How Good Are State-of-the-Art Automatic Speech Recognition Systems in Recognizing Dutch Diverse Speech?

An Evaluation of Meta MMS and OpenAI Whisper on Native and Non-Native Dutch Speech

Yiming Chen[1]

Supervisor(s): Odette Scharenborg[1]**, YuanYuan Zhang**[1]

[1]**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: Yiming Chen
Final project course: CSE3000 Research Project
Thesis committee: Dr. Odette Scharenborg, YuanYuan Zhang MSc., Dr. Catharine Oertel

Abstract

Automatic speech recognition (ASR) is increasingly used in daily applications, such as voice-activated virtual assistants like Siri and Alexa, real-time transcription for meetings and lectures, and voice commands for smart home devices. However, studies show that even state-of-the-art (SotA) ASR systems do not recognize the speech of everyone equally well.

To the best of my knowledge, this paper, for the first time, evaluates the performance of Meta's SotA ASR system, Massively Multilingual Speech (MMS), on Dutch native and non-native speech. Using the Jasmin Corpus dataset, which includes a diverse set of both native and non-native Dutch speakers, this study uses metrics such as word error rate (WER), character error rate (CER), and word information lost (WIL) to assess performance. Additionally, the same methodology is applied to the same data using OpenAI's ASR system, Whisper, to provide a comparative analysis.

The paper analyzes WER, CER, and WIL error metrics, processing time, and investigates the best-suited beam size for Whisper. It also lists out the types of errors made in terms of deletions, insertions, and substitutions made by each model across different age groups of Dutch speakers.

1 Introduction

Automatic speech recognition (ASR) systems have achieved impressive performance in various applications, enabling voice-activated virtual assistants, real-time transcription, and smart home device commands. However, even state-of-the-art (SotA) ASR systems do not recognize the speech of all users equally well. Their performance varies due to the diversity in speakers' characteristics such as race, gender, age, and nativeness. Specifically, research has shown that White speakers are more accurately recognized than Black speakers [5, 10], and there exists significant bias between native speakers and non-native speakers [11].

The unequal performance of ASR systems must first be identified and evaluated to contribute to a more equitable speech recognition future. This helps identify underlying issues causing bias in diverse speech. For example, ASR training data might primarily consist of speech from one specific demographic group, leading to poor performance for underrepresented groups. If the training data predominantly features native speakers, the system may perform poorly for non-native speakers due to the lack of non-native speech data. Similarly, if the data primarily includes male speakers, the system may not accurately recognize female speakers. Age-related biases can also occur if the training data is skewed towards adult speech, resulting in poor recognition of children's or elderly people's speech. Additionally, biases can arise from the underrepresentation of different accents or age groups, causing ASR systems to perform poorly with these groups.

According to recent data, the Netherlands had at least 2,412,344 residents with a first-generation migration background in 2022, accounting for approximately 13.72% of the total population [1]. This estimate is a lower bound, as children who migrate with their parents are classified as second-generation but are still non-native Dutch speakers. As shown in Figure 1, the portion of non-native Dutch speakers will at most take up 1/4 of the entire Dutch population. This large number of non-native Dutch speakers underscores the importance of developing ASR systems, which can cater to the linguistic needs of both native and non-native speakers. Additionally, this percentage only includes those with Dutch nationality; when accounting for foreign workers living in the Netherlands, the percentage of non-native Dutch speakers would be even higher.

