Linguistic Sources of Performance Disparities in Dutch ASR[1] for Non-Native Adults

Kiarash Karimi

EEMCS, Delft University of Technology, The Netherlands
Supervisors: Odette Scharenborg and YuanYuan Zhang

1. Introduction

- **ASR systems** are used more every day and have to perform equally well for everyone.
- Recent systems reach high accuracy in some settings, but **performance disparities** remain across speaker groups.
- Previous work shows lower recognition accuracy for non-native speech, including **Dutch non-native** speech.
- These disparities are often linked to pronunciation differences, phonemic variation, and speaker background.
- Less is known about **word-level and utterance-level linguistic factors** behind these errors.
- This study focuses on two factors: **utterance length** and **word category**.
- **Read** speech and **HMI**[2] speech are compared separately, because they differ in linguistic structure and spontaneity.

Research question

To what extent do utterance length and word category affect the performance of state-of-the-art ASR systems for Dutch speech from non-native adult speakers, and how do these effects differ between read speech and HMI speech?

2. Data & ASR Models

- Speech data comes from the **JASMIN corpus**, a Dutch speech corpus with read speech and human-machine interaction dialogues.
- This study focuses on **non-native adult speakers** of Dutch.
- Two speech styles are analyzed separately:
  - **Read speech**: more controlled and scripted
  - **HMI speech**: more spontaneous interaction with a system
- The same reference transcripts are compared against outputs from two strong multilingual ASR systems:
  - **Google Chirp 2**
  - **Whisper large-v3**
- All transcripts are normalized before scoring: lowercase text, removed punctuation, and Dutch textual forms for numerals.
- Errors are analyzed as **deletions**, **substitutions**, and **insertions**.

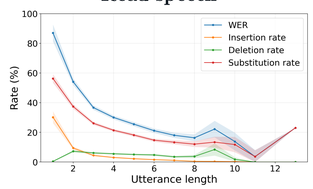
3. Experiment 1: Utterance Length

- Utterance length is measured as the word count of the reference utterance.
- WER[3] and insertion, deletion, and substitution error rates were computed:

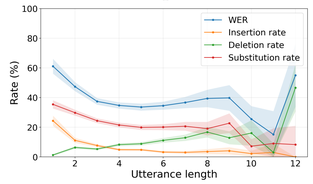
$$\text{WER} = \frac{I + D + S}{N}, \quad \text{IR} = \frac{I}{N}, \quad \text{DR} = \frac{D}{N}, \quad \text{SR} = \frac{S}{N}$$

  where I, D, and S are insertions, deletions, and substitutions, and N is the utterance length.
- Figures show **Whisper**; Google Chirp follows the same overall pattern.
- **Answer:** very short utterances are most error-prone.
- Read speech improves more consistently as utterances become longer.
- HMI speech improves at first, but WER rises from five words.
- Substitutions follow the overall WER pattern most closely.
- Deletion rate rises for HMI speech.

Read speech

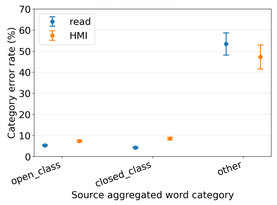


HMI speech

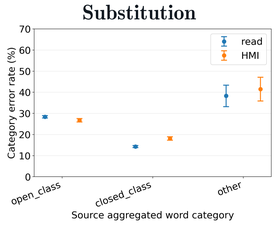


4. Experiment 2: Word Category

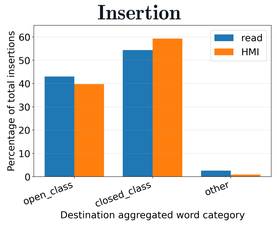
Deletion



Substitution



Insertion



- Reference and error words are PoS[4]-tagged with spaCy into nouns, verbs, determiners, ...
- PoS tags are grouped into **open-class**, **closed-class**, and **other** categories.
- Deletion and substitution counts are normalized by category frequency in the reference data, to calculate *category error rates*.
- Insertions are analyzed by the category of the inserted word.
- **Answer:** word category mainly affects substitutions and insertions.
- **Substitutions:** Open-class words are substituted more often than closed-class words; Proper nouns, nouns, adjectives, and verbs show relatively high PoS-level error rates; Many high-ranked substitutions preserve PoS category (see table).
- **Insertions:** mostly closed-class; pronouns, determiners, and verbs are inserted most often.
- **Deletions:** differences are smaller and depend more on speech style.

Top-10 substitutions preserving PoS category

PoS pair	Chirp Read	Chirp HMI	Whisper Read	Whisper HMI
NOUN → NOUN	2 (14.4)	2 (13.3)	2 (17.3)	2 (16.0)
VERB → VERB	4 (11.6)	9 (6.4)	4 (11.2)	7 (8.4)
ADJ → ADJ	6 (7.9)	5 (8.9)	5 (8.6)	4 (10.2)
PROPN → PROPN	–	3 (11.8)	8 (8.0)	8 (8.2)
PRON → PRON	–	8 (7.7)	–	6 (8.6)
DET → DET	–	10 (5.4)	–	9 (7.1)

Some of the substitutions that appear in the top 10 substitution pairs of each ASR model and speech type. Each cell shows the rank of the substitution, with the error rate in parenthesis.

5. Main Findings

- **Aggregate WER hides linguistic structure.** The same overall error rate can come from different utterance-level and word-level patterns.
- **Very short utterances are especially error-prone.** One-word utterances have the highest WER for both ASR models and both speech styles.
- **Speech style changes the effect of length.** Read speech improves more consistently as utterances become longer, while HMI speech becomes worse for longer utterances.
- **Substitutions drive much of the length effect.** Substitution rates closely follow the overall WER pattern across utterance lengths.
- **Open-class words are more vulnerable to substitutions than closed-class words.**
- **Inserted words are mostly closed-class.** Function-like words account for a large share of insertions.
- **Many substitutions preserve word category.**

6. Limitations & Future Work

Limitation	Future work
Noisy PoS tags for incomplete words and single-letter tokens	⇒ Improve preprocessing and manually validate ambiguous tokens
Low-frequency word categories give less reliable estimates	⇒ Use larger datasets or group rare categories more carefully
Utterance length and word category were analyzed separately	⇒ Test whether word-category effects change for short vs. long utterances
Different length effects in read and HMI speech remain partly unexplained	⇒ Analyze fillers, repetitions, self-corrections, syntactic complexity, and local context

Abbreviations

1. **ASR**: Automatic Speech Recognition
2. **HMI**: Human-Machine Interaction
3. **WER**: Word Error Rate
4. **PoS**: Part-of-Speech

TU Delft