Faculty of Mathematics and Computer Science

Adam Mickiewicz University, Poznań

Michał Junczyk, MSc

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

PhD thesis

Supervisor:

prof. dr hab. Krzysztof Jassem

Discipline of science:

Computer and information sciences

Field of science:

Natural sciences

Wydział Matematyki i Informatyki

Uniwersytet im. Adama Mickiewicza w Poznaniu

mgr inż. Michał Junczyk

Zastosowanie metod zarządzania zbiorami nagrań mowy do oceny jakości systemów automatycznego rozpoznawania mowy dla języka polskiego.

Rozprawa doktorska

Promotor:

prof. dr hab. Krzysztof Jassem

Dyscyplina naukowa:

Informatyka

Dziedzina nauki:

Nauki ścisłe i przyrodnicze

Declaration

I, **Michał Junczyk**, declare that the work in this dissertation titled *"Application of speech data sets management methods for the evaluation of Automatic Speech Recognition systems for Polish."* is carried out by me. This work has not been submitted to Adam Mickiewicz University or any other educational institution for the award of a degree or educational qualification. The information published in this dissertation has been obtained and presented in accordance with academic rules and ethical conduct. Information obtained from other sources has been referenced appropriately.

Dedication

I am grateful to many who supported me throughout this PhD journey. I greatly appreciate the unwavering support, insightful feedback, and patient guidance of my supervisor prof. dr hab. Krzysztof Jassem.

I thank the leadership and staff of the Department of Mathematics and Computer Science and the Doctoral School of Exact Sciences for a supportive and stimulating environment.

I am grateful for the feedback and support from my mentors and colleagues at Samsung and Allegro, especially dr Mikołaj Wypych, dr inż. Bartosz Broda, dr inż. Marcin Sowański, mgr inż. Ireneusz Gawlik, mgr inż. Robert Mroczkowski, dr Aleksander Wawer, dr inż. Paweł Zawistowski and last, but not least, mgr inż. Paweł Cyrta.

A heartfelt thank you to my parents for their lifelong support and belief in me. Lastly, to my beloved wife and children: your patience, sacrifices, and unwavering support made this journey possible.

Abstract

Automatic speech recognition (ASR) systems transform spoken language into written text, enabling virtual assistants, transcription tools and intelligent home control. These systems rely on large and diverse speech data sets that reflect the linguistic and acoustic characteristics of the target population and user group. The Polish language, spoken by over 50 million people, presents ASR with unique challenges and opportunities due to its rich phonetic and morphological structure.

Public domain speech datasets are often underutilized due to discoverability and interoperability issues. Limited access to evaluation datasets makes it difficult to verify and replicate the quality tests of ASR systems. Comprehensive assessment of multiple ASR systems requires an efficient data management structure. This study addresses these issues by creating comprehensive, accessible, and actively maintained datasets, promoting best practices in ASR benchmarking inspired by international standards.

The study examined and cataloged 53 publicly available speech datasets, organized the dataset from 24 sources, and developed a quality assessment process for ASR systems. The curated dataset includes nearly 400,000 recordings and over 800 hours of speech from 5,000 speakers. Selected recordings were used to compare 7 ASR systems and 25 models. The research revealed significant differences in the performance of ASR systems in various test scenarios. All resources and results have been made publicly available to promote transparency, peer review, and collaboration within the research community.

This study improved methods for data management and benchmarking of ASR systems. The comprehensive review and catalog increased the discoverability of Polish ASR speech datasets, and the curated BIGOS and PELCRA datasets provided an extensive resource of diverse speech recordings. The use of Polish ASR datasets for comparative purposes has increased threefold compared to previous studies. Improved documentation and analysis

of the understanding of the test data and the availability of data sets and assessment tools will positively impact the ability to validate and compare test results. The development of a data management methodology and a benchmarking system has improved reliable assessments and comparative analyzes of ASR systems and understanding of the strengths and weaknesses of ASR systems for Polish.

To sum up, the conducted research has a positive impact on the practical usefulness of Polish ASR datasets for academic and industrial applications. They also contribute to the promotion of methods, tools, and good practices used for the benchmarking of ASR systems.

