



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

Ansen Weng¹

Supervisor: Prof. dr. Odette Scharenborg^{1,2}

¹EEMCS, Delft University of Technology, The Netherlands

²Delft Inclusive Speech Communication (DISC) Lab, 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: Ansen Weng

Final project course: CSE3000 Research Project

Thesis committee: Prof. dr. Odette Scharenborg, Dr. Bernd Dudzik

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

Abstract

Automatic speech recognition (ASR) systems are used in voice assistants and games. Studies have shown that these ASR systems do not work well for all everyone. E.g., speech from regional accents is recognised worse than the native standard accent and speech from older adults are recognised worse than adults.

However, no research has been done in comparing humans and ASR systems on Dutch older adults' speech. This paper will fill this gap by finding out the transcription performance of humans and ASR systems (Google Telephony and a conformer model with XLSR-53 features) on Dutch older adults' speech and comparing the performances with each other. This paper also investigated whether regional accents, gender, and age had a significant effect on the performance of humans and ASR systems.

The results showed that humans and ASR systems perform similar to each other and that their performance are positively correlated. The results also showed that humans also do not recognise speech equally from older adults as well. These results suggests that there are speech characteristics for which both humans and ASR systems have trouble with.

1 Introduction

The use of automatic speech recognition (ASR) systems, better known as speech-to-text, has increased in the past few decades. These ASR systems have been used in voice assistants (like Siri) and games [1, Ch.1]. However, ASR systems do not work well for all speaker groups (e.g., children, older adults, and non-natives) [2]. Paired with the global population ageing [3], it is important that ASR systems should work well for older adults.

Previous work has shown that there are several reasons why speech for older adults is not well recognised. One of which is that older adults have less control over their speech organs because of ageing [4]. Another reason is due to the use of language, where older adults tend interact with ASR systems as humans and use complex commands [4, 5].

Other work uncovered and quantified bias in ASR systems. Bias in this context being that ASR systems do not perform equally well for different speaker groups of for example different age, gender, and regional accent [2, 6, 7]. Additionally, performance evaluations of state-of-the-art (SotA) ASR systems have been done on diverse/non-standard speech (e.g., child, teenage, older adults, and non-native) to inform e.g. researchers, which ASR system works well for what kind of speech, and which ASR system needs improvement [2, 8, 9].

Although current research and performance evaluations in speech transcription of ASR systems indeed indicate that Dutch older adults' speech is not recognised well [2, 6, 8, 5], no research has been done in comparing human performance against the SotA ASR systems for Dutch older adults' speech. Hence, there is no information on whether the SotA ASR systems are performing worse, comparably, or better than humans.

This study should fill this gap by comparing humans against SotA ASR systems on Dutch older adults' speech and finding out what kind of errors humans and ASR systems make on this speech. This should provide researchers and developers with insights into what can be improved in the ASR systems. Concretely, this study tries to answer the following research questions (RQ):

RQ1 What is the transcription performance in Dutch older adults' speech between humans and SotA ASR systems?

RQ2 What errors do humans make compared to SotA ASR systems?

The paper is organised as follows to answer the questions. First the research methodology will be described in section 2. Then ethical considerations for this study will be discussed in section 3. Afterwards, the results will be provided in section 4, which will then be discussed in section 5. Finally, the conclusion will be provided in section 6.

2 Methodology

First the listening experiment will be described. Then the speech corpus containing Dutch older adults' speech will be described. Then information on human participants will be described and afterwards information on the ASR systems. Afterwards, the experiment set-up will be described. Finally, the evaluation measure used in this study will be described along with the analyses that are done.

2.1 Listening experiment

The listening experiment can be described as follows: Participants listen to 40 audio samples, and they get an opportunity to take a break every ten audio samples. For each audio sample, participants will need to transcribe what they hear, and they only get to hear a sample once. This is done to make it more similar to real life situations.

To participate in the listening experiment, participants must sign a consent form. After signing the consent form, participants will get written and verbal instructions (see Figure A1) on the listening experiment. Afterwards, participants can fill in a questionnaire containing questions like age, gender, and region they come from.