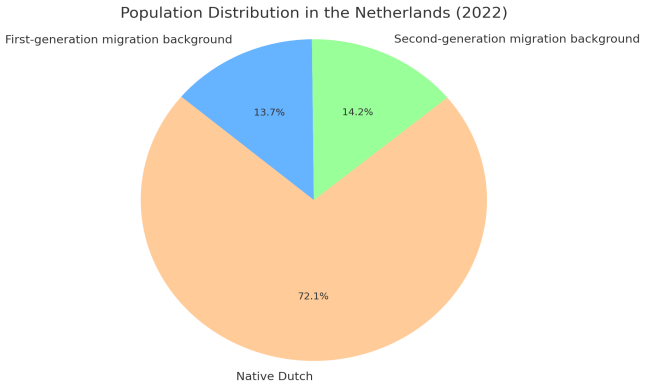

Figure 1: Population Distribution in the Netherlands (2022)

Several critical questions remain unanswered, highlighting the need for further investigation. Specifically, how effective are state-of-the-art ASR systems in recognizing native and non-native Dutch speech? This paper focuses on evaluating the performance of SotA ASR systems on native and non-native Dutch speech. I have chosen Meta's ASR system, Massively Multilingual Speech (MMS) [8], and OpenAI's Whisper [9] for comparison.

Although Whisper has been studied before [3], this research includes it again for several reasons. First, this study provides a comparative analysis using the same dataset and methodology as MMS. Second, this research introduces new metrics, Character Error Rate (CER) and Word Information Lost (WIL), to the evaluation of Whisper's performance. Additionally, I explore the optimal beam size for Dutch speech within Whisper, as adjusting this hyperparameter can significantly impact its performance. These additions provide new insights and enhance the comparative evaluation.

The Jasmin corpus is used in this study as it is the only easily accessible database that provides a comprehensive dataset of both native and non-native Dutch speech, making it an ideal resource to investigate the performance disparities of ASR systems. This study aims to contribute to the understanding of how well these systems perform with native and non-native Dutch speech and to identify areas for improvement in recognizing non-native Dutch speakers.

To answer that question, the following sub-questions are

addressed in this paper:

1. How accurately do the ASR systems recognize native and non-native Dutch speakers?
2. How does age affect the accuracy of the ASR systems?
3. What types of errors do each ASR system make, in terms of insertion, deletion, and substitutions, and what are the performance differences, including accuracy and execution time, between the OpenAI Whisper and Meta MMS ASR systems?

Answering these questions will help understand the performance disparities of these ASR systems on native and non-native Dutch speech.

2 Methodology

This section outlines the approach used to address the research questions, detailing the programs and systems employed, the processing steps, and the data used in this study.

2.1 Dataset

The dataset used for this study is the **Jasmin Corpus CGN** (Corpus Gesproken Nederlands) [2], which is an extension of the Spoken Dutch Corpus. The Jasmin Corpus focuses on contemporary Dutch spoken by various age groups, non-native speakers with different mother tongues, and elderly people in the Netherlands and Flanders. This study specifically uses the **Dutch (NL)** part of the corpus.

I have used the following data from the Jasmin Corpus:

- Native children (**NC**): 12 hours 21 minutes
- Native teenagers (**NT**): 12 hours 21 minutes
- Native elderly (**NE**): 9 hours 26 minutes
- Non-native teenagers (**NNT**): 12 hours 21 minutes
- Non-native adults (**NNA**): 12 hours 21 minutes

Each group contributes both read speech and extemporaneous speech recorded during human-machine interactions, aiming for a balanced representation. The corpus includes approximately 50% read speech and 50% spontaneous speech from human-machine dialogues. [2]

Read Speech and Human-Machine Interaction

The read-speech component involves speakers reading aloud from phonetically balanced texts. For children, texts are selected from educational materials, which includes texts of varying difficulty levels. For non-native speakers, texts are chosen from materials used in Dutch as a second language (L2) education.

The human-machine interaction component involves dialogues with a computer system. These dialogues are structured to induce states of mind like confusion, hesitation, and frustration, which are common in human-machine interactions.

2.2 Models

The ASR systems evaluated in this study are:

- OpenAI/Whisper-large-v3 [9]
- Meta/MMS-1b-all [8]
- Meta/MMS-1b-fl102 [8]

The MMS-1b-all model, also referred to as MMS-1b-l1162, is designed to handle 1162 languages using the largest datasets. In contrast, MMS-1b-fl102 is specifically trained on the Google/FLEURS dataset, covering 102 languages.

