. 2022). The authors report a WER of 23.9% on noisy and 15.6% WER on clean oral history interviews. Similar WER results (22.1% ) were reported on a study which used a combination of acoustic model (based on HuBERT method) and a language model (based on RoBERT method) trained on the Mozilla common voice dataset and tested on spontaneous speech data obtained from interviews conducted with older adults in long term care institutions (Hacking et al. 2023).

The performance of four ASR systems, namely Google Cloud Speech to Text API, Google Web Speech API, IBM Watson Speech to Text API, and Wit.ai Speech to Text API were compared in a longitudinal study (Siegert et al. 2020) that lasted several months. The datasets used in the study were collected by the authors themselves and consisted of spontaneous German speech data recorded in noisy environments. Two types of conversational settings were utilized: a living room setting where conversations between two humans and an Amazon Alexa voice assistant were recorded, with minimal background noise; and public conversations

at a German science fair between museum visitors and a voice assistant. Google Cloud Speech had a WER between 17-23% and Wit.AI had a WER between 23-24% in noisy environments. In quiet environments both Google and WIT.AI achieved around 7-11% WER. The authors did not note an improvement in performance in any of the above tools over the study period.

The viability of using manually corrected drafts of ASR transcripts of suspect police interviews as court evidence has been studied (Harrington 2023). The authors tested several different British accents and different testing environments (studio registered clean speech and speech shaped noise added to studio registered audio) on Amazon, Rev and Google transcription services. Amazon had a WER of 13.9% on studio registered speech and 26.4% on noisy speech data.

The feasibility of using ASR transcripts of a job interviewing dataset for further downstream natural language processing has been looked into (Pentland et al. 2022). Despite the fact that the interview data consisted of subjects responding to job interview questions in a laboratory based experiment, they report higher WER rates than normally reported by ASR vendors, 19.90% for Watson Speech to Text from IBM (as opposed to the 5% reported by this vendor (Tüske, Saon, Kingsbury 2021) on the SWITCHBOARD corpora).

The performance of the ASR tools in these settings indicates the level of ASR performance in real life situations is remarkably lesser than those reported with standard benchmark corpora.

## **2.6 Better corpora for benchmarking ASR**

Testing ASR exclusively on standard benchmark corpora can lead to an overestimation of ASR performance. State of the art ASR systems have improved to the extent that corpora like Librispeech and WSJ present minimal challenges to them. In contrast, challenging conversational scenarios, like those in the CHiME5 corpora, result in considerably higher WER, reaching between 46% and 73% for modern ASRs (Szymański et al. 2020). It seems like ASR systems have evolved quickly in the last decade but not the ASR benchmarks that are being used to test them.

A couple of studies have outlined criteria for developing ideal corpora to benchmark ASR performance (Aksénova et al. 2021; Szymański et al. 2020). An ideal corpus should encompass audio data from diverse speech styles, such as dictation, voicemail, conversations, meetings, podcasts, movies, TV, voice search, oration, and audio books. It should also account for technical aspects such as multi-domain transfer (as ASRs trained on general vocabulary may not perform well in specific domains like medicine), acoustic environments (including noisy and music-rich settings), and encoding formats (e.g., FLAC, MP3). Additionally, demographic characteristics play a crucial role, covering dialects, native vs. non-native speakers, gender, age, and speakers with speech impairments in the data set is important.

However, creating an ideal corpus that meets all these requirements can be a practical challenge. The YouTube corpus used in this study strives to address many of these considerations, featuring climate change-related content (domain-adapted), various speech types (conversations, webinars, interviews), diverse acoustic environments (both noisy and quiet), and speakers of different genders (though gender balance was not fully achieved, see Appendix I for more details). In cases where the features of the collected corpora (linguistic features like accents, native and non-native speakers, different speech styles, demographic



features like gender, age, etc) do not match with those of the targeted deployment setting the metrics that are used to evaluate the ASR can be re-weighted to reflect the deployment setting (Aksënova et al. 2021).

## **2.7 Inherent bias in ASR systems**

While benchmark corpora often inflate ASR performance, the training data used to develop these systems inherently introduce biases. AI systems are known to exhibit biases towards specific demographic groups, such as those with darker skin or women. One of the reasons is due to the fact that the training datasets do not accurately represent real-world diversity (Lazaro, 2023). In a study, three state-of-the-art ASR systems from Amazon, Google and OpenAI were evaluated using interviews with 42 white speakers and 73 black speakers. All five systems displayed racial bias, yielding an average Word Error Rate (WER) of 19% for white speakers and 35% for black speakers (Koenecke et al. 2020).

Another similar study investigated bias in Dutch state-of-the-art ASR systems concerning gender, age, regional accents, and non-native accents. The results indicated that female speech was more accurately recognized than male speech, and native Dutch speakers were better identified compared to non-native speakers. Furthermore, ASR systems struggled with transcribing children's speech compared to that of teenagers or seniors (Feng et al. 2021). These inherent biases in the ASR systems must be considered while benchmarking their performance and this underlines the importance of not depending on a metric that gives only the global performance of the system in question.

### 3. Research question

A researcher seeking transcription services is confronted with numerous options, ranging from commercially available ASR models to open-source ASR models. Vendors provide WER metrics for their models, but often lack transparency in specifying the datasets used for training and testing. As previously observed, state-of-the-art models may not perform consistently when applied to different knowledge domains or when dealing with challenges such as noisy data or speaking styles. Consequently, relying solely on published WER values is inadequate for model selection, necessitating a comprehensive evaluation of chosen models. Considering the numerous ASR tools available, we have selected three candidates: Google Speech-to-Text, Whisper from OpenAI, and Amazon Transcribe. These selections are based on their relatively low WER rates according to our literature review (Kim et al. 2019; Magalhães et al. 2022; Sampaio et al. 2021; Siegert et al. 2020) and their popularity.

Additionally, existing benchmark corpora are often insufficient for evaluating ASR tools, as performance varies across these benchmarks and may not align with specific use cases. To address this, researchers should either seek or create a corpus that mirrors their needs and serves as an effective benchmark. In the context of the forthcoming research project involving self-recorded climate change interviews in English, none of the existing benchmark corpora were suitable for achieving accurate results. Thus, we have curated our own climate change corpus from openly available YouTube videos and used it to test the performance of the three chosen models.

#### **Our research question is as follows:**

Among Google Speech-to-Text, Whisper from OpenAI, and Amazon Transcribe, which ASR tool performs the best on our YouTube climate change corpus?

## 4. Methods

This section will briefly describe the ASR tools used in the study and elaborate the methods that were used to generate the YouTube corpus. The process of transcription using the ASR systems and the statistical analysis that was carried out on the transcripts obtained will also be outlined in detail.

### 4.1 Climate change YouTube Corpus creation and transcript generation

Figure 4 gives a quick overview of the ASR performance evaluation pipeline used in this research. The process of generating the YouTube corpus and ASR evaluation consisted of the following steps.

#### Step1- Video selection from YouTube

To make sure that our corpus met with the criteria that defines an adequate benchmark (see [section 2.6](#)) we elaborated a set of search criteria given in Table 1 below. We searched for recent climate change discourses on YouTube using the below filters.

Filter	Chosen Settings
Duration of video	4 to 20 min
Language	English - US/UK
Period	2022-2023
Subtitles	Only videos with transcripts
Search terms	climate change + interviews; climate change + webinars; climate change + podcasts; climate change + conversations; climate change + opinions; climate change + survey/questionnaire; climate change + public opinion; what people on the street think about climate change
Speech Styles	Webinars, conversations, interviews

Table 1 : YouTube Search Criteria

#### Step 2: Corpus Creation

Corpus creation involved downloading both the mp4 and mp3 formats for each of the 16 videos that aligned with our search criteria. Subsequently, we attributed metadata to the acquired content. An overview of the corpus, including specific details, is provided in Table 2. For more information please consult [Appendix I](#).

