Google Chirp vs. Whisper: Evaluating ASR performance on Dutch Native vs. Non-Native Teenager Speech

Anish Jaggoe

Supervisors: Dr. Odette Scharenborg, YuanYuan Zhang

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Anish Jaggoe
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg, YuanYuan Zhang, Catharine Oertel

Abstract

Automatic Speech Recognition (ASR) systems have become increasingly important for society, yet their performance varies significantly across different diverse speaker groups. With a significant non-native population in the Netherlands, it is crucial that ASR systems accurately recognize diverse speech. Commercial state-of-the-art ASR systems are yet under-explored in their performance on Dutch diverse speech. This study evaluates the performance of two recently developed and affordable ASR systems, Google Chirp and OpenAI's Whisper, on speech from native and non-native Dutch teenagers. This research evaluates the recognition accuracy of these ASR systems and identifies common transcription errors. The results show slightly worse performance compared to previous research on non-native speech, and Whisper performing generally better than Google Chirp on the speaker groups.

1 Introduction

State-of-the-art (SotA) Automatic Speech Recognition (ASR) systems have become increasingly used in society through the use of intelligent voice assistants, search engines or by healthcare providers for medical documentation [1, 2]. Regarding the latter, Ajami [2] further concluded that the implementation of speech-to-text technology has improved the quality and efficiency of the documentation of health information and records.

With this increase and the importance of ASR technology in certain fields, the requirement for ASR systems to deal with many different types of speech is crucial for one of the most inclusive societies such as the Netherlands, according to the Kantar's Inclusion Index[3]. A large group of at least 2.6 million people in the Netherlands is born outside of the Netherlands, of which 1.7 million are born outside of non-European [4]. This group should also be able to use speech-to-text technology to use the benefits the ASR systems provide. While the ASR systems are currently equipped to transcribe the Dutch language well, ASR systems currently underperform different speaker groups like non-native Dutch speakers[5].

Park and Culnan [6] researched ASR performance on native and non-native English speakers and found that speech recognition accuracy is lower for non-natives than for natives. This was due to a mismatch in pronunciation between the two groups and specific pronunciation errors and the more considerable variability in non-native speech. The pronunciation errors were likely caused by their mother tongue [5]. The lower performance in this study opens up the possibility for more extensive research into SotA ASR systems on non-native speech.

Koenecke et al. [7] has also recently researched commercial ASR systems from Amazon, Google, IBM, Apple, and Microsoft on American English [7]. This study found that the speech of black Americans was consistently recognized worse than that of white Americans. While this research was an advancement in the field of ASR technology on diverse speech, Koenecke et al. [7] mentioned the problem of this performance being caused by disparities in the two separate databases used in the research. The difference in performance could be geographically related, rather than speaker group related.

Despite the limitation in the research on American English, this study helped discover that these ASR systems do not perform equally well for Dutch as well. Research into speech from Dutch diverse speaker groups has already been conducted. Research was conducted on locally developed ASR systems to study and attempt to quantify bias ASR systems using Dutch and Chinese speech corpora, like the JASMIN-CGN among other [8, 9]. Following Feng et al. [8, 9], a study was done on the recognition of diverse Dutch speaker groups and the performance on the models Wav2Vec2 and Whisper, developed by OpenAI[10], using the same dataset. This showed that Whisper outperformed Wav2Vec2 for all speech groups.

Evaluating the performance of these SotA ASR systems on diverse speaker groups can be beneficial for the improvement and further development of ASR technology, as this remains an understudied area. ASR performance can differ significantly in performance on certain speaker groups, as seen from previous works [7, 8, 9, 10]. The purpose of this research is to extend upon earlier studies and test the performance of commercial SotA systems on Dutch diverse speaker groups.

In this paper, the performance of two SotA ASR systems, Google Chirp [11] and Whisper, by OpenAI [12], are evaluated on Dutch diverse speech. Comparing the Chirp and Whisper systems is motivated by the costs and recency of the systems. The evaluation of the Chirp model remains relatively unexplored, as it was released recently in 2023. Due to affordability and being developed by Google with its widely used ecosystem, the Google Chirp model is bound to play a crucial for further ASR development. In contrast, Whisper's motivation lies in the rapidly increasing importance of OpenAI. Their importance can be attributed to their increasing user base, demonstrated by their renowned large language model, ChatGPT, with over 100 million weekly users within two years of its release[13]. This makes evaluating Whisper attractive, next to the ability to run the Whisper model at no cost.