2.3 Approach

The procedure is visualized in Fig. 2.

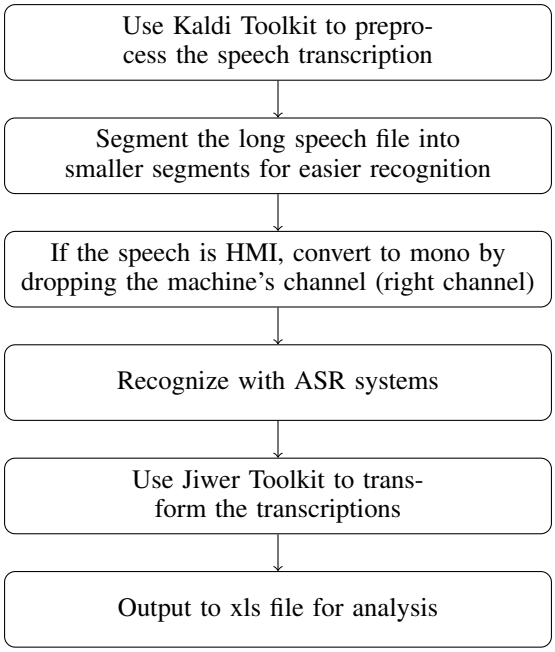

Figure 2: Procedure for speech segmentation and ASR system evaluation

Speech Segmentation

Since the audio lengths are around 10 minutes, processing them with both models requires a high amount of memory. Therefore, it is necessary to preprocess the speech files. To do so, I use the Kaldi toolkit [7]. For example, the preprocessed audio transcription contains lines that look like:

Table 1: Example of preprocessed audio transcription format

Speaker ID	Audio ID	Start	End	Transcription
speaker_1	audio_1	00:00	00:05	"Ik ben iKun."
speaker_2	audio_2	00:05	00:10	"Ik ook!"

Then, the speech can be segmented into smaller chunks, typically a few seconds long, using the start- and end-time information provided in the ground truth.

Speech recognition and post-process
The speech segments are fed into the ASR systems to infer the predicted transcriptions. The output from the ASR model, referred to as the recognized transcription, often includes punctuation and capitalization, which can differ from the true content of the audio segment, referred to as the true transcription. To standardize the transcriptions for analysis, the Jiwer toolkit [6] is used. Specifically, the following steps were taken:

1. Convert all text to lowercase.
2. Remove leading and trailing spaces.
3. Remove punctuation.

This processed information is then saved to an XLS file for result analysis.

The errors of each ASR system will be measured in terms of deletions, insertions, and substitutions made, and the performance will be analyzed using Word Error Rate (WER), Character Error Rate (CER), and Word Information Lost (WIL).

2.4 Types of Errors

Deletions, insertions, and substitutions are the primary types of errors in ASR systems:

- **Deletion (D)**: A word present in the ground truth transcription is omitted in the recognized transcription.
- **Insertion (I)**: An extra word not present in the ground truth transcription is added in the recognized transcription.
- **Substitution (S)**: A word in the ground truth transcription is replaced with a different word in the recognized transcription.

For example, consider the ground truth transcription and the recognized transcription:

Table 2: Examples of Deletions, Insertions, and Substitutions in Recognized Text

Original Text					
I	am		13	years	old.
Recognized Text					
I	am	not	12	years.	
Type of Errors					
		insertion	substitution		deletion

2.5 Performance Metrics

The performance of ASR systems is evaluated using the following metrics:

Word Error Rate (WER) and Character Error Rate (CER)
WER measures the rate of errors in the recognized transcription compared to the ground truth transcription. CER is similar but is calculated at the character level. They are calculated as:

$$\text{WER} = \frac{S + D + I}{N} \quad \text{and} \quad \text{CER} = \frac{\text{S} + \text{D} + \text{I}}{N}$$

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- N is the total number of words (for WER) or characters (for CER) in the ground truth transcription.