Feature	Details
Speaking Styles	Webinars, interviews, conversations
Environment	Quiet, Noisy
Size in minutes	206
Gender distribution	52 Male, 47 Female
Number of identified native speakers	63
Number of identified nonnative speakers	36

Table 2 : Details of the YouTube Climate Change corpus

### Step 3 – Assisted transcription:

Employing a professional human transcriber proved to be cost-prohibitive, leading us to undertake the transcription process internally. Given the impracticality of a full manual transcription from the ground up, we adopted the method of assisted transcription (Bazillon, Estève, Luzzati 2008). Our approach involved downloading the automatic captions provided by YouTube for each video and subsequently manually correcting them. 50% of the transcripts underwent correction by researcher A, while the remaining 50% were corrected by researcher B. We transcribed disfluencies (such as "um" and "uh") and back channels (e.g., "like," "yeah," "that's right") whenever feasible.

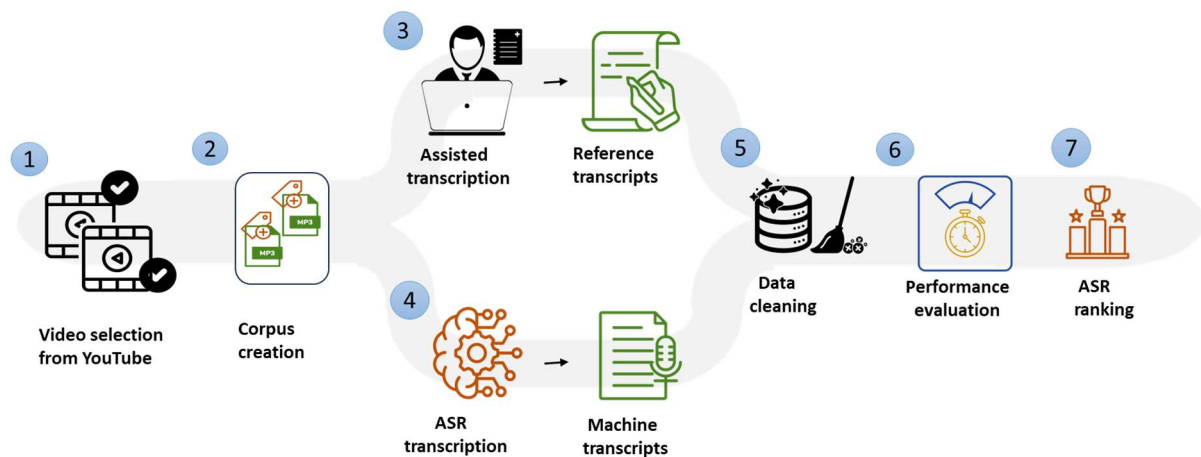


Figure 4: Pipeline of ASR performance evaluation on the YouTube Climate Change Corpus

### Step 4 - Automatic speech transcription:

The audio files were sent to the API's of the 3 automatic transcription services that were chosen for this study. We chose the basic model for all API's and set the language to "en-US".

## Step 5 – Data Cleaning

Before comparing and analyzing the downloaded transcripts, additional processing steps were implemented to address variations in formatting introduced by different ASRs during transcript generation. The reference transcripts, initially sourced from YouTube and subsequently corrected by our team, featured timestamps that were removed to facilitate accurate WER calculations. To ensure uniformity across all transcripts, we conducted a case conversion, converting all text to lowercase. Some ASR outputs contained annotations like [laughs] or [music], which were removed. Furthermore, the processing involved the elimination of speaker diarization and punctuations wherever present. Addressing discrepancies in the representation of numbers, which appeared in both spoken (e.g., "35") and written (e.g., "thirty-five") forms, we opted for consistency by converting all numbers to the written domain. Refer to Table 3 for an overview of these processing steps.

Feature of the transcript	Examples	Type of process
Punctuations	Comma, quotes, hyphen	Removed
String Case	Lower and upper	Converted to lower case
Numbers	32, 3.2	Converted to text
Time Stamp	00:31, 0:3	Removed
Annotations	[laughs], [music], [drum roll]	Removed
Postscripts	Video Credits:	Removed
Speaker Diarization	Rebecca:	Removed

Table 3 : Details of post processing of automatic transcripts

## Step 6: Performance Evaluation

We chose to use WER as the metric to evaluate ASR. WER was calculated using the standard formula for Levenstein's Distance (LD).  $LD = (I + S + D)$  where I is the number of insertions, S is the number of substitutions, D is the number of deletions.  $WER = LD / N$  where N is the number of words in the reference transcript.

## Step 7: ASR Ranking

All transcript processing and statistical calculations were done using the R statistical software (Version 4.3.2). The Levenstein distance between the pair of reference and machine transcripts were calculated using the stringdist package. We used the Wilcoxon signed rank test to calculate significant differences between pairs of ASR tools (`wilcox.test`). The z statistic

for the Wilcoxon signed test was calculated by using the wilcoxonZ function from rcompanion package.

## 4.2 Brief description of ASR models that were used in this study

### Google Speech-To-Text:

The Google transcription service was accessed at the following url: <https://cloud.google.com/speech-to-text?hl=en>. Asynchronous transcription of audio files was carried out as our corpus contained audio files that are longer than 60 seconds. The default model was chosen for transcription and the language was set to “en-US”. The resulting transcripts were without punctuation and time stamps.

### Amazon Transcribe:

The cloud-based transcription service was accessed at the following url: [https://aws.amazon.com/fr/transcribe/?nc1=h\\_ls](https://aws.amazon.com/fr/transcribe/?nc1=h_ls). The default model was chosen for transcription and the language was set to “en-US”. The resulting transcripts were with punctuation and without time stamps.

### Whisper from OpenAI:

The cloud-based transcription service was accessed at the following url: <https://platform.openai.com/docs/guides/speech-to-text/quickstart>. The default model was chosen for transcription and the language was set to “en-US”. The resulting transcripts were with punctuation and without time stamps.

## 4.3 WER Calculation

Mean WER values for each ASR system was calculated along with 95% confidence intervals using boot strapping given the constraints of non-normality of our WER samples.

### 4.3.1 Nuanced WER Calculation

The mean WER values for each tool and its associated confidence intervals provide us with a broad understanding of ASR performance. But this metric can be “sliced” further to reveal variations in performance based on inherent characteristics of the corpora. Table 4 below presents the linguistic and technical features that were used to “slice” the global WER values. We calculated sliced average WER values for each of the categories mentioned below. We also performed Wilcoxon signed ranked tests for pairs of ASR systems for each of these categories. For more information on each of these categories see [Appendix I](#).

Type of feature	Details
Video Type	Webinar, Conversation, Interview
Background	Quiet (office/home), Noisy (street)

Table 4 : Features that were used for sliced WER calculations and analysis

### **4.3.2 Fine-grained ASR transcript analysis**

Transcripts were processed into tokens (individual words) and analyzed further using natural language processing. We extracted proper nouns from deleted and inserted words using the udpipe language model in R and they were further analyzed for the presence of rare words, spelling convention differences etc., manually. The transcripts were also analyzed for the presence of disfluencies like “um” and “uh”.

## 5. Results

We chose three ASR models namely Amazon Transcribe, Google Speech-To-Text and Whisper from OpenAI and tested them on a corpus of 16 YouTube videos that treat the subject of climate change.