This introduces the main research question: **How well do Google Chirp and Whisper recognise speech of native Dutch teenagers compared to non-native Dutch teenagers?** To address this research question, the following sub-questions will be answered in this paper:

RQ1 How do Google Chirp and Whisper perform in recognizing speech from native and non-native teenagers?

RQ2 What are common transcription errors for native and non-native Dutch teenagers with Google Chirp and Whisper?

RQ3 How do age and gender influence the recognition accuracy of the ASR systems between native and non-native speakers?

RQ4 How does the performance of Google Chirp and Whisper compare to previous research on Dutch diverse speech?

Comparing the recognition accuracy of native speech to non-native speech for all age groups introduces data disparities, due to the age groups in the database aligning to each other. Comparing native and non-native teenager speech ensures that the age groups are roughly equal. Additionally, teenager speech resulted in the best recognition performance in previous works [8, 10], which helps with my evaluation accuracy for native and non-native speech.

Regarding the purpose of RQ3, the influence of gender and age on ASR performance has been explored before. Previous research has found that female speech is better recognized than male speech by ASR systems [9, 14]. Additionally, the age difference should be taken into account for the performance of the ASR systems, as ASR technology has seen challenges in recognizing child speech before [15].

The following section will cover the methodology, which will provide details on the database, ASR systems, and evaluation metrics that will be used in this research, along with the setup of the experiment conducted. Section 3 will summarise the results of the experiment. This will be analyzed and discussed in Section 4. Section 4 will also cover any limitations found during this research. Section 5 will answer the main research question and sub-questions separately and will provide recommendations for future research. Finally, Section 6 will discuss how the ethical aspects of this study will be addressed.

2 Methodology

This section will discuss the method and steps taken to answer the research questions. Section 2.1 describes the JASMIN-CGN, the dataset used for the study. This is followed by Section 2.2 which describes the ASR systems used in this study, Google Chirp and Whisper by OpenAI. Section 2.3 will cover the metrics used to evaluate the results from the Chirp and Whisper models. Lastly, Section 2.4 will explain the experiment conducted in this study.

2.1 JASMIN-CGN

The JASMIN-CGN corpus is an extension of the CGN corpus and will be the main dataset used for this research. The JASMIN corpus contains speech The data from this corpus is annotated by speaker groups, gender, age, nativeness, native language, proficiency in Dutch, region, and dialect. I use speech from the following groups:

- Dutch teenagers (DT), age 12-18 years, 12h 21m of speech
- Non-native teenagers (NNT), age 11-18 years, 12h 21m of speech

The DT group consists of 59 speakers (30 male, 29 female) and the NNT group of 52 speakers (25 male, 27 female). In Table 1 the distribution of the number of speakers per speaker group can be found per age.

Table 1: Distribution of speakers across different age groups. The table presents the number of speakers within two groups DT and NNT for the

Age	DT	NNT
11	0	1
12	9	2
13	8	11
14	9	11
15	6	13
16	6	9
17	1	3
18	2	3

The speech in the JASMIN corpus is divided into two types of speech: read speech and human-machine interaction speech, hereafter HMI speech. Both will be used in the experiment to have a reliable test of performance.
The JASMIN corpus provides annotation to the speech in the corpus, with intervals based on utterances. These intervals are less than thirty seconds, making it suitable to feed the ASR systems the audio of these intervals and compare them to the reference text. In Section 2.4 I will further explain the setup of the experiments in this study.

2.2 ASR Models

I will be evaluating the performance of two SotA ASR systems: Google Chirp and Whisper.

Google Chirp
The Google Chirp model is the latest speech-to-text model released in 2023. Chirp is developed by Google AI and integrated into Google Cloud's Speech API [11].

One of the key innovations of the Google Chirp model is its integration of self-supervised learning methods. This approach makes the model able to learn from unlabeled data, reducing the dependency on manually annotated datasets. As a result, Chirp can adapt to diverse linguistic contexts and accents, possibly providing more accurate transcriptions across hundreds of languages.

