



Utterance Length and ASR Performance Disparities

Native and Non-Native Dutch-Speaking Children in Read and Human-Machine Interaction Speech

Zev Comvalius¹

Supervisor(s): Odette Scharenborg¹, Yuanyuan Zhang¹

¹EEMCS, 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 21, 2026

Name of the student: Zev Comvalius

Final project course: CSE3000 Research Project

Thesis committee: Odette Scharenborg, Catharine Oertel

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.

Abstract

Automatic speech recognition (ASR) is increasingly used in speech-based services, but its performance is often worse for children and non-native speakers. This study examines whether utterance length helps explain the higher word error rates observed for non-native Dutch-speaking children compared with native Dutch-speaking children, and whether this relationship differs between read speech and human-machine interaction (HMI) speech. Using 42,415 utterances from 188 child speakers in the Dutch Jasmin corpus and ASR outputs from Google Chirp 2 and Whisper-large-v3, the analysis compares word error rates by utterance length, measures native/non-native performance gaps, standardizes both speaker groups to the same length distribution, and estimates uncertainty with speaker-level bootstrap resampling. The results show that shorter utterances have higher error rates for both groups, speech types, and systems. Utterance length explains a measurable part of the read-speech disparity, because the gap decreases with length and is reduced after length standardization. However, a substantial gap remains. For HMI speech, the evidence is less consistent. Overall, utterance length affects how ASR disparities are measured, but it does not fully explain the disadvantage for non-native child speech.

1 Introduction

Automatic speech recognition (ASR) supports voice assistants, dictation, spoken search, customer-service systems, and educational applications. When a system makes more errors for one speaker group than for another, its practical usefulness is also distributed unequally. Inclusive-ASR research has documented such performance disparities across demographic and sociolinguistic groups, including children and non-native speakers [4, 5, 13]. Measuring where these disparities occur is important for fair evaluation: a single corpus-wide error rate can hide conditions under which a system serves one group particularly poorly.

The present study focuses on native and non-native Dutch-speaking children. Children’s speech is challenging for ASR because articulation, speaking rate, pronunciation, and language use develop with age and differ from adult speech [10]. Non-native child speech also reflects second-language pronunciation and differences in fluency and proficiency [11]. Evaluations of Dutch ASR have reported higher word error rates (WERs) for non-native than for native speakers, including children [5, 11, 13]. Those studies establish an aggregate performance gap, but an aggregate gap does not reveal whether particular kinds of utterances contribute disproportionately to it.

Utterance length is one plausible source of variation for two reasons. First, a short utterance gives an ASR decoder less context with which to distinguish competing word sequences. Second, WER is coarse for short references: one error represents a large proportion of a one- or two-word utterance. Sentence structure, word complexity, and conversational speaking style are already known to influence intelligibility and recognition performance [6, 8]. Length does not capture those properties fully, but it provides a simple, reproducible measure of how much verbal material and potential context an utterance contains. The Dutch disparity studies cited above do not report how the observed native/non-native gap changes across reference-word counts. Consequently, it remains unclear whether part of the aggregate disparity reflects different utterance-length distributions or different performance at the same length.

The scientific contribution of this work is therefore not another corpus-wide comparison. It connects a fairness outcome (the native/non-native WER difference) to a measurable property of the evaluated utterances. This shows whether the composition of an evaluation set can partly change the reported disparity and identifies length ranges that deserve closer

evaluation. Read and HMI speech are considered separately because read and spontaneous speech differ in linguistic constraint and speaking style [2, 3].

The main research question is: *To what extent is utterance length associated with ASR performance disparities between native and non-native Dutch-speaking children, as measured by word error rate, and how does this differ between read speech and HMI speech?* The term *affect* in the broader project question is interpreted statistically rather than causally: the analysis tests whether observed WER and the observed native/non-native gap vary with utterance length, but it does not claim that length is the only cause of those differences.

The main question is divided into five subquestions:

1. What is the overall ASR performance disparity between native and non-native Dutch-speaking children in read and HMI speech?
2. How is utterance length associated with WER for native and non-native child speakers?
3. Do the two speaker groups differ in their distribution of utterance lengths?
4. Does the estimated native/non-native WER gap decrease when both groups are evaluated under the same utterance-length distribution?
5. Are the length-related patterns similar for Google Chirp 2 and Whisper-large-v3?

The study answers these questions using observed error counts, length-specific WERs, length standardization, and speaker-level bootstrap confidence intervals. The same bootstrap samples are retained across systems and speech types so that their length trends can also be compared directly.

