Comparing Human Listeners and Dutch ASR on Transcribing Child Speech

The Effect of Familiarity with Child Speech on Transcription Performance

Ilse Huisman[1]

Supervisor: Odette Scharenborg[1]

[1]EEMCS, 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
January 25, 2026

Name of the student: Ilse Huisman
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg

Abstract

Automatic Speech Recognition (ASR) systems are becoming increasingly common in day-to-day life. Yet, child speech remains challenging for ASR systems. This paper gives the first comparison of Dutch human listeners and Dutch ASR systems on Dutch child speech. It tests whether familiarity with child speech improves human transcription performance and sees if the age of the child speaker influences the transcription performance.

A balanced set of 40 utterances were taken from the JASMIN database (speakers aged 7-11), and were transcribed by 20 humans (10 familiar with child speech (parent/caretaker) and 10 unfamiliar). Transcripts were also gathered from two state-of-the-art ASR systems (Google Telephony and a Conformer model). These transcripts were evaluated against reference transcripts using Word Error Rate (WER). Statistical significance was tested.

Results show that overall ASR transcription performance was comparable to human performance, and in some cases slightly, but not significantly, better. Familiar listeners did not outperform unfamiliar listeners. In fact, there was no significant performance difference between the two groups of humans. Within the 7-11 age range, no clear relationship between speaker age and WER was shown, but results were sensitive to sentence difficulty outliers and "speaker effects".

1 Introduction

Automatic Speech Recognition (ASR) systems are becoming more and more popular for everyday use by people all over the world [1, 2]. Yet, the performance of these systems often varies across speaker groups (e.g., children, older adults, accents, etc.). Previous work has already shown that child speech is recognized worse than adult speech by all types of ASR systems [3, 4]. Because ASR is increasingly used in e.g. health care [5] or voice assistants [6], inclusivity is crucial.

For non-Dutch speech, ASR systems' transcription performance has already been compared to human listeners' transcription performance. The first person to ever compare this was Lippmann in 1997. He concluded that "human word error rates are often more than an order of magnitude lower than those of current" [7, p. 12] ASR systems. More recent studies found that often times humans only slightly outperform most ASR systems [8, 9]. Dutch ASR systems have also been compared to Dutch human listeners on Dutch conversational speech. Lopez and colleagues [10, p. 138] found that "ASR transcripts contained fewer words than their human counterparts with a 33% difference for Dutch".

Because Dutch ASR systems are becoming more common in devices used by children (e.g. smart speakers and voice assistants) [11], poor transcription of child speech reduces the usability of these devices. It is important that Dutch ASR systems work reliably for all potential users. To evaluate how large this limitation of Dutch ASR systems in accurately transcribing Dutch child speech currently is, this experiment focuses on a comparison between Dutch human listeners and Dutch ASR systems on Dutch child speech.

If this study finds differences between human and ASR transcription performance on Dutch child speech this will provide insight into the capabilities and limitations of Dutch ASR systems. If it is the case that human listeners transcribe Dutch child speech better than Dutch ASR systems, this will show a clear human-machine gap, which gives developers of ASR systems a concrete target for improvement. On the other hand, if humans do not clearly perform better than ASR systems, this indicates that child speech is challenging, even for experienced listeners, which provides an interesting starting point for further research, for example what makes certain stimuli difficult for both humans and ASR systems.

Children's voices, sentence structure and pronunciation differ substantially from adults' [12]. These differences can make child speech harder to transcribe. Previous work found that parents are better at transcribing their own child's speech than all other listeners [13]. This suggests that familiarity with speech improves performance. It remains unclear whether this familiarity advantage translates to child speech from unfamiliar children. To test this, the human listeners in this experiment were split into two groups: human listeners with familiarity with child speech (e.g. parents, caregivers) and naive (without familiarity with child speech) human listeners.

While children mature, their speech becomes more similar to adult speech. ASR systems transcribe child speech worse than adult speech [3, 4]. In addition to this, Feng and colleagues [14] have shown that ASR systems transcribe teenage speech better than speech of younger children. This supports the idea that transcription performance by ASR systems is better for older children in comparison to younger children. Therefore, this research also focuses on how the age of the child speaker influences transcription performance. This could provide developers of ASR systems with clear insight into which age range of child speakers to train their ASR system the most on.

As explained above, this research focuses on comparing human and ASR system performance on Dutch child speech. Specifically, it asks: **How well do Dutch human listeners transcribe Dutch child speech in comparison to Dutch ASR systems?** A second question examines whether familiarity with child speech matters: **To what extent does familiarity with child speech influence human transcription performance?** This study also investigates a related factor that may influence transcription performance: **How does the age of the child speakers affect the performance of Dutch ASR systems and human listeners?**

The remainder of this research paper is structured as follows. Section 2 describes the methodology. Section 3 discusses responsible research. Section 4 presents and discusses the results. Section 5 outlines the limitations and gives recommendations for further research. Section 6 concludes the paper by summarizing the main findings and implications.

2 Methodology

An experiment will be conducted with two groups of human listeners (familiar and unfamiliar). Each listener will be asked to listen to 40 different sentences of child speech during an online survey, and will then after each sentence be asked to transcribe it. The results from the two groups will be compared. The same 40 sentences will also be transcribed by two state-of-the-art Dutch ASR systems, the performance of which will be compared to that of the human listeners.

2.1 Selecting Speech Samples

The first step in setting up the experiment, was selecting the speech samples that will be used. The goal was to obtain a balanced set of stimuli of child speech from the JASMIN database [15]. The JASMIN database contains a collection of around 115 hours of Dutch speech of children, teenagers, elderly and non-natives, living in Flanders and The Netherlands. For all speech, transcriptions are present. These were made by trained transcribers, were then checked by a second transcriber after which a spelling check was conducted.

For this experiment, 40 stimuli were selected. Forty stimuli were chosen to keep the experiment duration reasonable, while still providing enough transcriptions per participant to ensure reliable results. Using too many stimuli with too few participants would increase noise because of individual listener differences and therefore the results would be less reliable.

To ensure the stimuli set was balanced for gender, 20 stimuli were obtained from speakers that are labeled as women in the JASMIN database and 20 stimuli were obtained from speakers that are labeled as men. Only child speakers were considered. For this experiment, people between the ages of seven and 11 were considered children. Younger children (below seven) were not part of the JASMIN database and children above the age of 11 were considered teenagers.