We decided that the whole procedure (signing consent form, reading the instructions, filling in the questionnaire, and doing the experiment) should take around 30 minutes. This decision was made to prevent exhausting and overburdening the participants so that participants can transcribe all audio samples the best they can. We determined that 40 audio samples can be reliably transcribed within that time. Hence, why participants listened to 40 audio samples.

2.2 Speech corpus

Jasmin-CGN corpus is a dataset containing speech of Dutch children, teenagers, and older adults and their corresponding transcriptions [10]. The speech in Jasmin-CGN corpus consists of read and human-machine-interaction (HMI) speech and totals around 9 hours and 26 minutes (4 hours and 43 minutes for each type of speech). Read speech is speech that has

been read aloud from a script and HMI speech is speech that occurs from humans interacting with machines (e.g., voice assistants). Human-machine-interaction speech of Dutch older adults is used from the Jasmin-CGN corpus to compare the human and ASR transcription performance. Only the HMI speech has been used because HMI speech resembles real life situations more than read speech.

Aside from the speech data, the Jasmin-CGN corpus also contains metadata of the speakers such as age, gender, and region. In Jasmin-CGN there are 68 different Dutch older adults from four different regions in the Netherlands, namely West (W), Transitional (T), North (N), and South (S) with their ages ranging from 58 to 95 (mean = 78.9, S.D. = 8.9). See Table 1 for the distribution of older adults for each gender and region.

Table 1: Speaker distribution of Jasmin-CGN per gender and region

Gender	Regions			
	W	T	N	S
F	13	9	13	10
M	5	8	4	6

Additionally, as described in [2], the speech data is pre-processed into smaller chunks such that each chunk contained one utterance (i.e., sentence) and silent chunks were removed.

2.3 Participants

A total of 16 adult participants were recruited for this study. Recruitment was done through personal connections of the author (specifically through sports club, friends, family, and peers). These participants are Dutch native speakers or have a C2-level proficiency, according to the Common European Framework (CEF) [11].

The participant data distribution is as follows:

- Seven male and nine female participants.
- All participants indicated to have lived the longest in the W region.
- Two female participants indicated a non-Dutch nationality.
- Minimum age and maximum age are 20 and 63, respectively. Mean and standard deviation are 37.8 and 17.3, respectively.

2.4 ASR systems

Two SotA ASR systems have been used for comparison with humans, namely Google Telephony (GT) and a conformer-based ASR model trained by Zhang et al. [12]. The results on the Jasmin-CGN corpus for both ASR systems were obtained by Zhang et al. and provided for this study.

GT is used, because it is an ASR model that specialises in telephone speech which is a type of conversational speech that is similar to HMI speech [13] and because it showed the best performance on Dutch HMI speech [12]. The results were obtained via the Google Cloud API using synchronous speech recognition.

Unlike GT, the conformer model was chosen because it is trained from scratch and because it had the best performance among the self-trained ASR systems. This conformer model is solely trained on the Corpus Gesproken Nederlands (CGN) using similar segmentation described in Section 2.2 with the addition of excluding chunks shorter than 0.1 seconds. Its training set consists of 424.6 hours of Dutch and 265.9 hours of Flemish and the validation set consists of 6.9 hours [12]. The conformer model is trained by following similar steps in [2, 7]. Additionally, the model is trained with XLSR-53 features, specifically "1024-dimensional fused representations from all 24 layers of XLSR-53" [12]. These features were extracted by following [14].

2.5 Listening experiment set-up

The questionnaire and listening experiment were both set up on Qualtrics, software designed for making surveys [15]. The experiment took place in person in a quiet room to ensure minimal distractions to the participant. Additionally, participants listened to the audio using a Sennheiser HD 200 Pro headphone during the experiment. The headphones were plugged in a HP Zbook Power G9 laptop via its jack port. The experiment was also done on the same laptop.

