How well can state-of-the-art ASR systems recognize Dutch diverse speech?

Exploring bias of SotA ASR systems on Dutch and Flemish regional dialects

Student: Simon Kasdorp Responsible professor: Odette Scharenborg Supervisor: YuanYuan Zhang

1 Background

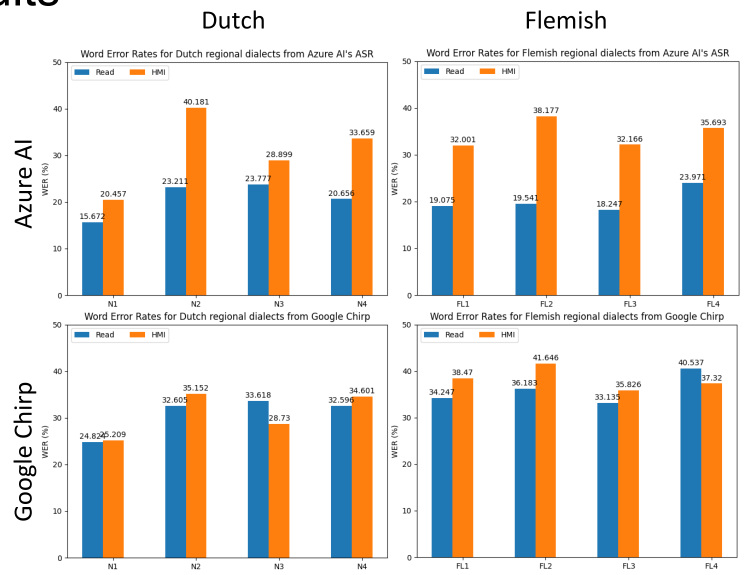
Dutch and Flemish speakers come from different regions, with different accents. It is important that speech from all speakers from these regions is recognized equally well. For this purpose, four dialect regions are identified for both Dutch and Flemish:

- N1- West-Dutch
- N2 - Transitional
- N3 - Peripheral
- N4 - Southern peripheral
- FL1 - West-Flemish
- FL2 - East-Flemish
- FL3 - Brabant
- FL4 - Limburg

The bias for regional accents has been explored for Wav2Vec2 and Whisper on Dutch and Flemish speech[1], as well as a self-trained ASR system on Flemish speech[2]. The research available in this domain for Microsoft's Azure AI and Google Chirp is much more limited. Since these systems are state-of-the-art and sometimes deployed in real-life applications, it is important these systems are unbiased.

4 Results

Overall, Azure AI's ASR system performed better on the test data. However, it tends to be more biased (particularly on HMI speech). Read speech tends to get recognized better than HMI speech. Both systems are biased towards speakers from region N1. These results are not wholely unexpected, as Herygers et al. found similar results for Flemish speech[2].



2 Research question

How well do state-of-the-art ASR systems perform on Dutch speech from different regional dialects?

In order to answer the main research question, some sub-questions are formed which each aim to answer a component of the research question. These are as follows:

1. Do SotA ASR systems perform better on Dutch speakers from certain regions?
2. Do Flemish speakers from different regions achieve a lower or a higher WER than Dutch speakers from different regions?
3. How do models differ in WER for speakers from different regions?

3 Methodology

The experiment set up is as follows. It involves the following systems and data:

- JASMIN corpus
- Two ASR systems (Azure and Chirp)

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

First, segment speech files into individual lines of speech based on annotated data. Perform speech recognition on segments. Then, group recognized segments by region. Next, calculate the WER for each segment individually. Finally, compute mean WER per region by taking the mean of all WERs in each region. Repeat for Dutch, Flemish, HMI speech, read speech and both ASR systems.

5 Future research

Since not all of the data in the JASMIN corpus was used, a first idea for future research is to explore the results when more data is used. This could help explain some outliers found in the current results.

Another option would be to investigate the bias of ASR systems that were not used in this research or other research, such as Amazon AWS Transcribe.

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/1031489

[2] A. Herygers, V. Verkhodanova, M. Coler, O. Scharen- borg, 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