The JASMIN database consists of two types of speech, human-machine interaction (HMI) and read speech. HMI speech is where a human has a conversation with a machine, which makes it a type of conversational speech. Read speech is where a human reads out a script. The research questions should be investigated on speech that best matches how children speak when interacting with an ASR system. In order to do so, only HMI speech is considered.

We want to create a test set balanced on sentence length, because short sentences have less context but fewer opportunities for mistakes, whereas long sentences give more context but add complexity and more opportunity for errors. Both are different and both should be tested. The HMI stimuli set of the JASMIN corpus is skewed towards short stimuli. Therefore, purely random sampling would lead to a test set dominated by short stimuli. To avoid this, the sampling was done by a stratified random sampling procedure.

An equal number of samples from all ages had to be guaranteed, to avoid any age group being overrepresented in the test set. Children's speech evolves as they grow older and this may impact the transcription performance. Because of stratified random sampling, it could be ensured that equal amounts of samples from all age groups were present in the test set.

The procedure consists of five steps, described below.

Step 1: Data preparation

Speaker metadata (gender, age and dialect region) were combined with stimulus metadata (audio filenames and transcriptions), by joining them on speaker ID. This resulted in a table containing each audio file and transcript together with its corresponding speaker information.

From this table, only records were considered that meet the following requirements:

- The word count of the transcription must be higher than four, because most shorter transcripts were not valid sentences or consisted mainly of filler words or sounds.
- Speakers between the ages of seven and 11.

In order to be able to do random stratified sampling, a column that contains a random decimal between zero and one was added to each row of the dataset. Later on, we sort ascending on these decimals to ensure a random order of the rows (stimuli) in the dataset.

Step 2: Sampling procedure

The dataset was divided into categories based on age of the speaker and word count.

Figure 1 visually displays the distribution of the 40 speech samples. First, all speech records were split by gender: 20 samples by men and 20 by women. Then, for each gender, they were split on age: seven, eight, nine, 10 and 11 years old. Each age group gets four samples.

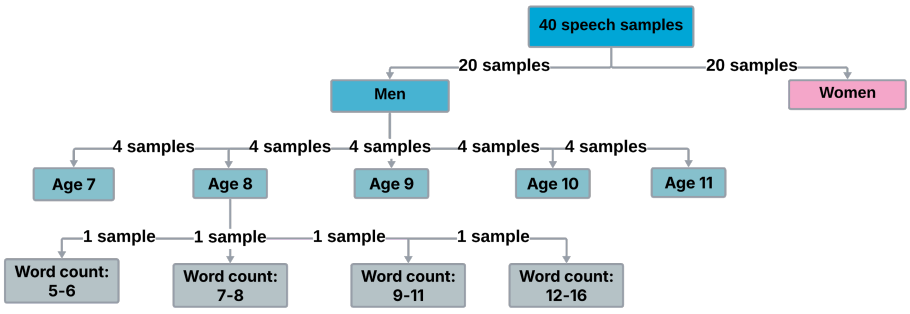

Figure 1: Distribution of speech files

Next, within these age bins all samples were further divided into four bins based on word count: 5-6 words, 7-8 words, 9-11 words and 12-16 words. From each age category one stimulus was selected for each word count category. This resulted in one stimuli per age-word count bin per gender, and 20 stimuli per gender in total.

After explaining the broad idea of the stratification scheme, now the actual sampling algorithm that was applied is described in Figure 2. The algorithm was applied independently for each gender.

1. **Filter by age bin (one bin at a time).**
   For each age (7-11) the dataset was restricted to records belonging to speakers of that age. They are processed from youngest to eldest.
2. **Filter by word count bin (one bin at a time).**
   For the current age group, records were filtered to only contain records whose transcription length falls within the current word-count bin.
3. **Sort by random number.**
   The selected records are sorted in ascending order on the random decimal column.
4. **Select the first item.**
   The top item in the sorted list was selected.

Figure 2: Random sampling algorithm

This process was repeated until one stimulus was selected from each age-word count combination for both genders.

Step 3: Redistribution rules

During the random sampling algorithm a couple redistribution rules are applied. Because younger speakers occur less frequently in the dataset, their stimuli were selected first. The same is the case for the higher word count bins. Nevertheless, it was still possible that a word count bin did not contain enough records. In that case, an extra stimulus was selected from the nearest lower word count bin for the same age bin. In case an age bin (not yet filtered on word count) contained less than the four required stimuli (one for each word count bin), an extra stimulus was collected from the nearest higher age bin.

No one child speaker should be over represented in the test set, because the test set should reflect the overall characteristics of child speech. Therefore, no more than two stimuli per speaker should be in the test set. If a stimulus is picked and the child speaker already has two stimuli in the test set, the first available stimulus from a different speaker from the same age-word count bin is picked. If there is no such stimulus available, the original stimulus is picked, since preserving the age distribution was considered more important than balancing the speaker distribution.

Step 4: Manual quality check

After the random selection of a record, a manual quality check was done. The purpose of this step was to ensure that the stimuli were (1) clearly audible and not obscured by background noise and (2) only contained existing Dutch words or filler words (e.g. uh, uhm, mm, etc.).

If a stimulus failed one of these criteria, it was replaced by the next item in its bin.

Step 5: Normalizing volume

The audio samples in the JASMIN database all have different volume levels. This is not desirable for a listening experiment, because abrupt changes in volume can startle participants when a louder clip follows a quieter one. Similarly, a quieter clip can become inaudible when it follows a loud one.

Thus, all audio clips were normalized to have the same volume. This was done using the loudnorm filter of FFmpeg [16]. The loudnorm filter ensures consistent perceived loudness. The loudnorm filter (EBU R128) in single-pass mode was used, using integrated loudness of -24 LUFS, loudness range of 7 LU, and maximum true peak of -2 dBTP.

2.2 Human Listener Experiment

A human listener experiment was conducted with 20 participants. Participants were recruited from the social circle of the researcher. Participants fell into two groups: familiar with child speech and unfamiliar with child speech. Participants familiar with child speech should fall into one of the following two categories: parents or caretakers of a Dutch speaking child. The child had to be a minimum of four years old and no maximum was put on the age of the child. Anyone who has spoken with children aged 2-11 on an (almost) daily basis for several years, like teachers, were not deliberately excluded from the experiment, but weren't part of the participant group for this particular experiment. For each group of participants, 10 people were recruited. An equal balance between men and women was maintained in both groups. A diverse age distribution was aimed for. Figure 3 shows the distribution of the participants for this experiment. It shows the number of participants per gender for the familiar and unfamiliar listener groups, as well as the number of participants in four age groups. Under the participant counts, the participants' exact ages are shown in light gray.

