Human vs AI: Recognising Teenage Speech

How good are humans at recognizing teenage speech samples compared to state-of-the-art AI-based automatic speech recognisers?

Garv Singh[1]

Supervisors: Yuanyuan Zhang[1]**, 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 23, 2026

Name of the student: Garv Singh
Final project course: CSE3000 Research Project
Thesis committee: Yuanyuan Zhang, Odette Scharenborg, Bernd Dudzik

Abstract

Automatic Speech Recognition (ASR) systems have achieved remarkable performance in recent years, however their robustness against diverse speech, such as teenage speech, remains an area of investigation and research. This study evaluates the performance of a state-of-the-art ASR system (Google Telephony) compared to human listeners in transcribing native Dutch teenage speech. Additionally, it investigates whether a listener's social exposure to teenage speech influences their recognition accuracy. A listening experiment was conducted using a dataset of 40 speech samples of Human Machine Interaction (HMI) speech from native Dutch speakers aged 14 to 16, curated from the JASMIN corpus. The audio samples were transcribed by the ASR model and a group of young adult participants (aged 20–24). Performance was evaluated using Word Error Rate (WER), with a specific focus on the impact of normalizing common Dutch contractions and clitics. The results demonstrate that the ASR system outperformed the average human listener, achieving a lower WER compared to both groups of participants, the one with exposure to teenage speech and the one without. However, human participants with regular social exposure to teenagers performed significantly better on average than those without, confirming that familiarity with the demographic improves recognition accuracy. Furthermore, the analysis reveals that orthographic inconsistencies regarding contractions significantly inflate WERs, with normalization reducing the WERs by quite an extent. These findings suggest that while current ASR models are highly robust for this demographic, human domain knowledge remains a relevant factor in understanding teenage speech speech patterns based on the lower WERs of the humans with exposure to teenage speech than those without.

1 Introduction

Automatic Speech Recognition (ASR) has become an increasingly pervasive technology. Having been deployed in critical applications ranging from emergency response centers to domestic voice assistants and healthcare systems [1] and given the importance of speech in human interaction [2], it is therefore essential and critical that these systems can effectively recognize the different varieties in human speech ranging from differences in speaker demographics, regional accents, age, gender etc. However, recent experimental evidence is unfortunately not that promising as it suggests that today's state-of-the-art ASR systems fail to recognize the different varieties of human speech equally well [3]. Significant performance disparities have been documented, where speech from non-native speakers, children, older adults and individuals with speech disorders is recognized less accurately than that of healthy, non-elderly adults [1, 4, 5].

A knowledge gap thus arises from this problem: **How does this documented AI bias compare to human performance?** There is a lack of data about whether ASR systems are performing substantially worse than a human listener would or if humans exhibit similar biases and struggles to recognize diverse speech groups to a comparable degree. This gap is particularly notable for the Dutch language, where resources like the Jasmin-CGN corpus [6] exist but have not been used to benchmark ASR against human listeners.

1.1 Research Focus and Motivation

This project aims to address this gap by focusing on one particular human demographic: **teenagers (14–16 years old)**. While there has been quite some research conducted on child speech, teenagers represent a unique challenge for ASR due to for instance specific acoustic mismatches (e.g., shifting formants and pitch during puberty) [7, 8]. This is also motivated by the fact that standard ASR models are typically trained on adult speech which lead to poor recognizing of speech for this transitional age group [9].

To evaluate ASR performance, **young adults (approx. 21 years old)** will be recruited for the human experiment. This study hypothesizes that young adults are particularly adept listeners for this task due to their unique social positioning. Research by Williams and Garrett [10] suggests that young adults form a distinct group in their communications with teenagers, differing significantly from other age groups like older adults. Teenagers often seem to perceive older adults as a more distant or authoritative groups. Conversely, young adults are viewed more favorably, often sharing a closer cultural and linguistic "ingroup" status. Due to the close shared social proximity that young adults have with teenagers, the hypothesis is thus that young adults with a higher social exposure to teenagers will better navigate the specific acoustics of teenage speech than those with low exposure, and as a result, be able to better recognize the speech of teenagers.

