# Uncovering bias in ASR systems

*Evaluating the performance of Wav2vec2 and Whisper for Dutch speakers*

**Author(s)**
Fuckner, Marcio; Horsman, Sophie ; Janssen, Iskaj; Wiggers, Pascal

**Publication date**
2024

**Document Version**
Final published version

**License**
Unspecified

DUTCH SPEECH TECH DAY 2024

# Uncovering bias in ASR systems

## Evaluating the performance of Wav2vec2 and Whisper for Dutch speakers

The study evaluated two speech recognition systems, Wav2vec2 and Whisper, for potential biases for Dutch speakers. Results obtained by evaluating on the JASMIN corpus revealed biases against non-native speakers, children, and the elderly, with (slightly) better performance for women. The study emphasizes the need for ASR systems to handle variations in speaking in order to reach equal performance among all users.

### AUTHORS

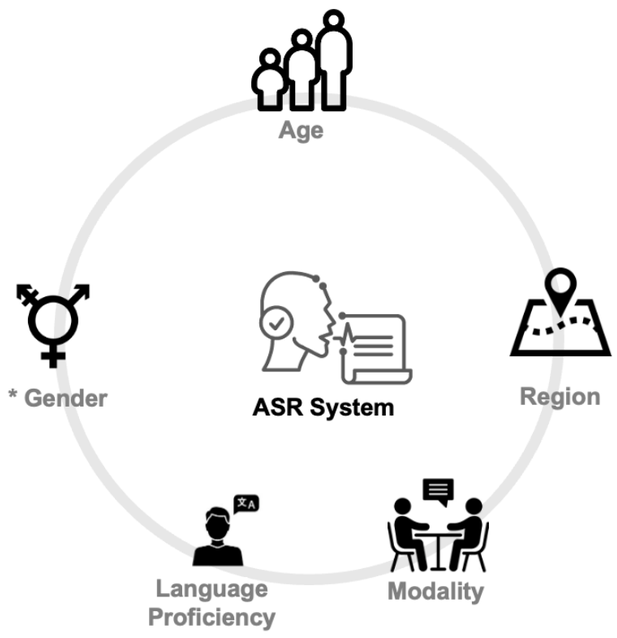
Marcio Fuckner[1], Sophie Horsman[1], Iskaj Janssen[2], and Pascal Wiggers[1].
(1) Responsible AI Lab, Amsterdam University of Applied Sciences
(2) RTL Nederland

### ACKNOWLEDGEMENT

This research is part of the RAAK DRAMA project: Designing Responsible AI for Media Applications. DRAMA is a collaboration between de University of Applied Sciences of Rotterdam, Utrecht and Amsterdam and VPRO, NPO, RTL and Beeld & Geluid.

### INTRODUCTION

The research builds on the approach used in [1] to investigate potential bias in speech recognition systems. We are interested in the potential quality-of-service harm that comes with the more widespread use of ASRs, which are increasingly used for downstream applications, such as virtual assistants, automatic interview transcriptions and automatic subtitling. Here we will evaluate and quantify bias for two state-of-the-art ASR systems, namely Wav2vec2 [2] and Whisper [3].



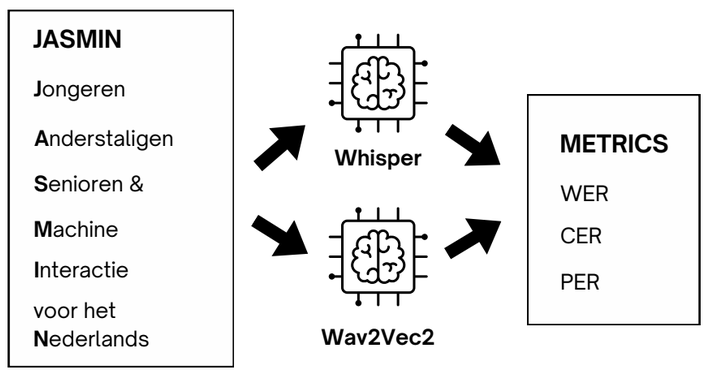

* note that the data does not include transgender and non-binary speakers

### METHODS

The foundation of the bias analysis is the Dutch speech corpus JASMIN, which was used as a test set to evaluate Whisper and Wav2vec2, measured with the WER. This test corpus contained data for different age groups (children, teenagers, elderly), regions (Dutch and Flemish) and non-native accents. The corpus comprises two types of speech, read speech and human-machine interaction (HMI) speech, used in the experimental evaluations.