Word Information Lost (WIL)
WIL is a metric that provides a more nuanced understanding of the ASR system's performance by taking into account the weighted errors. It is calculated as:

$$\text{WIL} = 1 - \frac{H}{N}$$

where H is the number of correctly recognized words and N is the total number of words in the ground truth transcription.

For example, consider the ground truth transcription "I have a cat" and the recognized transcription "I have the cat":

- Correctly recognized words: 3 ("I", "have", "cat")
- Total words in ground truth: 4

$$\text{WIL} = 1 - \frac{3}{4} = 0.25$$

This detailed metric helps in understanding the impact of each error type on the overall performance of the ASR system.

2.6 Beam Size

Beam size is a critical hyperparameter in Automatic Speech Recognition (ASR) systems that determines the number of hypotheses considered during decoding. A larger beam size explores more possible transcriptions, potentially improving accuracy by considering more possibilities. However, it also increases computational complexity and processing time.

Due to its architecture, only Whisper supports altering this parameter. In section 4.3, I examine the Whisper-large-v3 model with varying beam sizes to determine its optimal setting for Dutch native and non-native speech. I start with Whisper's default beam size of 5, as recommended by Radford et al. [9], and explore its impact on accuracy metrics (WER, CER, WIL) and processing time.

3 Results

The performance evaluations are summarized in Table 3, Table 4 and Table 5. Table 3 shows the Word Error Rate (WER), Character Error Rate (CER), and Word Information Lost (WIL) for each model, split by read speech and HMI (Human-Machine Interaction) speech. Table 4 shows the WER and processing time for different beam sizes of the Whisper model and the 2 MMS models. Table 5 presents the types of errors made by each ASR model across various age groups and speech types, including deletions (Del), insertions (Ins), and substitutions (Sub).

Table 3: Types of errors made by different ASR models across various age groups and speech types. WER = Word Error Rate, CER = Character Error Rate, WIL = Word Information Lost. Metrics are shown for both read speech and HMI (Human-Machine Interaction) speech. The lower the metric value, the better the performance. The best result for each metric is bolded.

		Read speech			HMI speech		
Group	ASR	WER	CER	WIL	WER	CER	WIL
Native Children	$\text{Whisper}_{\text{v3}}$	**19.6**	9.9	**29.7**	**33.4**	**18.5**	**47.6**
	$\text{MMS}_{\text{1b-all}}$	22.8	**9.0**	35.0	43.3	21.8	58.5
	$\text{MMS}_{\text{1b-fl102}}$	27.4	11.0	41.4	49.2	23.7	66.4
Native Teenagers	$\text{Whisper}_{\text{v3}}$	**10.5**	**5.4**	**15.8**	**27.4**	**14.4**	**40.1**
	$\text{MMS}_{\text{1b-all}}$	17.3	6.5	27.3	35.4	15.8	50.9
	$\text{MMS}_{\text{1b-fl102}}$	20.0	7.6	30.9	42.1	18.7	58.7
Native Elderly	$\text{Whisper}_{\text{v3}}$	**15.2**	**7.9**	**23.4**	**34.4**	**20.9**	**47.7**
	$\text{MMS}_{\text{1b-all}}$	24.2	9.7	37.3	43.7	22.5	59.3
	$\text{MMS}_{\text{1b-fl102}}$	21.3	8.1	33.4	48.1	23.7	64.7
Native Average	$\text{Whisper}_{\text{v3}}$	**15.1**	**7.7**	**23.2**	**33.1**	**19.5**	**46.3**
	$\text{MMS}_{\text{1b-all}}$	21.5	8.5	33.4	42.5	21.4	58.1
	$\text{MMS}_{\text{1b-fl102}}$	22.9	8.9	35.4	47.5	23.1	64.2
Non Native Teenagers	$\text{Whisper}_{\text{v3}}$	**33.1**	16.4	**47.2**	**44.8**	**24.7**	**60.0**
	$\text{MMS}_{\text{1b-all}}$	37.6	**16.1**	54.7	62.5	36.4	77.7
	$\text{MMS}_{\text{1b-fl102}}$	49.3	19.7	69.8	67.8	33.9	85.0
Non Native Adults	$\text{Whisper}_{\text{v3}}$	**34.1**	**16.4**	**48.2**	**49.3**	**31.1**	**62.4**
	$\text{MMS}_{\text{1b-all}}$	42.6	18.7	61.3	70.6	47.0	83.0
	$\text{MMS}_{\text{1b-fl102}}$	49.8	20.6	70.5	71.1	42.7	85.4
Non Native Average	$\text{Whisper}_{\text{v3}}$	**33.6**	**16.4**	**47.6**	**47.1**	**28.0**	**61.5**
	$\text{MMS}_{\text{1b-all}}$	40.0	17.3	57.8	68.4	43.9	81.6
	$\text{MMS}_{\text{1b-fl102}}$	49.5	20.1	70.1	70.2	40.2	85.3
All Group Average	$\text{Whisper}_{\text{v3}}$	**21.3**	**10.6**	**31.9**	**38.1**	**22.4**	**51.9**
	$\text{MMS}_{\text{1b-all}}$	27.7	11.4	42.1	51.8	29.2	66.7
	$\text{MMS}_{\text{1b-fl102}}$	31.9	12.7	48.4	55.6	29.0	72.3

