State-of-the-art Automatic Speech Recognition Systems on Dutch Regional Dialects

Exploring Bias in Dutch-trained Automatic Speech Recognition Systems

Simon Kasdorp[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: Simon Kasdorp
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg, YuanYuan Zhang, Catherine Oertal

Abstract

Automatic Speech Recognition is a field that has seen a strong increase in developments in recent years. In order to ensure objectivity and reliability in these systems, it is crucial they remain unbiased and treat speakers equally. This paper explores the bias of two state-of-the-art ASR systems in the domain of Dutch and Flemish speech, specifically towards regional dialects. Specifically, it explores Microsoft's Azure AI Speech Services ASR system and Google Chirp. It analyses the performance of these two systems on the JASMIN-CGN language corpus. The results show that speech from West-Dutch regions is recognized correctly significantly more often than other Dutch regions and speech from Brabant is recognized correctly significantly more often than other Flemish regions.

Index terms - speech recognition, bias, Dutch

1 Introduction

1.1 Background

Automatic Speech Recognition (ASR) systems are becoming increasingly prevalent in modern life. They are used in e.g. live transcription, voice assistants and accessibility options, providing a large component of interaction between humans and technology. For this reason, it is important these systems are unbiased and able to convert human speech to written word equally well for different groups of people. However, recent research has shown that state-of-the-art ASR systems do not recognise the speech of different speaker groups equally well. For example, state-of-the-art ASR systems trained on Dutch perform better on female speech than male speech [1]. Similarly, state-of-the-art ASR systems perform better on speech performed by teenagers and older adults than speech produced by children [2]. Furthermore, Dutch speech is exercised by people of several different dialects and regional accents. Speakers from northern areas pronounce certain words with different tone and emphasis than speakers from the west, for example. Because different speaker groups, including those from different regions, should be able to interact with modern technology equally well, ASR systems should recognize each speaker group at a similar quality. This forms the basis of this research.

1.2 Existing Research

Fuckner et al. found that "state-of-the-art models are overall less biased compared to earlier ASR systems, but are still biased against speakers with accents that deviate from standard Dutch." [1]. This indicates that state-of-the-art ASR models have improved in decreasing bias, although the models that they tested still show some bias in certain dialect regions. Likewise, Herygers et al. [3] found that some bias was present in Flemish speech for speakers from the region of Brabant and against speakers from West Flemish speakers and those from Limburg. More specifically, speech from Brabant was recognized correctly more often than speech from Limburg. However, not much research is present for Dutch speakers from the Netherlands or for models other than Whisper and Wav2Vec. It is therefore unknown what the bias for SotA ASR systems regarding regional dialects looks like outside of these models.

1.3 Research Question

This paper seeks to answer the following main research question: "How well do state-of-the-art ASR systems perform on Dutch and Flemish speech from different regional dialects?". This question will be answered through help of some subquestions, which will each analyse a small component of the main question. These questions are as follows:

1. "Do state-of-the-art ASR systems perform better on speakers from a certain region than on speakers from different regions?"
2. "Do Flemish speakers (from different regions) achieve a lower or higher WER than Dutch speakers (from different regions)?"
3. "How do models differ in WER for speakers from different regions?"

The ASR models I will be using are Microsoft's Azure AI's ASR system and Google's Chirp. This is partly because some research exists for e.g. Whisper on Dutch speech ([1]), but not as much for Chirp and Azure AI's ASR system.

A strong reason for choosing these systems is feasibility of research. Both these systems provide options for free trials, i.e. a number of hours where the systems do not require financial transactions. The corresponding times available on both systems are 200 hours for Azure AI and 300 hours for Chirp. Therefore, to explore the bias of different state-of-the-art ASR systems on regional dialects, this research shall focus on these two systems.

To answer the aforementioned research questions, the paper will discuss the method of research in section 2. In section 3, the results will be presented. Then, section 4 will contain an analysis of the results and discuss the results obtained. Section 5 will address some ethical concerns and how responsible research was applied. Section 6 will then conclude the paper and provide some suggestions for future work.