1.2 Research Questions and Contributions

The main research questions for this project are:

1. **How well do young adults (20–24 years old) recognize the speech of teenagers (14–16 years old) compared to state-of-the-art ASR systems?**
2. **Does a young adult's exposure to teenage speech influence their listening accuracy?**

This project will provide a quantitative comparison of ASR against human performance for Dutch teenage speech. The objective criteria for success for the first question will be a comparison of the ASR System's Word Error Rate (WER) against the human listeners' transcription error rate. For the second question, the objective criteria for success will be defined by a possible identification of a statistical correlation between self-reported interaction frequency with teenagers and transcription accuracy.

2 Methodology

To investigate the comparative performance of human listeners against the ASR systems in recognizing teenage speech,

a quantitative experiment was designed. The study involved benchmarking a state-of-the-art ASR system against human participants using a specific subset of speech samples from the overall database. This section goes over the source of the speech data, the selection process for the specific speech samples, the specific ASR model used, the experimental setup for the human participants and the metrics defined for evaluation to compare the performances of the human listeners and the ASR system.

2.1 Speech Database Corpora

The speech samples for this research were obtained from the Jasmin-CGN corpus [6]. This corpus was specifically chosen due to it being able to provide vast and valuable data for diverse Dutch speaker groups, including children, teenagers, elderly people, and non-natives. For this project, the specific subset of teenagers aged 14–16 is chosen for the audio samples to benchmark performance on this particular age group.

The audio samples for the experiment were curated from the database using a strict two-step filtering process to ensure linguistic accuracy (defined here as ensuring the speech is devoid of non-native speakers) and relevance (ensuring the speakers fall strictly within the target 14–16 age demographic). The primary filter applied to the database was **native language**. To ensure the validity of the research aims, only speakers identified as native Dutch speakers were considered. Any samples from speakers listed as non-native were excluded. The secondary filter was **demographic age**, specifically targeting the window of **14 to 16 years old**. Samples from other age groups were excluded.

2.2 Data Distribution

The final dataset consisted of **40 sentences** of **Human-Machine Interaction (HMI) speech** in total, spoken by **10 unique speakers**. To ensure a balanced analysis, **4 sentences** were selected from each of the **10 unique speakers**. A perfect gender balance was achieved to prevent gender-based bias, with exactly 20 sentences selected from male speakers and 20 from female speakers. Regarding age distribution, the dataset comprises 4 samples from 14-year-olds, 28 samples from 15-year-olds, and 8 samples from 16-year-olds. As observed, the sample size is slightly skewed towards 15-year-olds (70% of the data). To prioritize the authenticity of native Dutch speech, the selection maximized the available valid entries in the database, resulting in the current age distribution.

Table 1: Sentence Length Distribution

Word Count	Frequency
4	4
5	12
6	4
7	4
8	5
9	7
10	2
11	1
12	1

Table 1 illustrates that the samples display reasonable variation in length, ranging from 4 to 12 words, with the highest frequency of sentences containing 5 words.

2.3 ASR Model

To provide a state-of-the-art benchmark, the audio samples selected were transcribed using the **Google Telephony (GT)** model (via Google Cloud Speech-to-Text). This ASR system represents a widely deployed standard in modern speech recognition and was selected for this research based on two key criteria: architectural alignment with the dataset and proven performance on diverse demographics.

2.3.1 Model Architecture and Suitability

GT is explicitly optimized for processing audio from telephony and voice applications [11]. Unlike other models, which are typically trained on read speech, GT is designed to handle **conversational speech**. This is critical for this project, as the category of speech selected for the samples is HMI speech. HMI speech shares key linguistic features with conversational speech, such as spontaneous phrasing, informal register and conversational flow, making GT the most linguistically appropriate model choice for this specific type of speech [6].

