# How do ASR systems of Google and Microsoft compare when recognizing Dutch spoken by native speakers over the age of 60?

**Thomas de Valck**[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: Thomas de Valck
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg, YuanYuan Zhang, Catharine Oertel

## Abstract

Automatic Speech Recognition (ASR) systems are found in many places and are used by many people. Some groups of people, superficially older Dutch adults, are recognized less well by these systems. Given the aging population of the Netherlands, it would be beneficial to have ASR systems be more inclusive to allow for more independence of the older adults. By conducting tests on the ASR systems of Google and Microsoft, making use of the JASMIN dataset, I compared the two using word-error-rate (WER), word-information-lost (WIL) and character-error-rate (CER). Results show Microsoft outperforming Google with an average word error rate of 19.6% compared to 27.35%. However, Google is less biased on the topics of gender and age. Microsoft was slightly less biased in regards to region, but only by a small margin. Overall, the most notable findings from both systems are a small bias toward female speakers, and a strong bias against speakers from the southern regions of the Netherlands. These findings highlight the need for more inclusive ASR systems, enhancing the independence of older adults.

Keywords: Automatic Speech Recognition, Dutch, Google, Microsoft, Bias, Older Adults

## 1 Introduction

Nowadays, Automatic Speech Recognition (ASR) systems are found everywhere. They are the systems that turn your speech into words, allowing you to use things such as a voice assistant or some automated customer service. Most people won't have many issues using any of these systems, as they are able to speak a language clearly. There are however, a multitude of groups that have a rougher time, whether it's because of their gender, age, place of origin, or a disability. The aim of this research is to put the inclusiveness of state-of-the-art ASR systems to the test, in particular for older Dutch adults.

One issue with the Dutch population is the proportion of older people, and more importantly soon-to-be older people. This will inevitably cause problems sooner or later. Currently, the Netherlands already has a large group of adults and elderly in its population [1; 2]. Combining this with an average life expectancy of roughly 81.5 years [3] results in a society that is not sustainable, as a lot of these older people require more care than what is able to be provided. If ASR systems could be improved in regards to this group, more of them would be able to live independently for longer, which reduces the workload on the working population.

The ASR systems of choice are from Google and Microsoft, who both offer speech recognition as part of their AI Cloud Service programs. These two were chosen because they are some of the largest companies out there, with a very large share in the digital market. Both of them see millions of users worldwide, with their ASR systems being used a lot on both mobile phones and computers. Choosing specifically two ASR systems allows for a good analysis and comparison of their performance and bias.

Given this context, the main research question that will be topic of this paper is: **How do ASR systems of Google and Microsoft compare when recognizing Dutch spoken by native speakers over the age of 60?**

### 1.1 Existing Literature

There is little research on the topic of inclusiveness of ASR systems, especially in the context of the Dutch language, but I will go over some research that has been done on similar topics.

The majority of the currently available research on bias in Dutch speech recognition comes from Fuckner et al. [4] and Feng et al. [5]. Both state a strong bias against non-native speakers, a smaller bias against children and older speakers, and a slight bias toward female speakers. The former also specifically mentions the poor performance on older speakers from the southern regions of the Netherlands.

The bias toward female speakers is also found internationally, in both American English and French, as stated by M. Adda-Decker and L. Lamel [6]. They suggest the cause of this bias is because female speakers tend to adhere closer to conventional speech compared to male speakers.

It is also important to note what is lacking in current existing literature. With the ASR systems available nowadays, they vary wildly in performance, speed and cost. What has yet to be answered is, how do they vary in bias? We have some scores on Wav2Vec2 and Whisper from the aforementioned research by Fuckner et al. [4], but there are a lot more ASR systems out there. This research aims to shrink that knowledge gap, if only a little.

### 1.2 How do ASR Systems work?

As a little piece of background information, it is useful to dive a bit into the inner-workings of an ASR system. In the past, an ASR system had 3 major components which were used to decipher a piece of audio into legible text. It would turn the audio into phonetics, phonetics into possible words, and lastly pick the most likely possible words for the final result. This last part follows patterns in the language, i.e., some words are more likely to be followed by certain others.
However, modern state-of-the-art ASR systems are different, as they are now end-to-end models [7]. These models go directly from audio to text in neural networks that went through a process called Deep Learning. These models have to be trained on large datasets of speech and transcriptions in order to get to the level they are at today.

## 2 Methodology

Firstly I will go over the data used in this research. Following this I will discuss the ASR systems used in more detail.

I will then briefly go over the metric used for measuring performance of the ASR systems. Lastly I will discuss the actual experiments that will be performed.