### 5.1 General overview of ASR Performance

Figure 5 presents the boxplot of WER values for each of the chosen ASR tools in this study. The median WER values of Google Speech-To-Text and OpenAI's Whisper are similar and are 58.8% higher than the median WER value of Amazon Transcribe. Note the presence of outliers in the case of Amazon Transcribe and Google Speech-To-Text. Statistical normality of WER values were confirmed via a Shapiro-Wilk test. Whisper from OpenAI had normally distributed WER scores ( $p = .7364$ ). In contrast, Amazon Transcribe and Google Speech-To-Text have skewed distributions, as indicated by  $p = .03338$  and  $p = .01748$ , respectively.

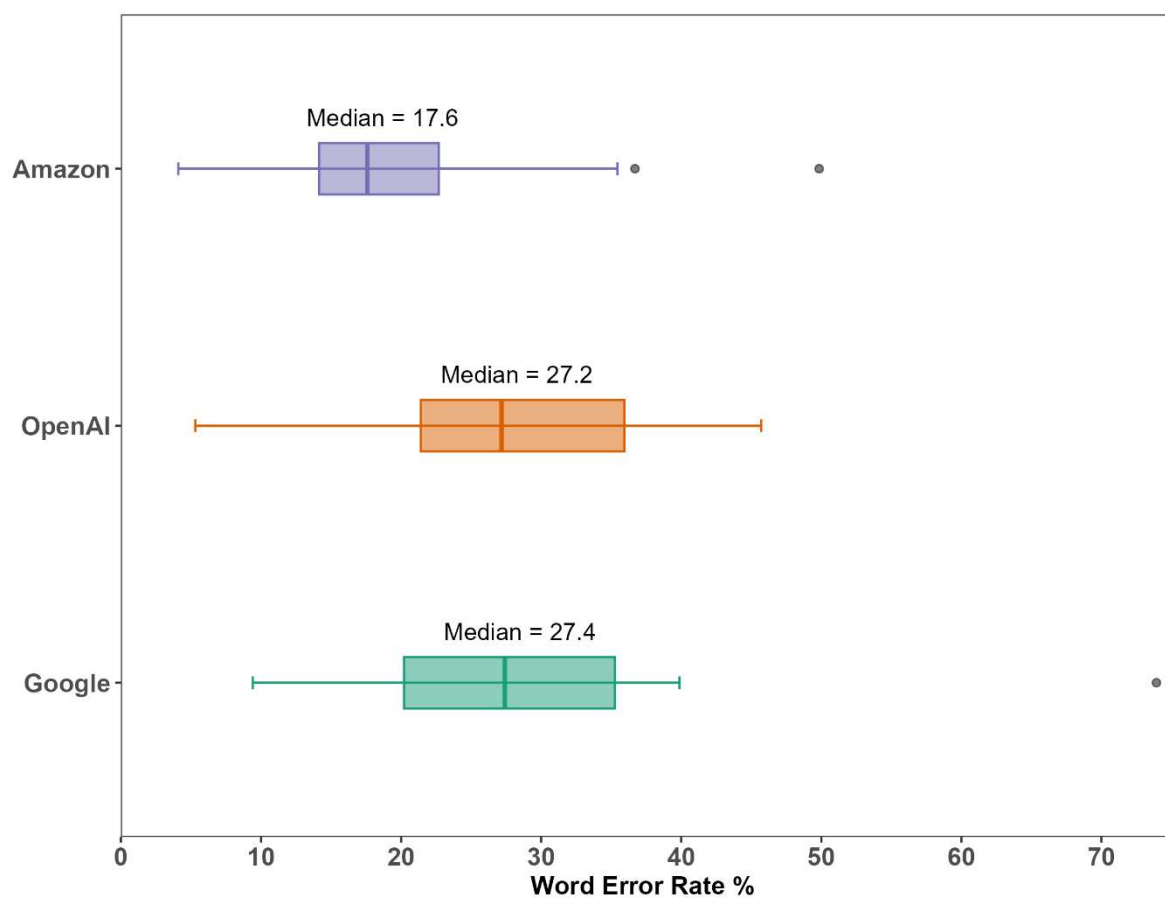


Figure 5: Distribution of WER values obtained for Google (Google Speech-To-Text), Amazon (Amazon Transcribe), OpenAI (Whisper from OpenAI). Amazon Transcribe is the best performing system and Google Speech-To-Text is the worst performing system.

We report the boot strapped confidence intervals for the mean WER values as our sample sizes are small and the distribution of WER values are not normal for Amazon Transcribe and Google Speech-To-Text. See Table 5 below. The mean WER rates reported here are within the wide range reported by others (Kim et al. 2019; Pentland et al. 2022; Harrington 2023) supporting the finding that performance varies with the audio source.



ASR	Average WER	Lower CI	Upper CI	SE
Amazon Transcribe	20.30	15.24	26.16	2.80
Whisper from OpenAI	27.49	22.01	32.80	2.76
Google Speech-To-Text	29.25	22.74	37.18	3.66

Table 5 : Bootstrapped 95% CI values of the mean WER values of each of the tools examined in this study. Amazon Transcribe has the lowest average WER. Google Speech-To-Text has the highest average WER value.

A Wilcoxon Signed Rank Paired test indicated that there were no significant differences in performances between the three pairs of tools. Even though Google Speech-To-Text's mean WER was 43.9% higher when compared to that of Amazon Transcribe's mean WER and we obtained a  $p = .0004272$ , the Cohen's  $d = .668608$  had a zero in its 95% CI [-0.073, 1.410]. A similar result was obtained for the comparison of mean WER between Whisper from OpenAI and Amazon Transcribe. We did not find any significant differences in performances between Whisper from OpenAI and Google Speech-To-Text. See Table 6 for more information.

ASR pair	p value	z statistic	Cohen's D	95 % CI for Cohen's D
Google Vs Amazon	.0004272**	2.11	0.668 (medium)	[-0.073, 1.410]
OpenAI Vs Amazon	.02496	2.15	0.629 (medium)	[-0.109, 1.369]
OpenAI Vs Google	.782	0.113	0.129 (negligible)	[-0.593, 0.852]

Table 6 : Results from a Wilcoxon Signed Ranked Paired test between pairs of ASR tools. Google Speech-To-Text (Google), Whisper from OpenAI (OpenAI) and Amazon Transcribe (Amazon). Results indicate that the tools perform equally well with no observable statistically significant differences. \*\* p values < 0.005 for Google Speech-To-Text vs Amazon Transcribe but note that the 95% confidence intervals of Cohen's d spans a zero.

## 5.2 ASR Performance Across Different Presentation Styles

Our corpus contained different types of video presentations such as webinars where one person speaks most of the time, interviews where at least two or more participants spoke together, and conversations where more than two participants engaged in talk about a topic, each taking turns. We investigated if ASR performance was different for these presentation styles. We distinguish 4 categories namely webinars, interviews recorded in noisy environments, interviews recorded in quiet environments, and conversations. For more details

on how we defined criteria to categorize the videos in the YouTube corpus into webinars, interviews, and conversations, see [Appendix II](#).

ASR performance was **optimal on webinars** which demonstrated the least WER amongst all presentation styles (Mean = 19.4, SD = 12.9). Performance was the **worst in interviews recorded in noisy environments** (Mean = 35.1, SD = 17.0). WER values were similar for interviews recorded in quiet environments (Mean = 23.7, SD = 10.5) and conversations (Mean = 23.5, SD = 6.36). Dialogue type presentations (interview and conversations recorded in quiet environments) had 21% higher WER than webinars (monologue type presentations recorded in quiet environments).