2.3.2 Performance on Diverse Groups

In a comprehensive benchmark of Dutch ASR systems, it was observed that GT achieved some of the lowest Word Error Rates (WER) across diverse speaker groups for HMI speech [12]. Most importantly for this research, the model was found to perform the best specifically for **native teenagers**, which aligns perfectly with the target demographic of this study (14–16 years old) [12].

2.4 Experiment Set-up

- **Participants:** 10 native Dutch young adults aged 20–24 years old, all male, were recruited for the experiment. The average age of the participants was $\sim$22. The participants were recruited by asking fellow students at TU Delft if they were Dutch and were willing to take part in the experiment. Once they agreed, the participants signed a consent form before beginning with the experiment. None of the participants were paid as they all agreed to do it for free.
- **Questionnaire:** Participants completed a questionnaire in **Qualtrics** to assess their social exposure to teenage speech, specifically establishing the frequency of their communication with teenagers. 6 participants reported exposure to teenage speech while 4 reported not having exposure. The questionnaire also asked the participant for their age (to ensure they fit the criteria of young adults between the ages of 20–24), their gender, their native language (to ensure they are native Dutch listeners) and whether or not they have any hearing problems. While initially participants who reported that they had hearing problems were to be removed from the analysis, only one participant reported to have hearing problems and despite that, that participant's result did not deviate significantly and thus their result was kept for the analysis.

- **Task:** The experiment was then conducted (also in **Qualtrics**) by having participants listen to the exact same subset of audio files (the 40 selected sentences) as the ASR system. The participants were presented with only one audio file at once which they could listen to only once following which they were required to provide the transcription for that particular file. The participant could then go to the next page for the next audio file and so on. After every 8 audio files, the participant was given the opportunity to take a short break to ensure they maintained their concentration and focus during the experiment and to prevent fatigue.

2.5 Evaluation and Analysis

The evaluation and analysis are carried out as follows:

- **Post-processing:** Before performing the evaluation, the transcriptions provided by the participants underwent a post-processing phase. This involved fixing obvious typos (e.g., spelling errors like 'ballonnen' spelt as 'ballonen') to ensure a fair scoring and comparison with the ASR system. Additionally, specific non-linguistic symbols and filler words were removed from the transcriptions as well to therefore compare only the actual words. The list of non linguistic symbols is as follows: *[LAUGH], [FIL], [UNK], ggg, ah, aha, ai, au, bah, boe, bwa, eih, eikes, goh, ha, haha, hé, hè, hei, ho, hu, hum, jee, mm-hu, mmm, oeh, oei, oesje, oh, o, oho, poeh, pst, sjt, sst, tut, uh, uhm, uhu, wauw, woh, zuh, zulle.*
- **Normalizing Transcriptions:** To account for the variability in written representations of spoken Dutch such as contractions, the transcriptions of both the ASR system and the humans were normalized to handle common contractions such as *'me'*, *'mijn'* and *'m'n'* which were standardized to a single stem (e.g., *mijn*). From a linguistic perspective, these variations represent the same underlying lexical item, differing only in their phonetic realization (strong vs. weak forms) or the speaker's habit, rather than a semantic difference [13]. Penalizing these variations could potentially inflate the error rates without reflecting a true failure in speech recognition. However, to ensure a comprehensive evaluation, the analysis is carried out for both the **original** and **normalized** transcriptions.
- **Word Error Rate (WER):** The primary metric for success is the Word Error Rate. WER is a standard metric used to evaluate the performance of an ASR system. It represents the ratio of errors in a transcript to the total number of words in the ground truth reference multiplied by 100%. The formula for the WER is defined as:

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

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the ground truth transcription. The ground truth transcriptions for this research are the transcriptions provided by the Jasmin database for each audio sample.