2 Data and Methodology

2.1 Corpus and speaker groups

The Jasmin corpus extends the Spoken Dutch Corpus with recordings from children, non-native speakers, and older adults. It contains prompted read speech and speech produced in a human-machine interaction (HMI) setting [2, 3]. Jasmin contains both Dutch and Flemish speech, but this study uses only the Dutch data. Jasmin’s child speech is divided into two native groups, aged 7–11 and 12–16 years, and one non-native-speaking group aged 7–16 years. The two native groups are combined and compared with the non-native group.

Proficiency is not included because Jasmin does not provide one comparable proficiency label for all three child groups [3]. Gender is available in the metadata but is not used to divide the data further. In an evaluation of the same corpus, Zhang et al. found similar WER for female and male speech in almost all system, group, and speech-type comparisons [13]. The present analysis therefore concentrates its available data on nativeness, speech type, and utterance length.

Table 1 summarizes the analyzed data. The subset contains 42,415 unique utterances from 188 speakers, 187 speakers appear in each speech condition because one speaker differs between the read and HMI subsets. The ASR systems decoded the same utterances, so the counts in the table describe the corpus once rather than once per system.

Table 1: Dutch child-speech data analyzed in this study.

Speaker group	Read speakers	HMI speakers	Read utt.	HMI utt.
Native	134	134	22,022	6,100
Non-native	53	53	11,444	2,849
Total	187	187	33,466	8,949

2.2 ASR outputs

The study analyzes Google Chirp 2 and Whisper-large-v3 outputs produced for the same Jasmin utterances in the Dutch disparity evaluation by Zhang et al. [13]. The scoring outputs contain the reference text, the ASR hypothesis, together with indications of which words are correct, and which ones require substitution, insertion or deletion to match. No ASR system was trained or decoded as part of the present project. Whisper-large-v3 is a multilingual encoder-decoder model developed with large-scale weak supervision [9]. Chirp belongs to Google’s Universal Speech Model family, which uses large-scale multilingual pretraining [12]. These systems were selected to compare two contemporary multilingual recognizers with different deployment settings: Chirp is a commercial Google Cloud system and was the best-performing system overall on the non-dysarthric Jasmin test sets in Zhang et al. [13], while Whisper-large-v3 is a widely used downloadable model from OpenAI. Comparing both systems tests whether the observed utterance-length patterns are specific to one, or appear across different modern ASR systems.

2.3 Utterance-length measurement

Utterance length is the number of words in the reference text. For each utterance this is $N = C + S + D$, where C , S , and D are the scorer’s correct, substitution, and deletion counts. It matches the whitespace-token count in every normalized reference transcription. The supplied scoring references contain no bracketed filled-pause, speaker-noise, or non-speech tokens, and this study applies no additional punctuation removal or text normalization. Contractions and hyphenated forms therefore follow the tokenization already used to produce the ASR scores.

The read-speech condition consists of prompted material, but it is not a shared fixed list of identical utterances for all speaker groups. The analyzed groups contain different prompted materials and utterance inventories [2, 3]. Their mean read lengths can consequently differ even though the speech is read rather than spontaneous.

2.4 WER and performance disparity

WER is calculated as

$$\text{WER} = \frac{S + D + I}{N} \times 100\%, \quad (1)$$

where S , D , I are the number of substitutions, deletions and insertions, and N is the utterance length. For every reported combination of system, speech type, speaker group, and length, errors and reference words are first summed across the included utterances. WER is then calculated from these pooled totals. This is done to ensure each utterance contributes in proportion to its number of reference words, the analysis does not average utterance-level WER percentages.

Following earlier work on inclusive ASR and Dutch speaker-group disparities [4, 13], performance disparity is defined as

$$\Delta(\text{pp}) = \text{WER}_{\text{non-native}} - \text{WER}_{\text{native}}. \quad (2)$$

A positive value means that the evaluated system has a higher WER for non-native child speech. The difference is reported in percentage points (pp).

2.5 Speaker-level bootstrap and statistical comparisons

The bootstrap quantifies how much the reported WERs and disparities could vary if a different sample of children from the same speaker groups had been recorded. Resampling individual utterances would incorrectly treat repeated utterances from one child as independent. Speakers are therefore the sampling units, following the motivation of cluster and blockwise bootstrap methods for dependent ASR observations [1, 7].