3.1 Error metrics

The results presented in Table 3 indicate that $\text{Whisper}_{\text{v3}}$ consistently performs better across various age groups and speech types compared to the MMS models. For native children, teenagers, and elderly groups, $\text{Whisper}_{\text{v3}}$ shows lower WER, CER, and WIL values in both read and HMI speech. The average performance for native speakers also favors $\text{Whisper}_{\text{v3}}$, which maintains the lowest error rates across all metrics. For non-native groups, including teenagers and adults, $\text{Whisper}_{\text{v3}}$ again demonstrates better performance with lower WER, CER, and WIL values. Overall, $\text{Whisper}_{\text{v3}}$ outperforms $\text{MMS}_{\text{1b-all}}$ and $\text{MMS}_{\text{1b-fl102}}$ in all measured categories, achieving the best results in terms of error rates. However, $\text{MMS}_{\text{1b-all}}$ has a slight advantage in two CER metrics by a small margin.

3.2 Processing time

The results presented in Table 4 show that $\text{Whisper}_{\text{v3}}$ outperforms the MMS models across all groups in terms of WER for both read speech and HMI speech. $\text{Whisper}_{\text{v3}}$ consistently achieves the lowest WER values. However, the processing time for $\text{Whisper}_{\text{v3}}$ is substantially higher compared to $\text{MMS}_{\text{1b-all}}$ and $\text{MMS}_{\text{1b-fl102}}$. While $\text{Whisper}_{\text{v3}}$ excels in accuracy, the MMS models demonstrate a significant advantage in processing efficiency, with much lower time values across all groups and speech types.

Table 4: Performance metrics for different ASR models across various age groups and speech types. WER = Word Error Rate, Time = Processing Time. Metrics are shown for both read speech and HMI (Human-Machine Interaction) speech. The lower the metric value, the better the performance. The best result for each metric is bolded.

