How do ASR systems of Google and Microsoft compare when recognizing Dutch spoken by native speakers over the age of 60?

Thomas de Valck - Supervisors: Odette Scharenborg, YuanYuan Zhang

Delft University of Technology

Introduction

- Automatic Speech Recognition (ASR) is everywhere, provides assistance in certain tasks
- Aging population of the Netherlands, older adults require more care
- Inevitable problems if things continue, can we make older adults live independently for longer?

- ASR systems of Google and Microsoft, two of the largest around, used by millions, publicly available
- Microsoft: Makes uses of multiple data sources, which include its own programs such as Skype and Teams

How do ASR systems of Google and Microsoft compare when recognizing Dutch spoken by native speakers over the age of 60?

Methodology

Data

- JASMIN-CGN [1], 10 hours of speech by older adults
- Human Machine Interaction (HMI) & readspeech
- 67 speakers: 23 male, 44 female
- Aged 59 to 96, average age of 79
- Several regions: North Holland (NH), Gelderland (G), Overrijssel (O) and Limburg (L)

Metrics

- Word Error Rate (WER)

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

- Word Information Lost (WIL)

$$WIL = 1 - \frac{H}{N} * \frac{H}{P} = 1 - \frac{H^2}{(H + S + D)(H + S + I)} * 100\%$$

- Character Error Rate (CER)

$$CER = \frac{S + I + D}{N} * 100\%$$

S, I, D - Number of substitutions, insertions, deletions
N - Number of words/characters in the reference solution
H - Number of 'hits', words that remained the same
P - Number of words in the resulting transcription

Experiment

- Run the data on the ASR systems of Google and Microsoft
- Calculate WER, WIL and CER
- Compare results, including specific fields like gender, region and age.

Results

Table 1. Error rates of Google and Microsoft on HMI and readspeech.

	HMI	Reading	Average
Google - WER	31.75%	22.95%	27.35%
Microsoft - WER	**25.61%**	**13.59%**	**19.60%**
Google - WIL	45.86%	34.24%	40.05%
Microsoft - WIL	**37.04%**	**21.23%**	**29.14%**
Google - CER	17.22%	13.08%	15.15%
Microsoft - CER	**13.69%**	**6.31%**	**10.00%**

Table 2. Error Rates of Google and Microsoft on Male and Female speech.

	Male	Female
Google - WER	29.70%	26.12%
Microsoft - WER	**21.60%**	**18.56%**
Google - WIL	43.28%	38.36%
Microsoft - WIL	**31.81%**	**27.74%**
Google - CER	16.19%	14.60%
Microsoft - CER	**11.05%**	**9.45%**

Table 3. Error Rates of Google and Microsoft per region.

	NH	G	O	L
Google - WER	24.44%	26.99%	23.57%	34.77%
Microsoft - WER	**17.63%**	**18.75%**	**17.39%**	**24.93%**
Google - WIL	36.00%	39.85%	35.35%	49.51%
Microsoft - WIL	**26.30%**	**28.23%**	**25.93%**	**36.50%**
Google - CER	13.53%	15.06%	12.83%	19.38%
Microsoft - CER	**8.99%**	**9.40%**	**8.83%**	**12.96%**

Table 4. Error Rates of Google and Microsoft per age group.

	60-69	70-79	80-89	90-99
Google - WER	21.88%	27.17%	27.40%	35.77%
Microsoft - WER	**15.55%**	**19.27%**	**19.57%**	**26.22%**
Google - WIL	32.69%	39.96%	40.32%	50.33%
Microsoft - WIL	**23.31%**	**28.86%**	**29.21%**	**37.77%**
Google - CER	11.54%	15.17%	15.03%	20.75%
Microsoft - CER	**7.61%**	**10.00%**	**9.71%**	**14.25%**

Discussion

Microsoft sees lower error rates compared to Google in every category, with every metric.

Gender Bias

Table 5. Relative increase in error when comparing Female to Male.

	WER	WIL	CER
Google	**13.4%**	**12.8%**	**10.9%**
Microsoft	16.4%	14.6%	16.9%

Regional Bias

Table 6. Relative increase in error when comparing Overrijssel to Limburg.

	WER	WIL	CER
Google	47.5%	**40.1%**	51.1%
Microsoft	**43.4%**	40.8%	**46.8%**

Age Bias

Table 7. Relative increase in error when comparing 90-99 years old, to 60-69 years old.

	WER	WIL	CER
Google	**63.5%**	**54.0%**	**79.8%**
Microsoft	68.6%	62.0%	87.3%

Conclusion

- Overall, Microsoft performs better than Google

- Google is less biased towards gender
- Microsoft is slightly less biased towards regions/accents
- Google is less biased towards age

- Both Google and Microsoft see significant bias on the grounds of region and age with error rates increasing by 40% to 60% from one group to another.
- The southern region (Limburg) and the oldest age group (90-99) are recognized particularly poorly.

References

[1] Catia Cucchiarini, Hugo Van hamme, Olga van Herwijnen, and Felix Smits.
JASMIN-CGN: Extension of the spoken Dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality.
In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).