# HUMAN VS AI: HOW WELL DO HUMANS RECOGNIZE TEENAGE SPEECH COMPARED TO STATE OF THE ART AUTOMATIC SPEECH RECOGNITION SYSTEMS

GARV SINGH
DEPARTMENT OF COMPUTER SCIENCE, EEMCS, DELFT UNIVERSITY OF TECHNOLOGY (TU DELFT), THE NETHERLANDS

## Introduction and Motivation

- Current Automatic Speech Recognition (ASR) systems show significant bias, underperforming on "diverse" speech (e.g., non-natives, children, elderly) compared to healthy adults.
- It is currently unknown how this compares to human perception, as no comprehensive Dutch benchmarks exist for these groups.
- The aim of this project is to benchmark state-of-the-art ASR against human listeners, specifically focusing on teenage speech.
- The hypothesis is that young adults are the ideal human baseline. Due to their social positioning, those with higher exposure to teenagers should better navigate the specific acoustic mismatches (e.g. pitch) that may confuse the ASR.

## Research Questions

The two research questions for this research project are 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?**
- **Does a young adult's exposure to teenage speech influence their listening accuracy?**

## Methodology

Data Curation -

- Source - Jasmin-CGN corpus, specifically the subset of teenagers aged 14-16.
- Selection - A subset of audio files (40 sentences in total) of Human Machine Interaction (HMI) speech with the corresponding ground truth transcriptions.

ASR Evaluation -

- Each audio file was transcribed by the ASR system Google Telephony.
- The word error rate (WER) was calculated by comparing the output to the ground truth value.
- 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.

Human Listener Experiment -

- Participants - 10 native Dutch young adults with an average age of ~22, all male.
- Screening - Each participant was asked about their exposure to teenage speech (whether they communicate with teenagers on a regular basis or not). 6 participants reported exposure to teenage speech while the other 4 reported not having exposure.
- Task - The participants listened to the exact same audio files as the ASR system and provided the transcriptions for them.

Analysis -

- Post processing of transcriptions carried out to fix typos and remove non linguistic symbols and filler words.
- Transcriptions were normalized to handle common contractions like 'me', 'mijn', 'm'n', where every variation was standardized to its stem.
- Average human WER was compared to the WER of the ASR system for both the original and normalized transcriptions to answer the first research question.
- The average WERs of both groups of humans were compared with each other to see whether having exposure to teenage speech had an effect on transcription accuracy to answer the second research question.
- Statistical testing was carried out to see against which humans the ASR performed significantly better and against which humans it had a comparable performance.

## Results and Discussion

| Group | Average WER (Original) | Average WER (Normalized) |
|---|---|---|
| ASR | 12.8% | 7.0% |
| Humans with Exposure | 16.6% | 11.6% |
| Humans without Exposure | 21.4% | 14.2% |

Table 1: ASR and Human Performance

| Group | Comparison Outcome | Count | p-value (Original) | p-value (Normalized) |
|---|---|---|---|---|
| Humans with Exposure | ASR Comparable | 3 | 0.154 | 0.088 |
| | ASR Significantly Better | 3 | 0.049 | 0.028 |
| Humans without Exposure | ASR Comparable | 1 | 0.101 | 0.052 |
| | ASR Significantly Better | 3 | 0.032 | 0.044 |

Table 2: Statistical Test of ASR against Individual Humans

- Contrary to expectations, the ASR system (Google Telephony) achieved a lower WER than the average young adult listener. This suggests current models may rival human proficiency in processing the specific acoustic characteristics of teenage speech. Statistical testing seems to support this claim but due to the low sample size (N = 10), strong claims about generalization cannot be made. So the answer for the first research question is that the ASR system recognized teenage speech better than young adults.
- Human performance was significantly influenced by domain knowledge. Participants with regular social exposure to teenagers outperformed those without, confirming that familiarity with adolescent sociolect and prosody provides a measurable listening advantage. So the answer for the second research is that yes, a young adult's exposure to teenage speech positively influences their listening accuracy.
- An additional finding was that both the ASR system and humans had substantially lower WERs in the normalized setting. A major source of error for both ASR and humans in the original setting were Dutch contractions (e.g., 't vs. het). Normalizing these text-based stylistic mismatches led to substantial performance improvements across all groups, indicating that many initial "errors" were orthographic rather than intelligibility failures.

## Conclusion

- Google Telephony was benchmarked against human listeners on native Dutch teenage speech using the JASMIN corpus.
- The ASR system surprisingly outperformed the average human transcriber, suggesting high robustness to teenage acoustic patterns.
- For humans, familiarity and exposure matters. Listeners with regular exposure to teenagers performed significantly better than those without, confirming that social context aids speech perception.

## Limitations and Future Works

Limitations -

- The speech dataset was age-imbalanced, with 70% of samples coming from 15-year-olds, limiting generalization across the full teenage range.
- The human listener group was small (N = 10) and lacked gender diversity (entirely male), reducing statistical power.
- Inconsistent spelling of Dutch contractions (e.g., 't vs. het) artificially inflated the WERs for both humans and ASR, reflecting stylistic rather than intelligibility failures.

Future Works -

- Validate these findings on a larger, age-balanced dataset to ensure broader applicability.
- Expanding the participant pool to include a diverse range of listeners would solidify statistical significance.
- Benchmark human performance against other ASR systems, such as OpenAI's Whisper, to determine if the "machine advantage" holds across different ASR technologies.