Uncovering Bias in ASR Systems

Evaluating Wav2vec2 and Whisper for Dutch speakers

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

Publication date
2023

Document Version
Author accepted manuscript (AAM)

Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers

Marcio Fuckner, Sophie Horsman, and Pascal Wiggers
Amsterdam University of Applied Sciences
Email: {marcio.fuckner, s.horsman, p.wiggers}@hva.nl

Iskaj Janssen
RTL
Email: iskaj.janssen@rtl.nl

***Abstract*—It is crucial that ASR systems can handle the wide range of variations in speech of speakers from different demographic groups, with different speaking styles, and of speakers with (dis)abilities. A potential quality-of-service harm arises when ASR systems do not perform equally well for everyone. ASR systems may exhibit bias against certain types of speech, such as non-native accents, different age groups and gender. In this study, we evaluate two widely-used neural network-based architectures: Wav2vec2 and Whisper on potential biases for Dutch speakers. We used the Dutch speech corpus JASMIN as a test set containing read and conversational speech in a human-machine interaction setting. The results reveal a significant bias against non-natives, children and elderly and some regional dialects. The ASR systems generally perform slightly better for women than for men.**

***Index Terms*—speech recognition, bias, Dutch**

I. Introduction

The accuracy of Automatic Speech Recognition (ASR) systems has drastically improved, with, among others, the introduction of end-to-end transformer-based models and the shift towards unsupervised training techniques [1]. However, the accuracy of speech recognition still varies depending on the task, the quality of the audio and the complexity of the language being spoken. For example, recognising speech in noisy environments or with non-native speakers is still challenging.

In this research, 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. This potential harm is a known problem for ASR systems since existing research shows that ASR systems do not work equally well for all speaker groups. Lower performances for certain speakers are due to differences in pronunciation and language use for different speaker groups that are not well-represented in the training data [2] [3]. The advancements of the past years put forward the question of to what extent current state-of-the-art ASR systems that were trained on very large datasets, mostly scraped from the internet, are prone to bias and how this compares to earlier systems.

Here we will evaluate and quantify bias for two state-of-the-art ASR systems, namely Wav2vec2 (Facebook; [4]) and Whisper (OpenAI; [1]). We will do so for the Dutch language and investigate bias against certain speaker groups (based on age, gender, speaker region and native language). To do so, we build upon a methodology introduced by [3] for quantifying bias in ASR systems. Our research shows 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.

979-8-3503-2797-7/23/$31.00

II. Related Work

There is a growing body of literature on bias in automatic speech recognition for various languages and different aspects of speech. In [5], the performance of ASR systems for speakers with different English accents was investigated by evaluating YouTube's automatic captions for isolated chunks or particular words. It was concluded that the performance was significantly worse for Scottish speakers than for speakers from New Zealand, New England or California. In [6], the performance of five state-of-the-art ASR systems was compared for audio snippets of black speakers and white speakers. On average, black speakers had a WER of 0.35, whereas white speakers had a WER of 0.19. Finally, [3] investigated the disparities between different groups of Dutch people for an HMM-based ASR system. Speech samples from non-natives were recognised more poorly compared those of native speakers. Out of the native speakers, those from Flanders and southern regions of the Netherlands were recognised most poorly.

Looking at whether ASR systems work equally well for different genders, there is contradictory evidence. In [3], we find that ASR systems work slightly better for females than males for the Dutch language. However, other studies concluded the opposite [7] [8]. As ASR systems entail large language models, there is another way in which downstream applications could be biased. Since word embedding models often represent stereotypical worldviews, this could affect the way that an ASR system makes its predictions.

Finally, it is known that ASR systems don't work equally well for different age groups. The most challenging age groups are children and elderly speakers. Children's speech is hard for ASR systems because of their shorter vocal tracts, slower and more inconsistent speaking pace, and less precise articulation compared to adults [3]. ASR systems face challenges in understanding the speech of elderly due to age-related alterations in their speech organs, more pronounced accents, and potential hearing impairments or health conditions that can further impact their speech patterns [9].

