Google Chirp vs. Whisper: Evaluating ASR performance on Dutch Native vs. Non-Native Teenager Speech

Author
Anish Jaggoe
a.s.h.jaggoe@student.tudelft.nl

Supervisor
YuanYuan Zhang

Responsible professor
Odette Scharenborg

01 Background Information

- State-of-the-art (SotA) ASR systems are becoming increasingly import for their use in voice assistants, search engines and medical documentation.
- ASR are improving, but do not recognise diverse speech well.
- ASR systems are prone to biases to diverse speech due to not well-represented datasets[1]

- In this research I focus on two commercial SotAASR systems, Google Chirp and Whisper by OpenAI
- Relatively affordable, accessible and large user bases
- Focus on native and non-native speakers, due to the Dutch inclusive society
- Similar research has been conducted before on Whisper and Wav2Vec[2]

02 Research Question

How well do Google Chirp and Whisper recognise speech of native Dutch teenagers compared to non-native Dutch teenagers?

1. Performance in Word Error Rate (WER) and Character Error Rate (CER) on native and non-native speech
2. Common transcription errors
3. Significant differences in performance between genders/ages

03 Data & ASR systems

- JASMIN-CGN: Speech corpus, contains speech annotated by speaker groups, gender, age, nativeness, native language, proficiency in Dutch, region, and dialect [3]. Speech from the following spaker groups will be used:
  - DT: Dutch teenagers, 12h of speech, 59 speakers (30 M, 29 F)
  - NNT: Non-native teenagers, 12h of speech, 52 speaker (25 M, 27 F)

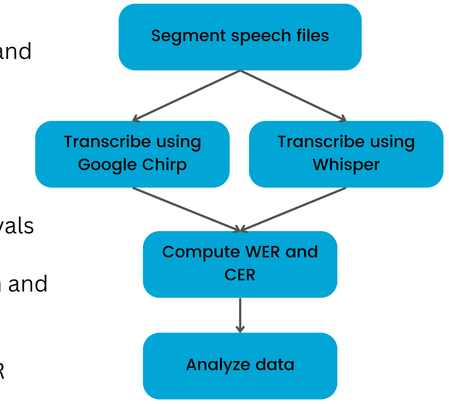
- Google Chirp: speech-to-text API developed by Google Cloud
- Whisper: open source speech-to-text model developed by OpenAI. The ‘whisper-large-v3’ model is used in this research.

04 Methodology

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

Number of (S)ubstitutions, (I)nsertions and (D)eletions, divided by (N)umber of words/characters in the reference transcription.

- Split speech into small utterance segments, based on utterance intervals in the reference text.
- 2 categories of speech: Read Speech and Human-Machine Interaction (HMI) Speech.
- Good transcription quality for a WER <10%, acceptable quality for WER 20%-30%, poor quality for WER >30%



05 Results

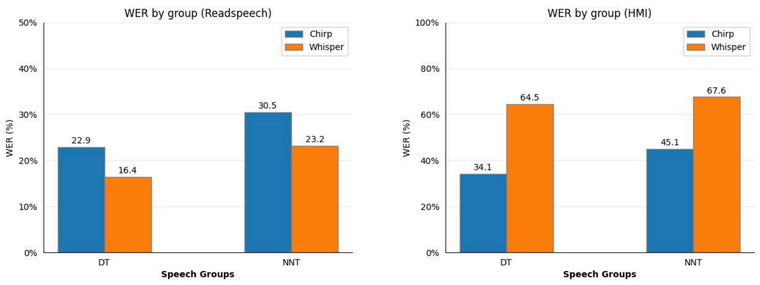


Figure 1: Average WER in percentage for Google Chirp and Whisper on Read speech (left) and HMI speech (right).Lower WER means better performance

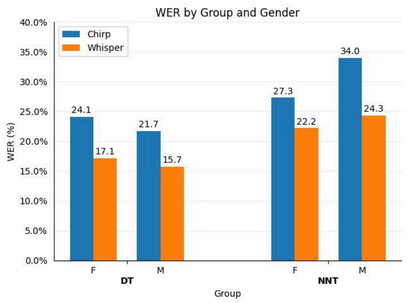


Figure 1: Average WER in percentage for Google Chirp and Whisper by gender on Read speech. Lower WER means better performance

Table 1: Average WER in percentage (%) for Google Chirp and Whisper by age of DT speaker group for Read speech

Age	Chirp	Whisper
12	31.9	**21.9**
13	30.5	**24.5**
14	32.5	**14.6**
15	15.8	**8.8**
16	**18.3**	20.1
17	62.4	**49.6**
18	27.5	**21.8**

Table 2: Average WER in percentage (%) for Google Chirp and Whisper by age of NNT speaker group for Read speech

Age	Chirp	Whisper
11	23.8	**20.3**
12	24.5	**14.1**
13	33.0	**21.9**
14	26.7	**25.2**
15	35.1	**24.1**
16	33.9	**22.0**
17	34.9	**27.4**
18	**6.2**	22.8

06 Analysis

- Whisper results on HMI speech deviates significantly from results of previous research by Fuckner et al. [2], possibly indicating some error in evaluation of Whisper HMI speech
- Google Chirp and Whisper perform better on native speech
- Better performance for native males,
- In contrast, better performance for non-native females
- There is no apparent correlation between age groups and nativeness

Table 3: Average transcription time for the speech files, in seconds, compared to average speech duration in seconds

Speech type	Avg. duration	Chirp	Whisper
RS	580.7	911.2	19.1
HMI	75.1	405.9	15.8

Table 4: Common error types in transcriptions. Reference transcriptions are compared to transcriptions by Google Chirp and Whisper

Error	Reference transcription	Chirp	Whisper
1	ravage	rafage	ravage
2	als je m	als je een	-
3	’t	het	het
4	half drie	2:30	half drie
5	uhm	-	hmm
6	herfst	härft	herfst

07 Conclusion and Future work

- Native speech gets recognized better than non-native speech by Google Chirp and Whisper
- Whisper outperformed Google Chirp on Read speech,
- Google Chirp outperformed Whisper on HMI speech, but the Whisper results indicate bad evaluation
- Both ASR systems have a form of analysis on the speech, resulting in correct but different transcriptionsm causing an increase in WER.
- Gender and age have no influence on the recognition accuracy between native and non-native speakers (on Read speech)

- Future work includes comparing more SotA ASR systems to native and non-native speech
- Phoneme Error Rate analysis on non-native speech

[1] Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021, March 28). Quantifying bias in automatic speech recognition. arXiv.org. https://arxiv.org/abs/2103.15122
[2] Fuckner, M., Horsman, S., Wiggers, P., & Janssen, I. (2023). Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers.
[3] Zhang, Y et al. (2023) Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages.

TU Delft