For example, consider the following sentence:

  - **Ground Truth:** *Zet de televisie aan* ($N = 4$)
  - **Transcription:** *Zet televisie nu aan*

Here, the word "de" is deleted ($D = 1$) and the word "nu" is inserted ($I = 1$). There are no substitutions ($S = 0$). Therefore:

$$WER = \frac{0 + 1 + 1}{4} \times 100\% = 50\%$$

Thus the WER of the above example sentence is 50%.

The WER is calculated for the ASR system for each audio sample. The WER is then calculated for the two groups of human participants, the group who had exposure to teenage speech and the group who did not have exposure to teenage speech.

- **Statistical Analysis:** A t-test was carried out to statistically determine against which human participants the ASR system performed significantly better and against which human participants the ASR system had a comparable performance.
- **Human vs. AI Comparison:** The average WER of the two human groups is then compared directly to the WER of the ASR system to answer the first research question.
- **Exposure Analysis:** The second analysis focuses on the influence of social exposure to teenage speech. This analysis explicitly compares the WER performance of the participant group with exposure to teenage speech against the participant group without exposure to teenage speech to answer the second research question.

3 Responsible Research

- **Participant Privacy:** Each participant's input was kept anonymous throughout the conduction of the experiment as their names were not inputted at any point thus ensuring confidentiality.
- **Participant Data Handling:** The data collected was strictly reserved for research purposes and will not be distributed to or shared with any third parties.
- **Use of Jasmin Database:** All audio samples downloaded from the Jasmin-CGN corpus for this experiment will be deleted from local storage upon the completion of the project.
- **Use of AI:** The generative AI owned by Google, Gemini, was utilized during the writing process to refine and improve the phrasing of certain sections of this research paper.

4 Results

4.1 Original Results

- **ASR Result:** The average Word Error Rate (WER) of the Google Telephony ASR system over the 40 sentences was $\sim$12.8%.
- **Human Results:** In terms of human performance, the group with high exposure to teenage speech achieved an average WER of $\sim$16.6%, whereas the group without

exposure achieved an average WER of $\sim$21.4%. Table 2 details the performance range (average, best and worst) for both groups, along with the standard deviations, highlighting the variance within the participants.

Table 2: Human Transcription Performance (Original)

Group	Avg. WER	Best WER	Worst WER	Std. Dev.
High Expos.	16.6%	13.9%	21.5%	2.88
Low Expos.	21.4%	15.0%	24.8%	4.54

- **Statistical Test:** A statistical comparison was conducted to determine if the ASR system performed significantly better than, or comparable to, individual human participants.

  As shown in Table 3, out of the 6 participants with high exposure, the ASR system was significantly better than 3 of participants, while performing comparably to the other 3. In contrast, for the low exposure group, the ASR system performed significantly better than 3 out of the 4 participants.

Table 3: Statistical Comparison: ASR vs. Individual Humans (Original)

Group	Comparison Outcome	Count	p-value
High Expos.	ASR Comparable	3	0.154
	ASR Significantly Better	3	0.049
Low Expos.	ASR Comparable	1	0.101
	ASR Significantly Better	3	0.032

4.2 Normalized Results

- **ASR Result:** After normalizing the transcriptions to account for contractions, the average WER of the ASR system dropped to $\sim$7.0%.
- **Human Results:** Normalization also improved human performance across the board. The high exposure group achieved an improved average WER of $\sim$11.6%, while the low exposure group achieved an average WER of $\sim$14.2%. Table 4 presents the detailed breakdown of the average, best and worst performances for both groups, along with the standard deviations in the normalized setting.

Table 4: Human Transcription Performance (Normalized)

Group	Avg. WER	Best WER	Worst WER	Std. Dev.
High Expos.	11.6%	7.7%	16.8%	3.63
Low Expos.	14.2%	8.0%	17.7%	4.59

- **Statistical Test:** The statistical outcomes for the normalized data mirrored those of the original data.