Age group	Familiar		Unfamiliar		**Totals**	
	Women	Men	Women	Men	**Familiar**	**Unfamilair**
20-29	-	1 (27)	1 (26)	-	**1**	**1**
30-39	3 (32, 34, 39)	1 (35)	3 (30, 30, 30)	2 (32, 37)	**4**	**5**
40-49	1 (47)	2 (41, 41)	-	2 (45, 48)	**3**	**2**
50-59	1 (51)	1 (58)	1 (54)	1 (55)	**2**	**2**
Total	**5**	**5**	**5**	**5**	**10**	**10**

Figure 3: Distribution of the participants

Participants were not paid and were asked to sign an informed consent form (Appendix A) before starting the experiment. All experiments were held in a quiet room and all participants used the same laptop (Dell Inspiron 14 5425) and headphones (Sennheiser HD 200 Pro). The experiment was

hosted in Qualtrics and one audio clip was shown per page. The order of the audio clips was randomized per participant to prevent order effects caused by practice or fatigue. So, to avoid participants performing better or worse simply because items were presented earlier or later in the experiment. Participants first read the instructions. Then they listened to 40 audio clips in random order. They were allowed to listen to each clip only once, because hearing it more than once can introduce learning effects [17]. An experiment was done in order to decide how many times a participant can listen to each audio sample, see more about this experiment in Appendix B. Participants were asked to type their transcriptions while or after listening to the audio. After every 10 questions, an instruction screen appeared asking participants to take a break. They could pause as long or short as they wanted and were instructed to navigate to the next page in order to continue the experiment. Correct transcriptions were not shown during the experiment, because this could introduce a learning curve [18], which would influence the results. Almost all participants wanted to compare their answers to the correct transcriptions once the whole experiment was conducted.

The setup of the Qualtrics survey can be seen in Appendix C. An example of a survey page with an audio file can be seen in Figure 4

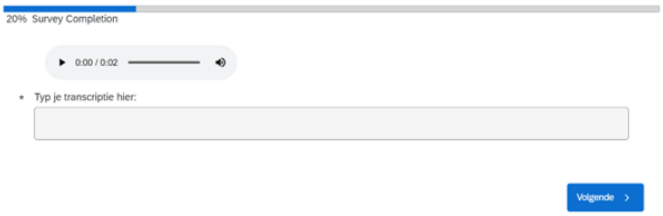

Figure 4: Survey page with audio file

From each participant, some personal information was gathered, that was deemed informative for later interpretation of the results. Each participant was asked the following things:

- Age: To ensure an as balanced as possible age distribution for the participants.
- Gender: To ensure an evenly balanced gender distribution for the participants.
- Familiarity with child speech. In particular, are they a parent of caretaker of a Dutch speaking child?: Used to assign familiarity groups described earlier.
- Dutch as native language?: Only native Dutch speakers were considered for this experiment, so differences in transcription performance could not be attributed to differences in Dutch proficiency.
- In what region of the Netherlands did they grow up?: Being born in a specific region of the Netherlands can correlate with exposure to specific regional accents, which may influence how easily a listener transcribes speech from a given region.
- In what region of the Netherlands did they live the longest?: Living somewhere for a long time also can correlate with exposure to specific regional accents, which may influence how easily they transcribe child speakers from specific regions.
- Do they have hearing problems?: To exclude participants whose severe hearing problems could affect transcription performance.

2.3 ASR System Experiment

For this experiment, the human transcriptions are compared to two state-of-the-art ASR systems. Namely, Google Telephony [19] and a Conformer [20] trained with XLSR-53 features [21].

Google Telephony was chosen for this experiment since it is optimized for telephone speech [19]. Telephone speech is a type of conversational speech. This project focuses on human machine interaction speech, which is also a type of conversational speech. Furthermore, the performance of Google Telephony is good compared to other systems (Zhang, et al., unpublished paper, email to the author, Dec. 13, 2025). Unfortunately, the training and validation data for Google Telephony are not transparent, so nothing can be said about this.

Zhang and colleagues [20] trained three Conformer-based models from scratch on data from the Corpus Gesproken Nederlands (CGN) [22] using different features. CGN consists of approximately 900 hours of speech. These 900 hours were split up into shorter speech clips, invalid/inaudible clips were removed. What remained was a training set of 690.5 hours and a validation set of 6.9 hours [20].

Zhang and colleagues [20, p. 9] found that "all three models achieved state-of-the-art performance on the CGN test sets compared to results reported in [14], [23], [24]." From these three, the Conformer trained with XLSR-53 features [21] was chosen, since it has the lowest WER on data from the JASMIN database, which is the same database from which the child speech stimuli were taken, making it the best ASR baseline to test the human-ASR transcription performance gap.

The Conformer trained with XLSR-53 features [21] was trained using "1024-dimensional fused representations from all 24 layers of XLSR-53" [20, p. 8].

Transcriptions by these ASR systems for all audio samples in the JASMIN database were provided to us. From these, the 40 transcriptions of the audio samples used in the human experiment were gathered.

2.4 Post processing

The goal of the post processing step was to correct any errors that can be made during typing and to match the style of the human transcriptions to that of the ASR transcriptions to ensure a fair comparison with each other.

ASR systems do not transcribe using letter cases or punctuation. Therefore, all transcriptions were converted to all lowercase and any punctuation was removed. Any additional (trailing) white spaces were also deleted.

Any newly observed filler words or other non-lexical sounds (sounds that are not actual words but have meaning

in conversation, such as uh, uhm, mm) in the human transcripts were added to the non-linguistic symbols list. This list is used to remove these words from the transcripts.

In case laughter was transcribed despite the instructions saying not to do so, these parts of the transcriptions were removed.

Obvious typing errors were corrected when the intended word was fully unambiguous (e.g. v an → van, hius → huis). When the intended word could not be determined with certainty, the transcription was left unchanged. Spelling mistakes were only corrected when they did not change the word(s) or the meaning of the word(s). For example: "afentoe" → "af en toe" or "er achteraan ga" → "erachter aan ga". Numbers and letter-number combinations in transcriptions should be compared in their spoken form (e.g. mpdrie). So any obvious variants of the same word were converted to one chosen format.

2.5 Evaluation

Transcription accuracy was evaluated using WER, the standard and most well known metric in ASR evaluation. WER compares a transcription with a "ground truth", which in this case is the transcription in the JASMIN database created by trained transcribers. WER is the proportion of substitutions, deletions and insertions required to transform the human/ASR transcription to the ground truth relative to the total number of words in the ground truth * 100%.