Chirp is designed as a generic, largely pre-trained model and is trained using mainly unlabeled YouTube-based audio and both labeled and unlabeled speech from public datasets. This means that all training on the model is done using in-domain datasets and no training is done using private data. Chirp offers a model adaptation boost feature, assigning a weight to certain provided phrases in the process. I will be testing the Chirp system as is, meaning using the model without the use of this feature.

Whisper
The Whisper model is an ASR system developed by OpenAI. Whisper also employs self-supervised learning methods, where the model learns representations from unlabeled audio data, like the Chirp model. Unlike Chirp, it makes use of deep

learning techniques and pre-training on diverse, multilingual datasets.

For Whisper, the *'whisper-large-v3'* model is used to transcribe the text. This is the latest model developed by OpenAI that consists of 1.55B parameters [12]. The model is executed on powerful a GPU to enhance computational performance and efficiency with parallel processing. To have a meaningful comparison, the Whisper system will also be used as is, without fine-tuning. This way, the generic performance of both Google Chirp and Whisper can be compared.

2.3 Evaluation Metrics

The performance is evaluated in terms of Word Error Rate (WER) and Character Error Rate (CER). These 2 different metrics are used to evaluate and compare the performance of the ASR models. WER is the percentage of words transcribed incorrectly by the ASR. The WER is calculated by adding the Substitutions (S), Insertions (I), and Deletions (D) together from a transcript and dividing this by the total number of words in the reference text (N), which looks as follows:

$$WER = \frac{S + I + D}{N} \times 100\%$$

WER is used as a generic metric for comparing the performance of the systems on diverse speech. The CER is, similar to WER, the percentage of characters that have been transcribed incorrectly by the ASR. The formula for CER is similar to the formula for the WER, instead looks at character level. The reason for evaluating the CER is for cases with short transcriptions. In the case that a transcription contains few words, the WER might be disproportionately high. The CER offers a more stable result with few words.

2.4 Experimental Setup

Following the works of [10] and [16], this research will follow a similarly structured experiment.

The speech files are split into speech segments and run these through the Google Chirp and Whisper systems. When recognizing speech using the Google Chirp system, a constraint is that a speech file must be under 60 seconds in duration if the file is processed from local storage rather than from Google Cloud Storage. By segmenting the speech files, the files are kept on local storage to ensure that the sensitive speech data remains protected during processing. The speech files will be segmented based on the utterance intervals given in the reference annotations in the JASMIN corpus. These segments are then processed by the Chirp and Whisper systems. During the experiment, both models were set with language code 'nl-NL', to prevent the models from automatically detecting the speech language and enforce the models to transcribe to Dutch.

The WER and CER are then computed for the transcriptions by both ASR systems. The transcriptions are first normalized, removing capital letters, punctuation and other (strings of) symbols. The transcriptions are then compared to the reference annotation for the WER. For the CER, the transcription is transformed to a list of characters with white spaces removed, and compared to the transformed list of characters for the reference text. The Average WER and CER will be computed per speech type and for each speaker group. For further analysis related to RQ3, the average WER is computed for male and female speakers separately and per age.

To address RQ2, there was no systematic method to evaluate common errors available in the short time available next to the main focus of this research. Therefore, many transcriptions were checked by hand to scan for deviations between the transcriptions and reference text. Any deviations that are apparent among multiple files and any inexplicable deviations are documented and analyzed. 50 out of 232 transcriptions are checked manually: 25 read speech, 25 HMI speech and for both speech types 12 DT speakers and 13 NNT speakers. This is to have ample data by checking at least 10% of all transcriptions, while still leaving room for other analysis.

3 Results

The average WER and CER of the experiment can be found in Table 2. The results of the research by Feng et al. [8] using their proposed TDNNF model and the results by Fuckner et al. [10] on Whisper are also presented in this table for comparison. Table 2 displays the results for the DT and NNT speaker groups for all mentioned ASR systems. Since the studies by Feng et al. [8] and Fuckner et al. [10] did not employ the CER as an evaluation metric, these missing results will be displayed with a dash symbol ('-') in the table.