  As detailed in Table 5, the ASR system performed significantly better than 3 of the high-exposure participants and comparable to the remaining 3. For the low-exposure group, the ASR system again outperformed 3 participants significantly and was comparable to 1.

Table 5: Statistical Comparison: ASR vs. Individual Humans (Normalized)

Group	Comparison Outcome	Count	p-value
High Expos.	ASR Comparable	3	0.088
	ASR Significantly Better	3	0.028
Low Expos.	ASR Comparable	1	0.052
	ASR Significantly Better	3	0.044

5 Discussion

The results of this study provide a benchmark for the performance of state-of-the-art ASR compared to human listeners when transcribing native Dutch teenage speech. The findings are discussed below in terms of overall performance and the linguistic challenges posed by Dutch orthography. The research questions formulated at the beginning of this paper are answered at the end of the section.

5.1 ASR vs. Human Performance

The primary objective was to determine if current ASR technology could compete with human hearing in this specific demographic domain. The analysis indicated that the ASR system is not only comparable but also superior to many human transcribers. In the original setting, the ASR system achieved a notably lower error rate than the average performance of both the human groups.

This suggests that for native teenage speech, the ASR system has seemed to reach a level of proficiency that rivals human listeners. While human listeners rely on acoustic cues and linguistic intuition, the ASR model's vast training data appears to provide a more robust model for handling the specific acoustic characteristics of this demographic. While the statistical testing seems to support this claim, it should be noted that due to the small sample size ($N = 10$), strong claims about generalization cannot be made.

Impact of Orthographic Normalization: A additional finding of this research involved the discrepancy between phonetic realization and orthographic convention in Dutch. A qualitative analysis of the transcriptions revealed that both the ASR and human participants frequently struggled to orthographically represent contractions correctly (e.g., *'t* vs. *het*), often defaulting to the full form or alternative phonetic spellings. Normalization of these forms led to a substantial improvement in performance across all groups, with error rates dropping substantially for both the ASR and human participants. This confirms that a large portion of the measured "errors" were stylistic mismatches rather than true failures in intelligibility.

5.2 Influence of Social Exposure

The secondary objective of this research was to see the impact of "domain knowledge" on human transcription accuracy. The results support the hypothesis that social exposure to teenage speech aids transcription accuracy. When comparing the two human groups, participants with regular social exposure to teenagers outperformed those without such exposure on average. The group with exposure to teenage speech achieved better average transcription accuracy than the group

without exposure in both the original and normalized settings. This performance gap suggests that familiarity with the specific sociolect, prosody and conversational habits of adolescents provides a measurable advantage.

5.3 Answering the Research Questions

Based on the experimental results, the research questions formulated at the start of this study can be answered as follows:

- **How well do young adults (20–24 years old) recognize the speech of teenagers (14–16 years old) compared to state-of-the-art ASR systems?**
  The results show that the ASR system (Google Telephony) recognized teenage speech better than the young adult listeners. This seems to showcase that this specific ASR model appears to rival human-level recognition for this specific demographic domain.
- **Does a young adult's exposure to teenage speech influence their listening accuracy?**
  Yes. The results suggest that people who communicate with teenagers regularly appear to be better at understanding them. Participants with regular social exposure performed significantly better on average than those who rarely had communication with teenagers. This indicates that simply being familiar with how teenagers talk can appear to give a listener a helpful advantage.

6 Conclusion

This research was aimed to evaluate how well current Automatic Speech Recognition (ASR) systems can handle the specific speech patterns of native Dutch teenagers. By comparing the Google Telephony model against human transcribers on a set of 40 sentences from the JASMIN corpus [6], this study provided a quantitative analysis of machine performance for this specific demographic domain. The experiment also looked at whether being socially familiar with teenagers helps humans understand them better. The results were that the ASR system performed better on average than humans and humans with exposure to teenage speech performed better on average than humans without exposure.