III. Experimental Setup

A. State-of-the-art ASR systems for the Dutch Language

In this section, we will elaborate on the ASR systems used in this experiment. Wav2vec2 and Whisper are widely-used neural network-based architectures that have delivered promising results. We compare their performance with the results from an earlier study that uses a TDNNF ASR architecture [3].

1) Wav2vec2: Wav2vec2 is a deep learning-based ASR system developed by researchers of Meta AI [4]. It is a self-supervised end-to-end architecture based on convolutional and transformer layers. The training of Wav2vec2 requires labelled data, but the amount of data needed is significantly reduced compared to traditional ASR systems due to the effectiveness of the self-supervised pretraining. For this experiment, we used a pre-trained model based on the original wav2vec2-xls-r-2b-22-to-16 model, fine-tuned on the CGN [10] and the Mozilla Common Voice 8 NL [11] datasets.[1]

2) Whisper: Whisper is an ASR system developed by researchers of OpenAI [1] that achieved substantial performance improvements in various speech recognition tasks. Contrary to Wav2vec2, which uses self-supervised techniques, Whisper uses a supervised approach, using up to 680k hours of labelled speech data from several sources. Whisper is based on an encoder-decoder Transformer architecture. This study employed the *Whisper-large-v2* model, which contains 1550 million parameters distributed across 32 layers and 20 attention heads. We conducted the evaluation in a zero-shot setting. The decoding was performed using greedy decoding with the best of 5 samplings. The following temperature weights were used for successive attempts: (0.2, 0.4, 0.6, 0.8, 1).[2]

B. Test Corpora

The JASMIN corpus [12] was used as our test set for both ASR systems. This dataset is an extension of the Corpus Spoken Dutch (Corpus Gesproken Nederlands, CGN) [10], augmenting the representativeness of different age groups and regional and non-native accents. The corpus has a balanced distribution between male and female speakers and comprises two types of speech—read speech and human-machine interaction (HMI) speech, both used in the experiments. The following speech samples were used for native speakers from both the Netherlands and Flanders:

- Dutch and Flemish children (7–11): 18h 31m
- Dutch and Flemish teenagers (12–16): 18h 31m
- Dutch and Flemish elderly adults (65+): 14h 31m

The participants in this study originate from five accent regions in the Netherlands and Belgium: West, Transitional, North, South, and Flanders. The dataset also includes two categories of non-native speakers residing in the Netherlands or Flanders, containing children and adults:

- Non-native children (7–16): 12h 21m
- Non-native adults (18–60): 12h 31m

[1]https://huggingface.co/FremyCompany/xls-r-2b-nl-v2_lm-5gram-os

[2]https://github.com/openai/whisper

C. Experiments and Evaluation

Our research builds on the approach used in [3] to investigate potential bias in speech recognition systems. Bias is understood as the WER gap between various diversity groups (based on age, gender and accent) using the same ASR system. To increase the reliability of our findings, we assess speech separately for read 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. By using the same techniques we enable a reliable comparison between the two studies.

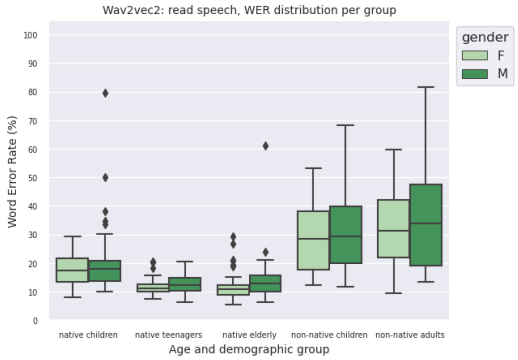
IV. Results

This section provides an overview of the word error rates (WERs) for Wav2vec2 and Whisper. Table I presents the results categorised by ASR system and age group, with a breakdown for female and male speech and an average across both genders (column *Avg*). The results are displayed separately for read speech and HMI speech. The upper portion of the table reports the WERs for native Dutch and Flemish speakers in each age group, while the lower portion details the results for non-native speakers in their respective age groups. Furthermore, the table includes the average WERs per gender across all age groups (row *Avg*), as well as for native (row *Avg natives*) and non-native (row *Avg non-natives*) speakers.