Table 2: Average Word Error Rate (WER) and Character Error Rate (CER) in percentage (%) across the Google Chirp and Whisper ASR systems. Categorized by the read speech (RS) and Human-Machine Interaction (HMI) speech types in JASMIN, for speaker groups Dutch Teenagers(DT) and Non-native Teenagers (NNT). Lower WER means better performance. Unavailable results are displayed with a '-'.

ASR		WER		CER	
		DT	NNT	DT	NNT
Chirp	RS	**22.9**	30.5	**15.0**	19.0
	HMI	**34.1**	45.1	**19.2**	29.0
Whisper	RS	**16.4**	23.2	**10.7**	16.4
	HMI	**64.5**	67.6	52.2	**51.7**
TDNNF[8]	RS	**14.0**	42.0	-	-
	HMI	**22.8**	42.5	-	-
Whisper[10]	RS	**8.0**	18.8	-	-
	HMI	**13.3**	25.0	-	-

All results show the two ASR systems performed better on the DT group than on the NNT group. A good performance of the ASR system can be defined as a WER between 0% and 10%. A 20% WER is perceived as an acceptable result and any WER >30% is considered poor recognition accuracy [17]. Following this statement, the off-the-shelf results for Whisper can be considered acceptable for read speech DT and NNT groups. In contrast, the results for Chirp for the DT and NNT speaker groups for both read

speech and HMI speech as well as the results for the DT and NNT speaker groups for Whisper results on HMI speech are quite poor. Whisper's results on HMI speech are a significant outlier amongst the other results. The average 66.1% WER on HMI speech for both DT and NNT speakers is quite a terrible result, which is not supported by any results of previous research using the JASMIN corpus.

Further evaluation is done based on gender and age. Table 3 shows the results for the WER per DT or NNT speaker group and gender. Table 4 displays the WER per speaker group and age group. The number of speakers per age group is repeated in this table for ease of reference. These results are only based on the read speech data, as the HMI speech results by Whisper would be outliers in these results and would show disproportionate results.

Table 3: WER per speaker group in percentage (%), grouped by (F)emale or (M)ale speaker on read speech, for Google Chirp and Whisper. The best WER between M or F per speaker group and ASR system is in **bold** and the best overall result is in ***bold and italic***.

Group	Gender	Chirp	Whisper
DT	F	24.1	17.1
	M	**21.7**	***15.7***
NNT	F	**27.3**	***22.2***
	M	34.0	24.3

The results in table 3 show better performance for male speakers than female speakers for the DT speakers. This difference differs from previous results, where female speakers were the foremost recognized speakers. For NNT, female speakers seem to be recognized better. The difference between the female NNT speakers to the male NNT speakers seems to be quite large for Google Chirp. This difference does not appear for the Whisper results.

The results for the WER per age display a difference per age group for the DT speakers, compared to the age groups for NNT speakers which have evenly distributed error rates. For the other age groups for DT speakers recognized by Google Chirp, the age groups 15, 16, and 18 have a WER of $<$30%, lower than the age groups 12, 13, 14, and 17 which have a WER of $>$30%. Results for Whisper show age groups 15 and 16 with a $<$20% WER and the remaining age groups for DT speakers with WER of $>$20%. A noticeable result is that of the sole 17-year-old speaker. Both ASR systems appear to have a bad performance on this speech. For the NNT speakers for Chirp, age groups of 11, 12, 14, and 18 have the best result with WER $<$30%. Age groups 14, 15, 16, and 17 have a WER $>$30%. For Whisper, all age groups outside of the 12-year-old speakers have a WER between 20 % and 30%.

Table 5 presents the average time each ASR system required in seconds to transcribe the speech from the DT and NNT speaker groups, categorized by read and HMI speech.