An insertion is counted when a word is added to the transcription in comparison to the ground truth. A deletion is counted when a word is deleted from the human/ASR transcription to match the ground truth. And a substitution is counted when a word in the human/ASR transcription is replaced by a different word to match the ground truth.

WER was calculated using the sclite scoring tool [25]. The sclite scoring tool compares a hypothesis (human/ASR transcription) with the reference (ground truth). The hypothesis and the reference are aligned, so all matching words align with each other. Then, for each sentence pair, the number of correct, substituted, inserted and deleted words are counted.

Now different WER are calculated, but the one useful for our calculations is the Sum/Avg WER. This WER is calculated by adding up all substitutions, deletions and insertions from all 40 sentences and dividing this total by the total word count of all 40 sentences.

In the remainder of this paper, WER is used to: (1) compare WER results of humans vs. ASR, (2) compare familiar with unfamiliar humans, and (3) interpret patterns using error-type breakdowns and analyses by speaker age and sentence length.

3 Responsible Research

Ethical approval for this study was granted by the Human Research Ethics Committee (HREC) with number 6233.

Participants for this experiment were recruited from the social circle of the researcher. They took part in a listening and transcribing experiment, before which they signed an informed consent form, which can be found in Appendix A. Participation was voluntary and participants could withdraw from the experiment at any time without having to provide reason.

Because participants were recruited from the author's social circle, the sample may not be representative of the population. This potential bias should be considered when interpreting the results.

From the participants, transcriptions were gathered as well as some non-identifiable personal information. Personal identifiers, such as name, were not gathered. Participants received a personal ID. During the study, data was stored on Qualtrics, following the guidelines the TU Delft has in place for this. After this project has succeeded all data will be removed from Qualtrics and will be moved to a secure storage provided by the TU Delft.

The speech samples used for this experiment were obtained from the JASMIN database. No children were recruited or recorded specifically for this project. The study only used previously collected speech data, that was made available for research use. The JASMIN database was created ethically, and all participants (or their guardians) have signed an informed consent form [15].

During this project, AI tools such as ChatGPT, were used for: (1) Writing support: improving structure, spelling checks, formality of writing. (2) Programming support: drafting debugging Python scripts for handling data. (3) Literature support: brainstorming search terms and finding related papers. (4) Visualization support: generating ideas how to present results. (5) Formatting help: use of LaTeX, structuring section, tables, figures and references. Examples of prompts used for the AI can be found in Appendix D. No participant data was ever uploaded to any AI system. All outputs of the AI systems were treated as suggestions. The author reviewed and edited all outputs before usage.

Results reflect the dataset and setup used for this experiment and should not be generalized to other ASR models or human listeners.

4 Results and Discussion

4.1 Humans vs. ASR

The WER for the humans and both ASR systems found can be seen in Table 1.

Table 1: WER humans vs. ASR

Group/System	WER (%)
All humans	17.6
Google Telephony	12.8
Conformer XLSR-53	18.5

From the table it can be seen that the Conformer model performs close, but slightly worse than humans. But no statistical difference can be found between the Conformer model and the humans after doing a two-tailed t-test with α=0.05 over the sentence WER, t(39)=0.566, p=0.574. The mean and standard deviation of all the humans, Google Telephony and the Conformer model can be seen in Table 2.

Table 2: Mean and standard deviation for each system/group ($n = 40$ sentences).

Group/System	Mean of sentence WER (%)	SD
Humans	16.974	16.504
Google Telephony	13.811	16.092
Conformer XLSR-53	18.554	20.016

This test compares the sentence WER. For the humans it uses the mean WER per sentence across all human listeners. For the ASR systems it uses the sentence WER as is.

Google Telephony on the other hand performed better than both the Conformer model and the humans. From a two-tailed t-test with α=0.05 it can be concluded that Google Telephony does not significantly outperform humans, t(39)=1.141, p=0.261, or the Conformer ASR model, t(39)=1.681, p=0.101.

These outcomes suggest that ASR systems exist that can slightly outperform human listeners in a transcription task such as in this experiment. But these systems do not yet significantly outperform humans.

Error type breakdown

It is interesting to look at what errors (substitutions, deletions, insertions) were made most by each group of humans and by each ASR system. This helps to understand in what way human transcriptions differ from ASR transcriptions and also gives developers of ASR systems a clear direction for improving their systems.

Figure 5 shows for each error type the (for humans average) number occurrences in all forty sentences for each human group/ASR system.

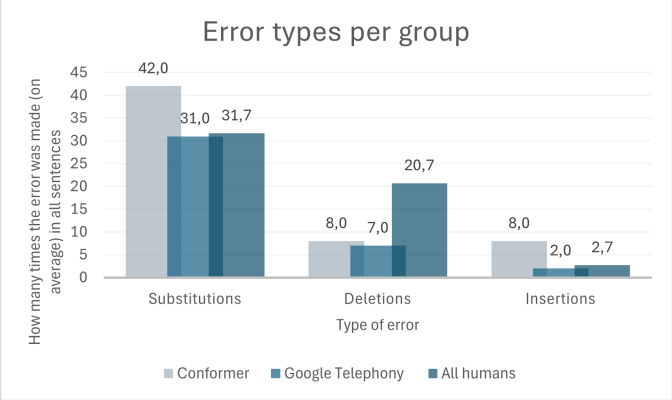

Figure 5: How many times each error type was made on average across all sentences for each group/ASR system.

Some observations can be drawn from this figure. It is very clear that humans have a lot more deletions in their transcriptions compared to the ASR models. This may be because people are more inclined to not write anything down when they don't fully understand or can't clearly hear part of the stimulus. On the other hand, most ASR systems always write down their best-guess transcription, rather than indicating when they are not sure [26].

It is noticeable that the Conformer ASR system had a lot more insertions and substitutions than both the humans and the Google Telephony ASR system. So this gives developers of the Conformer ASR model a clear objective to improve their model.

4.2 Familiar vs. Unfamiliar humans

The WER for the familiar and unfamiliar humans can be found in Table 3 as well as the mean and standard deviation of the statistical test over the sentence WER.

Table 3: WER, mean of sentence WER & SD of familiar vs. unfamiliar

Group/System	WER (%)	Mean of sentence WER (%)	SD
Familiar humans	18.0	17.354	17.472
Unfamiliar humans	17.2	16.593	16.251

From Table 3, it can be observed that for this experiment unfamiliar humans slightly outperformed familiar humans. However, a two-sided t-test with $\alpha = 0.05$ on the mean sentence WER found no significant difference between the groups, t(39)=0.686, p=0.496. The mean and standard deviation can be found in Table 3.