## 2.1 Data

To test the performance of the given ASR systems of Google and Microsoft, I will be using data from JASMIN-CGN [8]. This dataset is an extension of the CGN (Corpus Spoken Dutch) [9], which features adult speech. JASMIN-CGN has extended this dataset by adding Dutch spoken by children, elderly and non-native speakers. The dataset contains two forms of speech. Firstly, Human Machine Interaction (HMI), which is meant to resemble speech that takes place during a normal conversation and is used to provide extemporaneous speech. The second part is readspeech, where the speakers read aloud a pre-written piece of text.

**Details on Older Speakers**

To go into further detail on the data of older speakers in the JASMIN-CGN, there are a total of 67 speakers. Of these 67, 23 are men and 44 are women. The total length of all recordings amounts to 19 hours, of which 10.2 hours actually contains speech. Of this, 3.4 hours is spoken by men, whereas 6.8 hours is spoken by women. The ages range from 59 to 96 with the average age being 79. Although the dataset has divided the Netherlands into 4 main regions and 16 smaller sub-regions, only 4 of these sub-regions are present and are split as follows. There are 18 speakers from North Holland (NH), 17 speakers from the Gelders river area, including Arnhem and Nijmegen (G). Lastly, there are 16 speakers each from Overijssel (O) and Limburg (L).

## 2.2 Pre-processing

As the data in this research is sensitive user data, it should be handled with care. In accordance with the JASMIN-CGN, the data should never be in possession of a third party. This was an import criteria that was relevant when choosing fitting ASR systems for this research.

**Microsoft**

Microsoft allows you to choose your own storage location for the data, and is therefore not an issue. They also do not log any data on their servers, so there was no need for any changes.

**Google**

Google offers two separate services, for synchronous and asynchronous requests. The former is to be used for transcribing audio in-real-time. The other is used for transcribing prerecorded audio. This is split up in two groups, files shorter and longer than a minute. When a recording is longer than a minute, it is required to be stored in a google storage location, which would violate the conditions for JASMIN-CGN. The recordings from the JASMIN-CGN vary between 3 and 20 minutes, which is too long. To resolve this issue, recordings have been segmented based on utterances. This results in short fragments of a few seconds each, none longer than a minute while ensuring there are no words cut-off. Note that this change was made for Microsoft as well, to prevent this from having any consequences in the performance of either system.

## 2.3 ASR Models

Google and Microsoft both offer speech-to-text as part of their AI Cloud Services, with both supporting a large number of languages. For this research, I will specifically be using the nl-NL models for transcribing regular Dutch.
The key difference between the two models is their output. Google attempts to transcribe the speech, nothing more. Microsoft on the other hand, adds capitalization and punctuation. Any of this punctuation and capitalization is removed before calculating error rates.

## 2.4 Metric

To measure performance of the ASR systems, I will be making use of 3 different metrics, each of which is simple to calculate and revolve around similar principles. This metrics are all standard and used often in the context of speech recognition, however they are not always the best metric. One alternative that could be more fitting for this research would be semantic distance [10]. This involves seeing how well the meaning is preserved after the ASR system transcribes the audio. Given the context of voice assistants, this would be a better fit for measuring performance. However, this would require a lot more data, time and computational power, which is infeasible for this project.

All of the following metric revolve around the similar concept of substitutions, insertions and deletions. A substitution is the case where a word or character is replaced by a different word or character. An insertion is where a new word or character is inserted into the sentence. Lastly, a deletion occurs whenever a character or word is fully missing from the result. Each of the metrics will use these 3 things to calculate their corresponding error rates.

**Word Error Rate**

The Word Error Rate (WER) is one of the simplest metrics. The WER of a given sample is calculated as follows:

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

Here, S, I and D are substitutions, insertions and deletions of words, as were explained before. N refers to the word length of the reference transcription, the ground truth. The first part of this formula takes the total 'distance' between two sentences, i.e. how many words you need to add, change, or remove in order to get from one to another. Dividing this by the length of the actual sentence, and multiplying by 100% tells you how large of a proportion of the sentence is correct. One downside to this is the lack of an upper bound, as sentences can have an error rate of over 100%.