As mentioned in Section 2.1, 40 HMI audio samples were selected for the listening experiment. The selected audio samples have a word count of at least four, because the audio samples with word count three or less contained stop-words or non-linguistic tokens such as 'uh' and laughter which are not interesting to transcribe (mean word count = 6.3, S.D. = 1.840). The loudness of the selected audio samples was normalised using FFmpeg's loudnorm [16] that implements the EBU R128 loudness recommendation of the European Broadcasting Union [17]. FFmpeg's loudnorm makes the loudness of the audio samples more uniform. In addition, each audio sample comes from a different speaker. The speakers were selected using stratified sampling such that there is a balanced representation of age, gender, and region (see Table 2 for the final distribution) to minimise bias [7, 18, 19]. Additionally, the audio samples were ordered randomly for each participant to prevent the last samples to be transcribed better than the first due to acclimatisation of transcribing.

Table 2: Speaker distribution (in bold) in chosen audio samples (age-range in brackets for each region and gender)

Gender	Regions			
	W	T	N	S
F	2 (59-69)	1 (68-74)	3 (68-76)	1 (65-75)
	1 (70-80)	1 (75-81)	2 (77-85)	1 (76-86)
	2 (≥ 81)	3 (≥ 82)	1 (≥ 86)	3 (≥ 87)
M	2 (68-76)	1 (67-74)	2 (65-68)	2 (66-75)
	1 (77-85)	1 (75-82)	1 (69-72)	2 (76-85)
	2 (≥ 86)	3 (≥ 83)	1 (≥ 73)	1 (≥ 86)

2.6 Post-processing

Once all data were gathered from the participants. The following post-processing is done on the transcriptions:

1. Remove any punctuation except for apostrophe.
2. Convert numbers to its written form.
3. For each word in the transcription, check if it exists with the help of a lexicon.
4. Fix obvious spelling mistakes and typos.

Any words encountered in Item 3 that do not exist, were either corrected because it was categorised as Item 4 or left as it was, because it was not obvious what word the participant transcribed (e.g., kate). All typos and most spelling mistakes could be found this way and corrected. The rest of the spelling mistakes were in the lexicon, but were wrong in the given context (e.g., ik vindt instead of ik vind). These mistakes were corrected. Common typos and spelling mistakes were insertion of letters from nearby keys (e.g., werkken instead of werken, tweintig instead of twintig), wrong ordering of letters within a word (e.g., liefts instead of liefst), and concatenation of words that should not be concatenated (e.g., onderandere instead of onder andere).

2.7 Evaluation measure

The evaluation measure used in this study is the word error rate (WER). WER is calculated using the following Equation (1).

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

Where N is the total words in the ground truth, S the number of words substituted, D the number of words deleted, and I the number of words inserted. The lower the WER, the closer it is to the ground truth. The ground truth in this study are the transcriptions of the speech data in Jasmin-CGN corpus.

This WER is calculated using ESPnet2, specifically the scoring in [20]. The reason for using ESPnet2 is that the scoring method has an option to include a non-linguistic symbol list. The scoring will then ignore all symbols from that list in the WER calculations. Non-linguistic symbols are symbols of non-linguistic sounds like uh, uhm and laughter. With that list, no additional processing steps are needed to remove those non-linguistic symbols from the transcription.

To answer the first research question, we first evaluate the performance of humans and ASR systems and then compare the WER with each other. Additionally, we investigate the performance of humans and ASR systems for each of the regions separately.

To answer the second research question, we investigate the number of insertions, substitutions, and deletions. In addition, we investigate the correlation of the human WER and the ASR WER to see if humans and ASR systems make similar mistakes or not.

In addition to the performance evaluation of humans and ASR systems, we also include statistical tests to evaluate whether the differences between the WER of humans and ASR systems are significant. This is done using `scipy` and `sc.stats` (specifically `-mapsswe`) from the NIST Scoring Toolkit [21]. The same toolkit that ESPnet2 also uses under the hood for calculating the WER. In addition, one-way analysis of variance (ANOVA) was done to compare WER between more than two groups. Groups for example being the

four different regions, or the comparison of humans and ASR systems. If a significant difference is found via ANOVA, a post-hoc Tukey test is done to find out which groups differ significantly from each other.