Discussion

Familiar listeners did not outperform unfamiliar listeners. Instead, unfamiliar listeners achieved slightly lower, but not statistically significantly lower, WER than familiar listeners. This is not what was expected, namely familiarity with child speech improves transcription performance.

One possible explanation for there being no performance difference between familiar and unfamiliar humans, is that in short stimuli (1-2 seconds) there is a limited amount of context (sentence structure or conversational clues). This means transcription performance is mainly based on acoustic clarity. It is possible that a familiarity advantage would show when listeners can rely more on context.

It is also possible, that familiarity with child speech by having children, may not reflect familiarity with unknown children's speech. Understanding your own child's speech very well, might not directly imply very well understanding of all child speech.

Another interesting difference that was observed between familiar and unfamiliar listeners, was the difference in average completion time. Familiar listeners on average completed the experiment in 17 minutes and 40 seconds. On the other hand, unfamiliar listeners on average completed in 24 minutes. This could be explained by familiar listeners being more confident in their performance. Unfamiliar listeners may have compensated for their unfamiliarity by working slower and more careful. While familiar listeners may have responded faster, but at the cost of transcription accuracy.

Lastly, the JASMIN database consists of speech from children who are relatively old (7-11 years). At these ages, children's speech is closer to adult speech in comparison

to speech of children aged seven and below. This may reduce any advantage that caregivers have when transcribing younger children's (aged 2-7) speech. It is plausible that a familiarity benefit would be seen if the stimuli contained younger children's speech.

4.3 Effects on WER by age of the speaker

To find how the age of the child speaker affects the transcription performance of ASR systems and human listeners we have to look at the average WER per age group for the humans and for the average of the two ASR systems.

In Figure 6, for each speaker age, the average WER is shown for all humans versus the average of the two ASR systems. On the secondary axis the amount of unique speakers per age are shown.

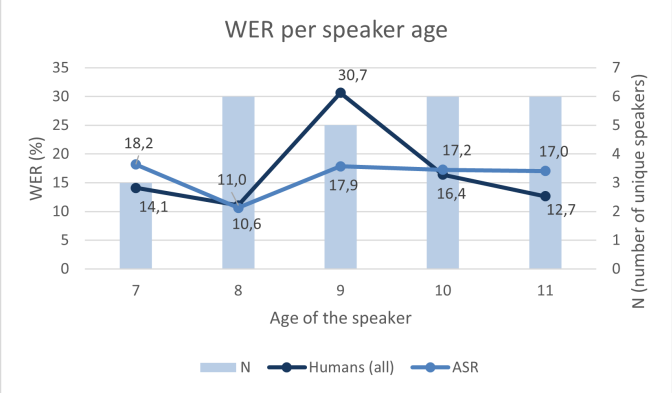

Figure 6: Average WER per speaker age group with number of unique speaker per age group.

From this graph, no trend can be seen between WER and age of the speaker. This suggests that speaker age alone is not a reliable indicator of stimulus difficulty.

For humans the average WER of the age group nine is surprisingly high. This could reflect that child speech becomes harder or easier to transcribe with speaker age, but in that case we would expect a gradual trend across speaker age. Rather, in the figure only age nine stands out, no age related trend can be seen otherwise.

Differences in stimuli difficulty (difficult words, longer sentences, difficult speakers) can lead to differences in average WER per speaker age that are not related to age. Especially here, where each bin only contains eight stimuli, so each stimulus has a relatively high influence on the average WER. To analyze these differences in stimuli difficulty, it is helpful to take a look at the stimuli on which humans performed the worst and see whether these sentences share any characteristics.

In Table 4 the five sentences with the worst average WER for humans are shown.

Table 4: Worst transcribed sentences by humans.

Worst sentence	Avg WER (%)	Speaker ID	Speaker age	Word count	Ground truth
1	81.1	N000027	9	9	binnenkort maar ik ben pas om achttien oktober jarig
2	55.5	N000027	9	10	hij doet een beetje hij doet het niet echt goed
3	42.8	N000057	9	9	kluiven en dat vind ik wel heel erg leuk
4	36.4	N000213	7	7	me vader me broer en me moeder
5	35.0	N000213	7	5	omdat 't daar koel is

Three out of the five worst transcribed sentences are by nine year olds. This clearly explains the unexpectedly high average WER for sentences from speakers aged nine.

We notice that the two worst performed on sentences originate from the same speaker aged nine. Number 4 and 5 are also by the same speaker aged seven. This clearly points to speaker effects. With so few unique speakers per age, speaker-specific characteristics can dominate the age groups average WER, which would explain the high average WER in the age nine group.

Overall, these results suggest that the surprisingly high average WER for the age nine group is largely driven by speaker-specific difficulty rather than an age trend.

Another interesting thing that stands out in the table is that the three worst performed on sentences have a word count of nine or higher. This makes it interesting to also analyze the effects of sentence length on the WER.

4.4 Effects on WER by sentence length

Figure 7 shows the average WER per sentence length bin for all humans and the average of the two ASR models. For each sentence length it shows the amount of sentences in this bin on the secondary axis.

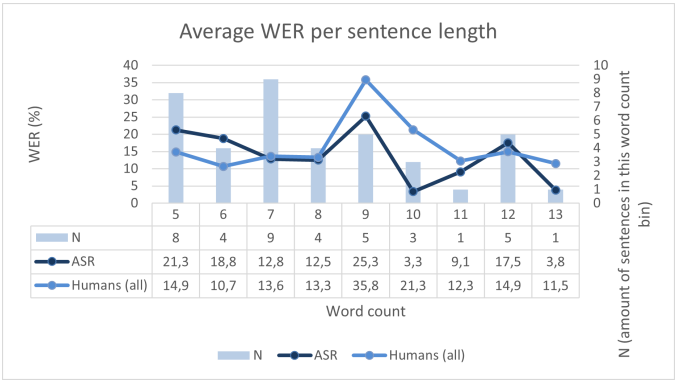

Figure 7: Average WER per sentence length

From this graph, no clear trend can be seen. This suggest that sentence length alone is not a strong indicator of the difficulty of a stimulus. Short sentences can still be hard, while some longer sentences might be easier to transcribe. So sentence length alone is unlikely to explain the WER results found.

The unexpectedly high WER for humans for the sentences with a word count of nine can again be explained by two of the five worst transcribed sentences by humans, which have a word count of nine. These two sentences have a WER of 81.1% and 42.8%. With an average WER of 17.6% and only five sentences in the word count nine bin, these two sentences disproportionately influence the mean WER for this bin, resulting in an outlier.

