Benchmarking and Research Infrastructures: Evaluating Dutch Automatic Speech Recognition

Dragoș Alexandru Bălan
University of Twente

Roeland Ordelman
University of Twente/NISV

Henk van den Heuvel
Radboud University Nijmegen

Motivation

- **Research Infrastructures (RIs)** in Arts and Humanities **provide tools**, identified via scholarly workflows, to facilitate digital research
- Assisting scholars in finding tools requires **optimizing** tool **findability**, **reuse**, and **understanding of** tool **limitations**
- An important but less provided information component: **performance**
- Our use case: **Automatic Speech Recognition (ASR)** for annotating **speech datasets and archives**
- **Goal:** create a Dutch Open Speech Recognition Benchmark

$$\mathbf{WER} = \frac{S+D+I}{N}$$

Where:

- **S** stands for substitutions,
- **I** stands for insertions,
- **D** stands for deletions,
- **N** is the number of words in the reference (that were actually said).

Figure 1: The formula for Word Error Rate (WER) [1]

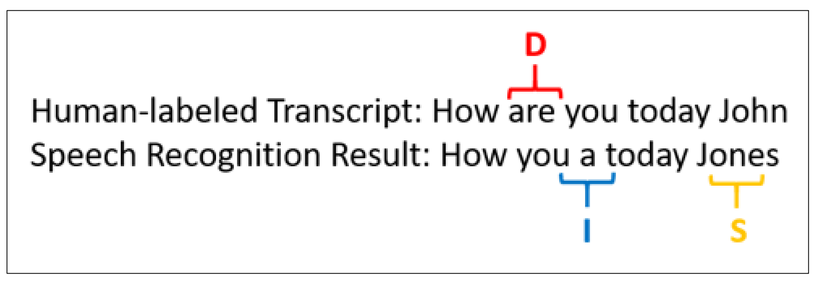

Figure 2: An example of errors that an ASR can make [2]

Benchmark setup

- 3 use cases: **Audiovisual Media (AV)**, **Oral History (OH)**, and **Medical domain conversations**
- **Datasets** for **AV** and **OH**: N-Best 2008 Eval. Corpus, JASMIN-CGN
- **Datasets** for **medical domain**: Medicijnjournaal (MJ), Medical Videos (MV), and sensitive patient-provider conversations
- **ASR systems**: **Kaldi_NL** (baseline), **Whisper**, and **wav2vec 2.0** based models
- **Metrics**: **Word Error Rate (WER)** and time to evaluate (not reported here)

Results

- **Whisper large-v2** with Voice Activity Detection (VAD) scores **the best for OH and AV data**
- **MMS**, on the other hand, performs **the worst for OH and AV data**

- For the **medical domain**, the end-to-end models (**Whisper** and **wav2vec2**) consistenly **outperform** the baseline (**Kaldi_NL**)

Model	The Netherlands		Flemish	
	Broadcast News	Conversations	Broadcast News	Conversations
Kaldi_NL	12.6%	38.6%	21.2%	59.4%
W2	10.6%	24.1%	**13.0%**	38.5%
W3	12.5%	25.5%	14.9%	38.4%
W2+VAD	**10.0%**	**23.9%**	13.6%	37.9%
W3+VAD	12.3%	25.1%	14.6%	**36.9%**
MMS	18.5%	42.7%	19.4%	57.7%

Model	The Netherlands				Flemish			
	Read		Conversational		Read		Conversational	
	N-Nat	E	N-Nat	E	N-Nat	E	N-Nat	E
Kaldi_NL	45.3%	20.9%	60.0%	44.0%	43.3%	24.7%	64.4%	47.4%
W2	30.6%	13.7%	77.7%	39.9%	21.0%	16.7%	67.3%	45.4%
W3	62.6%	27.6%	84.5%	51.4%	41.1%	38.7%	79.9%	68.3%
W2+VAD	**30.0%**	**12.8%**	**51.4%**	**26.8%**	**20.5%**	**14.4%**	**49.3%**	**30.6%**
W3+VAD	49.4%	25.2%	58.2%	33.6%	50.7%	33.6%	57.9%	44.6%
MMS	54.0%	28.3%	83.3%	59.9%	35.8%	22.3%	76.7%	60.8%

Model	MJ	MV	pat-prov_test	pat-prov_train
Kaldi_NL	16.1%	28.4%	71.2%	68.5%
Kaldi_NL fine-tuned	-	-	68.0%	-
W2	-	**10.9%**	**57.1%**	**34.1%**
wav2vec2	**12.8%**	24.2%	-	-

Table 1: WER results on N-Best dataset (top table), JASMIN-CGN dataset (middle table), and on medical domain (bottom table).

Conclusion

- **Benchmarking ASR engines** helps in **informed ASR selection and improvement**
- In the future: **expand the benchmark** with **more speech types** or **ASR configurations**
- **Together with RI providers**, discuss ways to **integrate results into tool registers**
- **Together with speech researchers,** investigate possible improvements of ASR models or optimize coverage of different speech types or topics.

Scan the QR for detailed benchmark results, along with descriptions of hardware used and data normalization steps taken

References
[1] B. Worthy. "Word Error Rate Mechanism, ASR Transcription and Challenges in Accuracy Measurement." GMR Transcription. Accessed: May 22, 2024. [Online]. Available: https://www.gmrtranscription.com/blog/word-error-rate-mechanism-asr-transcription-and-challenges-in-accuracy-measurement
[2] E. Urban and N. Mehrotra. "Test accuracy of a custom speech model." Microsoft Learn. Accessed: May 22, 2024. [Online]. Available: https://learn.microsoft.com/en-us/azure/ai-services/speech-service/how-to-custom-speech-evaluate-data?pivots=speech-studio#evaluate-word-error-rate-wer

SOUND & VISION
OH-SMArt
ORAL HISTORY – STORIES AT THE MUSEUM AROUND ARTWORKS
HoMed