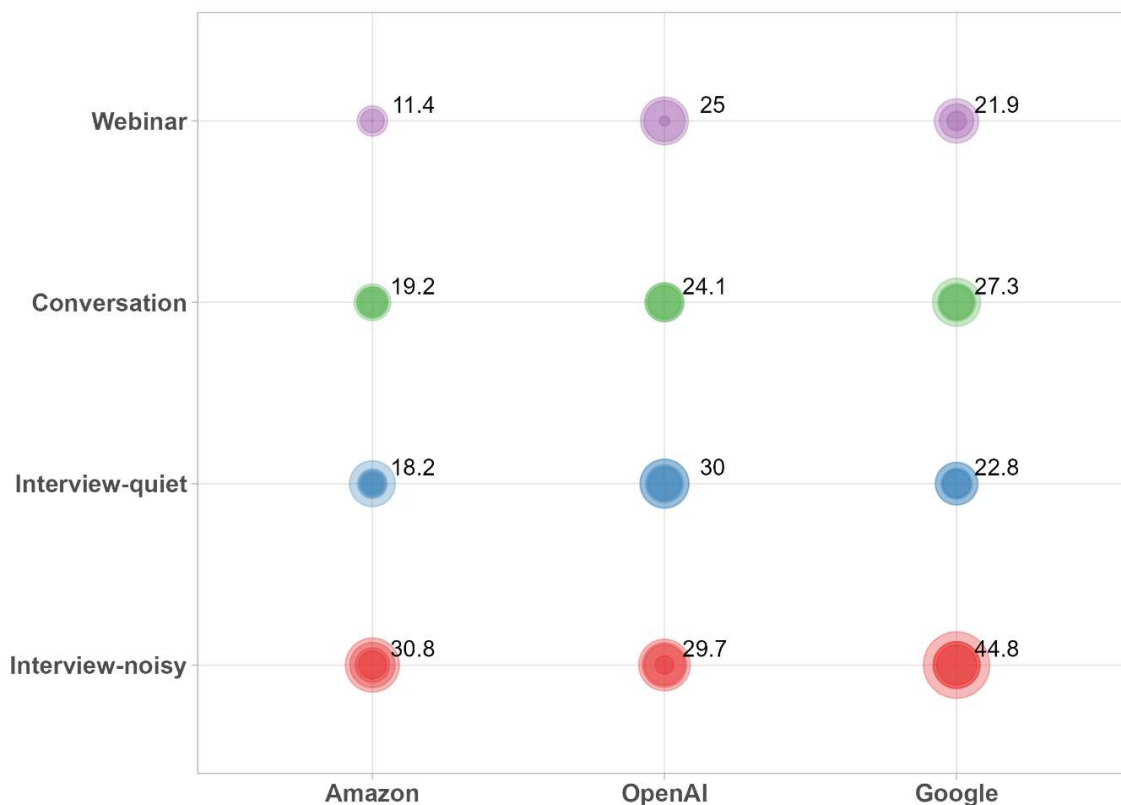


Figure 6 : Bubble chart shows the WER values of different tools for each category of video. Each circle represents a WER value. The rippling effect is produced by the superposition of the WER values in each category. The number of concentric rings represents the number of samples in each category. Average values for each category are marked in black. ASR performance is least affected by webinars and most affected by interviews recorded in noisy environments.

Amazon Transcribe had the lowest error rates in webinars (Mean = 11.4, SD = 6.98). It also showed the best performance in conversations (Mean = 19.2, SD = 4.01) and interviews recorded in quiet environments (Mean = 18.2, SD = 10.7). Whisper from OpenAI performed the best in interviews recorded in noisy environments (Mean = 29.7, SD = 15.5). Google Speech-To-Text performed the worst in all environments<sup>1</sup>. See Figure 6 for more details.

<sup>1</sup> It was not possible to obtain 95 % CI in the Wilcoxon Signed Rank Test that compared the performance of the pairs of tools (Amazon Transcribe Vs Google Speech-To-Text, Whisper from OpenAI Vs Amazon Transcribe and Whisper from OpenAI Vs Google Speech-To-Text) for each of the video categories mentioned above. This is probably due to the very small sample sizes in each category of videos. We could only calculate either 75% CI (for webinars) and 88% CI (for the other categories), results not shown here. For more details for sample number in each category of video please see [Appendix I](#).

Overall, these findings support the fact that has been well documented in ASR literature. That conversations and interviews pose challenges to ASR systems because of backchannels, disfluencies, repetitions, false starts etc (Żelasko et al. 2019).

### 5.3 Impact of noise on ASR performance

Figure 7 shows the performance of all three tools in noisy and quiet environments. ASR performance was **optimal for quiet** environments (Mean = 22.1, SD = 10.2). ASR systems are known to be affected by the presence of ambient noise and modern ASR systems apply several noise cancellation strategies (Yurtcan 2019). Not surprisingly, we obtained a **43.9 % higher WER** for **noisy** environments (Mean = 31.8, SD = 15.2). Amazon Transcribe had the best performance in quiet environments (Mean = 16.3, SD = 8.51). Whisper from OpenAI performed the equally well in quiet and noisy environments (Mean = 27.3, SD = 11.4 and M = 27.9, SD = 12.4 respectively). Google Speech-To-Text had the most difficulty in noisy environments ((M = 40.3, SD = 17.8).

We performed a Wilcoxon Signed Rank Test that compared the performance of the pairs of tools (Amazon Transcribe Vs Google Speech-To-Text, Whisper from OpenAI Vs Amazon Transcribe and Whisper from OpenAI Vs Google Speech-To-Text) for each of the conditions namely quiet and noisy environments. The results are summarized in Table 7.

Our results show that in quiet conditions, Amazon Transcribe significantly outperforms Whisper from OpenAI with a  $p = .01953$ , Cohen's  $d = 1.098$  (large), 95% CI [0.090, 2.106]. For noisy environments it's worth noting that Google Speech-To-Text had significantly higher mean WER values when compared to Amazon Transcribe  $p = .03125$ , a large effect size (Cohen's  $d = 0.841$ ) albeit with the 95% CI of Cohen's  $d$  spanning a zero (95% CI [-0.500, 2.183]). The rest of the comparisons were not significant. For more details on how we defined criteria to determine if the recording environments were noisy or quiet see [Appendix II](#).

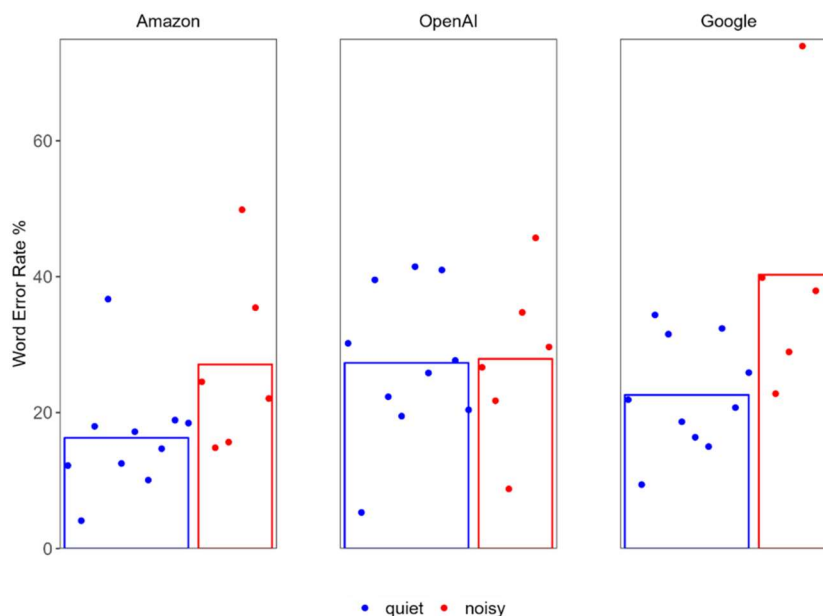


Figure 7 : Performance of Amazon Transcribe, Google Speech-To-Text and Whisper from OpenAI in quiet and noisy environments. The height of the colored bars represents the mean of the WER values for each category. The points show the individual WER values for each category, the width of the bar represents the size of the category.