5 Limitations and Future Research

First, the size of the stimuli set (40) makes the results very sensitive to outliers. A small set of particularly difficult stimuli can disproportionately influence the found WER. However, increasing the number of stimuli was not feasible for this experiment, because more stimuli would mean needing more participants in order to maintain sufficient participants per stimuli, which was not feasible in the 10 weeks allocated for this project. Hence, future research should increase the number of stimuli and ideally make sure each speaker occurs in the stimuli set once, so "speaker effects" can be minimized as much as possible. A larger stimuli set reduces the sensitivity to outliers.

Second, the speaker age related analysis is hard to interpret because of specific speaker effects. Some age bins contain only few unique speakers, which means "age effects" are likely partly originating from "speaker effects".

Third, the child speakers in this experiment were relatively old (7-11 years), because no speech from younger children was available in the JASMIN database. Speech at these ages is often closer to adult-like speech than that of younger children. This may have reduced the benefit of being a familiar human listeners. Because of this, future research could extend the age range to include younger children (below seven years). This will show whether the results shown in this experiment hold for a wider range of child speech or whether familiarity effects become noticeable for younger child speech.

Finally, familiarity with child speech was reduced to a binary grouping (familiar or unfamiliar), even though actual exposure to child speech likely varies a lot within both groups. For example, caretakers may differ in number of children, age of children but also in their amount of daily exposure to child speech. On the other hand, some participants in the unfamiliar group may communicate with children often even though they are not a parent or caretaker. To account for this, future research should refine the distinction between familiar and unfamiliar listeners. Familiarity could be measured using more levels, by gathering additional information from each participant. Such as, hours per week spent interacting with children, ages of the children they interact with. This research did not exclude parents or caretakers with children above the age of 11. It might be interesting to put parents or caretakers with only older or even adult children in a separate familiarity group. Measuring familiarity using more levels was not feasible during this experiment, because with only 20 participants, introducing multiple familiarity levels would have created very small subgroups, which makes for unreliable group comparisons because of potential outliers.

6 Conclusions

This paper compared Dutch human listeners and Dutch ASR systems on Dutch child HMI speech, and tested whether familiarity with child speech affects human transcription performance. Two ASR systems were tested: Google Telephony and a Conformer model.

Conclusion. This experiment shows that, for Dutch child speech from children between the ages of 7–11, current Dutch ASR systems perform comparably to Dutch human listeners on a transcription task like in this experiment. This suggests that (for this age range) child speech is not necessarily a hard task for Dutch ASR in the way it is often assumed, and that improving inclusivity may require more than just training more on child speech data. Familiarity with child speech did not improve transcription performance, and no clear effect of speaker age was observed, indicating that differences in performance are more likely driven by speaker characteristics and stimulus-specific factors.

A Consent form

ID participant:

Consent Form for Participation in the Comparing Human Listeners with Dutch ASR System Experiment

Thank you for your participation in our research project "Comparing Human Listeners and Dutch ASR: The Effect of Familiarity with Child Speech on Transcription Performance". This study is carried out by Student Ilse Huisman and Prof. Dr. Odette Scharenborg from the Delft Inclusive Speech Communication (DISC) Lab at the Technische Universiteit Delft (TU Delft).

The purpose of this study is to explore how well humans perform at recognizing Dutch child speech compared to an artificial intelligence (AI)-based automatic speech recognition (ASR) model. To this end, we are collecting the transcriptions and some general information from you, the participant.

In this experiment, you will listen to 40 spoken sentences. You will hear a sentence after which you are asked to type in what the speaker said. The experiment is expected to take approximately 30-40 minutes to complete. The typed responses will be stored at the Gitlab repository at TU Delft, and will be used for research purposes only. No identifying information of you will be stored.

You will be invited to fill in a questionnaire before the recording. The questionnaire asks about your gender, age, native language, region you live/grew up in, familiarity level with child speech and information about hearing problems. You can choose not to answer a question.

Please tick the appropriate boxes	**Yes**	**No**
Take part in the study		
1. I have read and understood the study information dated ____ / ____ / _____, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	☐	☐
2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions, and I can withdraw from the study at any time, without having to give a reason.	☐	☐
3. I understand that taking part in the study involves me listening to speech and typing in text. My responses will be compared with the correct answers ("ground truth") to assess accuracy. Additionally, my performance may be compared with that of an ASR model.	☐	☐
Potential risk of participating (including data protection)		
4. I understand that personal information that can identify me, such as my name, will only be used for the purpose of signing this informed consent form and will not be shared beyond the research team nor be part of the research.	☐	☐

ID participant:

5. I agree that the data collected as part of this study will be recorded by the researcher and stored electronically. All raw data (i.e., the typed responses and answers in the questionnaire) is stored, processed, and distributed using an anonymous subject number.	☐	☐
Use of the data in this study		
6. I understand I can request my data be removed from the dataset at any time. If this is done within 2 months of the completion of my participation in the experiment, my data will not be included in analysis. After these 2 months, my data can still be removed from future use of the dataset, however it is no longer possible to guarantee all older versions of the dataset are removed from circulation.	☐	☐
7. I understand and agree that all the data I provide will be used for academic publications and scientific reports produced by the research team, DISC Lab at TU Delft.	☐	☐
8. I understand and agree that the data collected from my participation in this study will be retained for future reuse for research by the research team and other interested researchers. The anonymized demographic data and created transcriptions will be archived and shared through the 4TU.ResearchData repository.	☐	☐

Signature

_____ _____

Signature Date

Study contact details for further information:

Ilse Huisman <I.N.Huisman@student.tudelft.nl>

Odette Scharenborg <O.E.Scharenborg@tudelft.nl>

B Experiment to decide how many times participants could hear audio samples

During the setup of the experiment, one of the big choices to make was how many times participants could listen to each audio sample. The hypothesis was made that hearing audio samples more than once, could drastically change transcription performance in humans. An experiment was conducted to find out whether this hypothesis is correct.

First, a participant was asked to do the listening experiment when listening a maximum of three times to each sample. This participant had a WER of 18.2%. Now, another participant was asked to do the listening experiment when only listening once. Right after, the participant did the same experiment again, but now listening the maximum of three times. When listening once, their WER was 20.4%. When listening three times, their WER was 13.7%. This is a difference of $\sim$6% which is considered a big difference.

