Benchmarking and Research Infrastructures. Evaluating Dutch Automatic Speech Recognition

Benchmarking and Research Infrastructures: Evaluating Dutch Automatic Speech Recognition

Dragoș Alexandru Bălan
University of Twente
d.a.balan@utwente.nl

Khiet Truong
University of Twente
k.p.truong@utwente.nl

Henk van den Heuvel
Radboud University
henk.vandenheuvel@ru.nl

Roeland Ordelman
University of Twente
roeland.ordelman@utwente.nl

Abstract

In this paper we present the setup and results of a Dutch Open Speech Recognition Benchmark strategy, initiated in the course of use cases in three research infrastructure projects in The Netherlands, to help scholars, infrastructure providers, and speech researchers make informed choices about which speech recognition engines, configurations and models to use, to facilitate or to improve upon. The obtained results are a tangible starting point to expand collaboration and optimize the coverage of the speech types and conditions addressed in the benchmark.

1 Introduction

With the aim to support digital scholarship, the development, preservation, and provisioning of (scientific) software commonly referred to as "tools" has been a central theme for Research Infrastructures (RIs) in the Arts and Humanities for long. Generally, tools are being identified by collecting scholarly use cases or research scenarios and extracting scholarly workflows: sets of infrastructure facilities, resources, and services that together serve the requirements of humanities scholars.By collecting use cases within research disciplines, and translating these to workflows consisting of infrastructure components, RIs are able to identify which workflows need to be facilitated, and to define a catalog of tools for an RI to provide. In order to help scholars discover the tools they need for their specific use scenario, RIs setup *registers* such as the CLARIN Virtual Language Observatory[1] (VLO) and the SSHOC Marketplace[2].

However, "helping scholars find tools they need", should be interpreted as something that requires more than listing the software a community has produced in online databases. Here, perspectives on optimizing *findability* and *reuse* of tools coincide with perspectives on scholarly *training*; not only on how a tool can or should be used properly in a research workflow, but specifically also on raising awareness of the limitations of tools, sometimes referred to as *tool criticism* (Koolen et al., 2018). Findability and reuse perspectives put emphasis on the information components that are needed to enable researchers to make a proper assessment of whether a tool can be attributed to their research or not, such as a tool's Technology Readiness Level (TRL), but also the "user readiness" of tools (for example, how a tool fits in with a research methodology or the digital context a researcher is used to work in).

An information component that is frequently mentioned to be an important element, but not so frequently provided, is *performance*. Implicitly, performance can be derived from reported scientific results accomplished by deploying a tool, or how frequently it is used in a research community. Often, however, explicit information on performance levels is required and asked for. In the case of automatic speech recognition (ASR), being able to assess the quality of the data enrichment process that decodes speech in a researcher's data set (e.g., audiovisual media archives, interviews, conversations) to a textual representation, is crucial for interpreting results from search and quantitative data analysis, as performance levels in ASR can vary drastically depending on the type of speech, acoustic condition, context (e.g., domain vocabulary), and characteristics of the speakers.

[1] https://vlo.clarin.eu/

[2] https://marketplace.sshopencloud.eu/

In providing ASR performance information, the different speech types and/or conditions that could be encountered in research data sets is one aspect, the other is the range of ASR engines, configurations and models that are available, especially since these are subject of rapid changes due to ongoing advances in the field of machine learning. Scholars want to know whether applying automatic speech recognition to their data would be helpful or not, and if so, which engine they should turn to. If performance measurements were below expectations, the scholar could decide to invest in the manual annotation (Gref et al., 2022) of (parts of) the data set, instead of wasting time making sense out of noisy data. But also for a research infrastructure supplier that would facilitate a (possibly large-scale) data enrichment workflow (e.g., data ingest, computational resources), performance levels are important to decide on the ASR tool(s) it should provide in its infrastructure or not. Finally, for researchers interested in improving available ASR (models), performance statistics would help to determine whether an improvement strategy would be feasible, for example given available data sets that can be used for training.

In the context of these stakeholders in need of ASR performance information, we decided to develop a *collaborative benchmarking strategy* that could feed into the general concept of tool criticism in the context of research infrastructures.