3 Responsible research

Due to the nature of this study, several ethical concerns have been addressed.

Jasmin-CGN has strict rules on the usage and distribution of its data. This study ensured that the data from Jasmin-CGN has only been used for research purposes and that only the necessary data have been used during the study. I.e., only the documentation, metadata of speakers, text information of the HMI speech of older adults, and audio files of the selected audio samples were used and downloaded.

Google Telephony was used to transcribe data from Jasmin-CGN corpus. It is ensured that the audio files were shorter than 60s and thus not saved in any way on Google Cloud.

The experiment was implemented on Qualtrics. This study ensured that responses from participants using Qualtrics are anonymised and that no identifying information was collected. Once this study is done, responses from the participants and a blank version of the survey will be exported and securely saved. The survey and all associating data will be deleted from Qualtrics [15].

Because data is obtained from humans, approval is needed from the Human Research Ethics Committee (HREC) [22]. This has been granted by HREC with the corresponding number 6233.

No AI tools like LLM's are used in the writing nor is it used for any other purposes anywhere during the study.

4 Results

4.1 Overall performance of human listeners and ASR systems

Table 3: Mean WER of human listeners and ASR systems

	Mean	S.D.
Human listeners	26.3	23.9
Conformer	23.0	31.8
GT	16.6	19.3

Table 3 shows the mean WER of the human listeners and the ASR systems along with the standard deviation of each mean. Although the average WER of human listeners are higher compared to the ASR systems, ANOVA shows no significant difference in the performance between the human listeners and the ASR systems.

It can be seen in Table 3 that the standard deviations are higher or almost equal to the mean, which suggests that the WERs are spread out. We further investigate this spread by investigating the mean WER of the speaker dimensions (e.g., region, gender, and age).

4.1.1 Performance of human listeners and ASR systems per region

Table 4: 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 4 shows the mean WER per region of all human listeners and ASR systems. Human listeners performed the best for region W and the worst for region S. Significance testing shows that for human listeners, there is a significant difference between the four regions ($F(2.87) = 4.26, p = 0.011$). Post-hoc Tukey test shows that the regions W and S differ significantly from each other ($Q(3.809) = 4.551, p = 0.0139$) and so is regions T and S ($Q(3.809) = 4.036, p = 0.0343$).

The conformer model performs the best for region W, but the worst for region N. ANOVA shows no significant difference between any of the four regions.

GT performs the best for region T, but the worst for region S. ANOVA shows no significant difference between any of the four regions.

Comparing human listeners and the ASR systems on each of the regions show no significant difference.

4.1.2 Performance of human listeners and ASR systems per gender

Table 5: 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

Table 5 shows the mean WER of the human listeners and ASR systems per gender. We can observe for human listeners and ASR systems that male speech is better recognised than female speech. No significant difference is found between the WER of female and male speakers for the human listeners, nor for the conformer model, nor for GT.

ANOVA shows no significant difference when comparing human listeners and ASR systems against each other on each of the genders.

4.1.3 Performance of human listeners and ASR systems per age

Table 6 shows the mean WER of human listeners and ASR systems per age group. We observe that the WER increases as the age also increases except for GT, where the last age group is between the middle and first age group. ANOVA shows there is a significant difference between the age groups for the human listeners ($F(3.252) = 4.550, p = 0.0171$). Post-hoc Tukey-Kramer test shows that the first and last group

Table 6: 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

differ significantly from each other ($Q(3.453) = 4.266, p = 0.125$). ANOVA shows no difference between the age groups for both the conformer model and GT.

ANOVA shows no significant difference when comparing human listeners and the ASR systems against each other on each of the age groups.