Table 4: WER per speaker group in percentage (%) for the read speech speech transcriptions, grouped by Dutch Teen (DT) and Non-native Teenager (NNT) and per age in years. The results are displayed for Google Chirp (left) and Whisper (right). Values under N display the number of speakers for the age group

Group	Age	N	Chirp	Whisper
DT	12	9	31.9	**21.9**
	13	8	30.5	**24.5**
	14	9	32.5	**14.6**
	15	6	15.8	**8.8**
	16	6	**18.3**	20.1
	17	1	62.4	**49.6**
	18	2	27.5	**21.8**
NNT	11	1	23.8	**20.3**
	12	2	24.5	**14.1**
	13	11	33.0	**21.9**
	14	11	26.7	**25.2**
	15	13	35.1	**24.1**
	16	9	33.9	**22.0**
	17	3	34.9	**27.4**
	18	3	**6.2**	22.8

This is displayed with the average duration of total speech from read and HMI speech segments. This is calculated for every second of speech, removing silences from audio. This shows that Whisper transcribes faster than Google Chirp on both read and HMI speech, on average requiring less time than the duration of speech.

Table 5: Average transcription time for the speech files, in seconds, compared to average speech duration in seconds. Displayed for RS and HMI speech files, for Google Chirp and Whisper.

Speech type	Avg. duration	Chirp	Whisper
RS	580.7	911.2	**19.1**
HMI	75.1	405.9	**15.8**

Table 6: Common error types in transcriptions, comparing original text to transcriptions by Google Chirp and Whisper. Unavailable data is noted by '-'.

Error type	Reference text	Chirp	Whisper
1	'ravage'	'rafage'	'ravage'
2	'als je m'	'als je een'	-
3	' 't '	'het'	'het'
4	'half drie'	'2:30'	'half drie'
5	'uhm'	-	'hmm'
6	'herfst'	'härft'	'herfst'

Table 6 shows a list of types of errors commonly found in

arbitrarily selected transcriptions. Each error is portrayed by an example found in the transcriptions. Error 1 is the most common type of error found. Speech is transcribed using different spelling depending on the articulation or tone of the speech. These errors are more common for NNT speech than DT speech. Error 2 is an error common for Whisper. The off-the-shelf Whisper system analyzes the speech to capitalize names and starts of sentences, and to add punctuation. Next to this, Whisper removes any repetition of speech found in a speech segment. This results in speech missing in transcription.

Error type 3 is found in Google Chirp and Whisper transcriptions. Both ASR systems transcribe the full word 'het' instead of the abbreviated " 't ". This error appears more often for DT speakers than for NNT speakers. Errors 4 and 5 are 6 are unique to Google Chirp. For error 4, Google Chirp analyzes speech similar to Whisper, transcribing time marks as timestamps, contrary to Whisper which transcribes this to text. Error 5 implies that stopwords like 'uhm' or 'hmm' do not get transcribed by Google Chirp. Finally, in some segments, speech is translated and transcribed to different languages despite being given an initial language code. This results in error 6, where speech is transcribed and translated to German.

4 Discussion

Comparing the results from this study for Whisper with the results by Fuckner et al.[10] results in unexpected differences. The difference in WER between the two results is quite significant. This difference in performance may appear because of the studies using two different Whisper models. This study made use of the *'whisper-large-v3'* model, while Fuckner et al [10] employed the *'whisper-large-v2'* model. Both models were trained using the same amount of parameters, the *v2* model, however, was trained on 680K hours of labeled speech [12] and the *v3* model on 1M hours of weakly labeled and 4M hours of pseudo labeled audio collected using *'whisper-large-v2'*[18]. This may have led to overfitting for the *v3* model, which in turn decreased the performance of the model.

This assumption could explain the nearly 8% difference in the Read speech performance between the two studies. However, a near 50% difference is quite significant. Next to this, the results between DT and NNT speech differ little, which does not follow previous results by Feng et al. [8] or Fuckner et al. [10]. This may indicate that something went wrong during the experiment with Whisper.

The results of the read speech transcriptions were used to better evaluate the results based on gender and age, instead of the results of both read and HMI speech. This was to help deliver an objective view of the results.

A correlation can be found between the WER and CER results in Table 2. This shows that there is a better CER compared to the respective WER for the ASR system and speaker group, while still reaching the same conclusion on performance that can be reached from the WERs. Atypical, Whisper on HMI speech has a better CER for NNT speech than DT speech, contrary to the WER which performs better on DT than NNT speech. However, the difference in results is not significant, unlike that of the other CER results. This does not follow the results of Google Chirp for CER on HMI speech. This further supports the error in the evaluation of Whisper HMI speech.

