



Which Word Categories Contribute to Higher ASR Error Rates for Native Dutch Children Than for Native Dutch Teenagers?

Adomas Valiukevičius
Delft University of Technology

CSE3000 Research Project
Supervisors: Odette Scharenborg and YuanYuan Zhang
June 2026

Abstract

Automatic speech recognition systems can be less accurate for children than for older speakers. For Dutch, previous evaluations found higher word error rates for native children than for native teenagers, but they did not show which types of words contribute to the difference. This paper studies whether that difference is concentrated in particular word categories. A word category is the part of speech assigned to a word in the reference transcript, for example noun, verb, pronoun, article, or preposition. The analysis uses native Dutch child and teenager speech from the JASMIN-CGN corpus in two speaking styles: read speech and speech from human machine interaction tasks. The analysis compares Google Chirp 2 and Whisper-large-v3 outputs with the reference transcripts and groups substitutions and deletions by word category. Insertions are reported separately because an inserted system word has no matching reference word category. Children have higher word error rates than teenagers in every condition. The difference is about 9 percentage points in read speech and about 4 to 5 percentage points in human machine interaction speech. In read speech, most of the substitution and deletion difference comes from nouns, verbs, and pronouns. In human machine interaction speech, errors are spread across categories, so no single category explains much of the difference. A control that compares utterances with the same counts of word categories reduces the verb and pronoun differences, while the noun difference remains largest. A check of the specific words behind these categories shows why this matters: many noun and verb errors come from words found only in the child read reference transcripts, whereas pronoun errors mostly involve words used by both groups. The results show that word categories are useful for locating where the error rate difference appears, but they are not a complete explanation without checking the specific words and speaking material behind each category.

1 Introduction

Automatic speech recognition (ASR) systems are increasingly used in voice assistants, dictation tools, automatic subtitling, customer service systems, and other speech applications [4, 10]. Because these applications are used by speakers of different ages and backgrounds, ASR systems should be evaluated on speech that differs by age, accent, language background, and speaking style. Previous work shows that current ASR systems do not perform equally well for all speaker groups. Koenecke et al. [10] found that ASR systems made almost twice as many errors for Black speakers as for white speakers in American English. For Dutch, Feng et al. [4], Fuckner et al. [5], and Zhang et al. [19] report performance disparities for regional accent, Dutch spoken as a second language, and age, while Feng et al. [4] and Fuckner et al. [5] report only small disparities for gender.

This paper focuses on the ASR performance disparity between native Dutch children and native Dutch teenagers. Reliable recognition of children’s speech is important because children interact with voice assistants and because educational applications such as reading tutors and language learning tools depend on it [6], yet children’s speech has long been among the most difficult speech for ASR systems [15, 6]. Children differ from older speakers in articulation, pronunciation consistency, speaking rate, and voice characteristics, and these properties keep changing during development [11]. In line with this, a recent evaluation of current ASR systems on Dutch found that the speech of native children is recognized considerably worse than the speech of native teenagers [19].

That evaluation, like most disparity studies, reported word error rate (WER), which counts substituted, deleted, and inserted words relative to a reference transcript. WER measures ASR performance on a test set. In this paper, the difference between the WERs of native children and native teenagers is used to describe the performance disparity between the two groups. That difference shows how large the disparity is, but not which reference words contribute to it. Vocabulary develops with age [1], so children and teenagers may use different words, and in a corpus the two groups may also be given different texts to read. A higher WER for children can therefore come from at least three sources: the ASR system may make more errors on the

same words when children say them, children may say more words that are difficult for ASR in general, or the texts assigned to children may contain different words from the texts assigned to teenagers.

One way to make ASR errors more interpretable is to analyze them per word category. In this paper, a *word category* is the grammatical category assigned to a word in the reference transcript. Examples are noun, verb, pronoun, article, preposition, conjunction, adjective, adverb, and interjection. This means that the analysis asks whether ASR errors are concentrated in particular kinds of reference words, not just whether one speaker group has a higher overall WER.

This definition is close to what linguistics and natural language processing usually call a part of speech (POS) category. Mansfield et al. proposed to analyze transcription errors beyond a single overall WER by separating substitutions, deletions, and insertions and by grouping words into function, content, and conversational words [13]. Goldwater et al. showed that recognition difficulty also depends on properties of individual words, such as frequency and prosody, and not only on word class [7]. For conversational speech, machine errors concentrate on short words, reductions, filled pauses, and backchannels [17]. For Dutch conversational speech, Lopez et al. showed that ASR systems frequently omit interactional words, such as short responses, filled pauses, and backchannels, which overall WER does not reveal [12]. However, Dutch speaker group disparity studies have reported WER by speaker group, speech type, or ASR system without decomposing the native child and teenager WER difference by reference word category [4, 5, 19]. Whether this WER difference is concentrated in particular word categories is therefore unknown.

Word category analysis is also relevant for comparing speaking styles. The JASMIN-CGN corpus [2] used in the Dutch disparity studies contains two speech types: read speech, in which speakers read assigned texts aloud, and human machine interaction (HMI) speech. HMI speech is conversational speech recorded while speakers answered questions from a dialogue system. It contains more short responses, hesitations, and interjections such as *eh*, *ja*, and *nee*. These are referred to in this paper as conversational words: short words that manage interaction rather than mainly naming objects, actions, or properties. ASR systems have different WERs on Dutch read and HMI speech [19]. Conversational words may also be handled differently across systems: some systems output them, while others omit them or replace them with a more standard written form [12, 13]. The two speech types are therefore analyzed separately.

The main research question of this paper is:

To what extent are ASR performance disparities between native Dutch children and native Dutch teenagers concentrated in particular word categories, and how does this differ between read speech and HMI speech across Google Chirp 2 and Whisper-large-v3?

The main question is answered through four subquestions:

1. How does the distribution of word categories in the spoken material differ between children and teenagers?
2. Which word categories contribute the largest parts of the WER difference between children and teenagers in read speech and HMI speech?
3. For the largest read speech categories, how much of the difference between children and teenagers remains after comparing utterances with the same counts of nouns, verbs, pronouns, and other POS categories?
4. For the largest read speech categories, to what extent does the difference come from words used by both groups or from words found only in the child read reference transcripts?

