Human vs AI: Comparing Transcription Performance In Dutch Older Adults' Speech

Author: Ansen Weng
a.weng@student.tudelft.nl

Responsible Professor & Supervisor:
Prof. dr. Odette Scharenborg

1 Background

- Automatic Speech recognisers (ASR) can convert speech to text.
- State-of-the-art (SotA) ASR systems do not work well for older adults' speech.
- More older adults due to global ageing.
- Research done in ASR comparisons with different speech, but no human-ASR comparisons have been made with Dutch older adult (65+) speech.

2 Research Questions

Main: How do the transcription performance of humans compare SotA ASR systems in Dutch Older Adults' Speech?

Sub question:

- *What errors do humans make compared to SotA ASR systems?*

3.1 Methodology

Experiment

- Collect general data from participants via questionnaire.
- Participants transcribe 40 audio fragments of Dutch older adults' speech.

Speech Database

- Jasmin-CGN, corpus contains Dutch older adult (65+)speech.
- Contains metadata like age, gender, and dialectRegion.
- Stratified sampling to choose audio fragments for balanced speaker representation in age, gender, and region.

ASR models

- Google Telephony (company ASR model).
- Conformer (trained by Zhang et al.).

3.2 Methodology

Post-processing

- Remove punctuations (except apostrophe).
- Convert numbers to written form.
- Fix obvious spelling and typos.

Evaluation Measure

- Word Error Rate (WER) =

$$\frac{Substitutions + Deletions + Insertions}{Number\ of\ words\ in\ correct\ transcript} * 100$$

Example:

Correct:	ik	wil	bananen	Correct:	ik wil bananen
Transcript:	ik		banaan	Transcript:	ik wil twee bananen
Error:		D	S	Error:	I

4.1 Results

Table 1: Mean WER of human listeners and ASR systems. Components of WER are also shown.

	Mean	S.D.	Sub	Del	Ins
Human listeners	26.3	23.9	14.4	9.6	2.3
Conformer	23.0	31.8	14.2	5.4	3.3
GT	16.6	19.3	11.3	3.5	1.9

Table 2: Mean WER per region

	Regions			
	W	T	N	S
Human listeners	15.1	18.3	26.8	45.7
Conformer	11.0	16.3	36.4	28.2
GT	10.8	8.2	21.6	25.9

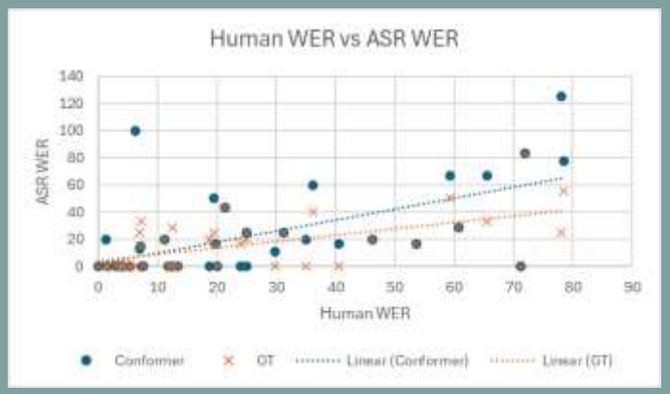
Table 3: Mean WER of human listeners and ASR systems per gender

	Female	Male
Human listeners	28.6	23.9
Conformer	29.1	16.2
GT	18.6	14.5

4.2 Results

Table 4: Mean WER of human listeners and ASR systems per age group

Age range	[59,72)	[72,84)	≥ 84
Human listeners	11.36	26.3	38.1
Conformer	12.5	25.5	28.5
GT	13.0	19.9	16.0



- No significant differences could be found at all for ASR systems, contrary to expectations.
- WER of humans and ASR systems are positively correlated.

5 Conclusion

- Humans perform comparably to ASR systems.
- Humans showed significant performance differences for region and age.
- Error rates of humans and ASR systems are comparable.
- Humans and ASR systems may struggle on same speakers.

Next steps

- More audio fragments should be transcsribed.
- Investigate deeper in the errors that humans and ASR systems make.