The motivation behind this experiment was to find out how well ASR systems can transcribe children's speech in conversation compared to human listeners. During conversation it is not possible to listen to what someone says twice.

This together with the big difference in transcription performance when listening multiple times, is what ultimately led to the decision to let participants listen to each audio sample only once.

C Qualtrics survey

C.1 Survey instructions

Welkom en bedankt voor uw deelname!

Voordat het experiment begint, wordt u vriendelijk verzocht **eerst de toestemmingsverklaring (consent form)** te lezen en te ondertekenen. Na het ondertekenen stellen we u **enkele algemene vragen** die nodig zijn voor de analyse van dit onderzoek.

Daarna volgt het eigenlijke experiment: U zult **40 korte geluidsfragmenten** te horen krijgen, **gesproken door kinderen**.

Belangrijke instructies voor het transcriberen

- Alle geluidsfragmenten bevatten **alleen echte woorden.**
- Geluidsfragmenten kunnen **grammaticaal incorrect** zijn.
- **Elk fragment kan maar 1 keer afgespeeld worden.**
- Geluiden zoals **"uh", "uhm", "mm", zuchten,** etc., maar ook **interpunctie** (punten, komma's, hoofdletters) **mogen opgeschreven worden, maar worden niet meegenomen in de beoordeling.**
- Let op dat sommige woorden op verschillende manieren uitgesproken kunnen worden en dan dus ook anders getranscribeerd moeten worden, zoals: **"me"** $\neq$ **"m'n"** $\neq$ **"mijn"**. **Noteer altijd wat u daadwerkelijk hoort.**
- Transcribeer alleen kindspraak en niet mogelijke andere spraak.
- Wanneer u helemaal niets hebt verstaan van een fragment, typ alstublieft een "," (komma) in het tekst veld.

Tot slot

Probeer rustig en geconcentreerd te werken. Er is geen tijdsdruk, neem alle tijd die u nodig heeft.

Hartelijk dank voor uw tijd en inzet! Uw bijdrage is ontzettend waardevol voor dit onderzoek.

C.2 General questions

1. Vult u alstublieft het persoonlijk ID in wat genoteerd staat op uw consent form.
   Open-ended
2. Wat is uw leeftijd?
   Open-ended
3. Wat is uw gender?
   ( ) Vrouw
   ( ) Man
   ( ) Anders, namelijk:
   ( ) Zeg ik liever niet
4. Bent u vertrouwd met het verstaan van Nederlands kindspraak (kinderen van 2-11 jaar)?
   ( ) Ja, ik ben vertrouwd met Nederlands kindspraak (ik ben ouder/verzorger van een Nederlands sprekend kind)
   ( ) Nee, ik ben niet vertrouwd met Nederlands kindspraak
5. Is uw moedertaal Nederlands?
   ( ) Ja
   ( ) Nee, namelijk:
6. In welke regio bent u opgegroeid?
   ( ) West Nederland (Zuid-Holland, Noord-Holland (excl West Friesland), West Utrecht incl stad)
   ( ) Transitie regio (Zeeland, Oost Utrecht (zonder stad), Gelders rivier gebied (incl Arnhem en Nijmegen), Veluwe (tot en met IJssel), West Friesland, Polders)
   ( ) Noord Nederland (Achterhoek, Overijssel, Drenthe, Groningen, Friesland)
   ( ) Zuid Nederland (Noord-Brabant, Limburg)
   ( ) Anders, namelijk:
7. In welke regio heeft u het langs geleefd?
   ( ) West Nederland (Zuid-Holland, Noord-Holland (excl West Friesland), West Utrecht incl stad)
   ( ) Transitie regio (Zeeland, Oost Utrecht (zonder stad), Gelders rivier gebied (incl Arnhem en Nijmegen), Veluwe (tot en met IJssel), West Friesland, Polders)
   ( ) Noord Nederland (Achterhoek, Overijssel, Drenthe, Groningen, Friesland)
   ( ) Zuid Nederland (Noord-Brabant, Limburg)
   ( ) Anders, namelijk:
8. Heeft u gehoorproblemen die het verstaan van spraak kunnen beïnvloeden?
   ( ) Nee
   ( ) Ja, lichte gehoorproblemen
   ( ) Ja, ernstige gehoorproblemen
   ( ) Weet ik niet/zeg ik liever niet

C.3 Instructions test audio and start of survey

We willen dat u de spraakfragmenten op een comfortabel geluidsniveau beluistert. U kunt het onderstaande fragment meerdere keren afspelen. Stel het geluidsniveau nu eerst in op een comfortabel niveau voordat u verdergaat.

Het experiment begint op de volgende pagina. We willen u nogmaals vragen om geconcentreerd en rustig te werken. Het geluidsfragment speelt af door op de play knop te drukken.

Na iedere 10 geluidsfragmenten krijgt u de mogelijkheid om kort te pauzeren.

Let op: U kunt maar 1 keer naar elk geluidsfragment luisteren.

D Prompts used for AI tools

Example of prompts that were used for AI tools:

D.1 Writing support

- Help me generate a thank you text at the end of my survey. In Dutch please
- Please rewrite this piece of text to make it a more formal writing style. Do not change the structure of the text and do not change any statements that are made. Please provide your changes by striking through what you deleted/altered and making bold what you added.

D.2 Programming support

- Please help me write a Python script that takes my speaker Ids. And then from this gets the utterances I could use. Make sure the utterances are between 3 and 7 seconds. Seconds can be calculated by: N000016-fn000029-1007.6059-1008.9621. Subtracting 1008.9621-1007.6059. That is the amount of seconds
- Please help me write a Python script that merges two files set up like this (files provided). I want to have the following columns: file name, transcript and gender.
- I have a lexicon. I have transcription files for all my participants named like hyp_u1.trn. For each word in my transcripts I want to see if it is in the lexicon list. If not it should be flagged as an non existing word. Please help me write a Python script

D.3 Literature support

- Are there already scientific papers that look at how well humans that have children are at recognizing child speech vs humans that do not have children vs ASR systems?
- I want to focus on child speech. And then I want to focus on comparing how well AI does versus 2 groups of humans. One group of humans that has children themselves. One group of humans that does not have children themselves. Help me with a search query to find research papers related to this.
- I provided you with a paper. Please summarize the main findings in this paper. If you find any parts of the paper that are especially interesting for my experiment, provide me these with clear directions where to find them in the paper.