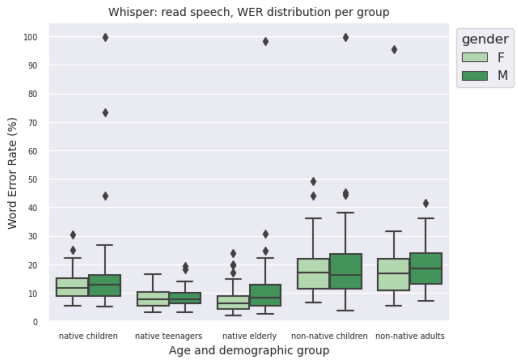
TABLE I: WER for read and HMI Speech on ASRs by gender and age group

	Age Group	ASR	Read speech			HMI speech		
			F	M	Avg	F	M	Avg
Dutch Natives	Children	TDNNF [3]	25.6	25.9	25.8	31.5	28.5	30.0
		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	TDNNF [3]	12.8	15.3	14.0	22.0	23.5	22.8
		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	TDNNF [3]	16.9	22.0	18.6	28.7	33.8	30.6
		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	TDNNF [3]	18.3	21.1	19.6	28.4	30.8	29.4
		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	TDNNF [3]	41.5	42.6	42.0	42.0	43.0	42.5
		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	TDNNF [3]	43.4	43.8	43.6	42.2	45.9	43.7
		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	TDNNF [3]	42.5	43.1	42.7	42.2	44.9	43.3
		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	TDNNF [3]	26.7	28.2	27.4	33.3	35.9	34.4
		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 significantly outperformed *Wav2vec2*, achieving an improvement of WER for both read and HMI speech. While *Wav2vec2* demonstrated average WERs of 17.3 for read speech and 27.6 for HMI speech, Whisper achieved even better results,

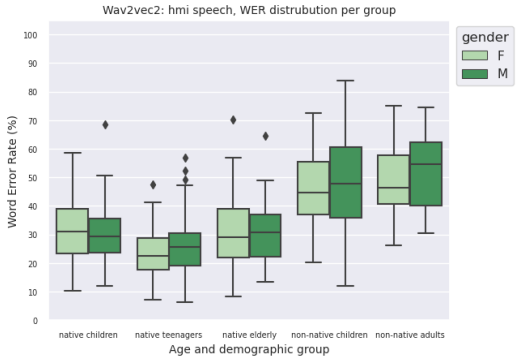


(a) Wav2vec2

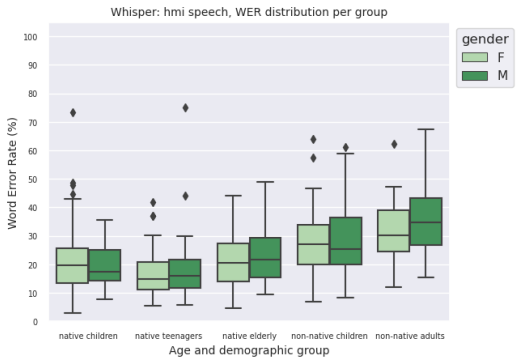


(b) Whisper

Fig. 1: Distribution of WERs for read speech



(a) Wav2vec2



(b) Whisper

Fig. 2: Distribution of WERs for HMI speech

with 13.3 for read speech and 20.5 for HMI speech. Compared to the baseline results reported in [3], which had WERs of 27.4 for read speech and 34.4 for HMI speech, both *Wav2vec2* and *Whisper* have made substantial progress in enhancing the overall performance. The results are indicators of the advancements of these state-of-the-art WER systems in a short period, significantly reducing errors by more than half in some cases compared to the previous benchmark.