Throughout the paper, word categories are treated as a way to identify where in the reference transcripts the WER difference appears, not as final explanations by themselves. A category can stand out because one group’s reference transcripts contain more words of that category, because the ASR systems make more errors on that category for one group, or because a few specific words within the category are misrecognized very often [7]. The analysis therefore moves from categories to the individual words behind them.

To answer these questions, this study analyzes Dutch read and HMI speech of native children and teenagers from the JASMIN-CGN corpus, using ASR outputs from the Dutch evaluation by Zhang et al. [19]. The two selected systems are Google Chirp 2 and Whisper-large-v3. They are used because they had low WERs for the native child and teenager groups in that evaluation and because they represent different system types: a commercial multilingual ASR service and an openly available multilingual ASR model. The analysis also relates WER to speakers’ exact age as a supporting check, to test whether the group difference is consistent with a gradual age pattern. The remainder of this paper is structured as follows. Section 2 describes the data, the ASR systems, and the analysis methods. Section 3 presents the results. Section 4 discusses the findings, their relation to previous work, and the limitations. Section 5 reflects on responsible research, and Section 6 concludes.

2 Data and Methodology

2.1 Dataset

This study uses the JASMIN-CGN corpus [2], a Dutch speech corpus that extends the Spoken Dutch Corpus with speaker groups that were underrepresented in it. It contains speech recorded in the Netherlands and Flanders from children, speakers of Dutch as a second language, and older adults. Each speaker contributed two speech types: read speech, in which the speaker read assigned texts aloud, and HMI speech, in which the speaker answered spoken questions of a simulated dialogue system. The HMI speech is conversational and contains more short answers, hesitations, and interjections than the read speech.

This work uses the speech of the native Dutch children (ages 7 to 11) and the native Dutch teenagers (ages 12 to 16) recorded in the Netherlands. The corpus group labels are not perfectly strict age filters in this subset: the exact age is missing for 25 of the 135 included speakers, and a few speakers fall slightly outside the nominal age range of their group. Among speakers with exact age, the child-labeled group ranges from 6 to 13 years and the teenager-labeled group from 12 to 18 years. The comparison in this paper therefore follows the corpus group labels rather than exact age. There are 72 child speakers and 63 teenager speakers in total; one child speaker appears only in read speech and one only in HMI speech, so Table 1 lists 71 child speakers per speech type. Table 1 summarizes the selected subset.

Table 1: Statistics of the selected subset of the JASMIN-CGN corpus. Reference words is the total number of words in the reference transcripts.

Speaker group	Speech type	Speakers	Utterances	Reference words
Children	Read	71	12,984	60,026
Children	HMI	71	3,526	11,296
Teenagers	Read	63	9,038	56,661
Teenagers	HMI	63	2,574	7,498

2.2 ASR systems

The analysis is based on the ASR outputs produced by Zhang et al., who evaluated twelve current ASR systems on the JASMIN-CGN speaker groups using the ESPnet toolkit [19, 18]. From those twelve systems, two were selected for the present analysis because they were among the best performing systems on the Dutch native child and teenager groups and because they represent different access and model types. Google Chirp 2 is treated here as a commercial multilingual ASR service. Whisper-large-v3 is an openly available multilingual model from the Whisper family [16]. Comparing both systems tests whether the word category pattern is specific to one ASR system or appears across two different kinds of current multilingual ASR systems.

2.3 WER source and computation

The ASR systems were not run again for this paper. The analysis starts from the ESPnet scoring outputs of Zhang et al. [19, 18]. Those outputs include the reference transcript, the ASR output, and the SCLITE scoring summary [14] for each included speaker group, speech type, and ASR system. The SCLITE summary is enough to report total WER, but not enough for the word category analysis: it gives the total number of substitutions, deletions, and insertions, but not which reference word was involved in each error.

WER is used because it is the standard metric in the source ASR evaluation and because it directly counts substituted, deleted, and inserted words. The formula is included because the later category contribution measure uses the same denominator as WER: all reference words in the condition, not only the words in one category.

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

where S , D , and I are the numbers of substituted, deleted, and inserted words, and N is the total number of reference words. The headline WER values in this paper follow the same definition and total word and error counts as the original ESPnet scoring with SCLITE. The extra matching step is needed because this analysis asks not only what the WER is, but which word categories contribute to the WER difference. SCLITE gives aggregate counts such as substitutions, deletions, insertions, and WER. To say that a noun was substituted or a pronoun was deleted, however, the analysis must know which reference word each error belongs to. The analysis therefore reconstructed a table with one row per reference word, showing whether that word was correct, substituted, or deleted. As a validation check, the recomputed word counts and total error counts were compared with the original scoring summaries and matched for all eight combinations of system, speaker group, and speech type. The split between substitutions, deletions, and insertions can differ slightly from SCLITE because ambiguous alignments can be resolved differently; the operation labels used for category attribution are therefore the reconstructed matching labels.

To build that table, each utterance is matched word by word: each reference word is matched to the corresponding ASR word, to a gap if the ASR system omitted it, or to an extra ASR word if the system inserted one. The reference transcript and the ASR output are lowercased, Unicode normalized, and split into words on whitespace. The two word sequences are then compared with edit distance. For example, if the reference transcript is *de hond loopt* and the ASR output is *de kat loopt*, the comparison marks *hond* as a substituted reference word. Because *hond* is a noun, that error is counted as a noun error. If the ASR output is *de loopt*, *hond* is a deleted reference word and is again counted as a noun error. If the ASR output is *de hond snel loopt*, *snel* is an insertion and receives no reference word category. When several matches have the same edit distance, this reconstruction chooses exact matches or substitutions before deletions and insertions [14].

2.4 Word category annotation

Operationally, each word category is a part of speech (POS) label assigned to a reference word before it is compared with the ASR output. POS tagging is performed with spaCy, an open source natural language processing library, using its small Dutch pipeline, `nl_core_news_sm` [8]. To keep the POS labels aligned with the WER tokens, the reference transcripts are passed to spaCy as word sequences that are already tokenized, rather than allowing spaCy to split the text independently.