4.2 Error results of human listeners and ASR systems

Table 7: Mean substitutions, deletions, and insertions of human listeners and ASR systems (in %)

	Sub	Del	Ins
Human listeners	14.4	9.6	2.3
Conformer	14.2	5.4	3.3
GT	11.3	3.5	1.9

Table 7 shows the mean substitution, deletion, and insertion error rates of the human listeners and ASR systems. We observe that the error rates across the human listeners and ASR systems are similar, except for the deletion error rate which shows that human listeners have a higher deletion error rate. ANOVA does not show any significant difference when comparing the human listeners and the ASR systems against each other on the substitution, deletion, and insertion error rates.

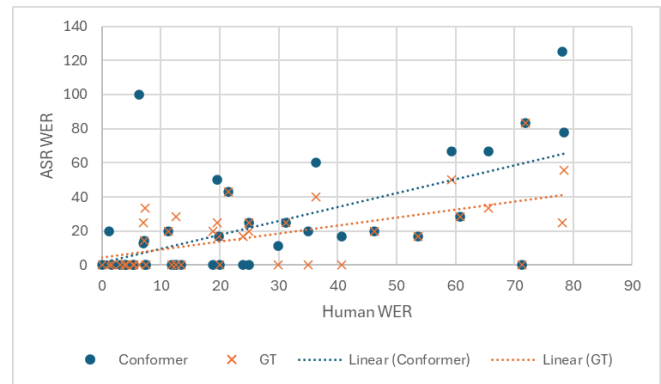


Figure 1: Human WER vs ASR WER

A closer look into where human listeners and ASR systems make errors can be seen in Figure 1. Figure 1 shows a scatter-plot of the human WER and ASR WER at the speaker-level. It can be observed that the human WER and ASR WERs are

related. The correlation coefficient between the human listeners and the conformer model is $\rho = 0.602$, and the correlation between human listeners and GT is $\rho = 0.550$. Both correlation coefficients show that the WER of humans and ASR systems are positively related. This suggests that there are speech characteristics that both humans and ASR systems have trouble with. We can also observe there are speakers for which humans have a high WER compared to the ASR systems, which may suggest there are speech characteristics that humans have trouble with but ASR systems not.

5 Discussion

The overall results showed that human performance is comparable to ASR performance indicating that ASR systems are improving and approaching human performance, which is in line with previous studies [23, 24, 25].

A closer investigation on the mean WER of different speaker dimensions (region, gender, age) showed that humans exhibit significant performance differences within each dimension, except for gender where no such conclusion could be made. And no reason could be found to explain this, although previous work has shown that, at least for ASR systems, the results for gender are conflicting [2, 18, 26]. For the other two dimensions plausible reasons could be found. For regional speech it could be because the recruited human listeners all lived the longest in the region with the lowest WER and are most familiar with speech from that region. As for the significant performance difference in the speaker age, one possible and obvious reason would be because of the effect of ageing [4]. These findings are similar to what [2] found for ASR systems, where ASR systems exhibit significant performance differences in different speaker dimensions.

Contrary to [2, 7], no significant differences could be found for the ASR systems in this study in the different speaker dimensions. An explanation could be that there were only one speech sample for each speaker and therefore the performance of the ASR system for a particular speaker is highly dependent on that singular speech sample.

Additionally, we could not find any significant difference between the human listeners and ASR systems for each of the speaker dimensions. Together with the results showing that the WERs of human listeners and ASR systems are positively correlated with each other, suggests that human listeners and ASR systems make similar mistakes and that there are speech characteristics for certain speaker groups that both humans and ASR systems have trouble with. This is in line with the findings in [25].

Some limitations of this study include time constraints, speech sample size, and representation of the human listeners. Particularly this study was conducted in, effectively, 9 weeks. So, some analyses such as phoneme analysis in [2, 25] could not be done. Additionally, a limited number of human listeners could be recruited, and a limited number of audio samples could be transcribed. Future research should take these limitations into account to make the results more generalisable.

Nonetheless, this study showed that humans show similar performance differences as ASR systems and that suggested that the errors that humans and ASR systems make might be

similar.