The findings corroborate those presented in [3]. For example, female speech is generally better recognised than male speech across all native and non-native groups, as well as the speech style (reading or HMI). A more pronounced discrepancy in WER between native female and male speakers was observed for native elderly adults but smaller for HMI speech for all ASRs. The group of teenage speakers is the one that attains the best WER performance in both read and HMI speech. In line with the reference paper, our results show that children exhibit higher word error rate (WER) performance than the elderly group, particularly in read speech where there is a more considerable discrepancy: a difference of 6.5% for Wav2vec2 and 5.2% for Whisper for the elderly group, respectively. For non-native speakers, the performance disparities between children and adults are relatively small.

To better understand the distribution of word error rates and pinpoint potential outliers, we plot the distributions for different age groups and genders per ASR system. Figure 1, focused on read speech, depicts that both ASR systems achieve lower word error rates for native teenagers, followed by native elderly and native children. There is a substantial gap between the observed WERs for native and non-native speakers. This gap is more pronounced for *Wav2vec2*, while *Whisper* shows a smaller disparity, indicating that the overall bias is smaller for *Whisper*. Moreover, the distribution of word error rates for both children and adults appears to be more dispersed in *Wav2vec2*, with a slight improvement observed for *Whisper*.

An aspect that merits attention for *Whisper* is the outliers reaching almost a 100% WER. These outliers were found to be caused by the so-called 'failure or repetition loop' [1], where the model predicts a nonsensical repetition of words, leading to this staggering error rate. Manual investigation for these cases showed no clear pattern or explanation as to why this phenomenon occurred. When not conditioning on the previous text, the failure loop is prevented.

In Figure 2, focused on HMI speech, we notice again that the ASR systems recognised native teenagers the best. A difference, however, is that while in the read speech setting the group of native elderly speakers was the second-best recognised group, in the dialogue setting, both ASR systems performed equally well for native children and native elderly groups. Furthermore, the difference in performance between all subgroups is smaller for Whisper in the HMI speech setting, indicating that the overall bias in Whisper is smaller compared to Wav2vec2. Finally, another remarkable observation in the HMI setting is that there are fewer outliers for both ASRs.

A Kruskal-Wallis analysis and post-hoc Dunn test showed that for both ASRs and in both settings, there was a significant difference in the distribution of the performance between all native speaker groups and non-native speaker groups ($p < 0.05$). Furthermore, in the read speech setting, there was a significant difference between native children and native teenagers and between native children and native elderly for both ASR systems. In the HMI setting, a significant difference was also found between native children and native teenagers, but in contrast also between native teenagers and native elderly for both ASRs. In other cases, disparities in performance between groups were not significant ($p > 0.05$).

A. Native and Non-native Word Error Rates

Table II presents WERs for different age groups and regions for Wav2vec2 and Whisper. The results show a significant improvement in speech recognition accuracy compared to the baseline paper in [3] for all age groups and accent regions. However, while the overall performance has improved, the WERs still vary depending on the age and regional accent of the speakers.

TABLE II: Average WER for native speakers (**W**est, **T**ransitional, **N**orth, **S**outh and **F**landers regions.)

(a) Read Speech

Group	ASR	WER				
		W	T	N	S	F
Children	TDNNF [3]	N/A	23.8	28.3.2	25.6	35.3
	Wav2Vec2	N/A	15.4	21.5	16.6	17.3
	Whisper	N/A	11.3	16.5	13.5	14.3
Teenagers	TDNNF [3]	14.0	15.7	13.7	14.0	30.1
	Wav2Vec2	9.8	10.3	9.2	9.3	12.9
	Whisper	9.5	8.3	6.2	6.7	8.6
Elderly	TDNNF [3]	17.2	19.0	13.3	25.0	22.5
	Wav2Vec2	10.0	11.6	8.3	15.3	10.4
	Whisper	6.5	8.2	5.9	12.0	9.7