The errors found in the transcriptions in table 6 by Google Chirp and Whisper affect the performance evaluation significantly. I would however argue that Error types 2 through 6 have a positive function in a general setting of using the ASR systems. Google Chirp for example makes relative time expressions like 'half drie'(translated 'half past 2') easier to read in transcriptions by converting these to absolute time expressions like '2:30'. Korvorst et al. [19] did research into how Dutch native speakers responded to relative and absolute time expressions. They found that absolute time expressions were processed the quickest with the least amount of errors.

Furthermore, Whisper transcriptions have an intriguing feature. Whenever a speech segment has inaudible parts for the ASR system, Whisper analyzes the inaudible parts and inserts a descriptive transcription of this sound in the predefined language. As this is an insertion, it is not considered a transcription error in this study. As an illustration, music is often transcribed as "Muziek" (translated "music") and noise as "GELUID VAN MACHINES"(translated "machine sounds"). While this is an interesting feature that can be solely researched, the descriptive transcriptions are additions to the full transcription and therefore affect the evaluation of the transcriptions.

Finally, running the Whisper model on a GPU allowed for fast execution and transcription time, but without access to the parallel processing a powerful GPU provides, the transcription time increases considerably. This means that Whisper performs better and significantly faster than Google Chirp under favorable conditions as was tested in this research. Nevertheless, in a generic situation, Whisper would perform better but would transcribe at a similar rate or slower than Google Chirp.

5 Conclusions and Future Work

In this paper, I have researched the off-the-shelf performance of two commercial state-of-the-art ASR systems, Google Chirp and Whisper, on native and non-native Dutch speech. The ASR systems of these developers were researched due to their large user base, their importance in the field of speech recognition technology and their relative affordability. Evaluating the performance of the ASR systems on (non-)native Dutch speech is beneficial for the further development of ASR technology.

To answer the main question of this research, both Google Chirp and Whisper recognize Dutch native teenager speech better than non-native teenager speech. Speech from native speakers has a better performance in both read speech and HMI speech than speech from non-native speakers. Overall, Whisper outperformed Google Chirp. Both ASR systems performed better on native Dutch speech than on non-native speech. Google Chirp and Whisper have some form of

analysis on the recognized speech, which causes differences in transcription, like repetition avoidance by Whisper or stop word omission and conversion of spoken time expressions by Google Chirp. While the performance of the ASR systems deviates per gender and age, there is no correlation between gender and age with the performance differences. The results on Google Chirp and Whisper in this paper contradict the results of previous works on ASR performance of Dutch diverse speech. However, this cannot be supported yet, given the recent development of the two systems and the limited research on ASR performance on Dutch diverse speech.

Improvements to this study could be to conduct research comparing other SotA ASR systems on native and non-native Dutch speech. This could provide a more comprehensive evaluation of how well off-the-shelf ASR systems perform on non-native speech. A meaningful addition and recommendation to this study would include analyzing the Phoneme Error Rate (PER) for the speech of the DT and NNT speaker groups. PER is the word error rate of the predicted phoneme sequence compared to the truth phoneme sequence [20]. This measures the accuracy of recognizing phonemes in speech. Analyzing the PER for the NNT speaker group, researchers can gain insights into how different native languages influence the pronunciation of the Dutch language. This can help further linguistic research into non-native speakers or further development of ASR technology.

Much research can be further conducted to assess the performance of SotA ASR systems on native and non-native Dutch speech, as the field of Dutch diverse speech remains heavily under-explored.

6 Responsible Research

An important consideration is to ensure that the research is making fair use of data and that this study is wholly reproducible. This section serves as a clarification on these topics and the measures taken to avoid ethical issues.

6.1 Data

All the data gathered for this study came from the JASMIN-CGN. The JASMIN-CGN is available at request for educational and research purposes and thus not publicly available. Therefore I had to make sure the data in the JASMIN database would be handled carefully. Luckily, it was possible to keep all speech files on the DelftBlue servers [21] and handle all files locally. Using the Whisper development model kept the recognition of these files on the servers and not processed by Whisper API. Using the Chirp model API required the model to upload the audio files to the Google Cloud servers for processing. By default, Google Cloud uses the data it processes to provide insights and recommendations[22]. I can opt out of this data processing, resulting in the audio only being processed by the Google Cloud servers and returning the results. This way I have taken the steps to ensure that data has been handled carefully.