Within the native and non-native groups, speakers are sampled with replacement. This means that each bootstrap resample is a new pseudo-dataset of the same size as the original, created by randomly drawing speakers from the observed speakers, allowing the same speaker to appear more than once and others not to appear. Every selected speaker contributes all of their relevant error and reference-word totals, after which pooled WER and the non-native-minus-native disparity are recalculated. The procedure uses 4,000 resamples and a fixed random seed to ensure reproducibility. The 2.5th and 97.5th percentiles form a 95% confidence interval (CI). For the disparity, the null hypothesis of no native/non-native difference corresponds to a disparity of zero. Therefore, if the 95% confidence interval for the disparity excludes zero, zero is not among the plausible values under the bootstrap estimate, and the disparity is treated as statistically significant at the 0.05 two-sided level.

The speaker-level bootstrap is used to answer the subquestions as follows.

2.5.1 Overall ASR performance disparity

For subquestion 1, the bootstrap is applied to all utterances combined within each system and speech type to estimate the overall native/non-native WER disparity. This produces the confidence intervals for the aggregate gaps reported in Table 2.

2.5.2 WER per utterance length

For subquestion 2, the bootstrap is applied separately at each reference-word count to estimate uncertainty around the length-specific WERs for each speaker group. These pointwise intervals are shown in Figure 1. The same length-specific bootstrap estimates are also used to calculate the native/non-native disparity at each word count, shown in Figure 3. These intervals localize where the observed gap occurs.

2.5.3 Distribution of utterance lengths

For subquestion 3, the same speaker-level bootstrap is used to estimate uncertainty in the difference between the native and non-native mean utterance lengths. This tests whether the two speaker groups differ in their utterance-length distributions before ASR performance is considered.

2.5.4 Disparity decrease under same utterance length distribution

For subquestion 4, length standardization is used to test whether the groups’ different length distributions contribute to the aggregate disparity. Within the one- to eight-word range, length-specific native and non-native WERs are weighted by a common reference-word distribution pooled across both groups. The resulting standardized gap is compared with the unadjusted pooled gap over the same range, and the bootstrap confidence interval is calculated for the reduction in the gap.

2.5.5 Similar patterns for Google Chirp 2 and Whisper-large-v3

For subquestions 2 and 5, a reference-word-weighted linear slope is computed over one- to eight-word utterances. This slope summarizes the average WER change per additional word for each speaker group. The difference between the non-native and native slopes gives the disparity slope; a negative disparity slope indicates that the gap narrows as utterance length increases. Because the same resampled speakers are used for both ASR systems and both speech types, paired bootstrap confidence intervals can also be computed for Chirp-versus-Whisper and read-versus-HMI slope differences.

3 Results

3.1 Overall native/non-native disparity

The first subquestion asks whether a baseline disparity is present before utterance length is considered. Table 2 shows that non-native WER is higher in all four system and speech-type combinations. Every speaker-bootstrap CI for the gap excludes zero. The evaluated systems therefore recognize the non-native child speech in this dataset less accurately than the native child speech in both read and HMI conditions. The table establishes the baseline gap.

Table 2: Overall observed WER and native/non-native disparity.

Model	Speech type	Native WER (%)	Non-native WER (%)	Gap (pp)	95% CI
Chirp	Read	13.0	28.2	15.2	[12.7, 17.7]
Chirp	HMI	21.0	33.9	12.9	[10.1, 15.7]
Whisper	Read	14.5	31.1	16.6	[14.2, 19.0]
Whisper	HMI	25.3	40.1	14.8	[11.6, 17.9]

3.2 WER by utterance length

The second subquestion asks how WER is associated with length within each speaker group. Figure 1 contains four panels: rows separate Chirp and Whisper, columns separate read and HMI speech, and each panel compares the native and non-native WER curves. One-word read utterances have the highest WER in all four group-system combinations. WER then decreases over the common length range. HMI speech shows the same overall decrease, but its longer-length estimates are less precise because the number of utterances falls rapidly, for example, the HMI data contain 1,689 native and 808 non-native one-word utterances, but only 146 and 44 eight-word utterances.