D.4 Visualization support

- What would be the best way to visualize the WER for the three groups, familiar, unfamiliar and ASR? Should I use a graph or a table? If using a graph, what type of graph would be best suited?
- What would be the best way to make a table to show the age distribution of my participants. I want to show 3 things. Familiar or unfamiliar. Man or Woman and then also the age distribution. How can I do this?
- I have to think about how I want to store the results I gather from the interviews I will do. What will be the easiest way to later compare them and do analysis.

D.5 Formatting help

- How do you create subsections in an appendix in latex?
- What are latex reference list options, for example APA but what else?

This is not an exhaustive list of prompts used for consulting LLM's during this project, but these examples should give a clear idea of the kind of prompts that were used and the way that LLM's were used as assistance during this project.

References

[1] A. Hannun, "The History of Speech Recognition to the Year 2030," 7 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2108.00084

[2] W. Seymour, X. Zhan, M. Coté, and J. Such, "A Systematic Review of Ethical Concerns with Voice Assistants," in *AIES 2023 - Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, Inc, 8 2023, pp. 131–145. [Online]. Available: https://doi.org/10.1145/3600211.3604679

[3] A. A. Attia, J. Liu, W. Ai, D. Demszky, and C. Espy-Wilson, "Kid-Whisper: Towards Bridging the Performance Gap in Automatic Speech Recognition for Children VS. Adults," in *AIES 2024*, San Jose, California, 2024, pp. 74–80. [Online]. Available: https://doi.org/10.48550/arXiv.2309.07927

[4] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying Bias in Automatic Speech Recognition," 4 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2103.15122

[5] A. R. Shour, R. Anguzu, and A. A. Onitilo, "Speech Recognition Technology and Documentation Efficiency," *JAMA Network Open*, vol. 8, no. 3, 3 2025. [Online]. Available: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2831872

[6] A. Brasoveanu, M. Moodie, and R. Agrawal, "Textual evidence for the perfunctoriness of independent medical reviews," in *CEUR Workshop Proceedings*, vol. 2657. CEUR-WS, 4 2023, pp. 1–9. [Online]. Available: https://ceur-ws.org/Vol-2657/paper1.pdf

[7] R. P. Lippmann, "Speech recognition by machines and humans," Lincoln Laboratory MIT, Lexington, Tech. Rep., 4 1997. [Online]. Available: https://doi.org/10.1016/S0167-6393(97)00021-6

[8] C. Patman and E. Chodroff, "Speech recognition in adverse conditions by humans and machines," *JASA Express Letters*, vol. 4, no. 4, 11 2024. [Online]. Available: https://doi.org/10.1121/10.0032473

[9] C. Spille, B. Kollmeier, and B. T. Meyer, "Comparing human and automatic speech recognition in simple and complex acoustic scenes," *Computer Speech and Language*, vol. 52, pp. 123–140, 11 2018. [Online]. Available: https://doi.org/10.1016/j.csl.2018.04.003

[10] A. Lopez, A. Liesenfeld, and M. 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)*. Potsdam, Germany: KONVENS 2022 Organizers, 2022, pp. 135–143. [Online]. Available: https://aclanthology.org/2022.konvens-1.16/

[11] H. Bradley, M. E. Yu, and E. K. Johnson, "Voice assistant technology continues to underperform on children's speech," *JASA Express Letters*, vol. 5, no. 5, 3 2025. [Online]. Available: https://doi.org/10.1121/10.0036052

[12] S. Ghai and R. Sinha, "Exploring the effect of differences in the acoustic correlates of adults' and children's speech in the context of automatic speech recognition," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2010, 1 2010. [Online]. Available: https://link.springer.com/article/10.1155/2010/318785

[13] P. Flipsen, "Speaker-listener familiarity: Parents as judges of delayed speech intelligibility," *Journal of Communication Disorders*, vol. 28, no. 1, pp. 3–19, 1995. [Online]. Available: https://doi.org/10.1016/0021-9924(94)00015-R

[14] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101567, 3 2024. [Online]. Available: https://doi.org/10.1016/j.csl.2023.101567

[15] Dutch Language Institute, "JASMIN-spraakcorpus Commercieel (Version 1.0)," 2008, [Data set]. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-e4.

[16] The FFmpeg developers, "FFmpeg loudnorm." [Online]. Available: https://ffmpeg.org/ffmpeg-filters.html

[17] R. Ruhm, C. Leitner-Jones, A. Kulmhofer, T. Kiefer, H. Mlakar, and U. Itzlinger-Bruneforth, "Playing the Recording Once or Twice: Effects on Listening Test Performances," *International Journal of Listening*, vol. 30, no. 1-2, pp. 67–83, 5 2016. [Online]. Available: https://doi.org/10.1080/10904018.2015.1104252

[18] A. R. Bradlow and T. Bent, "Perceptual adaptation to non-native speech," *Cognition*, vol. 106, no. 2, pp. 707–729, 4 2008. [Online]. Available: https://doi.org/10.1016/j.cognition.2007.04.005

[19] Google Telephony, "Compare transcription models." [Online]. Available: https://docs.cloud.google.com/speech-to-text/docs/transcription-model

[20] Y. Zhang, T. De Valck, and O. Scharenborg, "State-of-the-art speech recognition systems show bias against Dutch diverse speech," Ph.D. dissertation, TU Delft, private communication, in preperation.

[21] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-lingual Representation Learning for Speech Recognition," 12 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2006.13979

[22] Instituut voor de Nederlandse Taal, "Corpus Gesproken Nederlands (CGN)." [Online]. Available: https://taalmaterialen.ivdnt.org/download/tstc-corpus-gesproken-nederlands/?utm_source=chatgpt.com

[23] T. Patel and O. Scharenborg, "Improving End-to-End Models for Children's Speech Recognition," *Applied Sciences (Switzerland)*, vol. 14, no. 6, 3 2024. [Online]. Available: https://doi.org/10.3390/app14062353

[24] Y. Zhang, A. Herygers, T. Patel, Z. Yue, and O. Scharenborg, "Exploring data augmentation in bias mitigation against non-native-accented speech," 12 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.15499

[25] "Sclite," 7 2004. [Online]. Available: https://sources.debian.org/data/main/s/sctk/2.4.10-20151007-1312Z%2Bdfsg2-3.1~deb10u1/doc/sclite.htm

[26] T. Bäckström, O. Räsänen, A. Zewoudie, P. P. Zarazaga, L. Koivusalo, S. Das, E. G. Mellado, M. B. Mansali, D. Ramos, S. Kadiri, P. Alku, and M. H. Vali, *Introduction to Speech Processing*, 2nd ed. Aalto University, 2022. [Online]. Available: https://speechprocessingbook.aalto.fi