		Read speech		HMI speech	
Group	Model	WER(%)	Time(s)	WER(%)	Time(s)
Native Children	$\text{Whisper}_{\text{v3-b5}}$	**19.6**	6364	**33.4**	1864
	$\text{Whisper}_{\text{v3-b6}}$	19.9	6795	34.0	1881
	$\text{Whisper}_{\text{v3-b7}}$	20.1	7558	34.2	2123
	$\text{MMS}_{\text{1b-all}}$	22.8	**886**	43.3	273
	$\text{MMS}_{\text{1b-fl102}}$	27.4	934	49.2	**213**
Native Teenagers	$\text{Whisper}_{\text{v3-b5}}$	**10.5**	4890	**27.4**	1235
	$\text{Whisper}_{\text{v3-b6}}$	10.5	5237	27.5	1267
	$\text{Whisper}_{\text{v3-b7}}$	10.6	5836	28.5	1427
	$\text{MMS}_{\text{1b-all}}$	17.3	**695**	35.4	188
	$\text{MMS}_{\text{1b-fl102}}$	20.0	769	42.1	**177**
Native Elderly	$\text{Whisper}_{\text{v3-b5}}$	**15.2**	5507	**34.4**	1389
	$\text{Whisper}_{\text{v3-b6}}$	15.4	5887	34.7	1449
	$\text{Whisper}_{\text{v3-b7}}$	15.7	6560	35.5	1638
	$\text{MMS}_{\text{1b-all}}$	24.2	**825**	43.7	**208**
	$\text{MMS}_{\text{1b-fl102}}$	21.3	1184	48.1	225
Non Native Teenagers	$\text{Whisper}_{\text{v3-b5}}$	**33.1**	5244	**44.8**	3222
	$\text{Whisper}_{\text{v3-b6}}$	33.7	5610	44.8	3444
	$\text{Whisper}_{\text{v3-b7}}$	34.0	6260	45.1	3884
	$\text{MMS}_{\text{1b-all}}$	37.6	**948**	62.5	739
	$\text{MMS}_{\text{1b-fl102}}$	49.3	1036	67.8	**715**
Non Native Adults	$\text{Whisper}_{\text{v3-b6}}$	**34.1**	6064	**49.3**	4355
	$\text{Whisper}_{\text{v3-b6}}$	35.3	6497	49.3	4649
	$\text{Whisper}_{\text{v3-b6}}$	35.7	7234	50.0	5357
	$\text{MMS}_{\text{1b-all}}$	42.6	1394	67.8	**785**
	$\text{MMS}_{\text{1b-fl102}}$	49.8	**991**	71.1	795

Table 5: Number of type errors each ASR made, across various age groups and speech types. Where Del = deletion, Ins = insertion, and Sub = substitution. The number indicates the times the type of error occurs, the lower the better. The best result is bolded.

		Read speech			HMI speech		
Group	ASR	Del	Ins	Sub	Del	Ins	Sub
Native Children	$\text{Whisper}_{\text{v3}}$	**2173**	1987	**7645**	**894**	592	**2524**
	$\text{MMS}_{\text{1b-all}}$	4226	**574**	8955	2573	**100**	2621
	$\text{MMS}_{\text{1b-fl102}}$	4914	725	10864	2628	110	3280
Native Teenagers	$\text{Whisper}_{\text{v3}}$	**1611**	933	**3383**	**511**	339	**1365**
	$\text{MMS}_{\text{1b-all}}$	2721	**537**	6607	1078	**77**	1673
	$\text{MMS}_{\text{1b-fl102}}$	3136	772	7467	1328	97	1936
Native Elderly	$\text{Whisper}_{\text{v3}}$	**1742**	1592	**6027**	**3617**	2058	**7580**
	$\text{MMS}_{\text{1b-all}}$	4348	**516**	10057	8030	**337**	8560
	$\text{MMS}_{\text{1b-fl102}}$	3246	820	9111	8288	449	9866
Non Native Teenagers	$\text{Whisper}_{\text{v3}}$	**3505**	2411	**10161**	**990**	556	2471
	$\text{MMS}_{\text{1b-all}}$	6150	**406**	11701	3218	**46**	**2341**
	$\text{MMS}_{\text{1b-fl102}}$	5582	665	17662	2472	88	3517
Non Native Adults	$\text{Whisper}_{\text{v3}}$	**2738**	2508	**9204**	**3797**	1611	6026
	$\text{MMS}_{\text{1b-all}}$	5546	**391**	12148	11510	**77**	**5230**
	$\text{MMS}_{\text{1b-fl102}}$	4815	581	15748	9847	121	6974

