Evaluating the State-of-the-Art Automatic Speech Recognition systems for Dutch

Dragoș Alexandru Bălan
University of Twente

Roeland Ordelman
University of Twente/NISV

Khiet Truong
University of Twente

1. Motivation

- **Automatic Speech Recognition (ASR)** rapidly advanced in the past few years
- Currently: **End-to-end models** that are
  - **Robust** to a wide variety of speech types and conditions
  - **Multilingual** with some even supporting 1000+ languages **[1]**
  - **Can also do other tasks**, such as **TTS**, **LID**, or **speech translation [2]**
- Some researchers compared **3 ASR systems on 1 Dutch dataset using WER** as a metric **[3]**
- **However, no large-scale systematic comparison** of state-of-the-art **ASR** models exists for **Dutch**
- **Goal: Dutch Open Speech Recognition Benchmark,** where ASR systems are evaluated in terms of **accuracy** (WER, CER, etc.) and **computational performance** (memory usage, time to transcribe, etc.) on **several datasets**

2. Benchmark setup

- **Datasets:** N-Best 2008 Eval. Corpus, JASMIN-CGN, Common Voice
- **Text normalization:** numbers -> written out, remove punctuation, lowercase
- **Audio preparation: segment** according to **reference timestamps** (N-Best) or **silence** where **speech** is **not present** (JASMIN-CGN)
- **ASR systems: Kaldi_NL** (baseline), **Whisper** (Whisper large-v2, Whisper large-v3, with(out) VAD or Voice Activity Detection), and **wav2vec 2.0** based models (**XLS-R, MMS**)
- **Metrics: Word Error Rate (WER),** time to evaluate, and memory usage (which are not reported here)
- More details on benchmark website

Word Error Rates

Table 1: **N-Best 2008** evaluation corpus.

Model	The Netherlands		Flemish	
	Broadcast News	Conversational	Broadcast News	Conversational
Kaldi_NL	12.6%	38.6%	21.2%	59.4%
Whisper large-v2	10.6%	24.1%	**13.0%**	38.5%
Whisper large-v3	12.5%	25.5%	14.9%	38.4%
Whisper large-v2 + VAD	**10.0%**	**23.9%**	13.6%	37.9%
Whisper large-v3 + VAD	12.3%	25.1%	14.6%	**36.9%**
XLS-R	14.8%	33.5%	17.0%	51.7%
MMS - 1162 languages	18.5%	42.7%	19.4%	57.7%

Table 2: **JASMIN-CGN** corpus. **N-Nat**=Non-native adults; **E**=Native Elderly

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%
Whisper large-v2	30.6%	13.7%	77.7%	39.9%	21.0%	16.7%	67.3%	45.4%
Whisper large-v3	62.6%	27.6%	84.5%	51.4%	41.1%	38.7%	79.9%	68.3%
Whisper large-v2 + VAD	**30.0%**	**12.8%**	**51.4%**	**26.8%**	**20.5%**	**14.4%**	**49.3%**	**30.6%**
Whisper large-v3 + VAD	49.4%	25.2%	58.2%	33.6%	50.7%	33.6%	57.9%	44.6%
XLS-R	36.1%	17.2%	59.1%	47.0%	26.8%	16.4%	52.2%	47.8%
MMS - 1162 languages	54.0%	28.3%	83.3%	59.9%	35.8%	22.3%	76.7%	60.8%

Table 3: **JASMIN-CGN** corpus. **Ch**=Native children (7-11 y.o.); **Teen**=Native Teenagers (12-16 y.o.).

Model	The Netherlands				Flemish			
	Read		Conversational		Read		Conversational	
	Ch	Teen	Ch	Teen	Ch	Teen	Ch	Teen
Kaldi_NL	28.1%	16.2%	55.4%	62.4%	59.2%	33.5%	66.5%	49.8%
Whisper large-v2	20.3%	11.3%	58.9%	65.8%	42.4%	11.7%	87.6%	51.7%
Whisper large-v3	28.1%	25.2%	85.8%	68.3%	57.2%	30.6%	90.5%	65.2%
Whisper large-v2 + VAD	**19.1%**	**11.1%**	**28.2%**	**22.9%**	**41.8%**	**11.6%**	**28.7%**	**24.3%**
Whisper large-v3 + VAD	27.5%	22.4%	34.4%	28.6%	56.2%	26.7%	46.0%	37.7%
XLS-R	22.4%	13.3%	60.2%	62.2%	47.4%	13.3%	73.2%	62.2%
MMS - 1162 languages	28.9%	20.0%	82.4%	87.9%	49.2%	21.8%	86.1%	68.0%

Table 4: **Common Voice 17.0** - Test Set

Model	Common Voice 17.0 - Test Set
Kaldi_NL	20.7%
Whisper large-v2	5.6%
Whisper large-v3	**4.3%**
Whisper large-v2 + VAD	5.6%
Whisper large-v3 + VAD	4.4%
XLS-R	6.5%
MMS - 1162 languages	9.5%

3. Results & Discussion

- **Best overall:** Whisper large-v2 with Voice Activity Detection (VAD)
- **Worst overall:** MMS pre-trained on 1162 languages
- **Better performance on Netherlands Dutch** than **Flemish**
- **Age bias present**: Native **teenager** speech, which is close to adult speech, performs the best, followed by **the elderly**, and lastly **children** performing the worst
- **Non-native speech worse than native speech,** which was expected since non-natives may not have native proficiency and words may be pronounced incorrectly or used in wrong context
- **Common Voice** performance is **better with newer models** than with the baseline, **Kaldi_NL**, due to the fact that Common Voice data has been used to train these models
- **VAD has huge impact on performance**, especially for Flemish conversational Dutch where **WER drops** by an absolute **~20%**
- **Linguistical variance** in training data **does not help** with performance for a medium/high-resource language like Dutch, as demonstrated by MMS' low performance

4. Conclusion

- **ASR technology rapidly advanced**, being able to outperform the previous state-of-the-art (in our case, the hybrid system of Kaldi_NL)
- **New ASR models** are also **more robust**, being able to score relatively low Word Error Rates on a variety of speech, from read to conversational, native or non-native, or for different speaker age groups.
- **However,** there is still **room for future research**, especially for **less-represented speaker** groups or **accented speech**
- **Recommendation** for the **base ASR system** to build upon: **Whisper large-v2**, preferably with Voice Activity Detection (to reduce hallucinations, as well as evaluation time in certain cases)
- **In the future:** expand benchmark with future open ASR models released or new use cases for Dutch
- **We encourage** all **researchers/developers** to **add their results** on the website where the benchmark is hosted: https://opensource-spraakherkenning-nl.github.io/ASR_NL_results

Scan the QR for more benchmark results, along with descriptions of hardware used, script locations, and data normalization steps taken (**URL** in the last bullet, to the left)

References
[1] Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., ..., & Auli, M. (2023). Scaling Speech Technology to 1,000+ Languages.
[2] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.
[3] Fuckner, M., Horsman, S., Wiggers, P., & Janssen, I. (2023). Uncovering Bias in ASR Systems: Evaluating Wav2vec2 and Whisper for Dutch speakers.

OH-SMArt
ORAL HISTORY – STORIES AT THE MUSEUM AROUND ARTWORKS

UNIVERSITY OF TWENTE.