6.1 Limitations and Future Work

While these findings offer valuable insights, several limitations occurred along the way that must be acknowledged. First, the demographic distribution of the speech samples was slightly skewed, with 70% of the data originating from 15-year-old speakers due to database constraints. This limited the generalization of the findings across the full 14–16 age range. Second, the human participant pool was small ($N = 10$) and unfortunately lacked gender diversity, consisting entirely of male listeners. This small sample size reduced the statistical power of the human-machine comparison and prevented broader generalizations about human performance. Additionally, the analysis revealed that orthographic inconsistencies regarding contractions (e.g., *'t* vs. *het*) inflated error rates for both humans and ASR.

Future research should prioritize validating these findings on a larger, more balanced dataset that allows for instance, a uniform distribution of age. Expanding the human control group to include a diverse range of listeners would further solidify the statistical significance of the performance benchmarks. Additionally, future work could test humans against other ASR models (e.g., OpenAI's Whisper) to see if the machine advantage observed here holds true across different types of technology or if it is specific to the model used in this study.

7 Appendices

7.1 Consent Form

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

Thank you for your participation in our research project **"Human vs. AI: How good are humans at recognizing different kinds of speech compared to state-of-the-art AI-based automatic speech recognisers?"**. This study is carried out by Student Garv Singh 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 diverse 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 by Dutch teenagers aged 14-16. 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 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. External parties will not be allowed to use the responses.

You will be invited to fill in a questionnaire before the recording. The questionnaire asks about your gender, age, native language, whether you communicate with teenagers on a regular basis and whether you have any hearing problems or not. 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)		
5. 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.	☐	☐
6. 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		
7. 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.	☐	☐
8. 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.	☐	☐
9. 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.	☐	☐

Signatures

_____ _____

Signature Date

Study contact details for further information:

Garv Singh <G.SINGH-5@student.tudelft.nl>

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

7.2 Questionnaire Instructions

7.2.1 Dutch version for Participants

Welkom en bedankt voor uw deelname!
Lees en onderteken het toestemmingsformulier voordat het experiment begint. Na ondertekening stellen we u enkele algemene vragen die nodig zijn voor de analyse van deze studie.

Dan volgt het eigenlijke experiment:

U hoort 40 korte audiofragmenten, ingesproken door tieners van 14 tot en met 16 jaar.

Belangrijke instructies voor het transcriberen

- Alle audioclips bevatten alleen echte woorden.
- Audioclips kunnen grammaticaal onjuist zijn.
- Elk fragment kan slechts één keer worden afgespeeld.
- Klanken als “uh”, “umm”, “mm”, zuchten, etc., evenals leestekens (punten, komma’s, hoofdletters) mogen worden opgeschreven, maar worden niet meegenomen in de beoordeling.
- Houd er rekening mee dat sommige woorden anders kunnen worden uitgesproken en daarom anders moeten worden getranscribeerd, zoals: "me" ≠ "m'n" ≠ "mijn". Schrijf altijd op wat u daadwerkelijk hoort.
- Als u een passage niet begrijpt, typ dan een "," (komma) in het tekstveld.
- Na elke 8 samples verschijnt er een pauzescherm. Neem gerust een pauze als dat nodig is. Om na de pauze verder te gaan met het experiment, klikt u op het pictogram 'volgende pagina' rechtsonder in het pauzebericht.

Tot slot

Werk rustig en geconcentreerd. Er is geen tijdsdruk; neemt u alle tijd die u nodig heeft.

Hartelijk dank voor uw tijd en moeite! Uw bijdrage is ongelooflijk waardevol voor dit onderzoek.

7.2.2 Translated version in English

Welcome and thank you for participating!
Before the experiment begins, please read and sign the consent form. After signing, we will ask you some general questions necessary for the analysis of this study.

Then comes the actual experiment:
You will hear 40 short audio clips spoken by teenagers aged 14-16.