**Word Information Lost**

Word Information Lost (WIL) [11] is a little more complicated, but aims to calculate how much is lost from the original transcription. Meaning, a result which adds a bunch of words but still includes all of the original, will score better compared to a result which contains only correct words, but only half of the sentence. As the name implies, it approximates how much information is lost from the reference solution. The way this is calculated is with the following formula:

$$WIL = 1 - \frac{H}{N} * \frac{H}{P} = 1 - \frac{H^2}{(H+S+D)(H+S+I)}$$

Here S, I, D and N are the same as before, but new terms 'H' and 'P' are introduced. H refers to the number of 'hits', i.e. the number of words of the reference solution that are present in the given solution. P is the word length of the given solution. Unlike WER, WIL actually has an upper bound of 100%, which is the absolute worst a model could perform.

**Character Error Rate**

Lastly, there is another variant of the WER, but instead of counting words, it makes use of individual characters, named Character Error Rate (CER). In some sense this gives a more accurate representation of how close a model is to recognizing the original, as two different words can be a mere 1 letter apart. It should be noted that in practice, 1 character can make a big difference as to what is being said. Regardless of this, it is a good way of checking whether the pronunciation of characters was correctly recognized by an ASR system. The formula for the CER is as follows:

$$CER = \frac{S+I+D}{N} * 100\%$$

Again S, I, D and N are the number of substitutions, insertions, deletions and length of the reference solution, but this time in characters, rather than words.

## 2.5 Experiments

**Google and Microsoft**

Starting on the core part of the research, I will first run each experiment independently, and individually measure the WER of Google's and Microsoft's ASR systems. Per model I will take a look for patterns in things such as the gender of the speaker, the age of the speaker, or the region where the speaker is from. This should provide a good insight into how well the models can recognize older Dutch speech. Lastly, I compare the models against each other. In this comparison I take a detailed look at the overall performance, as well as the performance when looking at the aforementioned criteria such as gender and age.

**Transcription Markers**

In the transcriptions of the dataset, some transcription markers such as 'xxx' and 'ggg' are used to replace names and other sounds made by the speaker. I will be testing how much this affects performance by comparing the average error rates between the original text, and with any notes filtered out. As I expect the filtered version to give a more accurate representation of the actual error rates, I will be using this one on the aforementioned experiments.

# 3 Results

Firstly, the results of the comparisons between Google and Microsoft will be shown. Afterwards, I will quickly go over the experiment where transcription markers were filtered out of the transcriptions. Note that in any table shown, the best performer will be marked in **bold**.

## 3.1 Google VS Microsoft

Starting off with the basics, we compare Google and Microsoft on HMI and readspeech. Here, we also include results from Fuckner et al. [4], which tested Wav2vec2 and Whisper on the JASMIN corpus. As shown in table 1, Microsoft performs better than Google, but both Wav2vec2 and Whisper show even better performance.

Table 1: Error rates of Google and Microsoft on HMI and readspeech, including reference results of Wav2vec2 and Whisper.

| | HMI | Reading | Average |
|---|---|---|---|
| Google - WER | 31.75% | 22.95% | 27.35% |
| Microsoft - WER | 25.61% | 13.59% | 19.60% |
| Wav2vec2 - WER | 25.2% | 10.9% | 18.1% |
| Whisper - WER | **19.6%** | **8.7%** | **14.2%** |
| Google - WIL | 45.86% | 34.24% | 40.05% |
| Microsoft - WIL | **37.04%** | **21.23%** | **29.14%** |
| Google - CER | 17.22% | 13.08% | 15.15% |
| Microsoft - CER | **13.69%** | **6.31%** | **10.00%** |

Now to take bias into account, starting with gender. As visible in Table 2, female speech sees lower error rates across all metrics, with Microsoft performing better than Google.

Table 2: Error Rates of Google and Microsoft on Male and Female speech.

| | Male | Female |
|---|---|---|
| Google - WER | 29.70% | 26.12% |
| Microsoft - WER | **21.60%** | **18.56%** |
| Google - WIL | 43.28% | 38.36% |
| Microsoft - WIL | **31.81%** | **27.74%** |
| Google - CER | 16.19% | 14.60% |
| Microsoft - CER | **11.05%** | **9.45%** |

Continuing with regional bias, we see the following results in Table 3. The region of Limburg, sees significantly higher error rates across both ASR systems. Again, Microsoft sees better performance across all regions.

Table 3: Error Rates of Google and Microsoft per region, as previously defined in section 2.1.

| | NH | G | O | L |
|---|---|---|---|---|
| Google - WER | 24.44% | 26.99% | 23.57% | 34.77% |
| Microsoft - WER | **17.63%** | **18.75%** | **17.39%** | **24.93%** |
| Google - WIL | 36.00% | 39.85% | 35.35% | 49.51% |
| Microsoft - WIL | **26.30%** | **28.23%** | **25.93%** | **36.50%** |
| Google - CER | 13.53% | 15.06% | 12.83% | 19.38% |
| Microsoft - CER | **8.99%** | **9.40%** | **8.83%** | **12.96%** |

Lastly, when looking at the error rates for specific ages as seen in Table 4, we see gradually increasing error rates as the speaker gets older. Once more, Microsoft achieves lower error rates across all ages.

