Michał Junczyk
Poznań, 11.06.2024r.

Application of speech datasets management methods for the evaluation of Automatic Speech Recognition systems for Polish.

The aim of this dissertation was to improve the utility of available speech datasets and use them to evaluate the quality of ASR systems for the Polish language. The first task was to address issues related to the availability and interoperability of these datasets. To achieve this, a comprehensive catalog of Polish ASR speech data was created, encompassing 53 datasets described with 66 attributes. This catalog enabled the identification of speech datasets available under open licenses. The selected datasets were organized to facilitate their use by practitioners developing ASR systems. The curated datasets were named "BIGOS (Benchmark Intended Grouping of Open Speech)" and "*PELCRA for BIGOS*". PELCRA refers to a research group from the University of Łódź that agreed to provide their datasets for this study and the open challenge for the community. Together, the datasets contain over 800 hours and nearly 400,000 recordings from 5,000 speakers. Selected recordings from both curated datasets were used to evaluate the quality of 25 ASR models across 24 subsets with diverse characteristics, marking the largest study of this type for the Polish language. To increase the reliability of the analyses and facilitate result replication, a system for conducting tests and analyzing results was created and made available. The study also aimed to promote standard quality assessment methods for Polish ASR systems. This was realized by publicly sharing research results and organizing an open competition within the PolEval program.

The structure of the dissertation is described below.

Chapter 1 introduces the research goals and questions. It provides background on the problem, emphasizing the significance of speech datasets in the development and quality assessment of ASR systems. The chapter also outlines the challenges in managing datasets and testing ASR systems for the Polish language. The defined research objectives include:

- improving the availability and interoperability of existing Polish speech datasets
- preparing comprehensive and user-friendly datasets using publicly available resources under open licenses
- developing systematic quality assessment procedures (benchmarking)
- promoting the curated datasets and tools within the Polish ASR community.

Chapter 2 presents a detailed literature review, discussing previous work on the quality assessment of ASR systems and the management of speech datasets. It describes the challenges associated with managing and assessing the quality of ASR systems. The chapter also analyzes methods and tools for data management and quality testing used in international research and industrial environments. To further justify the need for standardizing ASR quality assessment methods for the Polish language, examples of good testing practices from other fields of machine learning are provided. The chapter also addresses common challenges in

machine learning quality assessment procedures, laying the foundation for understanding gaps and opportunities in managing and evaluating Polish ASR systems.

Chapter 3 describes the research methodology used in the study. It first details the data collection process and the attributes used to classify existing Polish datasets. The chapter then outlines the design assumptions for the organized dataset, the original sources of recordings, transcriptions, and metadata, as well as the organization process. It also describes the methodology for reviewing publicly reported ASR system comparisons for the Polish language. The chapter includes a description of the tools created for quality assessment, including definitions of evaluation metrics, protocols, and descriptions of the tested systems. Finally, it discusses the organization of the open competition aimed at promoting and encouraging the community to use the prepared datasets.

Chapter 4 starts with presenting the results of the review of Polish ASR speech datasets. It then provides a detailed description of the curated datasets using a set of standard metrics. These metrics, inspired by practices used in leading industrial centers dealing with speech recognition, represent dataset features such as recording length, speech rate, and the number of unique words. Metrics were obtained using tools developed in the course of the research. The chapter then presents the results of the review concerning comparative tests to assess the state of knowledge about ASR technology capabilities and testing procedures. The next section presents the results of ASR quality assessments for various test scenarios, including comparisons of system quality across all datasets, specific subsets, open and commercial systems, in relation to recording length, and for different socio-demographic groups. The analysis demonstrates the usefulness of the curated datasets and the effectiveness of the designed system in revealing differences between ASR systems based on the characteristics of the tested models and test data.

Chapter 5 discusses the research results. The strengths and weaknesses of the assessed systems are interpreted considering the knowledge of the test dataset characteristics. The discussion also covers the potential impact of the proposed quality assessment procedures on practices in science and industry. Potential improvements to datasets and quality assessment practices for future consideration are also discussed. The chapter concludes by emphasizing the importance of standardizing assessment methods and community involvement in the further development of ASR technology in Poland.

This dissertation makes several key contributions to the field of Polish ASR:

- A review and catalog of Polish ASR speech datasets to enhance their availability and utility.
- Increased accessibility of original datasets through a well-organized catalog, acknowledging the original authors.
- Organization of datasets for ASR quality assessment (BIGOS and PELCRA) in a convenient format.
- Survey of publicly reported benchmarks of Polish ASR systems.
- Creation of a system for systematic ASR quality assessment.

- Organization of an open challenge for the Polish ASR community.
- Improved documentation and analysis practices for a better understanding of ASR system quality and the datasets used for evaluation.