This study recommends recruiting more human listeners and increasing the number of audio samples to be transcribed by human listeners and ASR systems. Future research should also investigate deeper into the relation of the errors of humans and ASR systems at the phoneme-level.

6 Conclusions and Future Work

The aim of this study was to find the transcription performance of humans and to find out whether humans show similar performance differences like in ASR systems for Dutch older adults' speech, since no research has been done on this.

We found that the performance of human listeners and ASR systems are comparable to each other and that humans show significant performance difference for regional accents and speaker age. In this study, humans from the western region of the Netherlands recognised speech from the same region better than speech from the Southern region of the Netherlands. And humans recognised "younger" older adults (<72) better than the "older" older adults (≥ 84)

We could not find the exact type of errors that humans and ASR systems make due to time constraints. However, results showed that the performance of humans and ASR systems are positively related to each other which suggests that humans and ASR systems have trouble on the same speech characteristics.

Interestingly, the ASR systems in this study did not show any kind of significant performance differences for each of the different speaker dimensions (speaker age, gender, and regional accent). Which is not in line with the findings in [2, 7].

Some recommendations for future research are to increase the number of human listeners and the number of audio samples to be transcribed by both humans and ASR systems. Additionally, future research should also investigate the error types that humans and ASR systems make on the phoneme-level.

7 Acknowledgments

Research reported in this work was partially or completely facilitated by computational resources and support of the Delft AI Cluster (DAIC) at TU Delft (RRID: SCR.025091), but remains the sole responsibility of the authors, not the DAIC team.

I would like to thank my supervisor and responsible professor Prof. dr. Odette Scharenborg for the guidance and feedback during this study. I would also like to thank Yuanyuan Zhang for the technical help and guidance during this study. I would also like to thank Ilse Huisman and Rares Popa for their collaborations and discussions we held during the study.

References

- [1] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach* (Signals and Communication Technology), en. London: Springer London, 2015, ISBN: 9781447157786 9781447157793. DOI: [10.1007/](#)