ASR pair	Condition	p value	z statistic	Cohen's D	95 % CI for Cohen's D
Google Vs Amazon	Quiet	.01953	1.74	0.755 (medium)	[-0.216, 1.728]
OpenAI Vs Amazon	Quiet	.01953*	2.65	1.098 (large)	[0.090, 2.106]
OpenAI Vs Google	Quiet	.1934	1.06	0.472 (small)	[-0.480, 1.424]
Google Vs Amazon	Noisy	.03125	1.6	0.841 (large)	[-0.500, 2.183]
OpenAI Vs Amazon	Noisy	.6875	0.16	0.063 (negligible)	[-1.223, 1.349]
OpenAI Vs Google	Noisy	.09375	-1.28	-0.810 (large)	[-2.149, 0.527]

Table 7 : Results from the Wilcoxon Signed Rank Test for the three ASR models in quiet and noisy ambient conditions. \*Amazon Transcribe significantly outperforms Whisper from OpenAI in audios registered in quiet environments, indicated by its p value < 0.05, and large Cohen's d.

## 5.4 Difference between human and ASR transcripts

In our study some of the sources for the errors that ASR made lay in the differences in the way that humans and machines set transcription conventions, treat conversational elements of speech and handle rare or exotic words.

**Comparison of transcript lengths:** Both Google Speech-To-Text and OpenAI's Whisper yielded transcripts with fewer words compared to human transcriptions, which comprised 35,112 words. Specifically, Google Speech-To-Text produced a transcript of 33,691 words, marking a 4% reduction, while Whisper from OpenAI resulted in a 3.8% reduction with 33,748 words. Amazon Transcribe, on the other hand, had a slightly shorter transcript of 35,034 words, representing a 0.22% reduction.

**Transcript conventions:** By this we mean the guidelines adhered to by each transcribing party in generating transcripts. Spelling of proper nouns and compound word handling differed between humans and the three ASR models. Notably, there were disparities in how compound words were treated by humans and machines; for instance, "policymakers" was transcribed as a single word by ASRs, while humans segmented it into "policy makers." These words lack well-defined and universally agreed-upon written conventions. Consequently, when these terms gain prevalence in a corpus like our climate change corpus, they have the potential to

artificially elevate WER values. Furthermore, discrepancies in spelling for proper nouns were evident as seen in instances like Anderson, Andersson, Andersen, Ema, Emma, and Ima. While proper nouns like "Anderson," "Andersen," and "Andersson" were infrequent in this study and posed fewer challenges, their multiple acceptable spellings could present issues in different contexts, such as medicinal transcription or forensic applications.

### Treatment of disfluencies:

Both Google Speech-To-Text and Whisper from OpenAI produced shorter transcripts when compared to humans and this is partly explained by the fact that ASR systems tend not to transcribe conversational elements of speech like disfluencies ("um", "uh"). Excepting Amazon Transcribe the ASR systems used in this study failed to transcribe most of the disfluencies. Humans listeners although aware of disfluencies tend to miss conversational elements of speech during transcription (Mansfield et al. 2021). Amazon Transcribe had in fact transcribed 42 counts (13.7%) more of "um" when compared to humans. See Table 8 for more details.

Source of transcription	"um"	"uh"
Human	263	335
Amazon Transcribe	305	312
Google Speech-To-Text	5	0
Whisper from OpenAI	32	11

Table 8 : Differential treatment of disfluencies in transcription

### Treatment of reductions:

Reductions are also parts of conversational speech that we employ to facilitate pronunciation. Words like "gonna", "wanna" "gotta" are employed by humans instead of "going to", "want to", "got to" etc. Amazon Transcribe was the best at recognizing and transcribing reductions like "gonna" and was 19 (36.5%) counts higher than humans. See Table 9 below for more details.

Source of transcription	"Going"	"Gonna"	Total
Human	88	33	121
Amazon Transcribe	68	52	120
Google Speech-To-Text	98	23	121
Whisper from OpenAI	115	6	121

Table 9 : Comparison of reductions in transcription

### **Treatment of exotic names associated with climate change:**

ASRs had the most difficulty with exotic climate change related vocabulary. There were several words in our corpus that were associated with climate change phenomena that were “exotic”. Some of them like Turkana (Kenyan county) was transcribed as Trukana and Tukana by Whisper from OpenAI, and Tuana by Amazon Transcribe, Google Speech-To-Text failed to transcribe it. Garissa (Kenyan Village) was deleted by all systems except Whisper which came up with Karisa. Kehkashan, the first name of a climate change activist was transcribed as Kehkesha by Whisper from OpenAI and missed by the other systems. Marsabit another Kenyan county was missed by all three systems. Several other instances of these words are highlighted in [Appendix III](#).

### **Errors in spacing between words:**

Google Speech-To-Text had failed to introduce spaces between several words in its transcripts. We noticed this while analyzing transcripts, but we did not quantitate this.

## 6. Discussion

The present study investigated the performance of three ASR systems on a corpus of climate change discourses. We quantified performance using the popular metric used for ASR performance namely WER. Mean WER values indicate that Amazon Transcribe is the best performing model on the climate change corpus, followed by Whisper from OpenAI and Google Speech-To-Text. We find that factors like noise and type of video presentation had an impact on ASR performance. A fine-grained analysis of the transcripts revealed that ASRs differ in the way they treat linguistic elements like spelling and compound words with respect to humans. ASRs also treat conversational elements like disfluencies, reductions etc., differently. They struggle with “exotic” or rare climate change related words.

Our findings indicate that Amazon Transcribe has the best overall performance on the climate change corpus. It performed well across different presentation styles and showed robustness in noisy environments. It was the only system that transcribed conversational elements of speech like disfluencies and reductions. The superior performance with disfluencies exhibited by Amazon Transcribe might be due to Amazon’s experiment’s with the technique of adding synthetic disfluencies to their speech data to train their models to recognize them (Sen, Groves 2021). Conversational speech elements like reductions, disfluencies etc are important especially in ASR systems like virtual assistants because including them makes speech more human like (Lopez, Liesenfeld, Dingemanse 2022). Also, disfluencies might play an important role in expressing emotions in conversational speech. Studies have pointed to the inclusion of disfluencies while conducting emotional analysis of speech (Tian, Lai, Moore 2015; Johnson et al. 2023).

Google Speech-To-Text on the other hand has clearly emerged as the worst performing system in the model shown by lack of robustness in the presence of noise, failure to transcribe conversational elements of speech and its tendency to produce gross transcription errors. Whisper from OpenAI strikes a middle ground by performing consistently well across all categories of videos and robustness in noisy environments. It however like Google Speech-To-Text failed to transcribe conversational elements of speech.