6.2 Reproducibility

This study builds upon previous research done on the recognition accuracy of ASR models on Dutch diverse speech and the evaluation of these models. The goal is that further research can be done on this based on the work done in this study. With that in mind, I have attempted to thoroughly describe the experiments done. All data used can be requested free of charge for research purposes at the Instituut voor de Nederlandse Taal[1]. The Whisper model 'whisper-large-v3' used in this research can be used free of charge and downloaded from the Huggingface platform [18]

6.3 Large Language Model

ChatGPT, a Large Language Model(LLM) developed by OpenAI[23], has been used during this research. This was used to formalize the vocabulary and better structure the contents of this paper. This LLM has not in any way attributed to the contents or the methods used in this research.

7 Acknowledgements

I would like to thank my supervisor, YuanYuan Zhang, for her thorough guidance and support throughout this research and my professor, Dr. Odette Schareborg, for her wise life lessons and providing my peers and me with this experience in speech recognition technology. I want to express my gratitude to my fellow peer, Thomas, who provided me with insights which allowed me to finalize the last stage of my research. Finally, I would like to I thank my mother for her continued support and food provisions for the duration of this project.

References

[1] Albert Mosby. *69 Voice Search Statistics 2024 (Usage & Demographics)*. Yaguara. Mar. 16, 2024. URL: https://www.yaguara.co/voice-search-statistics/ (visited on 06/09/2024).

[2] Sima Ajami. "Use of speech-to-text technology for documentation by healthcare providers". In: *THE NATIONAL MEDICAL JOURNAL OF INDIA* 29.3 (2016).

[3] *Netherlands is home to the World's most inclusive workplaces*. URL: https://www.kantar.com/inspiration/sustainability/netherlands-is-home-to-the-worlds-most-inclusive-workplaces (visited on 06/17/2024).

[4] Centraal Bureau voor de Statistiek. *Hoeveel inwoners hebben een herkomst buiten Nederland*. Centraal Bureau voor de Statistiek. Last Modified: 17-04-2023T11:03:33. Mar. 1, 2023. URL: https://www.cbs.nl/nl-nl/dossier/dossier-asiel-migratie-en-integratie/hoeveel-inwoners-hebben-een-herkomst-buiten-nederland (visited on 06/09/2024).

[5] X. Wei et al. "Automatic Speech Recognition and Pronunciation Error Detection of Dutch Non-native Speech: cumulating speech resources in a pluricentric language". In: *Speech Communication* 144 (Oct. 1, 2022), pp. 1–9. ISSN: 0167-6393. DOI: 10.1016/j.specom.2022.08.004. URL: https://www.sciencedirect.com/science/article/pii/S016763932200108X (visited on 04/30/2024).

[1] https://taalmaterialen.ivdnt.org/download/tstc-jasmin-spraakcorpus-c/

[6] Seongjin Park and John Culnan. "A comparison between native and non-native speech for automatic speech recognition". In: *The Journal of the Acoustical Society of America* 145.3 (Mar. 1, 2019), p. 1827. ISSN: 0001-4966. DOI: 10.1121/1.5101679. URL: https://doi.org/10.1121/1.5101679 (visited on 06/20/2024).

[7] Allison Koenecke et al. "Racial disparities in automated speech recognition". In: *Proceedings of the National Academy of Sciences* 117.14 (Apr. 7, 2020). Publisher: Proceedings of the National Academy of Sciences, pp. 7684–7689. DOI: 10.1073/pnas.1915768117. URL: https://www-pnas-org.tudelft.idm.oclc.org/doi/10.1073/pnas.1915768117 (visited on 05/07/2024).

[8] Siyuan Feng et al. "Towards inclusive automatic speech recognition". In: *Computer Speech & Language* 84 (Mar. 1, 2024), p. 101567. ISSN: 0885-2308. DOI: 10.1016/j.csl.2023.101567. URL: https://www.sciencedirect.com/science/article/pii/S0885230823000864 (visited on 04/24/2024).