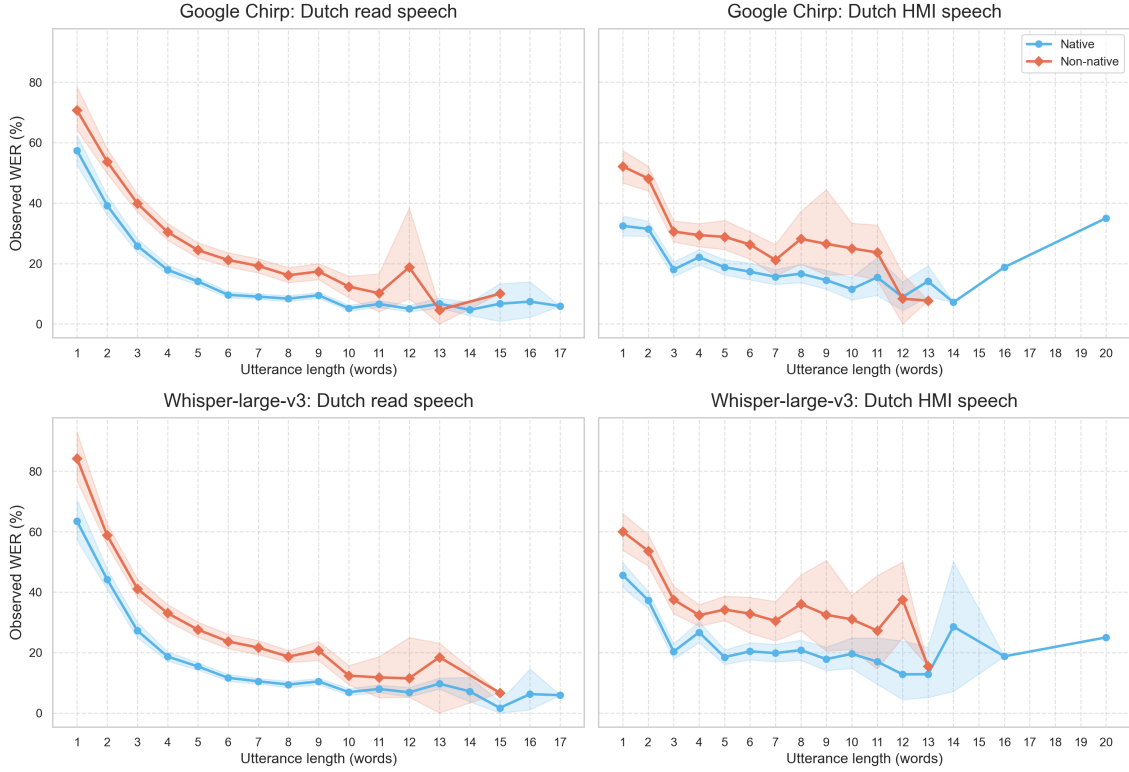


Figure 1: Observed WER by reference-word count and speaker group. Shaded areas are pointwise 95% speaker-bootstrap CIs.

Table 3 summarizes the average trend from one through eight words. Every group slope is negative and its CI excludes zero. Across both systems and speech types, each additional word is therefore associated with lower WER on average. The decrease is steeper for the non-native group in every comparison, which is why three of the four disparity slopes are significantly negative.

Table 3: Reference-word-weighted WER trends from one through eight words. Slopes are percentage-point WER changes per additional word.

Model	Speech	Native slope	Non-native slope	Gap slope	95% CI
Chirp	Read	-4.3	-6.2	-1.9	[-2.7, -1.2]
Chirp	HMI	-2.3	-4.2	-1.9	[-2.8, -1.0]
Whisper	Read	-4.5	-6.6	-2.1	[-2.8, -1.4]
Whisper	HMI	-3.2	-4.1	-0.9	[-2.1, 0.3]

3.3 Utterance-length distributions

The third subquestion concerns the composition of the dataset rather than ASR output. As Figure 2 shows, the non-native group contributes relatively more short read utterances. Its mean read length is 1.1 words shorter than that of the native group, and the speaker-bootstrap CI for this difference excludes zero. The HMI distributions are more similar, although non-native utterances are still 0.3 words shorter on average and that CI also excludes zero. These differences make length standardization necessary: an overall WER comparison otherwise combines performance differences at the same length with differences

in how frequently each length occurs. Native read utterances contain 5.3 ± 2.5 words (mean $\pm$ standard deviation), compared with 4.2 ± 2.0 for non-native read utterances. The corresponding HMI means are 3.1 ± 2.2 and 2.8 ± 1.8 words.

Figure 2 displays the complete distributions. Read speech is concentrated around four to six words, whereas HMI speech contains mostly one- to three-word utterances. The number of observations drops at longer lengths, especially for HMI speech; this reduction in sample size is considered when interpreting the uncertainty intervals.