(b) HMI Speech

Group	ASR	ASR				
		W	T	N	S	F
Children	TDNNF [3]	N/A	31.4	27.0	30.1	47.5
	Wav2vec2	N/A	22.0	23.0	24.0	26.3
	Whisper	N/A	18.5	17.7	14.9	16.2
Teenagers	TDNNF [3]	22.6	19.7	22.6	23.8	35.5
	Wav2vec2	18.9	15.8	15.7	18.0	18.4
	Whisper	11.8	5.8	11.4	14.7	15.1
Elderly	TDNNF [3]	29.0	29.3	24.3	37.4	36.4
	Wav2vec2	24.6	24.3	24.7	31.5	23.6
	Whisper	19.1	18.7	16.7	23.7	19.9

Table IIa displays the WERs for read speech, where Whisper outperformed Wav2vec2 in all cases. Interestingly, teenagers of all regions achieved lower WERs than children and elderly. Children from the northern region had higher WERs for Wav2vec2 and Whisper. In contrast, teenagers from Flanders had a lower performance with Wav2vec2, and teenagers from the Western region had a lower performance with Whisper. Elderly participants from the southern region of the Netherlands had the worst WER performance.

In Table IIb, the WERs for HMI speech were also lower for Whisper than Wav2vec2 across all regions. While there were differences in performance between regions, the differences were not significant for children and teenagers. However, the discrepancy was higher for elderly speakers, with participants from the southern region of the Netherlands achieving the worst WER performance for both read and HMI speech.

These findings suggest that while there has been a significant improvement in WER performance, further investigation is required to address the persistent differences in performance among age groups and regions. These issues are explored further in the error analysis section.

TABLE III: WERs for non-native speakers by CEF level, with separate "F/M" columns. Excludes B2 level (n=1)

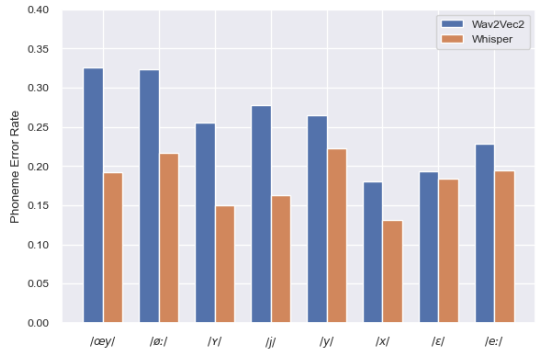
CEF	ASR	Read speech			HMI speech		
		F	M	Avg	F	M	Avg
A1	TDNNF [3]	44.6	44.4	44.5	43.7	47.6	47.0
	Wav2Vec2	24.3	29.1	26.5	35.8	47.5	40.1
	Whisper	17.1	21.7	19.2	27.5	40.6	33.3
A2	TDNNF [3]	44.9	38.7	43.3	44.4	41.4	43.5
	Wav2Vec2	24.3	25.7	24.7	35.8	37.4	36.3
	Whisper	22.3	17.9	20.9	33.2	32.1	32.9
B1	TDNNF [3]	37.6	51.5	42.6	38.4	44.7	40.4
	Wav2Vec2	21.0	24.3	22.0	29.7	33.3	30.8
	Whisper	14.3	16.6	15.0	27.8	28.7	28.1

In Table III, we present the word error rates of non-native speakers across Common European Framework of Reference for Languages (CEF) levels, with A1 representing the initial level of proficiency. The WERs for females are consistently lower than those for males across all proficiency levels and ASR systems. For instance, at the A1 level, the WERs of read speech for females using the Wav2vec2 and Whisper systems are 24.3% and 17.1%, respectively, while males are 29.1% and 21.7%, respectively. We found that the WERs are generally consistent with non-native speakers' proficiency levels regarding CEF levels. In other words, the more proficient the speaker, the lower the WER. This contrasts with the results reported in a reference paper [3], where no reduction in WER was observed with an increase in CEF level.