2 Methodology

2.1 Datasets

For the data, the JASMIN-CGN corpus is used[1]. The JASMIN-CGN is an extension of the spoken Dutch corpus (CGN), a language corpus consisting of spoken Dutch and Flemish from adults. The JASMIN-CGN corpus expands on this by adding several types of Dutch and Flemish speech, including, but not limited to, speech produced by older adults, non-native speakers and speech produced by children. Furthermore, the corpus contains speech from speakers from different regional dialects, both in Dutch and Flemish, making it an appropriate choice for this research. In total, the data used consists of 18 hours and 58 minutes of Dutch speech and 7 hours and 50 minutes of Flemish speech

[1] https://aclanthology.org/L06-1141/

The JASMIN-CGN corpus is labeled according to the speaker's corresponding speech files. For each speech file, the corpus contains information about the age, gender, and regional dialect of the speaker, as well as proficiency in Dutch in cases of non-nativity. Crucially, regional dialects are split into four major regions for both Flemish and Dutch. For Flemish, the corpus distinguishes the following four regions:

- FL1: West-Flemish (West-Flanders) (2h 03m of speech)
- FL2: East-Flemish (East-Flanders) (1h 56m of speech)
- FL3: Brabant (Antwerp and Flemish Brabant) (1h 44m of speech)
- FL4: Limburg (Limburg) (2h 5m of speech)

For Dutch, the corpus similarly distinguishes four regions:

- N1x: West-Dutch (North-Holland, South-Holland, West Utrecht) (4h 49m of speech)
- N2x: Transitional region (Zeeland, Eastern Utrecht, Gelders river area, Veluwe, West Friesland) (5h 0m of speech)
- N3x: Peripheral region (Achterhoek, Overijssel, Drenthe, Groningen, Friesland) (4h 29m of speech)
- N4x: Southern peripheral region (Noord-Brabant, Limburg) (4h 38m of speech)

In the Dutch version, x refers to more precise locations within each region (henceforth referred to as "sub-regions"). For example, N1a specifically refers to North-Holland, N3c refers to Drenthe etc. These will not be addressed thoroughly for the purposes of this research, but a note will be made about them in section 6. For the continuation of this paper, when referring to Dutch regions, they will be referred to as N1-N4.

Furthermore, the JASMIN-CGN corpus consists of two types of speech. The first of these is read speech, consisting of "phonetically rich sentences and stories or general texts to be read aloud"[4][2]. The second, newly introduced in the corpus compared to the original CGN Dutch language corpus, is human-machine interaction (HMI) speech. This type of speech consists of dialogues between a human and a machine.

The JASMIN-CGN corpus consists of speech files that often span several minutes and contain a lot of silence. Analyzing these large blocks of data is not only inconvenient, but also a concern for privacy, as elaborated on in section 5. For these reasons, the audio files are segmented according to the annotated data provided in the JASMIN-CGN corpus, which contains timestamps indicating when a line of speech is said. These segmented portions of speech will henceforth be referred to as "segments".

It is important to note that not all the data available in the corpus was used. This is due to time constraints in the research, and may explain some discrepancies that come up in the results. For Dutch and Flemish speech, the read and HMI speech from groups 1 and 2, as annotated in the JASMIN-CGN dataset, have been used.

[2]Quote retrieved from official documentation

2.2 ASR systems

As was mentioned in section 1, the ASR systems to be used are Microsoft's Azure AI Speech Services (henceforth referred to as Azure AI's ASR) and Google's Chirp. These systems have seen some research in non-Dutch speech[5], but in the Dutch domain of language, especially in regards to regional dialects, the research available is more limited.

In the case of both Dutch and Flemish, the Dutch version of the ASR system is used. This is to provide an even comparison of the same system for both languages. As a result of this, however, Dutch and Flemish cannot be directly compared, as the system may not be trained on Flemish data, or not to the same extent as it is trained on Dutch data.

2.3 Evaluation

The systems will be evaluated using the Word Error Rate (WER). To elaborate on how this is calculate, I first define a number of terms. An insertion is when a word is incorrectly recognised as appearing in an utterance when it is not there. A deletion is when a word in an utterance does not appear in the resulting text. A substitution occurs when a word in the resulting text is different from the word in the speech. The Word Error Rate can then be determined as follows:

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

where I is the number of insertions, D is the number of deletions, S is the number of substitutions and N is the total number of words in the line of speech (ground truth). N can also be expressed using the number of correct words (C) in the following way: $N = D + S + C$. Calculating the WER will be done using JiWER[3], a public python package designed for evaluating ASR systems.

