How good are humans at recognizing Flemish speech compared to AI-based ASR systems

Student: Rares Popa | Responsible Professor: Odette Scharenborg
EEMCS, Technical University of Delft, The Netherlands

Background

- **Automated Speech Recognition** (or ASR) are systems that convert speech to text, used in search engines or voice assistants.
- ASR systems need to account for:
  - Age
  - Gender
  - Regional Accent
  - Pathology
- **Purpose of the project:** investigate whether human listeners can perform better than AI-based ASR systems in recognizing Flemish Dutch speech, and understand the mistakes being made
- **Main focus:** Flemish speech, since Dutch ASR systems perform the worst on it
- **Motivation** : inclusive ASR systems that perform optimally on all types of speech

Research Question

- **How do the types of recognition errors made by Dutch listeners differ from those made by state-of-the-art ASR systems when processing Flemish Dutch speech?**
- **Sub-questions:**
  - How well do ASR systems transcribe speech compared to humans for Flemish Dutch?
  - What kinds of errors (substitutions, deletions, insertions) are most frequent for ASR compared to humans?
  - How does WER (word error rate) differ between Dutch listeners and ASR systems for Flemish Dutch?

Jasmin CGN

- The **JASMIN-CGN** is an extension of the **Spoken Dutch Corpus**
- Speech samples from all regions of Netherlands & Flanders.
- **4 regions** West Flanders, East Flanders, Brabant and Limburg.
- For this research we will use the following:
  - Native children between 12 and 16 (6h 10m)
  - Native adults above 65 (5h 5m)
- **2 types** of speech:
  - Read speech
  - **Human Machine Interaction**

Methodology

Experiment consists of participants listening to 40 speech samples and transcribe what they hear. The illustration on the side was created using AI[1].

State-of-the-Art **ASR: Google Telephony** and **Conformer**

Step 1 : **Speaker Selection**
- Each region gets **10 sentences**, equal split of sentences over **Age bracket** and **Gender** where it was possible
- Priority : **Oldest Teenagers** and **Youngest Elders**

Step 2: Speech Sample Selection per Speaker
- The speech samples have been **sorted by number of words**
- Priority over **shorter** samples
- Non-linguistic symbols and words, lowered priority still considered as valid sentences

Step 3: Experimental Setup
- The experiment is conducted in a quiet room
- Sentences are of **medium length** (6-15 words)
- Speech samples audio is normalized using **FFmpeg**
- Each speech sample can be listened **2 times**
- All participants will use the same laptop and headset
- **Data collected**:
  - **Transcriptions** of speech samples
  - **Age**
  - **Gender**

Step 4: Post Processing
- Lower Case transcription-> Remove Punctuation -> Remove non-linguistic symbols -> Check spelling
- Normalization step
- Use Word Error Rate (WER) as metric to evaluate post processed human and ASR performance

Results

- ASRs achieve **lower WER** than Dutch listeners for Flemish Dutch
- Error rate varies across speaker regions
- **Error Differences**:
  - **Humans**:
    - More misspellings
    - Context-aware guesses
  - **ASR** :
    - More unrelated words
    - Occasional use of non-linguistic symbols or words
- **Error Similarities**:
  - Incomplete transcriptions
  - Struggle with “het” and “‘t”, “m’n” and “mijn”, “ik” and “‘k”

	S	D	I	WER
Conformer	6,1	6,7	0,7	13,4
Google	5,7	5,2	0,2	11
Humans	9,6	6,3	1,4	17,3

Table 1: representing number of Substitutions, Deletions and Insertions and WER

	West-FL	East-FL	Brabant	Limburg
Conformer	6,2	13,8	29,9	8,8
Google	10,8	9,4	22,6	5,2
Humans	28,3	13,23	21,27	8,15

Table 2: representing WER over Speaker regions. FL stands for Flanders

Conclusion

- ASRs achieve lower error rates compared to humans for Flemish speech
- Similar recognition errors, differences appear in handling audio uncertainty
- Future improvements: more participants, phoneme analysis

[1] OpenAI, “ChatGPT (GPT-5.1),” OpenAI, San Francisco, CA, 2025. [Online]. Available: https://chat.openai.com