The spaCy tags are mapped to the word categories used in this paper: noun, verb, adjective, adverb, numeral, article, determiner, pronoun, preposition, conjunction, auxiliary, and interjection. The main category figures and tables report noun, verb, adjective, adverb, article, pronoun, preposition, and conjunction. Numerals, determiners, auxiliaries, and other labels are retained in the totals but are not interpreted as separate headline categories. This keeps the reported comparisons focused on categories that are frequent and directly relevant to the research question. Interjections and response words such as *ja* and *nee* are analyzed with conversational words following Mansfield et al. [13], because they mainly manage interaction in the HMI task.

Automatic POS tags can be wrong, especially for the lowercased proper names that occur in the child reading material. A small set of high impact story names was manually overridden as nouns; the remaining automatic POS labels were not fully audited, so the category assignments of rare words are treated as uncertain in the interpretation.

Substituted and deleted words are counted as errors of the POS category of the reference word. Inserted words are not assigned to reference word POS categories, because an inserted ASR word has no corresponding reference word. They are instead reported as overall insertion rates and discussed separately in Section 3.6.

2.5 Category measures

Two category measures are used because they answer different questions. A *category error rate* asks how often words of one category are misrecognized within a speaker group. It is the percentage of reference words in that POS category that were substituted or deleted. For example, if children speak 100 reference nouns and 20 of those nouns are substituted or deleted, the child noun category error rate is 20%.

A *category contribution to the WER difference* asks how much one POS category adds to the substitution/deletion part of the WER difference between children and teenagers. The denominator is all reference words, not only the words in that category, because WER itself is divided by all reference words. For each group, the substituted and deleted reference words in one category are divided by the total number of reference words for that group. The teenager value is then subtracted from the child value. This adapts Mansfield et al.’s frequency aware category analysis [13]: category errors are put on the same denominator as WER so that each value shows how much that category adds to the substitution/deletion part of the WER difference. A positive contribution means that substituted and deleted words from that category add more to child WER than to teenager WER. Both measures are needed: a category can have a high error rate but occur rarely, while a common category with a moderate error difference can make up a larger part of the substitution/deletion difference.

Table 2 shows the difference with a toy example. The rare category has the higher category error rate, but nouns contribute more to the WER difference because the noun errors make up a larger share of all reference words.

2.6 Control for word category counts

The raw category comparisons can be affected by the kinds of utterances children and teenagers speak. For example, suppose child read utterances contain more short utterances with pronouns

Table 2: Toy example showing why category error rate and category contribution answer different questions. Each group has 1,000 reference words in total; pp means percentage points.

Category	Child words / errors	Teenager words / errors	Category error rate	Contribution
Noun	200 / 40	200 / 20	20% child, 10% teenager	+2.0 pp
Rare category	10 / 5	10 / 1	50% child, 10% teenager	+0.4 pp

than teenager read utterances. A raw pronoun difference could then reflect the different utterance mix, not only a difference in how well the ASR system recognizes pronouns. The control in this section tests this alternative explanation.

The control represents each utterance by the number of words it contains from each word category. For example, *ze ziet de hond* and *de hond ziet ze* have the same category count pattern: one pronoun, one verb, one article, and one noun. The word order is different, but the category counts are the same. This is the sense in which the utterances are comparable for this control.

The controlled category error rate is computed in four steps. First, each utterance is assigned a category count pattern from the reference transcript. The ASR output is not used to decide this pattern; it is used later only to count errors. Second, the control includes only patterns that occur at least once in both the child data and the teenager data. Utterances with patterns found in only one group are not used in this controlled comparison, because there is no matching pattern in the other group. Third, for the category being analyzed, the category error rate is computed separately for children and teenagers within each shared pattern. Fourth, these pattern rates are averaged using the same weights for both groups. The weight of a pattern is the pooled number of child and teenager reference words from the analyzed category in that pattern. This is a direct standardization approach: both groups are compared under the same distribution of category count patterns [9].

The result is a category error rate after the control. The control also matches utterance length, because utterances with the same category counts have the same number of words. It does not match the exact words or word order. A stricter control would compare only utterances with the same exact reference text, but this is not usable for this corpus: fewer than 1% of read utterances have a reference text that occurs in both groups. By contrast, 55.1% of child read utterances and 52.6% of teenager read utterances share their category count pattern with the other group.

A category error rate after the control can be higher or lower than the raw category error rate because the control changes which utterance patterns receive weight. The important quantity is the difference between children and teenagers after the control. If that difference becomes smaller, then part of the raw difference was due to different utterance length or different category count patterns. If the difference remains large, the remaining difference is less likely to be only an utterance mix artifact. As an additional check, the category differences are also standardized over utterance length bins and over word length bins within each category, to test whether the results are explained by children producing shorter utterances or different word lengths.

2.7 Specific words and additional checks

The category count control compares utterance patterns, but it still does not compare the exact same words. The fourth subquestion therefore examines the specific words behind the main categories in read speech. Shared words are words that occur in both child and teenager reference transcripts. Words used only by children occur in the child read reference transcripts but not in the teenager read reference transcripts; in read speech, these mostly appear to be story words from the child read material. The analysis separates positive contribution gaps into the part from shared words and the part from words used only by children. It therefore shows where the child higher part of a category result comes from, not a full signed decomposition of

all individual word effects. Words that occur only in the teenager references are not counted here, because errors on those words increase the teenager side of the comparison and therefore reduce the difference in which children have higher WER.

As a second check, category error rates are recomputed over shared words only. This compares children and teenagers on identical written words, although the words can still occur in different sentences and frequencies. Because this restriction removes data, the share of child reference words kept by the shared word comparison is reported per category.

Three additional checks support the interpretation but are not separate research questions. First, some errors are spelling or style variants rather than clear misrecognitions, for example the reference word *me* aligned with the ASR output *mijn*. To quantify this, references and outputs are rescored after mapping known variant families (*éen/een*, *me/m'n/mijn*, *'t/het*, *'s/s*) to a single form, and the WER difference is recomputed. These rescored results estimate how sensitive the main results are to such variants. Second, uncertainty is estimated with bootstrap resampling over speakers [3]. Raw category contributions are resampled 5,000 times, while the length controlled support check is resampled 2,000 times. Third, exact age is available for 110 speakers overall. For each speech type and ASR system, Spearman's rank correlation between age and WER per speaker is computed for speakers that have both exact age and data in that condition. This age check supports the motivation for comparing corpus child and teenager groups; it is not one of the four main subquestions.