3.3 Types of errors

The results presented in Table 5 show that $\text{Whisper}_{\text{v3}}$ generally makes fewer deletion and substitution errors across all age groups and speech types compared to the MMS models. $\text{Whisper}_{\text{v3}}$ consistently has the lowest number of deletions and substitutions in both read speech and HMI speech. However, $\text{MMS}_{\text{1b-all}}$ achieves the best results for insertion er-

rors in most categories. For native children, teenagers, and elderly, $\text{Whisper}_{\text{v3}}$ makes fewer errors in deletions and substitutions, while $\text{MMS}_{\text{1b-all}}$ performs better in terms of insertions. For non-native speakers, $\text{Whisper}_{\text{v3}}$ also shows fewer deletion and substitution errors, but $\text{MMS}_{\text{1b-all}}$ maintains the lowest insertion error counts. These results shows that $\text{Whisper}_{\text{v3}}$ and MMS models tend to excel in different types of errors, with $\text{Whisper}_{\text{v3}}$ being better at minimizing deletions and substitutions, while $\text{MMS}_{\text{1b-all}}$ performs well in reducing insertions.

4 Discussion

The findings from the experiment revealed several key insights into the performance of state-of-the-art ASR systems, particularly Whisper-large-v3 and MMS models, in recognizing diverse Dutch speech.

4.1 Non-Native Speaker Performance

It is evident from Table 3 that both Whisper-large-v3 and MMS models perform worse on non-native Dutch speakers compared to native Dutch speakers. The higher error rates for non-native speakers suggest that these models may require further training with more diverse datasets to improve their performance across different speaker backgrounds.

4.2 Age-Related Performance

Another notable observation is that children (NC), adults (NNA) and elderly (NE) exhibit higher error rates compared to native teenagers (NT). This pattern is consistent within both native and non-native groups, indicating that the age of speakers can influence ASR performance. Teenagers (NT, NNT) consistently have the lowest error rates among both native and non-native speakers, which may be due to the resemblance of their speech to that used for training the models. This age-related discrepancy underscores the need for ASR systems to be more adaptive to various age groups.

4.3 Beam Size on Whisper

The Whisper-large-v3 model performs best with a beam size of 5, as shown in Table 4. This beam size offers a balance between accuracy and computational efficiency, as indicated by lower WER, CER, WIL values, and lower processing time. Increasing the beam size to 6 and 7 not only results in slightly higher error rates but also increases processing time. Consequently, the higher computational cost and time make larger beam sizes less suitable for real-time applications. My result aligns with the formal study by Kasai et al. [4], which highlights the "beam search curse," where larger beam sizes do not necessarily result in better generations.

4.4 Whisper and MMS Processing Time

The results presented in Table 4 show that while Whisper-large-v3 outperforms the MMS models across all groups in terms of WER for both read speech and HMI speech, it has a substantially higher processing time. This indicates that despite Whisper-large-v3's better performance in WER, the MMS models demonstrate a significant advantage in processing efficiency. The lower processing times of MMS models make them more suitable for scenarios where quick processing is crucial.

5 Conclusion and Future Work

The performance analysis of Whisper-large-v3 and MMS models indicates that both ASR systems face significant challenges with non-native Dutch speakers. Whisper-large-v3 performs significantly better overall compared to MMS models. However, both systems show higher error rates for non-native speakers and certain age groups, particularly children and adults.

Both ASR systems demonstrate that nativeness plays a crucial role in performance, with both state-of-the-art ASR systems performing better on native speakers. This disparity highlights the need for continued evaluation and improvement to address these biases. By testing, evaluating, and addressing the shortcomings of state-of-the-art ASR systems in terms of nativeness for Dutch diversity, I am making advancements toward more inclusive speech recognition technologies.