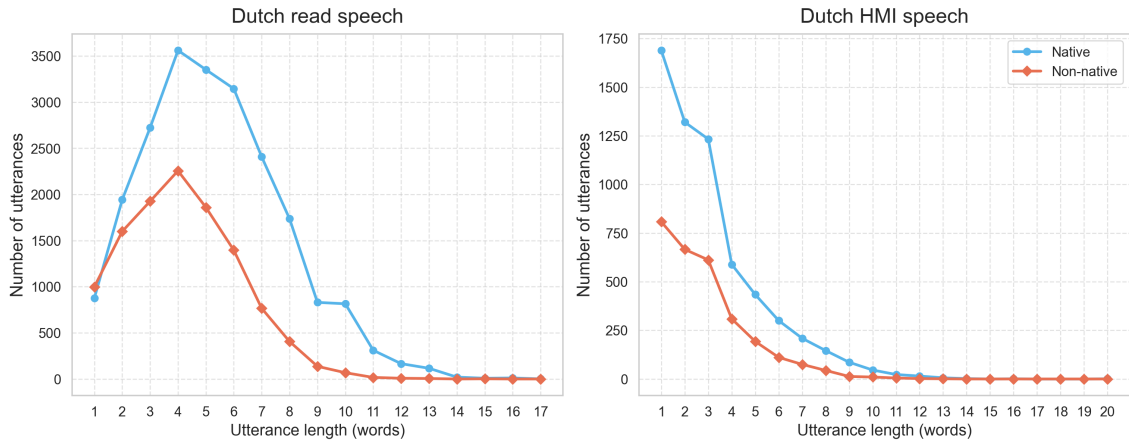


Figure 2: Number of corpus utterances by reference-word count, speech type, and speaker group.

3.4 Disparity by utterance length

Figure 3 directly shows the outcome in the main research question. Each point is the non-native WER curve minus the native WER curve from Figure 1; it is presented separately so that the size and uncertainty of the group gap can be read directly rather than inferred visually from two overlapping curves. Filled markers denote word counts for which the pointwise CI excludes zero.

For read speech, both systems show the largest gaps among the shortest utterances and smaller gaps at longer, well-populated lengths. Chirp decreases from a 13.3-point gap at one word to 7.8 points at eight words, while Whisper decreases from 20.8 to 9.3 points. The significant negative disparity slopes in Table 3 confirm that this is an overall trend rather than a conclusion based only on those two selected lengths.

In HMI speech, the individual point estimates fluctuate more. Chirp nevertheless has a negative average disparity slope over one through eight words, with a CI that excludes zero. Whisper’s estimated HMI gap slope is also negative, but its CI includes zero. The data therefore support a length-related narrowing of the HMI gap for Chirp but do not establish such a trend for Whisper.