2.8 Analysis overview

The results are ordered to match the research questions. Table 3 shows how each question is answered and which result to read for it. Conversational words, spelling variants, insertions, and exact age are treated as support checks rather than separate research questions.

Table 3: Connection between the research questions, the analysis steps, and the result sections.

Question	Analysis step	Main result
Overall group difference	Compare WER between children and teenagers for each speech type and ASR system.	Table 4
Which word categories occur more often in each group?	Compare POS category shares in the reference transcripts.	Figure 1
Which categories contribute most to the WER difference?	Compute category contributions to the substitution and deletion part of the WER difference.	Figure 2
How much remains after comparing similar utterance patterns?	Recompute category error rates after comparing utterances with the same word category counts.	Figure 3
Do category results come from shared words or words found only in the child read reference transcripts?	Split read speech category contributions by whether the specific words are shared by both groups or found only in the child read reference transcripts.	Tables 5 and 7
Support checks	Inspect spelling variants, conversational words, insertions, uncertainty, and exact age.	Section 3.6 and Figure 4

3 Results

3.1 Overall WER difference

This subsection establishes the overall WER difference between children and teenagers that the later word category analyses examine. Table 4 reports WER and its three error components for each speech type and ASR system. Children have higher WERs than teenagers in all four conditions. The WER difference is about 9 percentage points in read speech and 4 to 5 percentage points in HMI speech. HMI speech has higher WERs overall, but the group difference is larger in read speech. This means that the difference is not simply a result of HMI speech being more difficult for the systems. Substitutions and deletions are the main source of the read speech WER difference and can be assigned to POS categories of reference words; insertions are discussed separately in Section 3.6.

Table 4: Overall WER per speaker group, speech type, and ASR system. The WER totals match the original ESPnet scoring with SCLITE for aggregate word and error counts. Sub, Del, and Ins are substitution, deletion, and insertion rates from the local word matching used for category attribution, expressed as percentages of the reference words.

Speech type	ASR system	Group	Sub (%)	Del (%)	Ins (%)	WER (%)
Read	Google Chirp 2	Children	11.9	4.2	1.3	17.4
Read	Google Chirp 2	Teenagers	5.1	2.8	0.4	8.3
Read	Whisper-large-v3	Children	12.6	3.8	2.4	18.8
Read	Whisper-large-v3	Teenagers	6.1	2.8	1.1	10.0
HMI	Google Chirp 2	Children	15.6	5.5	1.8	22.8
HMI	Google Chirp 2	Teenagers	12.7	4.3	1.3	18.2
HMI	Whisper-large-v3	Children	17.8	5.1	4.2	27.0
HMI	Whisper-large-v3	Teenagers	15.1	3.9	3.7	22.7

3.2 Word category distribution

This subsection answers the first subquestion: how the POS category distribution differs between the child and teenager reference transcripts. Figure 1 shows the child share minus the teenager share for each POS category. Positive values mean that a category is more common in the child reference transcripts; negative values mean that it is more common in the teenager reference transcripts. In read speech, pronouns and verbs make up a larger share of the child transcripts, while nouns make up a larger share of the teenager transcripts. In HMI speech, the category share differences are smaller and no category strongly separates the two groups. These distribution differences matter because a category can contribute to the WER difference either because one group says more words of that category or because the ASR system makes more errors within that category.

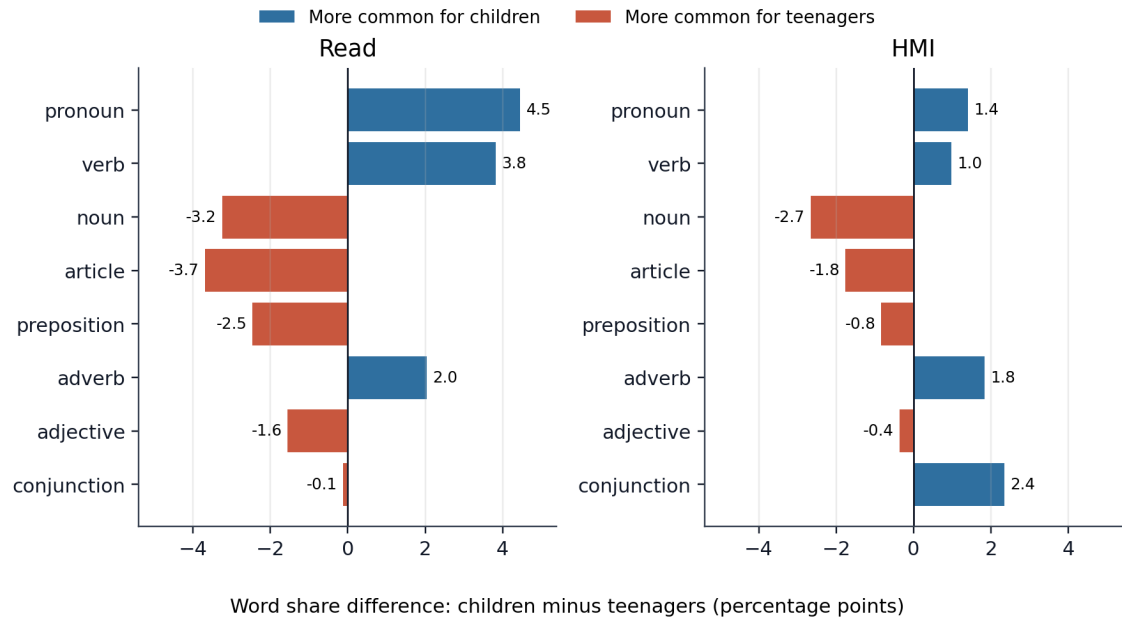


Figure 1: Difference in the share of each POS category between the child and teenager reference transcripts, in percentage points. Positive values mean the category makes up a larger share of the child transcripts; negative values mean a larger share of the teenager transcripts.