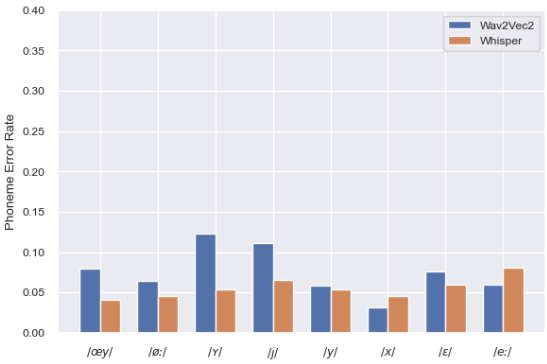
B. Error Analysis

We extend the analysis of the performance of Wav2Vec2 and Whisper with a focus on phoneme errors for different groups. Building upon the reference work in [3], we adopt a systematic approach to analyzing the sources of recognition errors by examining phoneme errors and conducting a qualitative dataset analysis. Our analysis reveals that the performance of Wav2vec2 and Whisper varies across different variables, such as non-native accents, age groups and regional accents.

For non-native speakers, specific phonemes such as /œy/, /ø:/, /ʏ/, /j/, and /y/ pose more challenges for Wav2vec2, with the highest error rates observed. In most cases, Whisper performed better than Wav2vec2, with lower error rates across all phonemes and a considerable reduction in error rates of /œy/ and /ø:/. Nonetheless, Whisper's performance improvement rate was lower for phonemes such as /y/ and /ɛ/. To visualise these findings, we present two charts in Figure 3: 3a highlights the most commonly misrecognised phonemes for both



(a) non-native speakers



(b) reference (native teenagers)

Fig. 3: Top misrecognised phonemes for non-native speakers and reference (native teenagers)

(a) native children

(b) native teenagers

Fig. 4: Top misrecognised phonemes for native children and teenagers (different phonemes per group)

Wav2vec2 and Whisper among non-native speakers, while 3b presents the same phonemes for the best-performing group of native teenagers. These results are consistent with those reported in previous studies [3], which found that specific vowels, such as /œy/, /øː/, and /ʏ/, remain consistently difficult for non-native speakers. Therefore, while Whisper generally outperformed Wav2vec2, improvements in performance for some challenging phonemes were not as significant.

Next, we examine how ASR systems perform for different age groups of native speakers. Figure 4 illustrates the performance of children and teenagers. For native children, certain phonemes like /ʃ/, /j/, /f/, /ʏ/, and /øː/ are the most difficult. These phonemes are known to be challenging for many groups, and both Whisper and Wav2vec struggled with them. In fact, both systems had error rates of around 35% when dealing with words that include the /ʃ/ phoneme, still high, as can be seen in Figure 4a. This problem was also noted in a previous study [3], where sibilant pronunciations caused confusion for the ASR systems. For example, both Wav2vec2 and Whisper substituted the word *chic* to *ziek* in some situations.

Moving on to native teenagers, ASR systems faced challenges with specific phonemes, including /ʏ/, /j/, /ʃ/, and /h/, as seen in Figure 4b. These phonemes were reported in the reference paper [3] for the same group. Error rates in common with the children group were significantly lower. For instance, for the phoneme /ʃ/, we noticed a considerable error rate reduction. While Wav2vec2 performed slightly better than Whisper, the differences were not significant - a pattern that was also observed for the same phoneme in children.

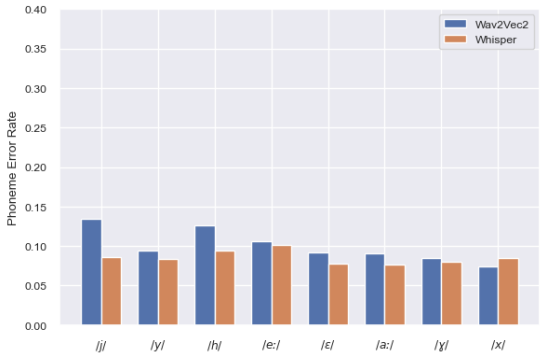
When we analysed the speech of elderly participants using our ASR systems, we noticed that both Whisper and Wav2vec2 tended to suppress the word *nou*. This word is often used as a filler in spontaneous conversations and does not contribute much to the meaning of the speech. Since this word appeared frequently, we chose to remove it from our analysis.