2.4 Experiment Setup

The experiment will take place as follows. First, the data is segmented based on annotated data. The resulting segments each have a duration of less than three seconds. Then, each segment is recognized by the two aforementioned ASR systems. The resulting recognized segments are then stored separately and compared to the ground truth. In order to evaluate the output of the ASR system, the output is first simplified. This is done by removing punctuation and converting all words to lowercase. Then, the WER is calculated according to the equation mentioned above.

To elaborate on this, the reference material is annotated fairly thoroughly. For example, when the speaker inhales, this is sometimes annotated as "ggg" in the ground truth for the data. Neither Azure AI's ASR system nor Google Chirp pick up on this, resulting in high Word error rates for segments containing these features. For this reason, these are first filtered out.

Then, the data (and their corresponding evaluations) are grouped by region for analysis. Additionally, the data as it is provided is divided into HMI speech and read speech. The data is also analyzed in respect to these categories, as it would

[3]https://github.com/jitsi/jiwer/

be informative to see how these differ for each region. The WER for a region is calculated by taking the mean WER of all the segments in that region.

Additionally, in order to analyse bias in each region more thoroughly, the difference between the smallest and largest WER per region is taken for each group (read speech and HMI speech, Dutch speech and Flemish speech).

3 Results

This section contains the performance of Azure AI's ASR system and Google Chirp on the JASMIN-CGN corpus. The experiments to obtain these results are outlined in section 2.4.

Table 1 shows the mean WER (%) of Azure AI's ASR system on Dutch and Flemish regional dialects, both on read speech and HMI speech. The regions with the lowest WER are made bold for convenience's sake in both categories. Regions are indicated as described in section 2.

Table 1: Performance of Azure AI's ASR on the JASMIN-CGN dataset per dialect region (in WER%)

Region	Read	HMI
$N1$	**15.672**	**20.457**
$N2$	23.211	40.181
$N3$	23.777	28.899
$N4$	20.656	33.659
$FL1$	19.075	**32.001**
$FL2$	19.541	38.177
$FL3$	**18.247**	32.166
$FL4$	23.971	35.693

The difference between the lowest and highest WER for each region is 7.539% for Dutch read speech and 19.724% for Dutch HMI speech. These numbers are 5.724% for Flemish read speech and 6.167% for Flemish HMI speech.

Table 2 shows the mean WER (%) of Chirp on Dutch and Flemish regional dialects, again on read speech and HMI speech.

Table 2: Performance of Chirp on the JASMIN-CGN dataset per dialect region (in WER%)

Region	Read	HMI
$N1$	**24.824**	**25.209**
$N2$	32.605	35.152
$N3$	33.618	28.730
$N4$	32.596	34.601
$FL1$	34.247	38.470
$FL2$	36.183	41.646
$FL3$	**33.135**	**35.826**
$FL4$	40.537	37.320

The difference between the lowest and highest WER for each region is 8.794% for Dutch read speech and 9.943% for Dutch HMI speech. For Flemish read speech, this is 7.402% and for Flemish HMI speech, this is 5.820%.

For comparison's sake, the following results are retrieved from Herygers et al.[3], who performed research on a Deep Neural Network-Hidden Markov Model (DNN-HMM) on Flemish speech. These results are as follows. Figure 1 shows the performance of the DNN-HMM on read Flemish speech. Figure 2 shows the performance of the DNN-HMM on HMI Flemish speech.

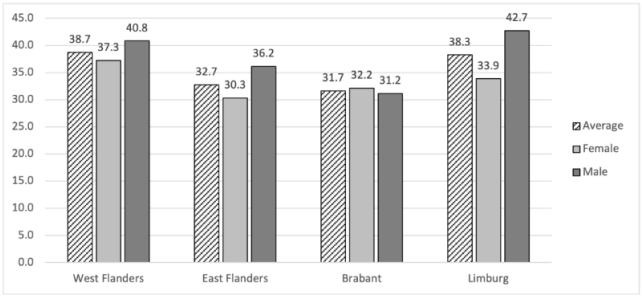

Figure 1: WERs (%) for read speech across region and gender. Obtained from Herygers et al.