3.3 Category contributions to the WER difference

This subsection answers the second subquestion: which POS categories contribute most to the substitution/deletion part of the WER difference between children and teenagers. Figure 2 shows category contributions to that part of the WER difference. A positive value means that substituted and deleted words from that category add more to child WER than to teenager WER. In read speech, nouns contribute the largest part of this difference, followed by verbs and pronouns, and this ranking is the same for both ASR systems. Together, these three categories contribute 6.5 of the 8.2 percentage point substitution/deletion difference for Google Chirp 2 and 5.4 of the 7.5 percentage point difference for Whisper-large-v3. In HMI speech, each category contributes a smaller part of the substitution/deletion difference; nouns, verbs, and pronouns together contribute about 1.8 percentage points, less than half of the substitution/deletion difference for either system. The interpretation is therefore different by speech type: in read speech, most of the category contribution comes from nouns, verbs, and pronouns; in HMI speech, the category contribution is distributed over several smaller categories. Speaker bootstrap intervals support that the noun, verb, and pronoun contributions are positive for both systems (Google Chirp 2: [3.0, 4.0], [1.8, 2.4], and [1.1, 1.5]; Whisper-large-v3: [2.0, 3.1], [1.6, 2.2], and [1.0, 1.5]).

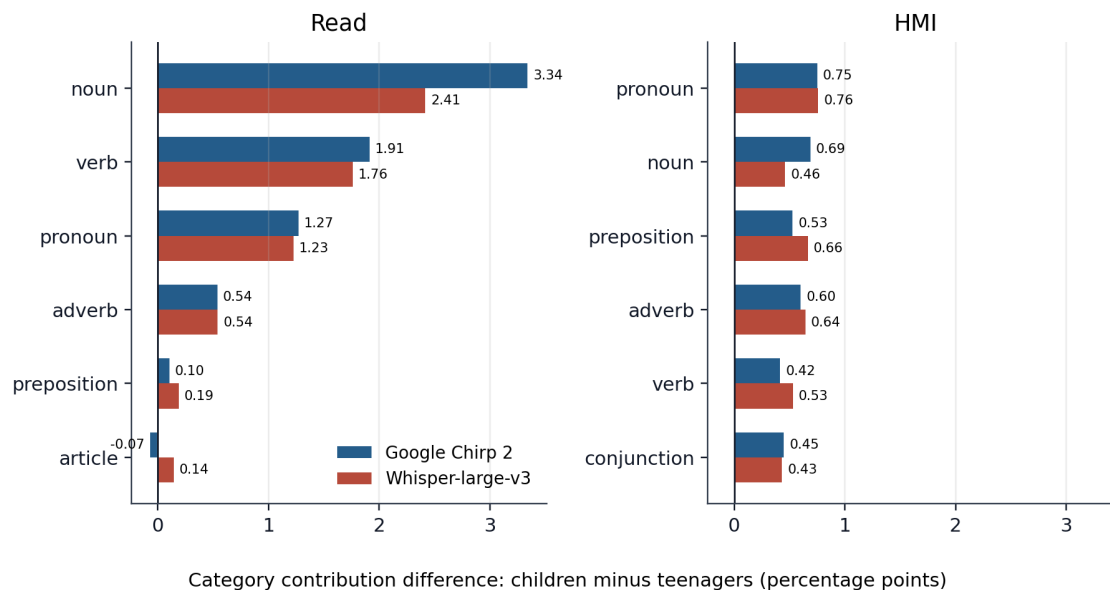


Figure 2: Category contribution to the substitution/deletion part of the WER difference between children and teenagers, in percentage points. Each bar shows the child value minus the teenager value for substituted and deleted reference words in that POS category. The figure shows the largest positive categories for each speech type; smaller and negative categories are omitted from the plot but included in the totals discussed in the text. Insertions are not assigned to reference word POS categories.

The read speech noun result needs careful interpretation. It is not caused by children reading more nouns: Figure 1 shows that nouns are more common in the teenager read transcripts. Instead, Figure 3 shows that the nouns read by children have much higher category error rates. Verbs combine two mechanisms, because verbs are more common in child read transcripts and have higher child category error rates. Pronouns look similar to verbs at this category level, but Section 3.5 shows that the specific words behind them are different.

3.4 Control for word category counts

This subsection answers the third subquestion: how much of the read speech noun, verb, and pronoun differences remains after comparing utterances with the same counts of word categories. Figure 3 shows the child minus teenager category error rate difference before and after this control. This is a category error-rate difference, not the category contribution measure in Figure 2. For verbs and pronouns, the difference becomes much smaller after the control, for both ASR systems. This means that part of the raw verb and pronoun results reflects differences in utterance length or in the mix of category count patterns being compared. The noun difference also becomes smaller, but it remains the largest difference after the control for both systems. The controlled comparison keeps the tokens in category count patterns shared by both groups: 44.9% of noun tokens, 45.7% of verb tokens, and 38.8% of pronoun tokens. The noun result is therefore less likely than the verb and pronoun results to be only an utterance structure artifact, but it is still based on the shared part of the category count distribution; the next subsection examines which specific nouns contribute to it.