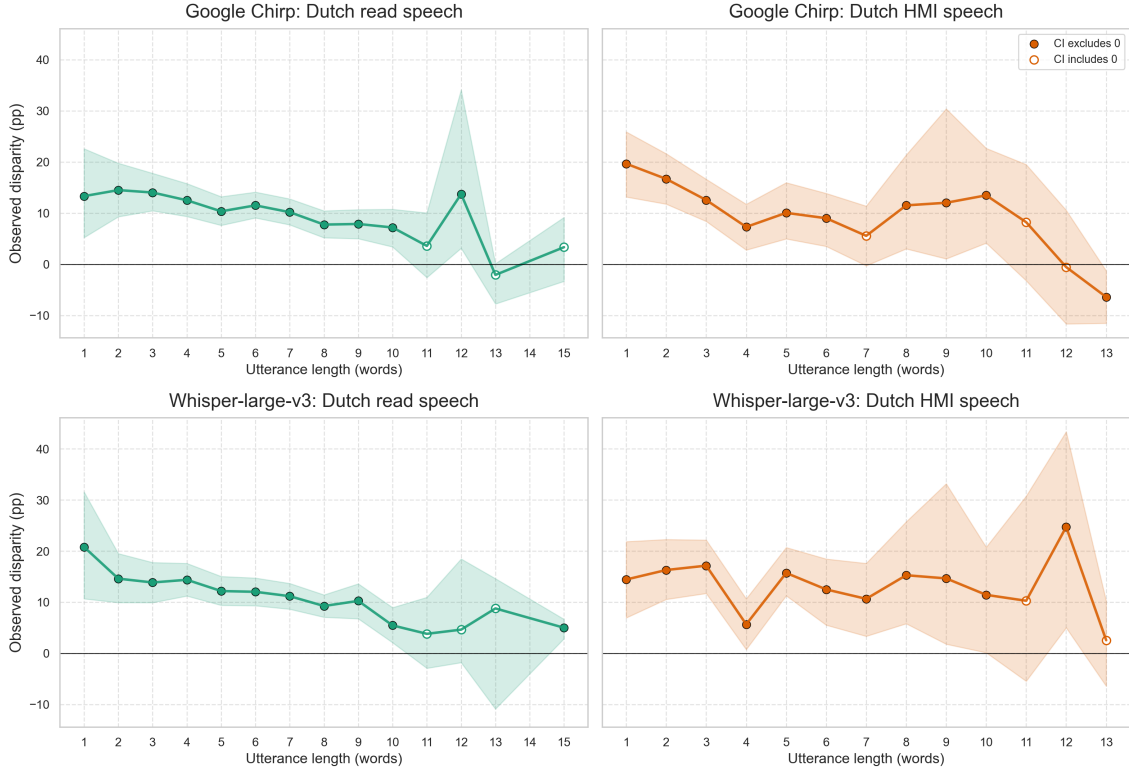


Figure 3: Observed non-native-minus-native WER disparity by reference-word count. Filled markers indicate that the pointwise 95% speaker-bootstrap CI excludes zero.

3.5 Gap after length standardization

The fourth subquestion asks whether the gap becomes smaller after accounting for the groups' different length distributions. Table 4 compares the unadjusted gap with a gap that gives both groups the same reference-word distribution over one through eight words. In read speech, standardization reduces the estimated gap by 3.2 points for Chirp and 3.3 points for Whisper; both reduction CIs exclude zero. Utterance-length composition therefore explains a measurable part of the overall read-speech disparity. The standardized gaps remain 11.2 and 12.4 points, so length does not explain most of the difference.

For HMI speech, the estimated reductions are only 0.6 and 0.5 points, and both CIs include zero. The HMI length distributions are already more similar, and the analysis does not show that standardization changes the HMI gap reliably.

Table 4: Native/non-native gap before and after standardizing both groups to one common utterance-length distribution over one through eight words.

Model	Speech	Unadjusted gap (pp)	Standardized gap (pp)	Reduction (pp)	95% CI
Chirp	Read	14.4	11.2	3.2	[2.0, 4.3]
Chirp	HMI	12.5	11.9	0.6	[-0.2, 1.5]
Whisper	Read	15.7	12.4	3.3	[2.1, 4.5]
Whisper	HMI	14.3	13.8	0.5	[-0.3, 1.5]

3.6 Comparison across speech types and ASR systems

The fifth subquestion asks whether the systems show similar length-related associations. The paired bootstrap compares their disparity slopes while retaining the same sampled speakers for both systems. In read speech, the Whisper-minus-Chirp slope difference is -0.2 points per word (95% CI: [-0.6, 0.3]); in HMI speech it is 1.0 (95% CI includes zero). There is therefore no statistically reliable evidence that Chirp and Whisper differ in how the disparity changes with length.

The analogous HMI-minus-read contrast also has a CI that includes zero for each system. Thus, the condition-specific results are clearer for read speech, both read trends and both standardization reductions are significant, but the paired slope comparison does not establish a statistically significant difference between the read and HMI length effects themselves.

4 Discussion

The main research question asks to what extent utterance length is associated with the native/non-native ASR performance disparity and whether this differs between read and HMI speech. The answer has three parts. First, WER decreases with length for both speaker groups in every evaluated system and speech type. Second, the native/non-native gap narrows with length for both systems in read speech and for Chirp in HMI speech; the Whisper HMI trend remains uncertain. Third, giving the groups the same length distribution reduces the read-speech gap by about three percentage points but does not remove it, whereas the HMI reductions are small and not statistically significant. Utterance length is therefore associated with part of the read-speech disparity, but it is not a complete explanation of the group difference. Evidence for an HMI relationship depends on the system and on whether the trend or the distribution-adjusted gap is considered.

4.1 Why WER decreases with length

WER already divides the number of errors by the number of reference words, so the higher WER of short utterances cannot be explained simply by saying that longer utterances offer more opportunities for errors. Two more specific explanations remain plausible. First, WER is granular for short references: a single substitution or deletion corresponds to 100% of a one-word reference but only 10% of a ten-word reference. Pooling errors and words avoids averaging these extreme utterance-level percentages, but it does not remove the limited denominator at each short length. Second, a short utterance provides less lexical and syntactic context to the decoder. Prior work has shown that linguistic structure, word complexity, and conversational style affect recognition and intelligibility [6, 8].

