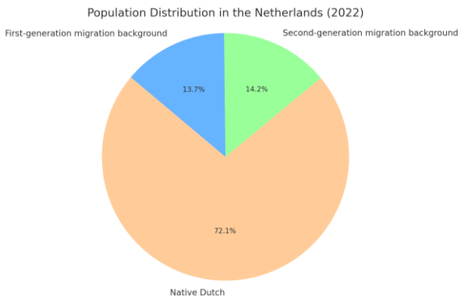
How Good Are State-of-the-Art Automatic Speech Recognition Systems in Recognizing Dutch Diverse Speech?

An Evaluation of Meta MMS and OpenAI Whisper on Native and Non-Native Dutch Speech

Author: Yiming Chen

Supervisors: Dr. Odette Scharenborg, YuanYuan Zhang MSc.

Introduction



Automatic speech recognition (ASR) systems, like those powering virtual assistants and transcription services, often exhibit performance disparities based on the speaker's demographic. This study evaluates the performance of Meta's Massively Multilingual Speech (MMS) and OpenAI's Whisper on recognizing Dutch speech from native and non-native speakers. Using metrics such as Word Error Rate (WER), Character Error Rate (CER), and Word Information Lost (WIL), the study addresses biases in ASR performance.

Research Questions

1. How accurately do the ASR systems recognize native and non-native Dutch speakers?
2. How does age affect the accuracy of the ASR systems?
3. What types of errors do each ASR system make, in terms of insertion, deletion, and substitutions, and what are the performance differences, including accuracy and execution time, between the OpenAI Whisper and Meta MMS ASR systems?

Methodology:

Dataset:

Jasmin Corpus CGN:

- Native Children **(NC)**
- Native Teenagers **(NT)**
- Native Elderly **(NE)**
- Non-Native Teenagers **(NNT)**
- Non-Native Adults **(NNA)**

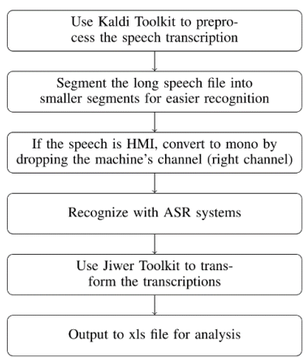
Models:

- OpenAI Whisper-large-v3
- Meta MMS-1b-all
- Meta MMS-1b-fl102

Speech Segmentation:

Speaker ID	Audio ID	Start	End	Transcription
speaker_1	audio_1	00:00	00:05	”Ik ben iKun.”
speaker_2	audio_2	00:05	00:10	”Ik ook!”

Workflow:



Performance Metrics

Types of Errors

Deletion (**D**) Insertion (**I**) Substitution (**S**)

Original Text					
I	am		13	years	old.
Recognized Text					
I	am	not	12	years.	
Type of Errors					
		insertion	substitution		deletion

Word Error Rate **(WER)** and Character Error Rate **(CER)**

$$WER = \frac{S+D+I}{N} \quad \text{and} \quad CER = \frac{S+D+I}{N}$$

N is the total number of words (for WER) or characters (for CER) in the ground truth transcription.

Word Information Lost **(WIL)**

$$Wil = 1 - \frac{H}{N}$$

where *H* is the number of correctly recognized words and *N* is the total number of words in the ground truth transcription.

Example:

Ground truth transcription	I	have	a	cat.
Recognized transcription	I	have	the	cat.

Results:

WER, CER and WIL

Group	ASR	Read speech WER	Read speech CER	Read speech WIL	HMI speech WER	HMI speech CER	HMI speech WIL
Native Children	Whisper$_{v3}$	**19.6**	9.9	**29.7**	**33.4**	**18.5**	**47.6**
	MMS$_{1b\text{-}all}$	22.8	**9.0**	35.0	43.3	21.8	58.5
	MMS$_{1b\text{-}fl102}$	27.4	11.0	41.4	49.2	23.7	66.4
Native Teenagers	Whisper$_{v3}$	**10.5**	**5.4**	**15.8**	**27.4**	**14.4**	**40.1**
	MMS$_{1b\text{-}all}$	17.3	6.5	27.3	35.4	15.8	50.9
	MMS$_{1b\text{-}fl102}$	20.0	7.6	30.9	42.1	18.7	58.7
Native Elderly	Whisper$_{v3}$	**15.2**	**7.9**	**23.4**	**34.4**	**20.9**	**47.7**
	MMS$_{1b\text{-}all}$	24.2	9.7	37.3	43.7	22.5	59.3
	MMS$_{1b\text{-}fl102}$	21.3	8.1	33.4	48.1	23.7	64.7
Native Average	Whisper$_{v3}$	**15.1**	**7.7**	**23.2**	**33.1**	**19.5**	**46.3**
	MMS$_{1b\text{-}all}$	21.5	8.5	33.4	42.5	21.4	58.1
	MMS$_{1b\text{-}fl102}$	22.9	8.9	35.4	47.5	23.1	64.2
Non Native Teenagers	Whisper$_{v3}$	**33.1**	16.4	**47.2**	**44.8**	**24.7**	**60.0**
	MMS$_{1b\text{-}all}$	37.6	**16.1**	54.7	62.5	36.4	77.7
	MMS$_{1b\text{-}fl102}$	49.3	19.7	69.8	67.8	33.9	85.0
Non Native Adults	Whisper$_{v3}$	**34.1**	**16.4**	**48.2**	**49.3**	**31.1**	**62.4**
	MMS$_{1b\text{-}all}$	42.6	18.7	61.3	70.6	47.0	83.0
	MMS$_{1b\text{-}fl102}$	49.8	20.6	70.5	71.1	42.7	85.4
Non Native Average	Whisper$_{v3}$	**33.6**	**16.4**	**47.6**	**47.1**	**28.0**	**61.5**
	MMS$_{1b\text{-}all}$	40.0	17.3	57.8	68.4	43.9	81.6
	MMS$_{1b\text{-}fl102}$	49.5	20.1	70.1	70.2	40.2	85.3
All Group Average	Whisper$_{v3}$	**21.3**	**10.6**	**31.9**	**38.1**	**22.4**	**51.9**
	MMS$_{1b\text{-}all}$	27.7	11.4	42.1	51.8	29.2	66.7
	MMS$_{1b\text{-}fl102}$	31.9	12.7	48.4	55.6	29.0	72.3