Table 4: Error Rates of Google and Microsoft per age group.

| | 60-69 | 70-79 | 80-89 | 90-99 |
|---|---|---|---|---|
| Google - WER | 21.88% | 27.17% | 27.40% | 35.77% |
| Microsoft - WER | **15.55%** | **19.27%** | **19.57%** | **26.22%** |
| Google - WIL | 32.69% | 39.96% | 40.32% | 50.33% |
| Microsoft - WIL | **23.31%** | **28.86%** | **29.21%** | **37.77%** |
| Google - CER | 11.54% | 15.17% | 15.03% | 20.75% |
| Microsoft - CER | **7.61%** | **10.00%** | **9.71%** | **14.25%** |

### 3.2 Transcription Markers

When comparing the results of the original and filtered transcriptions, visible in Table 5, we see slightly lower error rates when the transcription markers are filtered out. Note that the average was taken between both Google and Microsoft.

Table 5: Average Error Rates when removing transcription markers, compared to the original transcriptions.

| | WER | WIL | CER |
|---|---|---|---|
| Original | 24.36% | 35.49% | 13.38% |
| Filtered | **23.48%** | **34.60%** | **12.58%** |

## 4 Responsible Research

### 4.1 Fair Use of Data

The data used from the JASMIN-CGN dataset was used appropriately and handled with care, as previously explained in section 2.2. No data was ever stored by a third party, and any data I had on my personal device has been deleted.

## 5 Discussion

### 5.1 Findings

As was shown in Table 1, the reference results from Wav2vec2 and Whisper [4] perform better than Google and Microsoft. In the case of Whisper this is even by quite a significant margin. As the data the ASR systems were run on is the same, this suggests both Google and Microsoft perform worse at recognizing speech from older native Dutch speakers than Wav2Vec2 and Whisper. This is, of course, assuming there were no errors in the setup of either experiment.

In Table 2 we see a bias towards female speakers. This lines up with some previous works [4] [6]. The bias is slightly larger on the side of Microsoft, with a relative increase of 16.4% in the WER for male speakers, compared to 13.4% on the side of Google. This also holds for WIL and CER, with 14.6% and 16.9% higher error rates for Microsoft, where Google sees increases of only 12.8% and 10.9% respectively.

Continuing with regional bias, as visible in Table 3, we see significantly higher error rates for region N4a. This was also reported by Fuckner et al. [4], and has a much stronger effect on older speakers compared to children and teenagers. Furthermore, Gelderland sees a small increase in error rates, whereas North Holland and Overrijssel are nearly equal. The higher error rates are an indication the accents spoken in Limburg deviate further from average Dutch and are therefore harder to recognize by ASR systems.
Taking the best performing region with the worst performing region, Overijssel and Limburg respectively, we compare the relative bias for each ASR system. Google sees increases of 47.5%, 40.1% and 51.1% for WER, WIL and CER. Meanwhile, Microsoft finds error rates that are 43.4%, 40.8% and 46.8% higher, indicating it is slightly less biased in regards to region, but only by a small margin.

Taking a look at bias towards age in Table 4, we see gradually increasing error rates as the speakers get older. One interesting point to note is a large increase in error rates between the ages of 60-69 and 70-79, as well as 80-89 and 90-99. However, there is hardly any difference between 70-79 and 80-89. It can be theorized that around the age of 70 and 90, some changes take place that cause the higher error rates. However, it is hard to point to a specific cause to this, and would require further testing to confirm.
To calculate how biased either ASR system actually is, I will compare the increase in error rates between the best and worst performing group, which are the groups of 60 to 69 years old, and 90 to 99 years old. Google finds differences of 63.5%, 54.0% and 79.8% against Microsoft's 68.6%, 62.0% and 87.3%. From this we see Google is less biased towards older speakers compared to Microsoft.

In regards to the extra experiment surrounding transcription markers, both systems performed better with the markers removed from the transcription and therefore giving a more accurate error rate. Although this doesn't have any specific implications when only using the JASMIN dataset, this does allow for more accurate comparisons to other data which makes no use of transcription markers.