Important instructions for transcribing

- All audio clips contain only real words.
- Audio clips may be grammatically incorrect.
- Each fragment can only be played once.
- Sounds such as “uh”, “umm”, “mm”, sighs, etc., as well as punctuation (full stops, commas, capital letters) may be written down, but will not be included in the assessment.
- Note that some words can be pronounced differently and therefore need to be transcribed differently, such as: "me" ≠ "m'n" ≠ "mijn". Always write down what you actually hear.
- If you did not understand a passage at all, please type a "," (comma) in the text field.
- After every eight samples, a pause screen will appear. Feel free to take a break if needed. To continue the experiment after the pause, simply click the "next page" icon in the bottom right corner of the pause message.

Finally

Please work calmly and with focus. There's no time pressure; take all the time you need.

Thank you so much for your time and effort! Your contribution is incredibly valuable to this research.

7.3 Individual Results

Table 6: Individual Participant Results (Original)

Participant ID	Age	Exposure	Hearing Problems	WER
1	21	Yes	No	14.2%
2	22	Yes	No	13.9%
3	23	No	No	24.8%
4	20	Yes	No	21.5%
5	20	No	No	21.5%
6	22	No	No	15.0%
7	20	Yes	Yes	15.0%
8	21	Yes	No	17.3%
9	22	No	No	24.4%
10	24	Yes	No	17.6&

Table 7: Individual Participant Results (Normalized)

Participant ID	Age	Exposure	Hearing Problems	WER
1	21	Yes	No	9.6%
2	22	Yes	No	7.7%
3	23	No	No	17.7%
4	20	Yes	No	16.8%
5	20	No	No	13.5%
6	22	No	No	15.0%
7	20	Yes	Yes	8.0%
8	21	Yes	No	15.2%
9	22	No	No	14.2%
10	24	Yes	No	11.2&

References

[1] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[2] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Speech technologies in healthcare: A review," *IEEE Reviews in Biomedical Engineering*, 2021.

[3] R. Raes, S. Lensink, and M. Pechenizkiy, "Everyone deserves their voice to be heard: Analyzing predictive gender bias in asr models applied to dutch speech data," *arXiv preprint arXiv:2411.09431*, 2024.

[4] R. Vipperla, S. Renals, and J. Frankel, "Ageing voices: The effect of changes in voice parameters on asr performance," in *EURASIP Journal on Audio, Speech, and Music Processing*, Springer, 2010.

[5] S. Feng, B. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 80, p. 101496, 2024.

[6] C. Cucchiarini, O. van Herwijnen, F. Smits, and L. Boves, "JASMIN-CGN: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality," in *Proc. LREC*, 2006.

[7] L. M. Arrieta-L, C. E. Viviescas-M, *et al.*, "A review of asr technologies for children's speech," *Revista Facultad de Ingeniería Universidad de Antioquia*, no. 71, pp. 147–160, 2014.

[8] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme, "Child speech recognition in human-robot interaction: evaluations and recommendations," in *Proc. of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 81–88, 2017.

[9] Y. Zhang, S. Feng, *et al.*, "Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems," in *Proc. S4SG*, pp. 6–11, 2022.

[10] A. Williams and P. Garrett, "Teenagers' perceptions of communication and 'good communication' with peers, young adults, and older adults," *Language Awareness*, vol. 21, no. 3, pp. 267–283, 2012.

[11] Google Cloud, *Google Cloud Speech-to-Text: Transcription Models*, 2025. Accessed: 2025-09-08.

[12] Y. Zhang, T. De Valck, and O. Scharenborg, "State-of-the-art speech recognition systems show bias against dutch diverse speech," *Phonetica (Under Review)*, 2025. Multimedia Computing Group, Delft University of Technology.

[13] G. Booij, *The Phonology of Dutch*. Oxford: Oxford University Press, 1995. See specifically the chapter on Cliticization and Prosodic Structure.