Performance and time

Group	Model	Read speech WER(%)	Read speech Time(s)	HMI speech WER(%)	HMI speech Time(s)
Native Children	Whisper$_{v3\text{-}b5}$	**19.6**	6364	**33.4**	1864
	Whisper$_{v3\text{-}b6}$	19.9	6795	34.0	1881
	Whisper$_{v3\text{-}b7}$	20.1	7558	34.2	2123
	MMS$_{1b\text{-}all}$	22.8	**886**	43.3	273
	MMS$_{1b\text{-}fl102}$	27.4	934	49.2	**213**
Native Teenagers	Whisper$_{v3\text{-}b5}$	**10.5**	4890	**27.4**	1235
	Whisper$_{v3\text{-}b6}$	10.5	5237	27.5	1267
	Whisper$_{v3\text{-}b7}$	10.6	5836	28.5	1427
	MMS$_{1b\text{-}all}$	17.3	**695**	35.4	188
	MMS$_{1b\text{-}fl102}$	20.0	769	42.1	**177**
Native Elderly	Whisper$_{v3\text{-}b5}$	**15.2**	5507	**34.4**	1389
	Whisper$_{v3\text{-}b6}$	15.4	5887	34.7	1449
	Whisper$_{v3\text{-}b7}$	15.7	6560	35.5	1638
	MMS$_{1b\text{-}all}$	24.2	**825**	43.7	**208**
	MMS$_{1b\text{-}fl102}$	21.3	1184	48.1	225
Non Native Teenagers	Whisper$_{v3\text{-}b5}$	**33.1**	5244	**44.8**	3222
	Whisper$_{v3\text{-}b6}$	33.7	5610	44.8	3444
	Whisper$_{v3\text{-}b7}$	34.0	6260	45.1	3884
	MMS$_{1b\text{-}all}$	37.6	**948**	62.5	739
	MMS$_{1b\text{-}fl102}$	49.3	1036	67.8	**715**
Non Native Adults	Whisper$_{v3\text{-}b6}$	**34.1**	6064	**49.3**	4355
	Whisper$_{v3\text{-}b6}$	35.3	6497	49.3	4649
	Whisper$_{v3\text{-}b6}$	35.7	7234	50.0	5357
	MMS$_{1b\text{-}all}$	42.6	1394	67.8	**785**
	MMS$_{1b\text{-}fl102}$	49.8	**991**	71.1	795

Result of type of errors

Group	ASR	Read speech Del	Read speech Ins	Read speech Sub	HMI speech Del	HMI speech Ins	HMI speech Sub
Native Children	Whisper$_{v3}$	**2173**	1987	**7645**	**894**	592	**2524**
	MMS$_{1b\text{-}all}$	4226	**574**	8955	2573	**100**	2621
	MMS$_{1b\text{-}fl102}$	4914	725	10864	2628	110	3280
Native Teenagers	Whisper$_{v3}$	**1611**	933	**3383**	**511**	339	**1365**
	MMS$_{1b\text{-}all}$	2721	**537**	6607	1078	**77**	1673
	MMS$_{1b\text{-}fl102}$	3136	772	7467	1328	97	1936
Native Elderly	Whisper$_{v3}$	**1742**	1592	**6027**	**3617**	2058	**7580**
	MMS$_{1b\text{-}all}$	4348	**516**	10057	8030	**337**	8560
	MMS$_{1b\text{-}fl102}$	3246	820	9111	8288	449	9866
Non Native Teenagers	Whisper$_{v3}$	**3505**	2411	**10161**	**990**	556	2471
	MMS$_{1b\text{-}all}$	6150	**406**	11701	3218	**46**	**2341**
	MMS$_{1b\text{-}fl102}$	5582	665	17662	2472	88	3517
Non Native Adults	Whisper$_{v3}$	**2738**	2508	**9204**	**3797**	1611	6026
	MMS$_{1b\text{-}all}$	5546	**391**	12148	11510	**77**	**5230**
	MMS$_{1b\text{-}fl102}$	4815	581	15748	9847	121	6974

Conclusion:

Whisper-large-v3 consistently performs better across various age groups and speech types compared to the MMS models. *MMS-1b-all* has a slight advantage in two CER metrics by a small margin, Teenager performs the best among others.

MMS models demonstrate a significant advantage in processing efficiency, with much lower time values across all groups and speech types with acceptable performance.

Whisper-large-v3 and MMS models tend to excel in different types of errors, with *Whisper-large-v3* being better at minimizing deletions and substitutions, while *MMS-1b-all* performs well in reducing insertions.