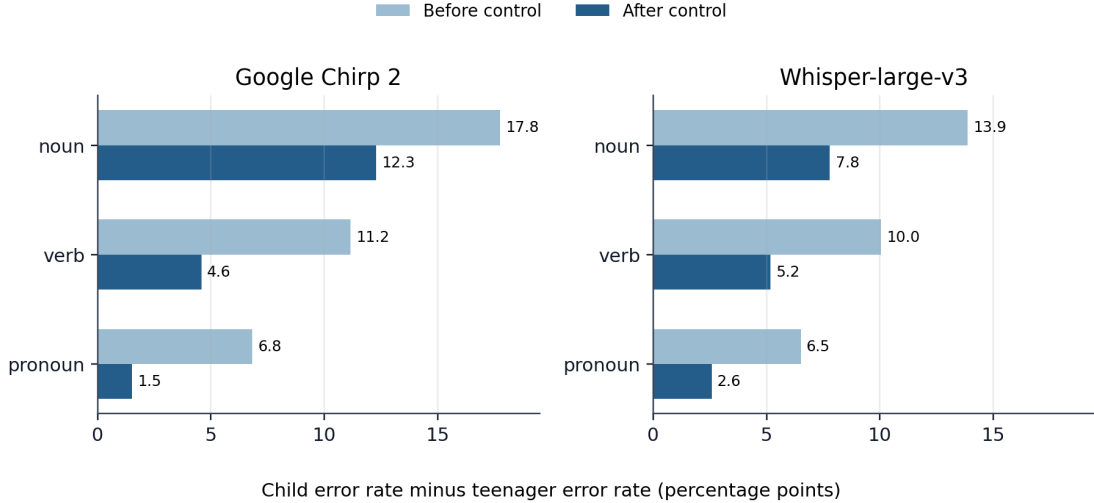


Figure 3: Read speech child minus teenager category error rate differences before and after comparing utterances with the same counts of word categories. This figure uses category error-rate differences, not the category contribution measure in Figure 2. Values are percentage point differences. The verb and pronoun differences shrink strongly after the control, while the noun difference remains largest.

3.5 Specific words behind nouns, verbs, and pronouns

This subsection answers the fourth subquestion: whether the read speech noun, verb, and pronoun differences come from words used by both groups or from words found only in the child read reference transcripts. Table 5 decomposes the positive category contributions into those two parts. Words used only by children make up almost all of the noun contribution and most of the verb contribution. Shared words make up most of the pronoun contribution. This means that the noun and verb results mainly reflect words found only in the child read reference transcripts, while the pronoun result is more directly comparable between the two groups.

Table 5: Decomposition of the positive contribution from specific words within the read speech category contribution differences. Shared words occur in both child and teenager reference transcripts; words used only by children occur in child read reference transcripts but not in teenager read reference transcripts. Rows sum to 100%.

ASR system	Category	Shared words (%)	Child-only words (%)
Google Chirp 2	Noun	7.5	92.5
Whisper-large-v3	Noun	7.1	92.9
Google Chirp 2	Verb	18.8	81.2
Whisper-large-v3	Verb	19.4	80.6
Google Chirp 2	Pronoun	86.6	13.4
Whisper-large-v3	Pronoun	88.6	11.4

Table 6 lists words used only by children with the highest Google Chirp 2 error counts. They are names and story words in the child read reference transcripts, such as *maartje*, *robbie*, and *juf*. The noun and verb results in read speech are therefore largely about specific words: the child read reference transcripts contain story words that Google Chirp 2 misrecognizes often, and these words are mostly nouns and verbs. The similar noun and verb decomposition for Whisper-large-v3 in Table 5 supports the same interpretation for that system.

The pronoun result comes mostly from words that both groups say. For example, with

Table 6: Read speech words used only by children with the highest Google Chirp 2 error counts. Occurrences is the number of times the word appears in the child reference transcripts, Errors is the number of substitutions or deletions for that word, and Speakers is the number of child speakers who read it.

Word	Category	Occurrences	Errors	Speakers
<i>juf</i>	Noun	293	138	22
<i>maartje</i>	Noun	155	109	12
<i>robbie</i>	Noun	241	104	12
<i>jess</i>	Noun	88	81	14
<i>zegt</i>	Verb	437	89	71
<i>denkt</i>	Verb	78	36	56

Google Chirp 2, the pronoun *ze* is misrecognized in 17% of its child occurrences and 6% of its teenager occurrences; *haar* in 18% and 12%; and *we* in 19% and 6%. Because these are the same words, the difference cannot be caused by words found only in the child read reference transcripts. It can still reflect other corpus factors besides speaker age, such as sentence context or recording differences.

To extend the shared word comparison to all categories, Table 7 reports read speech category error rates after keeping only words used by both groups. Children still have higher category error rates than teenagers in every category and for both systems. This shows that the difference between children and teenagers is not only about words found only in the child read reference transcripts. However, the coverage column explains why shared words do not determine the size of the noun and verb contributions: shared words cover only 20.2% of child noun reference words and 34.3% of child verb reference words, but 92.5% of child pronoun reference words. For nouns and verbs, most child reference words are words found only in the child read reference transcripts; for pronouns, most child reference words are shared with teenagers.

Table 7: Read speech category error rates (%) computed over shared words only. Coverage is the percentage of child reference words in that category that are shared words. Children and Teenagers columns show category error rates for those shared words.

Category	Coverage (%)	Google Chirp 2		Whisper-large-v3	
		Children	Teenagers	Children	Teenagers
Noun	20.2	18.7	6.5	20.2	10.4
Verb	34.3	13.8	5.3	13.8	6.8
Pronoun	92.5	13.5	6.5	13.4	6.4
Article	100.0	10.9	8.1	12.1	7.2
Preposition	94.2	7.7	5.6	9.2	5.9
Adverb	79.2	9.6	5.5	9.8	6.3
Adjective	42.4	9.1	6.2	9.9	6.5
Conjunction	95.4	10.3	5.6	8.7	5.8

3.6 Additional checks

This subsection reports additional checks that support the interpretation of the main results. First, although HMI speech is conversational, the HMI reference transcripts contain few conversational words such as *ja*, *nee*, and *nou*: about 4% of the child reference words and 5% of the teenager reference words. These words also have lower category error rates than HMI speech overall, between 6% and 11% for both groups and both systems. The conversational words

present in the references therefore do not explain the WER difference between children and teenagers in HMI speech.

Second, some errors are caused by the ASR systems outputting a standard written form where the reference contains a spoken variant. Rescoring after mapping such variant families to a single form changes the WER differences only slightly: mapping *éen* to *een* in read speech and *me* and *m'n* to *mijn* in HMI speech each reduce the corresponding WER difference by about 0.4 percentage points for both systems. Variant scoring therefore explains some individual errors, especially in HMI speech, but it does not explain the WER difference between children and teenagers or the POS category pattern.