The slow processing time of Whisper-large-v3 should be taken into consideration when applying ASR in daily applications, as this can cause huge delays. However, it is important to note that Whisper has other smaller variant models, though these models come with weaker performance. Future research could explore the potential of fine-tuning these smaller models to achieve a balance between accuracy and processing speed.

Future research should explore the impact of more balanced training datasets that incorporate a diverse range of speech patterns to achieve more equitable performance of ASR systems. Testing Whisper's smaller models to evaluate their performance compared to MMS at similar processing speeds would be a valuable area of research. Evaluating whether Whisper's smaller variants can match or exceed MMS's performance at comparable processing speeds could provide insights into optimizing ASR systems for both accuracy and efficiency. By understanding and addressing these limitations, future ASR systems can become more inclusive and better suited to diverse populations, ultimately leading to more equitable and effective communication technologies.

6 Responsible research

This section evaluates the ethical aspects of the research, ensuring adherence to the Code of Conduct for Research Integrity. I aim to discuss the fairness, reproducibility, and ethical implications of my study.

Firstly, the dataset used in this research is the Jasmin corpus, which includes a diverse set of native and non-native Dutch speakers across different age groups. The data was used in accordance with fair use policies, ensuring that it was sourced from an open and publicly available platform. However, due to licensing restrictions, the dataset itself cannot be shared online by me. Consequently, I have removed the result folder and dataset folder from the publicly accessible code repository.

All the models were run locally on the same machine with the following specifications:

- CPU: Intel i9-12900k
- GPU: NVIDIA RTX 3070

- CUDA version: 12.1
- Transformers library version: 4.42.0
- PyTorch version: 2.3.0

This consistent setup ensures that the reported processing times are accurate and comparable. During the execution of the code, the PC was dedicated solely to running the experiments to minimize external influences on processing time.

The code used for this project is openly available on GitHub. This repository includes all the scripts necessary to reproduce the experiments, barring the dataset due to the aforementioned restrictions. GitHub's version control ensures that the state of the repository at any given time can be viewed and replicated, enhancing the reproducibility of my work.

To ensure the robustness and accuracy of the results, I used established libraries and tools such as the Kaldi toolkit for preprocessing, and the Jiwer toolkit for evaluating the performance of the ASR systems. These tools are well-documented and widely used within the research community, which adds to the reliability of my methodology.

Additionally, this research has been reviewed by peers and supervisors at various stages, incorporating their feedback to refine and improve the study. This peer review process is crucial for maintaining scientific integrity and ensuring that the research adheres to high ethical standards.

Throughout the research process, I used tools like ChatGPT to assist with grammar and LaTeX formatting. These tools were used to enhance the clarity and presentation of the paper, ensuring that the focus remains on the scientific content.

References

[1] Centraal Bureau voor de Statistiek. Bevolking; kerncijfers, 2022. Accessed: 2024-05-30.

[2] Catia Cucchiarini, Hugo Van hamme, Olga van Herwijnen, and Felix Smits. Jasmin-cgn: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In *International Conference on Language Resources and Evaluation*, 2006.

[3] Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. 2023. 2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), SpeD ; Conference date: 25-10-2023 Through 27-10-2023.

[4] Sakaguchi K. Le Bras R. Radev D. Choi Y. Smith N. A. Kasai, J. A call for clarity in beam search: How it works and when it stops, 2023. arXiv preprint arXiv:2204.05424.

[5] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[6] Matthew Morris. jiwer: Similarity measures for automatic speech recognition. https://github.com/jitsi/jiwer, 2021. Accessed: 2024-05-30.

[7] Daniel Povey et al. Kaldi asr toolkit. https://github.com/kaldi-asr/kaldi, 2015. Accessed: 2024-05-30.

[8] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *arXiv*, 2023.

[9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[10] Rachael Tatman and Charlie Kasten. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Proc. INTERSPEECH*, pages 934–938, 2017.

[11] Yuhan Wu et al. See what i'm saying? comparing intelligent personal assistant use for native and non-native language speakers. In *22nd International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 1–9, 2020.