Humans vs. ASR: Transcription Performance on Child Speech

The Effect of Familiarity with Child Speech on Transcription Performance

Ilse Huisman – I.N.Huisman@student.tudelft.nl Supervisor and responsible professor: Odette Scharenborg

TU Delft

1. Introduction

- Prior work shows that child speech is transcribed worse than adult speech by Automatic Speech Recognition (ASR) systems.
- Humans generally transcribe better than ASR systems.
- Children's voices, sentence structure and pronunciation very different from adults.
- Older children's speech transcribed better than younger children's speech.

This study compares Dutch human listeners and a Dutch ASR system on Dutch child speech. It also examines whether familiarity with child speech improves human transcription, and how child speaker age affects performance.

Example Child Speech Transcription

🎧 Utterance: "I wanta go there."
✗ ASR output: "I want to go deer."
✓ Human listener output: "I want to go there."

vs.

ASR often wrongly transcribes child speech, while humans recover meaning even with unclear articulation.

2. Research Questions

1. How well do Dutch human listeners transcribe Dutch child speech in comparison to Dutch ASR systems?
2. To what extent does familiarity with child speech influence human transcription performance?

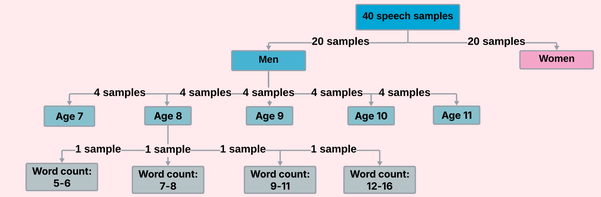
Sub question

How does the age of the child speakers affect the performance of Dutch ASR systems and human listeners?

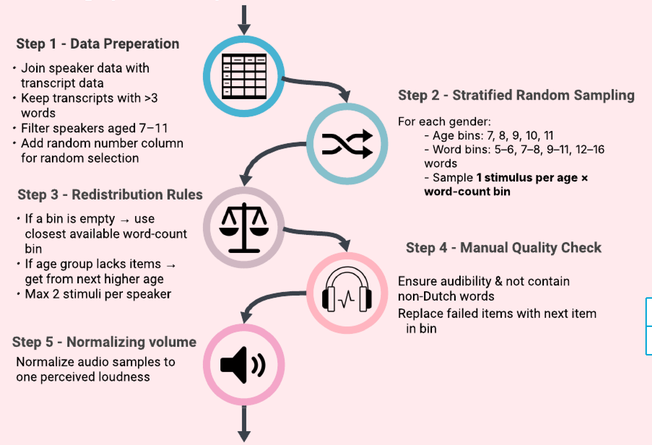
3. Methodology

Speech Material Selection

First, a balanced set of 40 stimuli of child speech was obtained from the Jasmin database. The set was distributed like this:



Selecting Speech Samples



Human Listener Experiment

20 participants, divided into:
- Familiar with child speech (parents/caretakers)
- Unfamiliar listeners

Balanced for gender and diverse age distribution.
Conducted in a quiet room with identical listening equipment.
40 audio clips in random order per participant.
1 listen max per clip to avoid learning effects.
Break after every 10 items.

ASR Experiment

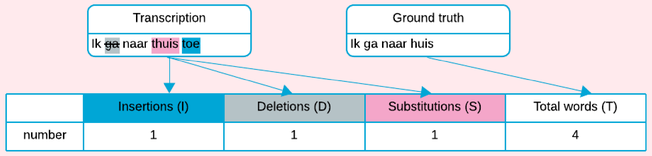
Same 40 audio stimuli processed through 2 Dutch ASR systems.

Post-Processing

Casing/punctuation normalized, obvious typos corrected.

Evaluation Metric: Word Error Rate (WER)

Measures how closely the transcription matches the ground truth.



$$\text{Word error rate} = \frac{I+D+S}{T} * 100\% = \frac{1+1+1}{4} * 100\% = 75\%$$

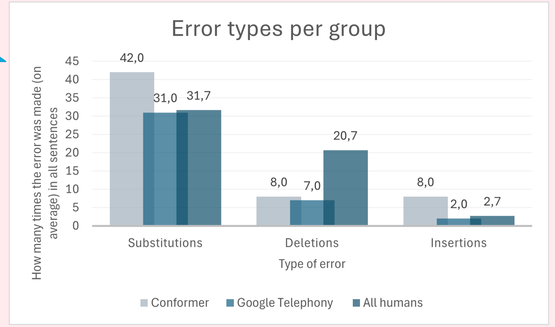
4. Results

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

Conformer model slightly worse than humans, but not significant (p=0.574). Google Telephony better than humans, but not significant (p=0.261).

Group/System	WER (%)
Familiar humans	18.0
Unfamiliar humans	17.2

Unfamiliar slightly better than familiar humans, but not significant (p=0.496).



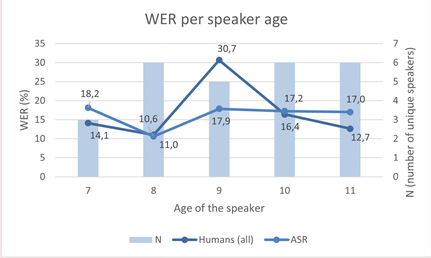
Humans did way more deletions than ASR models.
ASR models use best-guess.

Conformer model did a lot of substitutions and insertions in comparison to other groups.

5. Discussion

No trend between WER and speaker age.

WER of the humans for speaker age group 9 extremely high. → Take a look at worst transcribed sentences by humans.

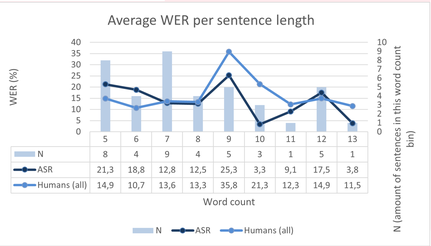


Worst sentence	Avg WER (%)	Speaker ID	Speaker age	Word count	Ground truth
1	81.1	N000027	9	9	binnenkort maar ik ben pas om achttien oktober jarig
2	55.5	N000027	9	10	hij doet een beetje hij doet het niet echt goed
3	42.8	N000057	9	9	kluiven en dat vind ik wel heel erg leuk
4	36.4	N000213	7	7	me vader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

3 worst sentences by age 9 → speaker effects → explains high WER.
3/5 have word count of 9 or more. → Look at WER per word count.

No trend between WER and sentence length.

Extremely high WER for word count 9. → Explained by 2/5 of the worst sentences.



6. Conclusion & Future Research

The Dutch ASR systems perform comparably to Dutch human listeners on Dutch child speech.
Familiarity with child speech has no influence on performance.
The age of the child speaker has no effect on the WER.
Differences in performance likely driven by speaker characteristics and stimulus-specific factors.

- Bigger stimulus set (>40).
- Each speaker only 1 stimuli. → Minimize speaker effects.
- Speech from children under the age of 7.
- Measure familiarity using more levels (hours per week, age of children).