We investigated the speech recognition of elderly participants from different regions. Our findings indicate that speakers from the southern region displayed higher error rates for phonemes like /j/, /y/, /h/, /eː/, and /ɛ/. While most native elderly encounter challenges with these phonemes, the discrepancies were notably evident, as Figure 5 shows. Interestingly, the performance of Whisper remained relatively consistent across most problematic phonemes, except for the /j/ phoneme.

V. Discussion and Conclusion

The results demonstrate that ASR systems can perpetuate biases for specific groups, which aligns with previous research. We found that female speech was recognised better than male speech. Furthermore, teenagers' speech was recognised most accurately, followed by elderly speakers and children's speech. The worse recognition of older adults' speech was due to less

(a) Native elderly - southern region of NL



(b) Elderly participants of other regions

Fig. 5: Top misrecognised phonemes for native elderly

well articulation, highlighting the importance of considering age-related differences in ASR system design.

Concerning non-native speakers, we found that the speech of native Dutch speakers was significantly better recognised than that of non-native speakers, regardless of age. A slight positive correlation between performance and CEF proficiency level was found, in contrast with previous research, highlighting the improvement of both ASRs in capturing minor improvements among non-native speech levels. 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. This aligns with previous research demonstrating that regional accents can be challenging for ASR systems.

Comparing the ASR systems, we found that Whisper overall outperformed Wav2Vec2. Furthermore, the relative gaps in performance between groups, indicating the biases of the ASR, were overall smaller for Whisper. However, for both ASRs it is still possible to detect cases of performance disparity between groups, emphasising non-natives, children, and elderly participants, and critical cases from the southern regions.

Our study highlights the importance of considering diversity in ASR system design and the need to mitigate bias. Overall accuracy is not necessarily indicative for the performance of an ASR system for different subgroups. We think that it is important to do research on performance measures for ASR that address inclusion as well as accuracy. Measuring the WER gap between various diversity groups and the results of the phoneme analysis can already be helpful in selecting and curating training data. An important step to move forward is to be more transparent about the training data used in the development of ASR systems. Furthermore, state-of-the-art ASR systems could be finetuned with curated speech datasets and synthetic speech data to mitigate biases and further increase the inclusivity of these models. Acknowledging the biases and limitations of ASR systems and striving for equitable, inclusive, and accurate models is essential.

ACKNOWLEDGMENT

This research is part of the RAAK DRAMA project: Designing Responsible AI for Media Applications.

REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv:2212.04356*, 2022.

[2] M. K. Ngueajio and G. Washington, "Hey asr system! why aren't you more inclusive? automatic speech recognition systems' bias and proposed bias mitigation techniques. a literature review," in *HCI International 2022–Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings*. Springer, 2022, pp. 421–440.

[3] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech and Language*, vol. 84, p. 101567, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230823000864

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[5] R. Tatman, "Gender and dialect bias in youtube's automatic captions," in *Proceedings of the first ACL workshop on ethics in natural language processing*, 2017, pp. 53–59.

[6] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[7] M. Garnerin, S. Rossato, and L. Besacier, "Gender representation in french broadcast corpora and its impact on asr performance," in *Proceedings of the 1st international workshop on AI for smart TV content production, access and delivery*, 2019, pp. 3–9.

[8] M. Garnerin, S. Rossato, and L. Besacier, "Investigating the impact of gender representation in asr training data: A case study on librispeech," in *3rd Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 86–92.

[9] L. Werner, G. Huang, and B. J. Pitts, "Automated speech recognition systems and older adults: a literature review and synthesis," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 2019, pp. 42–46.

[10] N. Oostdijk, "The spoken Dutch corpus. overview and first evaluation," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*. Athens, Greece: European Language Resources Association (ELRA), May 2000.

[11] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222.

[12] 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*. Genoa, Italy: European Language Resources Association (ELRA), May 2006.