We note here that all ASRs had difficulty with exotic climate change related vocabulary. This could have been because of two reasons, one the fact that these words were mostly spoken by non-native speakers of English, the second could have been because these words were not a part of the vocabulary on which these systems were trained on. We note that Whisper from OpenAI tends to guess these difficult words more than other systems. Our speculation is that this might be linked to the fact that Whisper is known to hallucinate when there is little or no audio at the end of speeches (Radford et al., 2022)

We note several limitations in this study. The small sample size hinders the generalizability of the study’s findings. The YouTube corpus has imbalances in terms of gender distribution and the representation of native and non-native speakers. We performed assisted transcription in this study. An inherent challenge in assisted transcript correction tasks is the potential for transcriber priming. Editors tend to trust the ASR system’s output and their corrections might be influenced by this, potentially leading to errors going unnoticed. Moreover, the YouTube corpus was transcribed by two different researchers from our team. Different transcribers may interpret and correct ASR output differently, introducing inconsistencies in the ground truth.

## 7. Conclusions

We conducted a comprehensive evaluation of three Automatic Speech Recognition (ASR) systems, assessing their performance on a YouTube corpus dedicated to the subject of climate change. All three ASR systems exhibited comparable levels of performance while exhibiting fine differences in the way they transcribe.

Amazon Transcribe performs best in semi-spontaneous speech registered in quiet settings and treats disfluencies and reductions best. Whisper's OpenAI performs consistently in quiet and noisy backgrounds. Google Speech-To-Text clearly offers the least advantages because it was the most affected by noise and produced several gross transcription errors like the omission of spaces between words.

Given the anticipated nature of our future study, primarily involving recorded self-interviews in quiet environments, the empirical data positions Amazon Transcribe as an advantageous selection.



## Bibliography

- AKSËNOVA, Alëna et al., 2021. How Might We Create Better Benchmarks for Speech Recognition? In : *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pp. 22–34. Online : Association for Computational Linguistics. August 2021. DOI 10.18653/v1/2021.bppf-1.4.
- ALDARMAKI, Hanan et al., 2022. Unsupervised Automatic Speech Recognition: A review. *Speech Communication*. Vol. 139, pp. 76–91. DOI 10.1016/j.specom.2022.02.005.
- ALLISON, Linn, 2016. Microsoft researchers reach human parity in conversational speech recognition. *The AI Blog* [online]. 18 October 2016. Retrieved from : <https://blogs.microsoft.com/ai/historic-achievement-microsoft-researchers-reach-human-parity-conversational-speech-recognition/> [accessed 29 April 2023].
- AMAZON, 2023. Amazon Transcribe – Speech to Text - AWS. *Amazon Web Services, Inc.* [online]. 2023. Retrieved from : <https://aws.amazon.com/transcribe/> [accessed 4 November 2023].
- ASSEMBLYAI, 2023. AssemblyAI | About. *AssemblyAI* [online]. 2023. Retrieved from : <https://www.assemblyai.com/> [accessed 4 November 2023].
- BANERJEE, Satanjeev and LAVIE, Alon, 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In : *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72 [online]. Ann Arbor, Michigan : Association for Computational Linguistics. June 2005. Retrieved from : <https://aclanthology.org/W05-0909> [accessed 16 April 2023].
- BAZILLON, Thierry, ESTÈVE, Yannick and LUZZATI, Daniel, 2008. Manual vs Assisted Transcription of Prepared and Spontaneous Speech. In : *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)* [online]. Marrakech, Morocco : European Language Resources Association (ELRA). May 2008. Retrieved from : [http://www.lrec-conf.org/proceedings/lrec2008/pdf/277\\_paper.pdf](http://www.lrec-conf.org/proceedings/lrec2008/pdf/277_paper.pdf) [accessed 29 April 2023].
- BROWN, Jennifer L., VANABLE, Peter A. and ERIKSEN, Michael D., 2008. Computer-assisted self-interviews: A cost effectiveness analysis. *Behavior Research Methods*. Vol. 40, no. 1, pp. 1–7. DOI 10.3758/BRM.40.1.1.
- CHIU, Chung-Cheng et al., 2018. *State-of-the-art Speech Recognition With Sequence-to-Sequence Models* [online]. arXiv:1712.01769. arXiv. arXiv:1712.01769. Retrieved from : <http://arxiv.org/abs/1712.01769> [accessed 10 April 2023]. arXiv:1712.01769 [cs, eess, stat]
- ERRATTAHI, Rahhal, EL HANNANI, Asmaa and OUAHMANE, Hassan, 2018. Automatic Speech Recognition Errors Detection and Correction: A Review. *Procedia Computer Science*. Vol. 128, pp. 32–37. DOI 10.1016/j.procs.2018.03.005.
- Fast Fourier transform, 2023 *Wikipedia* [online]. Retrieved from : [https://en.wikipedia.org/w/index.php?title=Fast\\_Fourier\\_transform&oldid=1179531012](https://en.wikipedia.org/w/index.php?title=Fast_Fourier_transform&oldid=1179531012) [accessed 27 October 2023]. Page Version ID: 1179531012

FENG, Siyuan et al., 2021. *Quantifying Bias in Automatic Speech Recognition* [online]. arXiv:2103.15122. arXiv. arXiv:2103.15122. Retrieved from : <http://arxiv.org/abs/2103.15122> [accessed 5 November 2023]. arXiv:2103.15122 [cs, eess]

FERRARO, Antonino et al., 2023. Benchmarking open source and paid services for speech to text: an analysis of quality and input variety. *Frontiers in Big Data* [online]. Vol. 6. Retrieved from : <https://www.frontiersin.org/articles/10.3389/fdata.2023.1210559> [accessed 18 November 2023].

GAIDA, Christian et al., 2014. Comparing Open-Source Speech Recognition Toolkits \*. .

GAROFOLO, John S et al., 1993. *DARPA TIMIT:: acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1*. Gaithersburg, MD : National Institute of Standards and Technology. NIST IR 4930. DOI 10.6028/NIST.IR.4930.

GEITGEY, Adam, 2020. Machine Learning is Fun Part 6: How to do Speech Recognition with Deep Learning. *Medium* [online]. 24 September 2020. Retrieved from : <https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a> [accessed 19 March 2023].

GEORGESCU, Alexandru-Lucian et al., 2021. Performance vs. hardware requirements in state-of-the-art automatic speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing*. Vol. 2021, no. 1, p. 28. DOI 10.1186/s13636-021-00217-4.

GILLICK, Larry and COX, Stephen, 1989. *Some statistical issues in the comparison of speech recognition algorithms*. journalAbbreviation: Proceedings of ICASSP

GODFREY, J.J., HOLLIMAN, E.C. and MCDANIEL, J., 1992. SWITCHBOARD: telephone speech corpus for research and development. In : *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 517–520 vol.1. March 1992. DOI 10.1109/ICASSP.1992.225858.

GOOGLE, 2023. Speech-to-Text: Automatic Speech Recognition. *Google Cloud* [online]. 2023. Retrieved from : <https://cloud.google.com/speech-to-text>, <https://cloud.google.com/speech-to-text> [accessed 4 November 2023].

GRAF, Michael et al., 2022. *Human and Automatic Speech Recognition Performance on German Oral History Interviews* [online]. arXiv:2201.06841. arXiv. arXiv:2201.06841. Retrieved from : <http://arxiv.org/abs/2201.06841> [accessed 29 April 2023]. arXiv:2201.06841 [cs, eess]

GUO, Jinxi et al., 2020. Efficient Minimum Word Error Rate Training of RNN-Transducer for End-to-End Speech Recognition. In : *Interspeech 2020*, pp. 2807–2811. ISCA. 25 October 2020. DOI 10.21437/Interspeech.2020-1557.

HACKING, Coen et al., 2023. The development of an automatic speech recognition model using interview data from long-term care for older adults. *Journal of the American Medical Informatics Association*. Vol. 30, no. 3, pp. 411–417. DOI 10.1093/jamia/ocac241.