- 978-1-4471-5779-3. Accessed: Jan. 14, 2026. [Online]. Available: <https://link.springer.com/10.1007/978-1-4471-5779-3>.
- [2] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," en, *Computer Speech & Language*, vol. 84, p. 101567, Mar. 2024, ISSN: 08852308. DOI: [10.1016/j.csl.2023.101567](https://doi.org/10.1016/j.csl.2023.101567). Accessed: Nov. 10, 2025. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0885230823000864>.
 - [3] World Health Organization, *Ageing and health*, en, Oct. 2025. Accessed: Nov. 11, 2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/ageing-and-health>.
 - [4] R. Vippera, S. Renals, and J. Frankel, "Ageing Voices: The Effect of Changes in Voice Parameters on ASR Performance," en, *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–10, 2010, ISSN: 1687-4714, 1687-4722. DOI: [10.1155/2010/525783](https://doi.org/10.1155/2010/525783). Accessed: Jan. 21, 2026. [Online]. Available: <https://link.springer.com/article/10.1155/2010/525783>.
 - [5] W. Klaassen, B. Van Dijk, and M. Spruit, "A Review of Challenges in Speech-Based Conversational AI for Elderly Care," in *Studies in Health Technology and Informatics*, E. Andrikopoulou et al., Eds., IOS Press, May 2025, ISBN: 9781643685960. DOI: [10.3233/SHTI250481](https://doi.org/10.3233/SHTI250481). Accessed: Nov. 18, 2025. [Online]. Available: <https://ebooks.iospress.nl/doi/10.3233/SHTI250481>.
 - [6] M. Fuckner, S. Horsman, P. Wiggers, and I. Janssen, "Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers," in *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania: IEEE, Oct. 2023, pp. 146–151, ISBN: 9798350327977. DOI: [10.1109/SpeD59241.2023.10314895](https://doi.org/10.1109/SpeD59241.2023.10314895). Accessed: Nov. 12, 2025. [Online]. Available: <https://ieeexplore.ieee.org/document/10314895/>.
 - [7] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, *Quantifying Bias in Automatic Speech Recognition*, 2021. DOI: [10.48550/ARXIV.2103.15122](https://doi.org/10.48550/ARXIV.2103.15122). Accessed: Nov. 12, 2025. [Online]. Available: <https://arxiv.org/abs/2103.15122>.
 - [8] D. Balan, K. P. Truong, H. v. d. Heuvel, and R. Ordelman, "Benchmarking and Research Infrastructures. Evaluating Dutch Automatic Speech Recognition," English (eng), in *CLARIN Annual Conference 2024*, Barcelona, Spain: CLARIN, 2024, pp. 140–143. Accessed: Jan. 5, 2026. [Online]. Available: <https://repository.ubn.ru.nl/handle/2066/312579>.
 - [9] D. Balan and C. Tejedor-García, *Dutch Open Speech Recognition Benchmark*, en-US, Github Pages. Accessed: Nov. 11, 2025. [Online]. Available: https://opensource-spraakherkenning-nl.github.io/ASR_NL_results/.
 - [10] C. Cucchiari, H. Van hamme, O. Herwijnen, and F. Smits, "JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives in the Human-Machine Interaction Modality," in *LREC 2006*, Genoa, Italy, 2006. Accessed: Nov. 12, 2025. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2006/pdf/254_pdf.pdf.
 - [11] Council of Europe, *Global scale - Table 1 (CEFR 3.3): Common Reference levels*. Accessed: Jan. 14, 2026. [Online]. Available: <https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale>.
 - [12] Y. Zhang, T. De Valck, and O. Scharenborg, *State-of-the-art speech recognition systems show bias against Dutch diverse speech*, In review. Accessed: Jan. 5, 2026.
 - [13] Google Cloud, *Compare transcription models — Cloud Speech-to-Text*, en. Accessed: Jan. 5, 2026. [Online]. Available: <https://docs.cloud.google.com/speech-to-text/docs/transcription-model>.
 - [14] A. Hernandez, P. A. Pérez-Toro, E. Noeth, J. R. Orozco-Arroyave, A. Maier, and S. H. Yang, "Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition," en, in *Interspeech 2022*, ISCA, Sep. 2022, pp. 51–55. DOI: [10.21437/Interspeech.2022-10674](https://doi.org/10.21437/Interspeech.2022-10674). Accessed: Jan. 22, 2026. [Online]. Available: https://www.isca-archive.org/interspeech_2022/hernandez22_interspeech.html.
 - [15] Qualtrics, *Qualtrics & GDPR Compliance*, en. Accessed: Jan. 15, 2026. [Online]. Available: <https://www.qualtrics.com/support/survey-platform/getting-started/qualtrics-gdpr-compliance/>.
 - [16] FFmpeg, *Audio Volume Manipulation*. Accessed: Jan. 14, 2026. [Online]. Available: <https://trac.ffmpeg.org/wiki/AudioVolume>.
 - [17] European Broadcasting Union, *R 128 LOUDNESS NORMALISATION and PERMITTED MAXIMUM level OF AUDIO SIGNALS*, pdf, Geneva, Nov. 2023. Accessed: Jan. 21, 2026. [Online]. Available: <https://tech.ebu.ch/docs/r/r128.pdf>.
 - [18] A. Koenecke et al., "Racial disparities in automated speech recognition," en, *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, Apr. 2020, ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1915768117](https://doi.org/10.1073/pnas.1915768117). Accessed: Jan. 21, 2026. [Online]. Available: <https://pnas.org/doi/full/10.1073/pnas.1915768117>.
 - [19] A. Bera and A. Agarwal, "Bias Detection and Mitigation Framework for ASR System," en, in *7th International Conference on Signal Processing and Information Communications*, C.-C. Wang and R. G. B. Sangalang, Eds., Cham: Springer Nature Switzerland, 2025, pp. 13–27, ISBN: 9783031899935 9783031899942. DOI: [10.1007/978-3-031-89994-2_2](https://doi.org/10.1007/978-3-031-89994-2_2).