Third, insertions show a system difference. Table 4 shows that Whisper-large-v3 inserts more words than Google Chirp 2 for both groups, especially in HMI speech. In HMI speech, 13% of the Whisper-large-v3 insertions and 8% of the Google Chirp 2 insertions are filled pauses such as *eh* and *ehm*, which the references do not contain at those positions. In read speech, filled pauses make up less than 1% of the insertions. Part of the higher HMI insertion rate therefore comes from the systems, especially Whisper-large-v3, writing hesitations that the reference transcription omits.

Finally, age is checked as a supporting analysis, not as a separate research question. The check asks whether the group difference in Table 4 is consistent with a gradual relation between age and WER. Exact age is available for 110 speakers overall; each correlation uses 109 speakers with exact age and data in that speech type. WER per speaker is lower for older speakers, especially in read speech: Spearman’s rank correlation between age and WER is -0.74 for Google Chirp 2 and -0.72 for Whisper-large-v3. In HMI speech the corresponding correlations are weaker, -0.37 and -0.38 . This supports treating the corpus labels for children and teenagers as part of a broader pattern related to age, while also showing that the age pattern is weaker for HMI speech.

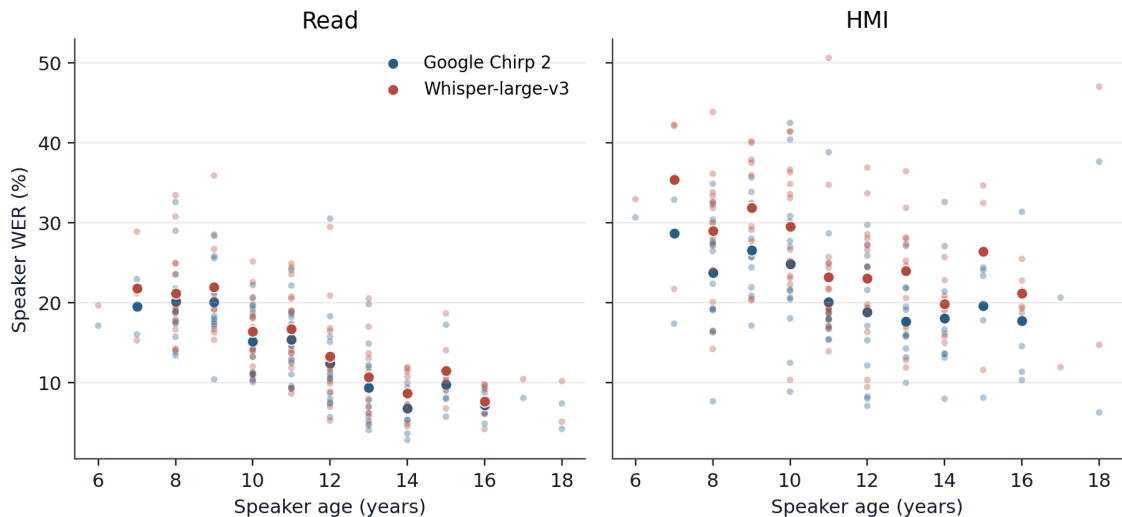


Figure 4: WER per speaker by exact age. Small points are individual speakers. Larger markers show the mean speaker WER for ages with at least three speakers. The age check is descriptive because the corpus labels are not strict age filters and exact age is missing for 25 speakers.

4 Discussion

The main research question asked whether ASR performance disparities between native Dutch children and native Dutch teenagers are concentrated in particular word categories, and how this differs between read and HMI speech. In read speech, a small set of POS categories contributes

most of the substitution/deletion part of the WER difference between children and teenagers: nouns, verbs, and pronouns. This ranking is the same for Google Chirp 2 and Whisper-large-v3. In HMI speech, POS categories are less informative because no single category contributes much of the substitution/deletion difference. The central interpretation is therefore not that POS categories explain the disparity on their own. Instead, POS categories identify which parts of the reference transcripts need closer inspection.

This distinction matters most for the noun result. The read speech noun difference remains after comparing utterances with the same counts of word categories, so it is not only a consequence of comparing different utterance lengths or different mixes of category count patterns. However, the analysis of the specific words shows that most of the noun contribution comes from names and story words that occur only in the child read reference transcripts. The result should therefore not be read as evidence that ASR systems are generally worse at recognizing children’s nouns. A more accurate interpretation is that, in this corpus subset, many frequently misrecognized child-only read words are nouns. This is consistent with ASR work on individual words showing that word classes can summarize error locations while individual word properties still explain much of the recognition difficulty [7].

The child-only read words are not the whole explanation. When the analysis keeps only words used by both groups, children still have higher category error rates than teenagers in every category. WER per speaker also decreases with age, especially in read speech. These two findings are consistent with a component related to speaker group, in line with acoustic and developmental changes that continue throughout childhood [11, 15]. They do not isolate age from sentence context, recording conditions, or other corpus factors. The overall picture is therefore a combination of two patterns: children are recognized worse than teenagers even on shared words, and words found only in the child read reference transcripts add large noun and verb contributions on top of that shared word difference.

The HMI results connect to previous findings on conversational speech, but mainly through scoring behavior rather than through conversational reference words. Errors on reference words such as *ja* and *nee* are rare, so those words do not explain the HMI WER difference between children and teenagers. Instead, the conversational character appears in insertions and spelling variants: Whisper-large-v3 often writes filled pauses that the references omit, and some HMI errors involve colloquial forms such as reference *me* aligned with ASR output *mijn*. This fits prior work on conversational ASR showing that interactional material and transcription conventions affect how conversational speech is scored [17, 12, 13]. In this study, however, rescored spelling variants changes the WER differences only slightly, so these effects explain individual errors rather than the main group difference.

The two ASR systems show the same read speech category ranking, even though Whisper-large-v3 has higher WERs and inserts more words, especially in HMI speech. The read speech category pattern is therefore not a quirk of one system. The higher insertion rate of Whisper-large-v3 partly reflects the filled pauses discussed above. The remaining insertions may reflect system specific decoding behavior, but this paper does not test that mechanism.