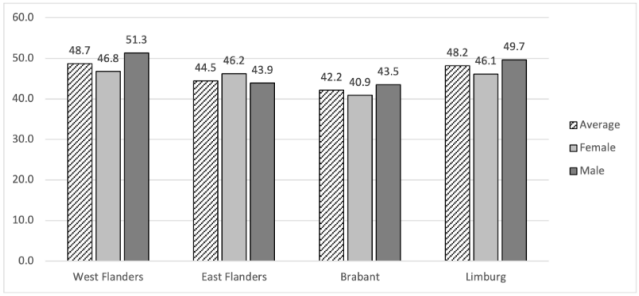

Figure 2: WERs (%) for HMI speech across region and gender. Obtained from Herygers et al.

These figures shall be addressed more thoroughly in section 4.

4 Discussion

This paper presented the setup and results of an experiment to analyse and compare the performance of two state-of-the-art ASR systems, namely Azure AI's ASR and Google Chirp. In this section, the results of these two systems, as outlined in section 3, will be analysed.

4.1 Azure AI's ASR

For Dutch read speech, Azure AI's ASR performs significantly better on region N1 than other regions, while it performs worse on region N3. Interestingly, for HMI speech, the system seems to perform better somewhat better on region N3 than the other regions (except N1) and N2 performs significantly worse. This indicates a strong bias towards West-Dutch speakers, revealing that a large portion of training data was potentially from West-Dutch speakers. The worse performance on HMI speech from region N2 could be explained by the age of the speakers, as a large amount of

speech from region N2 used in the experiment is produced by children, which state-of-the-art ASR systems perform worse on[6]. This is not entirely certain, however, as not all speakers have their age annotated in the corpus.

For Flemish read speech, Azure AI's ASR performs slightly better on region FL3 compared to other regions, while it yields worse results on region FL4. For Flemish HMI speech, there is similar pattern to Dutch HMI speech: the system gives better results for FL3 and FL1, but worse results for FL2. The results for read speech are not wholely unexpected, as Herygers et al.[3] showed that state-of-the-art ASR systems are biased towards Flemish speakers from FL3. However, we would expect to see the system perform about equally well on HMI speech produced by speakers from FL2. A possible explanation for this discrepancy is use of limited data. As was mentioned in section 2, not all of the data in the JASMIN-CGN corpus was used. It is therefore possible that more data was used from e.g. children compared to teenagers and older adults.

Furthermore, Azure AI's ASR system performs better on read speech than HMI speech. This is not unexpected, as ASR systems tend to perform worse on HMI speech than on read speech[2].

4.2 Google Chirp

For Dutch read speech, similarly to Azure AI's ASR, Google Chirp performs significantly better on region N1 compared to other regions. For HMI speech, a similar pattern emerges where Google Chirp yielded better results for N1 than other regions. Interestingly, however, Google Chirp performed better on HMI speech produced by speakers from region N3 than read speech from speakers from region N3. A possible reason for these results is that Google Chirp is trained on data similar to that found in the corpus under N3 HMI speech, but not on data similar to N3 read speech contained in the corpus.

For Flemish read speech, there are again similar results to Azure AI's ASR. The system performs better on read speech produced by speakers from region FL3, but worse on speakers from region FL4. On Flemish HMI speech, Chirp's results remain similar to those from Azure AI's ASR: Chirp performs better on speech from category FL3 and worse on speech from category FL2. This is again in contrary to the results from Herygers et al., although my theory for why this is, is identical to the one for Azure AI's ASR system. There is also a similar discrepancy to the one found in Google Chirp's performance on Dutch speech: Chirp seemingly performs better on HMI speech from region FL4 than read speech from the same region. A possible explanation is again similar to the one for Chirp's performance on Dutch speech: the data found in the corpus corresponding to Flemish HMI speech from region FL4 is more similar to Chirp's training data than Flemish read speech from region FL4.

Despite the aforementioned discrepancies, Google Chirp also generally yields better results for read speech than HMI speech.

4.3 Future Research

The first path to explore for future research is to analyse the performance of the ASR systems on the entire JASMIN-CGN Dutch language corpus. This is to explore whether the discrepancies mentioned in section 4.1 and 4.2 are in fact a result of under-representation in the test data.

As mentioned in section 2.1, Dutch dialect regions are divided into further sub-regions from which speakers originate. It would be interesting to determine whether systems differ in bias between these sub-regions as well.