4.2 Why the disparity changes

The non-native WER decreases more steeply than the native WER in both read-speech comparisons. This produces the negative disparity slopes and shows that the gap is not constant across word counts. The length-standardization result adds a distinct interpretation. Non-native read utterances are shorter on average, and the shortest read utterances have the highest WER and some of the largest gaps. A corpus-wide comparison therefore places more weight on this difficult region for the non-native group. Applying one common length distribution reduces that compositional contribution. However, the remaining standardized

gap is still substantial, which means that performance differences at the same length account for most of the read-speech disparity.

The HMI result is less uniform. Chirp shows a negative average disparity slope, while the Whisper slope CI includes zero and neither system shows a reliable reduction after length standardization. HMI utterances are shorter for both groups, their group distributions are more similar, and the number of observations falls quickly at longer word counts. These properties limit how much distributional adjustment can change the gap and make the shape of the length-specific curves less stable.

4.3 Speech-type and system comparisons

The paired contrasts do not show a statistically reliable difference between Chirp and Whisper in their disparity slopes. The broad read-speech pattern is therefore replicated across two systems with different development approaches. The system-specific HMI conclusions differ, but the paired HMI system contrast still includes zero; the present data cannot establish that the systems truly respond differently to length.

Likewise, the paired HMI-minus-read slope contrasts include zero. It would therefore be too strong to claim a proven statistical difference between speech types. A more accurate conclusion is that the evidence is more consistent for read speech: both systems have significant negative disparity slopes and significant reductions after standardization. For HMI speech, only one of the two slopes is significant and neither standardized reduction is. This distinction between condition-specific evidence and a direct between-condition test prevents visual differences in the curves from being overinterpreted.

4.4 Contribution and implications

Earlier Dutch evaluations showed an aggregate ASR disadvantage for non-native speakers [5, 11, 13]. The present study shows that the reported size of that disadvantage depends partly on utterance length and on the length composition of the evaluation data. This matters scientifically because fairness conclusions based only on one corpus-wide WER can change with the mixture of short and long utterances. It also matters practically because short child utterances are common in interactive applications and combine high WER with a sizeable native/non-native gap.

ASR evaluations should therefore report performance by meaningful utterance characteristics in addition to an overall WER. For system development, the results motivate targeted error analysis of short child utterances and evaluation sets that do not unintentionally compare speaker groups under different length distributions. The persistent standardized gap also shows that equalizing length alone is insufficient; improving recognition for non-native child speech requires attention to other acoustic and linguistic differences.

4.5 Limitations

Word count is a limited description of an utterance. Duration, speaking rate, lexical frequency, syntactic structure, pronunciation, and disfluency may differ at the same word count and may account for part of the remaining disparity. This is especially relevant because the read-speech groups do not contribute an identical fixed list of reference utterances. Length standardization controls the word-count distribution, but it does not make the lexical content or recording characteristics identical.

The analysis is observational and cannot show that changing an utterance’s length would cause its WER to change. The linear slopes summarize an average trend over one

through eight words; Figure 1 remains necessary because the relationship is not perfectly linear. Longer bins are displayed to describe the available data, but their pointwise intervals are based on fewer speakers and utterances. Because the intervals at individual lengths are not corrected for multiple comparisons, the paper relies on the pre-specified one- to eight-word trend for its overall inference and uses the pointwise results only to locate the observed pattern.

5 Conclusion and Future Work

This study finds a clear overall ASR performance disparity: non-native Dutch-speaking children have higher WER than native children in both read and HMI speech for Google Chirp and Whisper-large-v3. Utterance length is also associated with recognition performance, because WER decreases as reference-word count increases for both groups in every evaluated condition.

For the main research question, utterance length accounts statistically for part of the disparity most clearly in read speech. The gap narrows with length for both systems, and evaluating both groups under the same length distribution reduces the estimated read-speech gap. A substantial gap remains, so differences in utterance length do not account for most of the observed performance disparity. In HMI speech, Chirp shows a narrowing gap but Whisper does not provide statistically conclusive evidence of the same trend, and length standardization does not reliably reduce either system’s HMI gap. The paired comparisons do not establish that the length associations differ between the two systems or between the two speech types.