### 5.2 Limitations

The largest limitation to this research is the lack of data. Although 10 hours is quite some data, it is very little compared to the entirety of the CGN which has over 900 hours. This coincides with the number of speakers being a mere 67. Had there been more speakers, the experiment regarding age would have been a lot more precise. Currently, the age group of 90-99 consists of only 7 speakers, whereas the other groups feature between 15-20 speakers each. Of course, more data would improve the quality and certainty of the results of all of the experiments.

Additionally to the size of the data, there is also an imbalance in the types of speakers. Although JASMIN aimed for a 50-50 balance between male and female speakers, the older speakers feature 34% male against 66% female speakers. This imbalance also holds for the regions, as only 4 out of the 16 defined sub-regions were used.

## 6 Conclusion and Future Work

### 6.1 Conclusion

To answer the question of how Google and Microsoft compare when recognizing Dutch spoken by native speakers over the age of 60, we can first look at the raw numbers. On every aspect and with every metric, Microsoft performs better than Google by some margin.

Diving deeper into the topic of inclusiveness, Google is less biased on the topic of gender, achieving worse overall performance but having a smaller relative difference between error rates of male and female speakers. On the contrary, Microsoft is slightly less biased when looking at regions. Lastly, when dividing the speakers in age groups, we see that Google is less biased towards the oldest speakers compared to the less old speakers.

### 6.2 Future Work

One major point to improve on is the data itself, both in amounts of data, as well as how recent the data is. The original JASMIN dataset is from 2008. Throughout the years there has both been a change in language, as well as a change in the Dutch population itself.

To get a better overall overview of how biased some ASR systems are, more ASR systems should be tested. Right now we see how biased Microsoft and Google are relative to each other, but there is no reference point to check against.

When trying to look at why some speech from older speakers is recognized as well as it is, phoneme error rate could be used and analysed. This metric uses individual pronunciations rather than letters or words. If this were used, you would be able to highlight which specific sounds make the biggest difference, similar to what Fuckner et al. [4] showed for <sh> being recognized particularly badly for native children.

## References

[1] CBS, “Age distribution in the netherlands,” 2022. [Online]. Available: https://www.cbs.nl/en-gb/visualisations/dashboard-population/age/age-distribution

[2] ——, “Population pyramid of the netherlands,” 2023. [Online]. Available: https://www.cbs.nl/en-GB/visualisaties/dashboard-bevolking/bevolkingspiramide

[3] World BankThe World Bank, “Life expectancy at birth, total (years) - netherlands,” 2021. [Online]. Available: https://data.worldbank.org/indicator/SP.DYN.LE00.IN?locations=NL

[4] M. Fuckner, S. Horsman, I. Janssen, and P. Wiggers, “Uncovering bias in asr systems: Evaluating the performance of wav2vec2 and whisper for dutch speakers,” Feb. 2024, 2nd Dutch Speech Tech Day : Towards inclusive speech technology in research and applications, Dutch Speech Tech Day ; Conference date: 19-02-2024 Through 19-02-2024. [Online]. Available: https://sites.google.com/view/dutchspeechtechday/home

[5] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, “Quantifying bias in automatic speech recognition,” 2021. [Online]. Available: https://arxiv.org/abs/2103.15122

[6] M. Adda-Decker and L. Lamel, “Do speech recognizers prefer female speakers?” in *Proc. Interspeech 2005*, 2005, pp. 2205–2208.

[7] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, “End-to-end speech recognition: A survey,” 2023.

[8] C. Cucchiarini, H. Van hamme, O. van Herwijnen, and F. 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 *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, Eds. Genoa, Italy: European Language Resources Association (ELRA), May 2006. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2006/pdf/254_pdf.pdf

[9] H. Kloots, “”een beetje babbelen onder elkaar”: Verzameling, verwerking en studie van spontane spraak uit het corpus gesproken nederlands,” *Handelingen - Koninklijke Zuid-Nederlandse maatschappij voor taal- en letterkunde en geschiedenis*, vol. 60, pp. 53–70, 1 2006. [Online]. Available: https://openjournals.ugent.be/kzm/article/id/72135/

[10] S. Kim, A. Arora, D. Le, C.-F. Yeh, C. Fuegen, O. Kalinli, and M. L. Seltzer, “Semantic distance: A new metric for asr performance analysis towards spoken language understanding,” 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2104.02138

[11] R. Errattahi, A. El Hannani, and H. Ouahmane, “Automatic speech recognition errors detection and correction: A review,” *Procedia Computer Science*, vol. 128, pp. 32–37, 2018, 1st International Conference on Natural Language and Speech Processing. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050918302187