HAIN, T. et al., 2005. Automatic transcription of conversational telephone speech. *IEEE Transactions on Speech and Audio Processing*. Vol. 13, no. 6, pp. 1173–1185. DOI 10.1109/TSA.2005.852999.

HARRINGTON, Lauren, 2023. Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. *Frontiers in Communication* [online]. Vol. 8. Retrieved from : <https://www.frontiersin.org/articles/10.3389/fcomm.2023.1165233> [accessed 7 October 2023].

JOHNSON, Sheri L. et al., 2023. Emotion-Triggered impulsivity relates to speech dysfluency during high arousal states. *Journal of Research in Personality*. Vol. 105, p. 104397. DOI 10.1016/j.jrp.2023.104397.

KĚPUSKA, Veton, 2017. Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). *International Journal of Engineering Research and Applications*. Vol. 07, no. 03, pp. 20–24. DOI 10.9790/9622-0703022024.

KIM, Joshua Y et al., 2019. A Comparison of Online Automatic Speech Recognition Systems and the Nonverbal Responses to Unintelligible Speech. .

KOENECKE, Allison et al., 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*. Vol. 117, no. 14, pp. 7684–7689. DOI 10.1073/pnas.1915768117.

LAZARO, Gina. Understanding Gender and Racial Bias in AI. *ALI Social Impact Review* [online]. Retrieved from : <https://www.sir.advancedleadership.harvard.edu/articles/understanding-gender-and-racial-bias-in-ai> [accessed 5 November 2023].

LOPEZ, Alianda, LIESENFELD, Andreas and DINGEMANSE, Mark. Evaluation of Automatic Speech Recognition for Conversational Speech in Dutch, English, and German: What Goes Missing? *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*.

MAGALHÃES, Regis Pires et al., 2022. Evaluation of Automatic Speech Recognition Approaches. *Journal of Information and Data Management*. Vol. 13, no. 3. DOI 10.5753/jidm.2022.2514.

MANSFIELD, Courtney et al., 2021. Revisiting Parity of Human vs. Machine Conversational Speech Transcription. In : *Interspeech 2021*, pp. 1997–2001. ISCA. 30 August 2021. DOI 10.21437/Interspeech.2021-1908.

MICROSOFT, 2023. Speech to Text – Audio to Text Translation | Microsoft Azure. [online]. 2023. Retrieved from : <https://azure.microsoft.com/en-us/products/ai-services/speech-to-text> [accessed 4 November 2023].

Nyquist frequency, 2023 *Wikipedia* [online]. Retrieved from : [https://en.wikipedia.org/w/index.php?title=Nyquist\\_frequency&oldid=1145284140](https://en.wikipedia.org/w/index.php?title=Nyquist_frequency&oldid=1145284140) [accessed 24 October 2023]. Page Version ID: 1145284140

OLAH, Christopher. Understanding LSTM Networks -- colah's blog. [online]. Retrieved from : <https://colah.github.io/posts/2015-08-Understanding-LSTMs/> [accessed 27 October 2023].

O'SHAUGHNESSY, Douglas, 2008. Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognition*. Vol. 41, no. 10, pp. 2965–2979. DOI 10.1016/j.patcog.2008.05.008.

PALOMÄKI, Kalle, 2014. PPT - Automatic speech recognition PowerPoint Presentation, free download - ID:1590800. *SlideServe* [online]. 10 July 2014. Retrieved from : <https://www.slideserve.com/lona/automatic-speech-recognition> [accessed 22 October 2023].

PAPINENI, Kishore et al., 2001. BLEU: a method for automatic evaluation of machine translation. In : *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, p. 311. Philadelphia, Pennsylvania : Association for Computational Linguistics. 2001. DOI 10.3115/1073083.1073135.

PENTLAND, Steven J. et al., 2022. Does accuracy matter? Methodological considerations when using automated speech-to-text for social science research. *International Journal of Social Research Methodology*. pp. 1–17. DOI 10.1080/13645579.2022.2087849.

RADFORD, Alec et al. Robust Speech Recognition via Large-Scale Weak Supervision. .

REV, 2023. Speech to Text API | Speech Recognition Service. *Rev AI* [online]. 2023. Retrieved from : <https://www.rev.ai/> [accessed 4 November 2023].

SAMPAIO, Matheus Xavier et al., 2021. Evaluation of Automatic Speech Recognition Systems. In : *Anais do Simpósio Brasileiro de Banco de Dados (SBBD)*, pp. 301–306. SBC. 4 October 2021. DOI 10.5753/sbbd.2021.17889.

SAON, George et al., 2015. *The IBM 2015 English Conversational Telephone Speech Recognition System* [online]. arXiv:1505.05899. arXiv. arXiv:1505.05899. Retrieved from : <http://arxiv.org/abs/1505.05899> [accessed 16 April 2023]. arXiv:1505.05899 [cs]

SEN, Priyanka and GROVES, Isabel, 2021. Semantic Parsing of Disfluent Speech. In : *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1748–1753. Online : Association for Computational Linguistics. 2021. DOI 10.18653/v1/2021.eacl-main.150.

SHOR, Joel et al., 2023. *Clinical BERTScore: An Improved Measure of Automatic Speech Recognition Performance in Clinical Settings* [online]. arXiv:2303.05737. arXiv. arXiv:2303.05737. Retrieved from : <http://arxiv.org/abs/2303.05737> [accessed 7 October 2023]. arXiv:2303.05737 [cs, eess]

SIEGERT, Ingo et al., 2020. Recognition Performance of Selected Speech Recognition APIs – A Longitudinal Study. In : KARPOV, Alexey and POTAPOVA, Rodmonga (eds.), *Speech and Computer*, pp. 520–529. Cham : Springer International Publishing. Lecture Notes in Computer Science. ISBN 978-3-030-60275-8. DOI 10.1007/978-3-030-60276-5\_50.

SPEECHMATICS, 2023. AI Speech Technology | Speech-To-Text API. *Speechmatics* [online]. 2023. Retrieved from : <https://www.speechmatics.com> [accessed 4 November 2023].

STANFORD ONLINE, 2015. *Stanford Seminar - Deep Speech: Scaling up end-to-end speech recognition* [online]. 5 February 2015. Retrieved from : <https://www.youtube.com/watch?v=P9GLDezYVX4> [accessed 10 March 2023].

SZYMAŃSKI, Piotr et al., 2020. WER we are and WER we think we are. In : *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3290–3295. Online : Association for Computational Linguistics. 2020. DOI 10.18653/v1/2020.findings-emnlp.295.

TIAN, Leimin, LAI, Catherine and MOORE, Johanna D, 2015. RECOGNIZING EMOTIONS IN DIALOGUES WITH DISFLUENCIES AND NON-VERBAL VOCALISATIONS.

TRABELSI, Asma et al., 2022. Evaluation of the efficiency of state-of-the-art Speech Recognition engines. *Procedia Computer Science*. Vol. 207, pp. 2242–2252. DOI 10.1016/j.procs.2022.09.534.

TÜSKE, Zoltán, SAON, George and KINGSBURY, Brian, 2021. *On the limit of English conversational speech recognition* [online]. arXiv:2105.00982. arXiv. arXiv:2105.00982. Retrieved from : <http://arxiv.org/abs/2105.00982> [accessed 28 October 2023]. arXiv:2105.00982 [cs, eess]

WHISPER, 2023. Introducing Whisper. [online]. 2023. Retrieved from : <https://openai.com/research/whisper> [accessed 4 November 2023].