Future work should separate the two main explanations that word count cannot distinguish. Adding utterance duration and speaking rate would test whether short timing or reduced verbal context matters, while lexical and syntactic annotations would test whether the content of short utterances is intrinsically harder. These variables could also explain the disparity that remains after length standardization. Repeating the analysis on an evaluation set with matched reference content across speaker groups would further distinguish a speaker-group performance difference from differences in the linguistic material assigned to each group.

6 Responsible Research

This project is a secondary quantitative evaluation and fairness audit: it reanalyzes an existing corpus and existing ASR outputs rather than collecting new speech or developing a new recognizer. The research question concerns system performance, not speaker ability. A higher WER for non-native children means that the evaluated system is less successful on their speech, and it should not be interpreted as a deficiency of the children.

The study uses only the speaker metadata needed to define the planned groups and to preserve repeated-measure dependence. Speaker identifiers are used as bootstrap cluster labels and are not reported in the results. No attempt is made to identify participants or infer attributes beyond the supplied metadata. The decision not to subdivide by gender is based on the negligible gender disparities reported for the same Dutch data [13], rather than on an unsupported assumption that gender can never matter.

The computational workflow is reproducible from the corpus data, ASR output files, python scripts and the methodology section. The scripts are accessible via the project

repository¹, so that the exact steps can be reproduced. The bootstrap procedures are random, but they use a fixed seed to ensure identical results every time.

Reproduction still depends on authorized access to Jasmin and on the archived versions of the Chirp and Whisper outputs used by Zhang et al. [13].

AI assistance was used for both coding the Python scripts and writing the report. OpenAI's Codex was used to help with writing the Python scripts used to execute the steps described in the methodology. The pieces of code created by Codex were manually inspected and corrected where necessary. The outputs produced by the scripts were audited for errors. OpenAI's ChatGPT was used to help write the paper, correcting grammar, punctuation and terminology. ChatGPT's suggestions were manually read and integrated.

¹https://github.com/ZevComvalius/Research_Project_TUdelft

References

- [1] Maximilian Bisani and Hermann Ney. Bootstrap estimates for confidence intervals in ASR performance evaluation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I-409–I-412, 2004. doi: 10.1109/ICASSP.2004.1326009.
- [2] Catia Cucchiarini, Hugo Van hamme, Olga van Herwijnen, and Felix 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)*, Genoa, Italy, 2006. European Language Resources Association (ELRA). URL <https://aclanthology.org/L06-1141/>.
- [3] Catia Cucchiarini, Joris Driesen, Hugo Van Hamme, and Eric Sanders. Recording speech of children, non-natives and elderly people for HLT applications: the JASMIN-CGN corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08)*, Marrakech, Morocco, 2008. European Language Resources Association (ELRA). URL <https://aclanthology.org/L08-1195/>.
- [4] Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567, 2024. doi: 10.1016/j.csl.2023.101567.
- [5] Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj 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.
- [6] C. T. Justine Hui, Sahil Jain, and Catherine I. Watson. Effects of sentence structure and word complexity on intelligibility in machine-to-human communications. *Computer Speech & Language*, 58:203–215, 2019.
- [7] Zhe Liu and Fuchun Peng. Statistical testing on ASR performance via blockwise bootstrap. In *Proceedings of Interspeech 2020*, pages 596–600, 2020. doi: 10.21437/Interspeech.2020-1338.
- [8] Courtney Mansfield, Sara Ng, Gina-Anne Levow, Richard A. Wright, and Mari Ostendorf. Revisiting parity of human vs. machine conversational speech transcription. In *Proceedings of Interspeech 2021*, pages 1997–2001, 2021. doi: 10.21437/Interspeech.2021-1908.
- [9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023. URL <https://proceedings.mlr.press/v202/radford23a.html>.
- [10] Martin Russell and Shona D’Arcy. Challenges for computer recognition of children’s speech. In *Workshop on Speech and Language Technology in Education*, 2007.
- [11] Simone Wills, Yu Bai, Cristian Tejedor-García, Catia Cucchiarini, and Helmer Strik. Automatic speech recognition of non-native child speech for language learning applications. In *12th Symposium on Languages, Applications and Technologies (SLATE 2023)*,

volume 113 of *Open Access Series in Informatics (OASICS)*, pages 7:1–7:8. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi: 10.4230/OASICS.SLATE.2023.7.

- [12] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, 2023. doi: 10.48550/arXiv.2303.01037.
- [13] Yuanyuan Zhang, Thomas De Valck, and Odette Scharenborg. Speech recognition performance disparities between dutch diverse speaker groups. *Phonetica*, 2026. doi: 10.1515/phon-2025-0061.