To increase the reliability of our findings, we assess speech separately for reading and HMI (human-machine interaction) speech and analyse differences in WER for different speaker groups. In addition, we analyse phoneme-level error rates using a post-hoc approach that converts word-level transcripts to phoneme representations and aligns them using the Levenshtein distance.

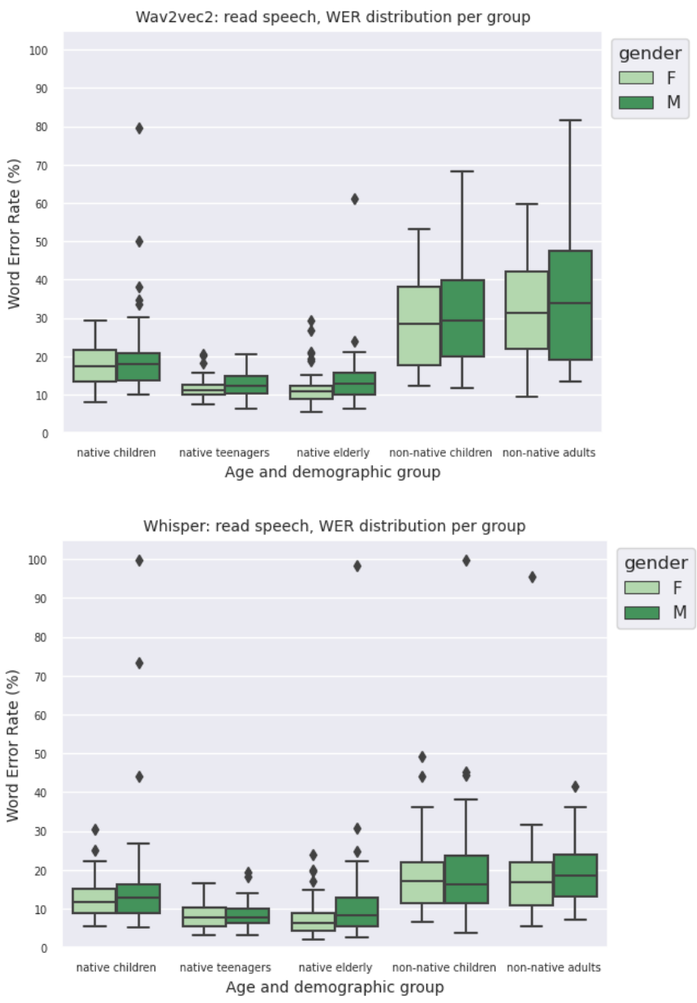
### EVALUATION PIPELINE



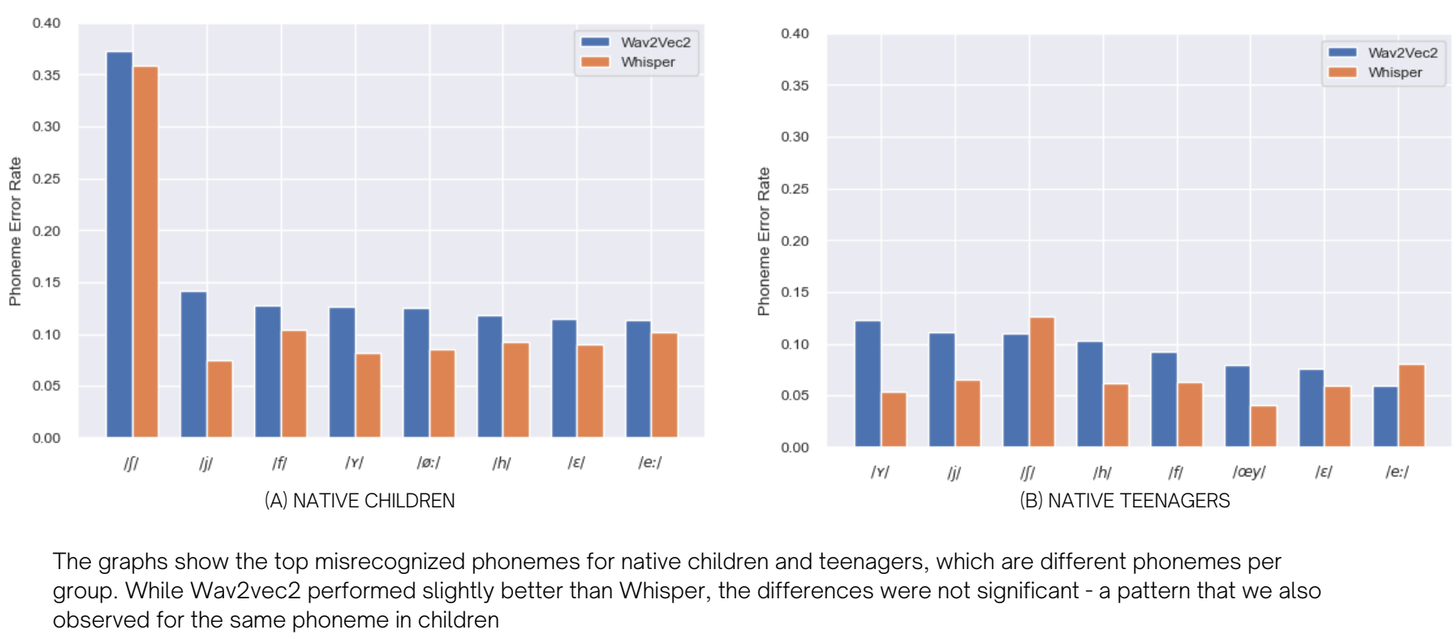

### RESULTS

The table below presents the results categorised by ASR system and age group, with a breakdown for female and male speech, read speech and HMI speech, and an average across both genders. Whisper significantly outperformed Wav2vec2, improving WER for both read and HMI speech. Compared to the baseline results in [1], with WERs of 36.2 for read speech and 47.5 for HMI speech, both Wav2vec2 and Whisper have made substantial progress in enhancing the overall performance.