WIT, 2023. Wit.ai. [online]. 2023. Retrieved from : <https://wit.ai/> [accessed 4 November 2023].

XIONG, W. et al., 2018. The Microsoft 2017 Conversational Speech Recognition System. In : *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5934–5938. April 2018. DOI 10.1109/ICASSP.2018.8461870.

YU, Dong and DENG, Li, 2015. *Automatic Speech Recognition: A Deep Learning Approach*. London : Springer London. Signals and Communication Technology. ISBN 978-1-4471-5778-6.

YURTCAN, Yaser, 2019. *Performance evaluation of real-time noisy speech recognition for mobile devices* [online]. Master Thesis . Middle East Technical University. Retrieved from : <https://open.metu.edu.tr/handle/11511/28055> [accessed 4 April 2023]. journalAbbreviation: Mobil cihazlarda gerçek zamanlı gürültülü konuşma tanıma performans değerlendirilmesi.

ŻELASKO, Piotr et al., 2019. *Towards Better Understanding of Spontaneous Conversations: Overcoming Automatic Speech Recognition Errors With Intent Recognition* [online]. arXiv:1908.07888. arXiv. arXiv:1908.07888. Retrieved from : <http://arxiv.org/abs/1908.07888> [accessed 25 February 2023]. arXiv:1908.07888 [cs]

ZHANG, Tianyi et al., 2020. *BERTScore: Evaluating Text Generation with BERT* [online]. arXiv:1904.09675. arXiv. arXiv:1904.09675. Retrieved from : <http://arxiv.org/abs/1904.09675> [accessed 7 October 2023]. arXiv:1904.09675 [cs]

## Appendix 1: Metadata of the YouTube Corpus

	Linguistic Features					Demographic Features		Technical Features					Other Metadata	
Name of the MP3 Files	Video Presentation	Style of Speech	Accent	Number of Native Speakers	Number of non native speakers	Male Speakers	Female Speakers	Equipment	Background	Number of Channels	Number of Minutes	Size of the MP3 File	Date of upload of video	Topic
webinar_1	Webinar	Semi spontaneous	American	1	0	1	0	self recorded	quiet	single	19:52	18.1MB	Nov 7, 2022	Climate change litigation and the right to a healthy environment
webinar_2	Webinar	Semi spontaneous	American	1	0	1	0	self recorded	quiet	single	11:41	10.7MB	Nov 17, 2023	Are we optimistic about climate change? with Paul S.
webinar_3	Webinar	Semi spontaneous	American	2	1	1	2	self recorded	quiet	multi	11:52	16.3MO	Oct 18, 2023	Climate change AI community, with David Rolnick
interview_Studio_1	Interview-quiet	Spontaneous	American	2	0	1	1	self recorded	quiet	two	09:26	8.64MB	Oct 6, 2022	Climate Change and Hurricanes: An interview with climate scientist Dr. Daniel Gilford
interview_Studio_2	Interview-quiet	Spontaneous	American	2	0	2	0	studio equipment	quiet	multi	13:31	12.3MB	Apr 8, 2023	Is climate change destroying the Amazon? Paul Rosolie and Lex Fridman
interview_Studio_3	Interview-quiet	Spontaneous	American	2	0	1	1	self recorded	quiet	single	07:35	10.9MO	02.juin.23	Navigating climate change with Dougald Hine : making
interview_Studio_4	Interview-quiet	Spontaneous	American	3	0	1	2	studio equipment	quiet	multi	13:46	19.8MO	16.juin.23	Lionel Shriver : The weather isn't climate change. SpectatorTV
interview_Noisy_1	Interview-quiet	Spontaneous	American	2	0	2	0	studio equipment	quiet	two	18:01	24.7MB	Mar 14, 2023	President Biden on Mobilizing Youth, Climate Change & Human Rights : The daily show

	Linguistic Features					Demographic Features		Technical Features					Other Metadata	
Name of the MP3 Files	Video Presentation	Style of Speech	Accent	Number of Native Speakers	Number of non native speakers	Male Speakers	Female Speakers	Equipment	Background	Number of Channels	Number of Minutes	Size of the MP3 File	Date of upload of video	Topic
interview_Noisy_2	Interview-noisy	Semi spontaneous	Various	13	14	8	19	self recorded	noisy	multi	14:26	20,8MO	08.août.22	Youth voices on climate - ActNowFilm
interview_Street_1	Interview-noisy	Spontaneous	British	10	0	5	5	studio equipment	noisy	multi	07:20	6.72MB	Jun 16, 2019	Street interview with climate protesters in London : Eco travel with David
interview_Street_2	Interview-noisy	Spontaneous	American	10	11	17	4	studio equipment	noisy	multi	04:58	7,2MO	14.févr.17	Jesse Watters asks the folks about the climate on 'The O'Reilly Factor'
interview_Street_3	Interview-noisy	Spontaneous	Foreign	0	10	5	5	studio equipment	quiet	multi	24:06	34,7MO	09.déc.22	Street debate : the impact of climate change on rural Kenya. "77% street debate"
conversation_1	Conversation	Spontaneous	American	2	0	1	1	studio equipment	quiet	two	16:41	15.2MB	Jun 7, 2023	The implications of climate change - author conversation with Dr. Frederica Perera
conversation_2	Conversation	Spontaneous	American	4	0	3	1	self recorded	quiet	multi	08:51	8.11MB	Dec 5m 2022	A conversation on climate change: Student opinions
conversation_3	Conversation	Spontaneous	American	4	0	2	2	studio equipment	noisy	multi	11:57	17,2MO	23.janv.23	Creating climate change content : "there's misinformation & disinformation"
conversation_4	Conversation	Spontaneous	American/Indian	5	0	1	4	studio equipment	noisy	multi	16:21	23,6MO	26.mai.23	Are climate doomers right ?

## Appendix 2 : Definition of metadata

How metadata was defined for each video that was used in this study :

### 1. Background

- **Clean:** No background noise OR music played during the video
- **Noisy:** Some background noise and/or music played during the video/indigenous language in the background (for which translation was provided)
- **Very Noisy:** Recorded in the street during a climate change strike

### 2. Equipment

- **Self-Recorded:** Recorded directly on the laptop using its inbuilt mike (No visible microphone)
- **External Microphone:** Microphones used for recording, visibly pinned to clothes or held by hand

### 3. Number of channels

- **Single Channel :** If one speaker dominated the conversation it was marked as single channel, if there was a second speaker he or she spoke almost to very little
- **Two Channel:** Two speakers only
- **Multi Channel :** More than two speakers

### 4. Speech Style

- **Semi spontaneous :** We considered webinars to be semi spontaneous speech as speakers had a visual presentation before them while talking
- **Spontaneous :** Conversations and interviews were considered to be spontaneous if no visible material was used to assist the speaker while talking

### 5. Video Presentation

- **Webinars:** One person presenting with or without the presence of a moderator
- **Interview:** One person asking questions and other answering
- **Conversation:** More than two people speaking to each other

### 6. Total number of speakers

We counted the total number of speakers for the entire length of the video

### 7. Ratio of male to female speakers

Ratio of male to female speakers. This is strictly a ratio of male to female speakers, it does not take into account the amount of speech contributed by each sex in the video. This tended to



vary a lot. There were videos in which women were present only as moderators and hence participated very minimally.

### Appendix 3 : Examples of rare words and spelling convention differences

Examples of Rare Words that were mis transcribed	Examples of Spelling convention differences
Garissa	Anderson, Andersson, Andersen
Kehkashan	Cos (because)
Kobe	Fredricka, Fredrika
La Rose	Ema, Emma
Lodwar	
Marsabit	
Marthur	
Trukana	