2 Experimental Setup

Benchmarking is a widely used practice to critically and quantitatively assess the performance of various (AI) models or algorithms on a specific topic or dataset. Research groups, both in humanities and computer science, are interested in ASR performance on speech types, depending on their specific research focus. Our research aims to create a Dutch Open Speech Recognition Benchmark initiative, providing a matrix of ASR performances relative to speech types. In this paper, we discuss the essential parts of the setup and results and refer to the official benchmark website[3] for more detailed information. The benchmark was initiated in the context of Dutch research infrastructure projects focusing on research use cases on specific speech types: audiovisual media (Ordelman et al., 2018), Oral History (OH-SMArt[4]) and conversations in the medical domain (Tejedor García et al., 2022).

Benchmark data: As a reference or baseline dataset, we choose the N-Best 2008 evaluation corpus (Van Leeuwen et al., 2009). N-Best contains broadcast news speech and telephone conversations, in Netherlands-Dutch and Flemish, and is reasonably representative for the data researchers encounter in an audiovisual media archive. This dataset was also used for the evaluation of the ASR system that was initially provided a number of years ago in the CLARIAH research infrastructures (Ordelman & van Hessen, 2018). We also evaluated the JASMIN-CGN corpus, an extension of the Spoken Dutch Corpus (CGN) that contains speech from native Dutch/Belgian children, the elderly, as well as non-native speakers. In this paper, the results of ASR on native elderly and non-native adult speech are reported, which occur often in Oral History. As for the speech in the medical domain, 3 datasets have been used: Medicijnjournaal (MJ) (Tejedor García & van der Molen, 2022), Medical Video (MV) material, and sensitive patient-provider conversations. For all datasets, the average performance is reported.

Data Preparation: Standard normalization procedures have been applied such as converting numbers into words, removing punctuation, and removing case distinctions. All normalization steps were stored on the benchmark website as a reference for collaborators and to support fair comparisons between datasets. As for the audio data, they have been either segmented according to the timestamps present in the reference files (in the case of N-Best) or silenced in the regions where no speech is present (in the case of JASMIN-CGN). As for the medical domain, MJ data has been manually annotated according to a protocol detailed in Tejedor García et al. (2022).

ASR Systems and Configurations: The Kaldi_NL system that has been used most frequently in Dutch RIs, is regarded as our baseline system. Kaldi_NL is a collection of DNN-HMM ASR models with speaker diarization developed using the Kaldi toolkit (Povey et al., 2011). Due to the rapid advancements in the field of ASR, it is important to compare its performance with newer systems to see if and for which types of data it should be replaced or updated. For the medical domain data, a fine-tuned Kaldi_NL was

[3] https://opensource-spraakherkenning-nl.github.io/ASR_NL_results/

[4] https://www.uva.nl/en/discipline/conservation-and-restoration/research/research-projects/oh-smart/oh-smart.html

made available, and trained on in-domain data to improve performance.

Given its current popularity and the frequent questions from researchers we receive about its performance, OpenAI's Whisper (Radford et al., 2022) was considered an important ASR system to evaluate. It is a multilingual ASR model that has become popular over the last year due to its high performance across several languages, without additional model training (fine-tuning) needed. The results show the performance of the 'large-v2' and 'large-v3' pre-trained models. Each one is either combined with Voice Activity Detection (VAD) or not. For the medical domain data, only Whisper 'large-v2' without VAD has been evaluated at the moment.

Also, we report on the Massively Multilingual Speech (MMS) engine from Meta AI (Pratap et al., 2023), the most recent development that uses wav2vec 2.0 as the underlying architecture. The version evaluated has been trained on more than 1000 languages. The medical domain data is evaluated instead using a version of wav2vec 2.0 fine-tuned on Dutch (Grosman, 2021).

Metrics: In this paper, only the commonly used Word Error Rate (WER) metric is reported, calculated as the sum of error words output by the model divided by the number of words in the reference text. The lower the metric is, the better the performance. Evaluation time has also been measured for N-Best and JASMIN-CGN, which can be found on the official benchmark website.