| | Age Group | ASR System | Read speech | | | HMI speech | | |
|---|---|---|---|---|---|---|---|---|
| | | | F | M | Avg | F | M | Avg |
| Dutch Natives | Children | Kaldi [5] | 34.8 | 35.7 | 35.3 | 43.5 | 43.3 | 43.4 |
| | | Wav2vec2 | 16.4 | 18.6 | 17.4 | 24.5 | 23.6 | 24.1 |
| | | Whisper | 12.3 | 15.5 | 13.9 | 18.2 | 15.5 | 16.9 |
| | Teenagers | Kaldi [5] | 16.5 | 20.1 | 18.4 | 34.4 | 36.2 | 35.3 |
| | | Wav2vec2 | 10.4 | 11.4 | 10.9 | 16.4 | 19.3 | 17.8 |
| | | Whisper | 7.8 | 8.1 | 8.0 | 12.9 | 13.8 | 13.3 |
| | Elderly | Kaldi [5] | 22.3 | 27.9 | 24.2 | 37.8 | 42.5 | 39.5 |
| | | Wav2vec2 | 9.8 | 12.9 | 10.9 | 24.6 | 26.5 | 25.2 |
| | | Whisper | 6.9 | 11.7 | 8.7 | 18.9 | 20.8 | 19.6 |
| | Avg natives | Kaldi [5] | 24.4 | 28.1 | 26.1 | 38.4 | 41.7 | 39.8 |
| | | Wav2vec2 | 12.1 | 14.4 | 13.2 | 22.1 | 22.8 | 22.4 |
| | | Whisper | 8.9 | 11.8 | 10.2 | 16.9 | 16.2 | 16.6 |
| Non-natives | Children | Kaldi [5] | 54.3 | 55.9 | 55.1 | 60.9 | 62.1 | 61.6 |
| | | Wav2vec2 | 26.0 | 30.2 | 28.0 | 35.8 | 42.1 | 38.8 |
| | | Whisper | 18.1 | 19.6 | 18.8 | 24.1 | 25.8 | 25.0 |
| | Adults | Kaldi [5] | 57.3 | 56.1 | 56.9 | 61.2 | 61.5 | 61.3 |
| | | Wav2vec2 | 25.6 | 32.2 | 28.1 | 35.8 | 44.7 | 39.3 |
| | | Whisper | 18.4 | 19.8 | 18.8 | 30.3 | 35.2 | 32.1 |
| | Avg non-natives | Kaldi [5] | 55.8 | 56.0 | 55.9 | 61.1 | 61.7 | 61.4 |
| | | Wav2vec2 | 25.8 | 31.1 | 28.1 | 35.8 | 43.2 | 39.0 |
| | | Whisper | 18.3 | 19.6 | 18.9 | 27.1 | 29.0 | 28.0 |
| All | Avg | Kaldi [5] | 35.4 | 37.2 | 36.2 | 46.5 | 49.0 | 47.5 |
| | | Wav2vec2 | 16.2 | 18.6 | 17.3 | 26.7 | 28.8 | 27.6 |
| | | Whisper | 12.3 | 14.5 | 13.3 | 20.4 | 20.7 | 20.5 |

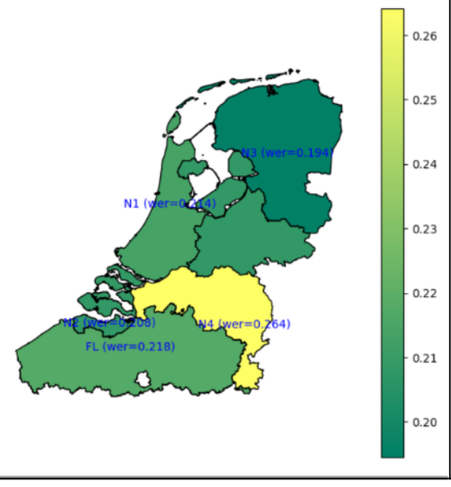



Whisper has a significant decrease in terms of WER across the board compared to Wav2vec2. Similarly to previous research [1], a substantial gap is reported between natives and non-natives as well as each age group.




The graphs show the top misrecognized phonemes for native children and teenagers, which are different phonemes per group. While Wav2vec2 performed slightly better than Whisper, the differences were not significant - a pattern that we also observed for the same phoneme in children

The graph shows the WER for Whisper for the elderly age group (65+), HMI speech. What can be observed is that the WER is higher for the south of the Netherlands, compared to the other regions. A similar discrepancy is observed when testing Wav2Vec2.




### CONCLUSION

ASR systems can perpetuate biases for specific groups, which aligns with previous research [1]. Whisper outperformed Wav2Vec2 in many tasks. However, it is still possible to detect cases of performance disparity between groups, emphasising non-natives, children, and elderly participants, and critical cases in the southern regions. The speech of native Dutch speakers was much better recognised than that of non-native speakers, regardless of age. Additionally, regional accents seemed stronger for older people than for children and teenagers, with elderly participants in the southern region of the Netherlands exhibiting lower performance. Acknowledging the biases and limitations of ASR systems and striving for equitable, inclusive, and accurate models is essential.

[1] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, “Towards inclusive automatic speech recognition,” Computer Speech & Language, vol. 84, p. 101567, Mar. 2024, doi: 10.1016/j.csl.2023.101567.
[2] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems 33, 12449–12460 (2020)
[3] A. Radford, “Robust speech recognition via Large-Scale Weak Supervision,” arXiv.org, Dec. 06, 2022. https://arxiv.org/abs/2212.04356