[9] Siyuan Feng et al. *Quantifying Bias in Automatic Speech Recognition*. Apr. 1, 2021. DOI: 10.48550/arXiv.2103.15122. arXiv: 2103.15122[cs, eess]. URL: http://arxiv.org/abs/2103.15122 (visited on 04/24/2024).

[10] Marcio Fuckner et al. "Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers". In: *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). Bucharest, Romania: IEEE, Oct. 25, 2023, pp. 146–151. ISBN: 9798350327977. DOI: 10.1109/SpeD59241.2023.10314895. URL: https://ieeexplore.ieee.org/document/10314895/ (visited on 04/25/2024).

[11] Yu Zhang et al. *Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages*. Sept. 24, 2023. arXiv: 2303.01037[cs,eess]. URL: http://arxiv.org/abs/2303.01037 (visited on 04/30/2024).

[12] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. Dec. 6, 2022. DOI: 10.48550/arXiv.2212.04356. arXiv: 2212.04356[cs,eess]. URL: http://arxiv.org/abs/2212.04356 (visited on 05/08/2024).

[13] Jon Porter. *ChatGPT continues to be one of the fastest-growing services ever*. The Verge. Nov. 6, 2023. URL: https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference (visited on 06/21/2024).

[14] Martine Adda-Decker and Lori Lamel. "Do speech recognizers prefer female speakers?" In: *Interspeech 2005*. Interspeech 2005. ISCA, Sept. 4, 2005, pp. 2205–2208. DOI: 10.21437/Interspeech.2005-699. URL: https://www.isca-archive.org/interspeech_2005/addadecker05_interspeech.html (visited on 06/21/2024).

[15] Martin Russell and Shona D'Arcy. "Challenges for computer recognition of children²s speech". In: *Speech and Language Technology in Education (SLaTE 2007)*. Speech and Language Technology in Education (SLaTE 2007). ISCA, Oct. 1, 2007, pp. 108–111. DOI: 10.21437/SLaTE.2007-26. URL: https://www.isca-archive.org/slate_2007/russell07_slate.html (visited on 06/21/2024).

[16] Yuanyuan Zhang et al. "Mitigating bias against non-native accents". In: *Interspeech 2022*. Interspeech 2022. ISCA, Sept. 18, 2022, pp. 3168–3172. DOI: 10.21437/Interspeech.2022-836. URL: https://www.isca-archive.org/interspeech_2022/zhang22n_interspeech.html (visited on 04/30/2024).

[17] eric-urban. *Test accuracy of a custom speech model - Speech service - Azure AI services*. Jan. 19, 2024. URL: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-custom-speech-evaluate-data (visited on 06/21/2024).

[18] OpenAI. *Whisper Large v3 Model*. 2024. URL: https://huggingface.co/openai/whisper-large-v3 (visited on 06/03/2024).

[19] Marjolein Korvorst, Ardi Roelofs, and Willem J.M. Levelt. "Telling Time from Analog and Digital Clocks: A Multiple-Route Account". In: *Experimental Psychology* 54.3 (Jan. 2007), pp. 187–191. ISSN: 1618-3169, 2190-5142. DOI: 10.1027/1618-3169.54.3.187. URL: https://econtent.hogrefe.com/doi/10.1027/1618-3169.54.3.187 (visited on 06/23/2024).

[20] Bradley He and Martin Radfar. *The Performance Evaluation of Attention-Based Neural ASR under Mixed Speech Input*. Aug. 2, 2021. arXiv: 2108.01245[cs, eess]. URL: http://arxiv.org/abs/2108.01245 (visited on 06/19/2024).

[21] Delft High Performance Computing Centre (DHPC). *DelftBlue Supercomputer (Phase 2)*. 2024. URL: %5Curl%7Bhttps://www.tudelft.nl/dhpc/ark:/44463/DelftBluePhase2%7D.

[22] *Opt out of data processing — Recommender Documentation — Google Cloud*. URL: https://cloud.google.com/recommender/docs/opt-out (visited on 06/21/2024).

[23] *Introducing ChatGPT*. URL: https://openai.com/index/chatgpt/ (visited on 06/23/2024).