Accessed: Jan. 21, 2026. [Online]. Available: https://link.springer.com/10.1007/978-3-031-89994-2_2.

- [20] *EspNet/espnet*, original-date: 2017-12-13T00:45:11Z, Nov. 2025. Accessed: Nov. 21, 2025. [Online]. Available: https://github.com/espnet/espnet/blob/master/egs2/TEMPLATE/asr1/README_multitask.md.
- [21] National Institute of Standards and Technology, *NIST Scoring Toolkit Version 0.1*, Documentation. Accessed: Jan. 14, 2026. [Online]. Available: <https://people.csail.mit.edu/joe/sctk-1.2/doc/sctk.htm>.
- [22] Delft University of Technology, *HREC Approval 1: Application*, nl. Accessed: Jan. 14, 2026. [Online]. Available: <https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics/hrec-approval-1-application>.
- [23] C. Spille, B. Kollmeier, and B. T. Meyer, “Comparing human and automatic speech recognition in simple and complex acoustic scenes,” en, *Computer Speech & Language*, vol. 52, pp. 123–140, Nov. 2018, ISSN: 08852308. DOI: [10.1016/j.csl.2018.04.003](https://doi.org/10.1016/j.csl.2018.04.003). Accessed: Nov. 12, 2025. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0885230817301705>.
- [24] C. Mena, A. Padilla-Ortiz, and F. Orduña-Bustamante, “Automatic speech recognition in the presence of babble noise and reverberation compared to human speech intelligibility in Spanish,” en, *Computer Speech & Language*, vol. 95, p. 101 856, Jan. 2026, ISSN: 08852308. DOI: [10.1016/j.csl.2025.101856](https://doi.org/10.1016/j.csl.2025.101856). Accessed: Nov. 14, 2025. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S0885230825000816>.
- [25] A. Stolcke and J. Droppo, “Comparing Human and Machine Errors in Conversational Speech Transcription,” en, in *Interspeech 2017*, ISCA, Aug. 2017, pp. 137–141. DOI: [10.21437/Interspeech.2017-1544](https://doi.org/10.21437/Interspeech.2017-1544). Accessed: Nov. 13, 2025. [Online]. Available: https://www.isca-archive.org/interspeech_2017/stolcke17_interspeech.html.
- [26] R. Tatman, “Gender and Dialect Bias in YouTube’s Automatic Captions,” en, in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Valencia, Spain: Association for Computational Linguistics, 2017, pp. 53–59. DOI: [10.18653/v1/W17-1606](https://doi.org/10.18653/v1/W17-1606). Accessed: Jan. 25, 2026. [Online]. Available: <http://aclweb.org/anthology/W17-1606>.

A Instructions

Welkom en bedankt voor uw deelname!

Voordat het experiment begint, wordt u vriendelijk verzocht **eerst de toestemming** ondertekenen. Na het ondertekenen stellen we u **enkele algemene vragen** die

Daarna volgt het eigenlijke experiment:

U zult **40 korte geluidsfragmenten** te horen krijgen, **gesproken door ouders** **voor het transcriberen**

- **Alle geluidsfragmenten bevatten alleen echte woorden.**
- Geluiden zoals “**uh**”, “**uhm**”, “**mm**”, **zuchten**, en andere niet-lexicale **worden niet meegenomen in de beoordeling.**
- Let op dat sommige woorden op verschillende manieren uitgesproken **moeten worden**, zoals: “**m’n**” ≠ “**mijn**”.
Noteer altijd wat u daadwerkelijk hoort.
- Elk fragment kan maar **1 keer** afgespeeld worden.
- Wanneer je helemaal niets hebt verstaan van een fragment, typ alstublie

Tot slot

Probeer rustig en geconcentreerd te werken. Er is geen tijdsdruk, neem alle tij

Hartelijk dank voor uw tijd en inzet! Uw bijdrage is ontzettend waardevol vo

Figure A1: Instructions used during the listening experiment.