5 Responsible Research

5.1 Data Handling

The data in the JASMIN-CGN corpus is comprised of speech produced by people who did not consent to their data being used for commercial use. As such, this data is sensitive and it is crucial it is not provided to third parties who may use it for personal gain. For this reason, it is important that the ASR systems used in this research do not store the data.

In the case of Azure Speech Services, the data is not logged or stored in any way as stated on the website of Azure Speech Services[4]. In the case of Google's Chirp, the situation is similar, but there is a subtle difference for larger data pieces. Namely, in order to provide the system with an audio file longer than 60 seconds, the data should be uploaded to Google's Cloud, which is a clear breach of privacy as outlined above. Therefore, when Chirp is to be used, I ensure to only use audio files shorter than 60 seconds and segment longer audio files to be several shorter files, as outlined in section 2.1.

Furthermore, to ensure the data is only used for research purposes, the data, including audio files, transcriptions and metadata, shall be removed from any devices other than the source by the end of this research, and shall not be distributed in any way.

5.2 Environmental Concerns

Other students may have used the same data on the same ASR systems, or I may have data I ran on ASR systems another student uses too. Should this be the case, we will then exchange results from this in order to save power (provided it is useful to do so), as running the same data on the same ASR system yields identical results. However, when this is done, it shall be acknowledged appropriately in section 7.

6 Conclusion

Despite the limited data used, there are still some logical conclusions to be drawn from the data. The sub-questions from section 1.3 shall be revisited here in order.

1. Do state-of-the-art ASR systems perform better on speakers from a certain region than on speakers from different regions? Given the results found in section 3, as well as the results from Herygers et al., it is clear that both Azure AI's ASR and Google Chirp exhibit significant bias towards speakers from region N1, both for read and HMI speech. For Flemish, this bias is not as clear, though still present as both ASR systems show slightly better performance on speakers from region FL3.

[4]https://azure.microsoft.com/en-us/products/ai-services/ai-speech

2. Do Flemish speakers (from different regions) achieve a lower or higher WER than Dutch speakers (from different regions)? To answer this question, the difference between the lowest and highest WER from each region is analysed. For both Google Chirp and Azure AI's ASR, the difference for Flemish speech is lower than that for Dutch speech. However, it is important to keep in mind that Flemish speech sees lower performance than Dutch speech.

3. How do models differ in WER for speakers from different regions? Azure AI's ASR performed better on both Dutch and Flemish read and HMI speech. This is evident from a direct comparison between the results for each model.

The answers to these questions lead to the main research question: "How well do state-of-the-art ASR systems perform on Dutch and Flemish speech from different regional dialects?" This paper has shown that state-of-the-art ASR are biased towards Dutch speakers from West-Dutch areas and Flemish speakers from Brabant. Despite it not being possible to point out bias against a specific group with certainty based on the results, it is clear that improvements need to be made in order to reduce bias in these systems.

7 Acknowledgements

I would like to thank fellow students and researchers T. de Valck and G. van Dijk for providing me with transcribed versions of some of the data in the JASMIN-CGN corpus. T. de Valck provided me with Chirp's transcriptions of the data and G. van Dijk allowed me to use his subscription key in order to transcribe the data from Azure AI's ASR system.

References

[1] M. Fuckner, S. Horsman, P. Wiggers, and I. Janssen, "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. [Online]. Available: https://ieeexplore.ieee.org/document/10314895

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

[3] A. Herygers, V. Verkhodanova, M. Coler, O. Scharenborg, M. Georges, and A. Bavaria, "Bias in flemish automatic speech recognition." in *Proceedings of the ESSV Konferenz Elektronische Sprachsignalverarbeitung*, 2023. [Online]. Available: http://resolver.tudelft.nl/uuid:5051d4fd-36d8-4362-9f56-ef6588cd11ba

[4] 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

[5] Q. A. Obaidah, M. E. Za'ter, A. Jaljuli, A. Mahboub, A. Hakouz, B. Alfrou, and Y. Estaitia, "A new benchmark for evaluating automatic speech recognition in the automatic call domain," 2024. [Online]. Available: https://arxiv.org/abs/2403.04280

[6] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice conversion based data augmentation to improve children's speech recognition in limited data scenario." in *Interspeech*, 2020, pp. 4382–4386.