The main limitation is that children and teenagers did not read the same texts in the JASMIN-CGN corpus. The analysis therefore cannot fully isolate speaker age from the words assigned to each group in the reading task. The shared word analyses reduce this problem but do not remove it, because even shared words can occur in different sentence contexts and at different frequencies for the two groups. This means that the paper does not determine how much of the read speech gap is caused by age or acoustic properties alone. Three smaller limitations follow from the method. Automatic POS tags can move individual words into the wrong category, especially rare words and lowercased names. Missing exact ages mainly threaten the age trend check, not the corpus label comparison. Tie breaks when matching reference words to ASR output words can affect operation labels and word category attribution, but not the headline WER totals.

These limitations point to future work. Recovering which reading text or story each read utterance comes from would allow separating reading text effects from group effects directly, for example by comparing speakers of different ages on the same story. Completing the manual POS audit would strengthen the category assignments that the analysis rests on. Finally, rescored with references that include filled pauses, or with scoring that tolerates hesitations, would show how much of the HMI insertion behavior is a transcription convention effect rather than a recognition error.

5 Responsible Research

This project compares ASR performance between speaker groups, including children, so its findings bear on fairness towards those groups. The results are therefore framed as system performance differences on a specific evaluation corpus, not as deficits of the speakers. In particular, the higher WER for children is interpreted as the joint outcome of model behavior, words in the read material, transcript conventions, and speech development; the errors are not attributed to the children themselves. The analysis of the specific words supports this framing by showing that a large part of the read speech difference characterizes the reading material rather than the speakers.

The JASMIN-CGN corpus is distributed for research purposes, and the ASR outputs from Zhang et al. [19] were used with permission of the authors; neither is redistributed with this paper. The analysis code is shared with the supervisors via GitLab, as required by the course.

All reported tables and figures were produced from the reference transcripts, ASR outputs, POS annotations, and the table matching reference words to ASR output words with command line scripts. The code records the filtering, category mapping, matching, and aggregation steps, so the reported results do not depend on manual table editing. The base ASR table was rebuilt from ESPnet reference and ASR output files and checked against the original ESPnet aggregate word and error totals. The main reproducibility risks are automatic POS tagging, tie breaks when matching reference words to ASR output words, and incomplete exact age metadata; these risks are stated as limitations rather than hidden from the interpretation.

5.1 Use of generative AI

Generative AI assistance, primarily OpenAI ChatGPT and Codex, was used for research and engineering support: drafting and debugging analysis code, summarizing intermediate results, suggesting paper structure, and reviewing draft text for clarity and overclaims. All AI output was verified against the analysis results and the cited literature before inclusion.

6 Conclusion

This study asked whether the higher WER for native Dutch children than for native Dutch teenagers is concentrated in particular word categories, and whether this differs between read and HMI speech for Google Chirp 2 and Whisper-large-v3. The answer is yes for read speech, but much less so for HMI speech. Children had higher WERs than teenagers in every condition, but the difference was roughly twice as large in read speech as in HMI speech. WER per speaker also decreased with age, especially in read speech. In read speech, nouns, verbs, and pronouns contributed most of the substitution/deletion part of the child-teenager difference for both ASR systems. In HMI speech, no single POS category contributed much of the substitution/deletion difference; conversational words in the references were recognized relatively well, while filled-pause insertions and spelling variants affected individual errors more than the overall group difference.

The read speech result is not simply a finding that ASR systems are worse at recognizing children’s nouns or verbs. After comparing utterances with the same counts of word categories, the verb and pronoun differences became much smaller, while the noun difference remained the largest. Checking the specific words changed the interpretation further: the noun and verb contributions came mostly from story words that occurred only in the child read reference transcripts, while the pronoun contribution mostly involved words used by both groups. At the same time, children still had higher category error rates on words used by both groups. The read speech disparity therefore combines a difference on shared words with a large contribution from words found only in the child read reference transcripts.

Word categories are therefore useful for locating where the WER difference appears, but they are not a complete explanation by themselves. The results should not be read as a causal estimate of age alone, because the corpus does not fully separate speaker group from reading text, sentence context, recording conditions, and other corpus factors. Future work should link read utterances to their assigned texts, so that speaker group and reading material can be separated more directly. For HMI speech, future work should also test references that include hesitations, because some filled-pause insertions may reflect transcription conventions rather than recognition errors.

References

- [1] Jeremy M. Anglin. Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development*, 58(10), 1993.
- [2] Catia Cucchiaroni, 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*, 2008.
- [3] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.
- [4] Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84, 2024.
- [5] Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In *International Conference on Speech Technology and Human-Computer Dialogue*, 2023.
- [6] Matteo Gerosa, Diego Giuliani, Shrikanth Narayanan, and Alexandros Potamianos. A review of asr technologies for children’s speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, 2009.
- [7] Sharon Goldwater, Dan Jurafsky, and Christopher D. Manning. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200, 2010.
- [8] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. *spacy: Industrial-strength natural language processing in python*, 2020.
- [9] Evelyn M. Kitagawa. Standardized comparisons in population research. *Demography*, 1(1):296–315, 1964.
- [10] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.
- [11] Sungbok Lee, Alexandros Potamianos, and Shrikanth Narayanan. Acoustics of children’s speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, 105(3):1455–1468, 1999.
- [12] Alianda Lopez, Andreas Liesenfeld, and Mark Dingemanse. Evaluation of automatic speech recognition for conversational speech in dutch, english and german: What goes missing? In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 135–143, Potsdam, Germany, 2022. KONVENS 2022 Organizers.
- [13] 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*, pages 1997–2001, 2021.
- [14] National Institute of Standards and Technology. Nist speech recognition scoring toolkit (sctk), 2021.
- [15] Alexandros Potamianos and Shrikanth Narayanan. Robust recognition of children’s speech. *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616, 2003.

- [16] 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*, pages 28492–28518, 2023.
- [17] Andreas Stolcke and Jasha Droppo. Comparing human and machine errors in conversational speech transcription. In *Proceedings of Interspeech 2017*, pages 137–141, 2017.
- [18] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. Espnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211, 2018.
- [19] Yuanyuan Zhang, Thomas De Valck, and Odette Scharenborg. Speech recognition performance disparities between dutch diverse speaker groups. *Phonetica*, 2026.