3 Results & Discussion

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%**
MMS - 1162 languages	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%
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%
MMS - 1162 languages	54.0%	28.3%	83.3%	59.9%	35.8%	22.3%	76.7%	60.8%

Model	**Medicijnjournaal**	**Medical Videos**	**pat-prov_test**	**pat-prov_train**
Kaldi_NL	16.1%	28.4%	71.2%	68.5%
Kaldi_NL fine-tuned	-	-	68.0%	-
Whisper large-v2	-	**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). **N-Nat**=Non-native; **E**=Elderly; **pat-prov**=patient-provider conversations.

As the benchmark is a collaborative initiative, the performance versus dataset matrix is *sparse*: not all ASR engines/configurations could be tested on all available data sets. The results can be found in table 1.

Overall, for the Oral History domain, Whisper demonstrates robustness on various styles of speech, different categories of speakers, and on both Flemish and Netherlands Dutch. In contrast, MMS performs the worst overall, indicating a lack in balancing Dutch training material. For the medical domain, both end-to-end models (Whisper and wav2vec2) outperform Kaldi_NL, emphasizing that end-to-end models manage to improve upon the previous state-of-the-art for the medical domain annotation task.

4 Conclusion

Benchmarking ASR engines, configurations and models, and providing evaluation results online such as we are currently doing on GitHub, helps scholars, infrastructure providers and speech researchers in making informed choices about which ASR to use, to facilitate or to improve upon. In order to maximize its potential, first of all, we aim to expand collaborations in the field in terms of providing annotated data sets for speech types that are not evaluated yet, and optionally, running evaluations with new or alternative ASR configurations or models. Especially the provisioning of (small amounts of) annotated data in combination with a semi-automatic evaluation procedure, could be an approach for researchers to obtain performance measures for their data set relatively quickly. Secondly, together with RI providers we will investigate methodologies to incorporate benchmark results into tool registers in a structural, replicable and transparent manner. Finally, together with speech researchers we will investigate how we could optimize the coverage of evaluated speech types and conditions in our benchmark, and for which (typically less common) speech it is required to improve on results provided by current systems and models.

Through this initiative, we encourage researchers and developers of Dutch and Flemish ASR to collaborate by contributing to the Dutch Open Speech Recognition Benchmark with their results. Additionally, we would welcome collaborations with similar initiatives for other languages on a more European level, within the CLARIN context, as well as beyond.

Acknowledgments

This work has been done as part of the CLARIAH-PLUS project funded by NWO (grant 184.034.023), as well as part of *PDI-SSH* projects ”Oral History – Stories at the Museum around Artworks (*OH-SMArt*)” and ”Homo Medicinalis. Recognising the voices of patients to retrace the opinions about medicines (*HoMed*)”.

References

Gref, M., Matthiesen, N., Schmidt, C., Behnke, S., & Köhler, J. (2022). Human and automatic speech recognition performance on german oral history interviews. *arXiv preprint arXiv:2201.06841.*

Grosman, J. (2021). Fine-tuned XLSR-53 large model for speech recognition in Dutch.

Koolen, M., van Gorp, J., & van Ossenbruggen, J. (2018). Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*, *34*(2), 368–385. https://doi.org/10.1093/llc/fqy048

Ordelman, R., Melgar, L., Van Gorp, J., Noordegraaf, J., et al. (2018). Media suite: Unlocking audiovisual archives for mixed media scholarly research. *Selected papers from the CLARIN Annual Conference*, *159*, 133–143.

Ordelman, R., & van Hessen, A. J. (2018). Speech recognition and scholarly research: Usability and sustainability. *CLARIN 2018 Annual Conference*, 163–168.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). The kaldi speech recognition toolkit [IEEE Catalog No.: CFP11SRW-USB]. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., ..., & Auli, M. (2023). Scaling Speech Technology to 1,000+ Languages.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision.

Tejedor García, C., & van der Molen, B. (2022). *Homed Transcriptions Medicijnjournaal* (Version 1). Radboud University. https://doi.org/10.34973/dpjc-0v85

Tejedor García, C., van der Molen, B., van den Heuvel, H., van Hessen, A., & Pieters, T. (2022). Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 1032–1039.

Van Leeuwen, D., Kessens, J., Sanders, E., & van den Heuvel, H. (2009). Results of the N-Best 2008 Dutch Speech Recognition Evaluation, 2571–2574. https://doi.org/10.21437/Interspeech.2009-677