Contents

Glossary **xix**

1 Introduction **1**

1.1 Problem background . . . 3

1.1.1 The role of datasets in the training and evaluation of machine learning systems . . . 3

1.1.2 The role of speech datasets in the training and evaluation of ASR systems . . . 4

1.1.3 Challenges in ASR speech dataset management . . . 5

1.1.4 Challenges in ASR evaluation . . . 7

1.1.5 State of the ASR speech datasets and ASR evaluation for Polish . . 9

1.2 Research aim . . . 10

1.3 Research hypothesis . . . 10

1.4 Research objectives and questions . . . 10

1.5 Research scope . . . 12

1.6 Limitations . . . 13

1.7 Methodology adopted . . . 14

1.8 Contributions . . . 15

2 Literature Review **17**

2.1 Introduction . . . 17

2.2 Benchmarking of Machine Learning Systems . . . 17

2.2.1 Challenges in ML benchmarking . . . 17

2.2.2 Examples of methods for curating ML benchmarking datasets 19

2.3 Benchmarking of Automatic Speech Recognition Systems . . . 24

2.3.1 Introduction 24

2.3.2 Overview of ASR benchmark design considerations 25

2.3.3 ASR use scenarios 26

2.3.4 Technical challenges 27

2.3.5 Performance metrics 31

2.3.6 Evaluation results analysis 41

2.4 ASR speech datasets management methods and tools 44

2.4.1 Introduction 44

2.4.2 ASR speech dataset lifecycle 45

2.4.3 Overview of the ASR dataset management methods 46

2.4.4 Challenges related to ASR speech datasets management 48

2.5 ASR speech datasets and benchmarks for Polish 50

2.5.1 ASR speech datasets for Polish 50

2.5.2 ASR speech benchmarks for Polish 50

2.6 Overview of tools for dataset management and ASR evaluation 52

2.6.1 ASR speech datasets management tools 52

2.6.2 ASR evaluation tools 53

3 Methodology 56

3.1 Overview 56

3.2 RO1: Survey of ASR speech datasets for Polish 56

3.2.1 Research objectives and questions 56

3.2.2 Research methodology 57

3.3 RO2: Design and curation of ASR benchmark dataset for Polish 61

3.3.1 Research objectives and questions 61

3.3.2 Research methodology 61

3.3.3 Dataset analysis process 75

3.3.4 Dataset release 75

3.4 RO3: Survey of ASR benchmarks for Polish 77

3.4.1 Research objectives and questions 77

3.4.2 Research methodology 77

3.5 RO4: Design and implementation of a system for ASR systems benchmarking 80

3.5.1 Research objectives and questions 80
3.5.2 Research methodology 80
3.6 RO5: Use of curated dataset for benchmarking ASR systems for Polish . . . 85
3.6.1 Research objectives and questions 85
3.6.2 Research methodology 85
3.7 RO6: Organization of competition for the ASR community 93
3.7.1 Research objectives and questions 93
3.7.2 Research methodology 93
3.8 Summary 93
3.8.1 Overview of the data management framework 93

4 Results 96

4.1 RO1: Survey of ASR speech datasets for Polish 96
4.1.1 Introduction 96
4.1.2 ASR speech datasets survey results overview 96
4.1.3 ASR speech data survey results 97
4.1.4 Survey availability 103
4.2 RO2: Design and curation of ASR benchmark dataset for Polish 104
4.2.1 Introduction 104
4.2.2 Datasets features derived from the documentation 105
4.2.3 Datasets features derived from the analysis of datasets contents . . . 108
4.2.4 Availability of curated datasets 114
4.3 RO3: Survey of ASR benchmarks for Polish 115
4.3.1 Introduction 115
4.3.2 Results 116
4.4 RO5: Use of curated dataset for benchmarking ASR systems for Polish . . . 126
4.4.1 Introduction 126
4.4.2 Evaluation setups 126
4.4.3 Evaluation scenarios 127
4.4.4 Reference and ASR Transcripts Normalization 145
4.4.5 Evaluation results sharing 147
4.5 RO6: Organization of open competition for the ASR community 148

4.5.1 Introduction . . . 148
4.5.2 Program selection and task creation . . . 149
4.5.3 Comparison of community ASR solutions with other systems for Polish149

5 Discussion 151

5.1 Overview . . . 151
5.2 RO1: Survey of ASR speech datasets for Polish . . . 152
5.2.1 Results overview . . . 152
5.2.2 Observations from community feedback . . . 155
5.2.3 Implications . . . 157
5.2.4 Limitations . . . 157
5.2.5 Future research directions . . . 158
5.3 RO2: Design and curation of ASR benchmark dataset for Polish . . . 160
5.3.1 Results overview . . . 160
5.3.2 Observations . . . 163
5.3.3 Implications . . . 164
5.4 RO3: Survey of ASR benchmarks for Polish . . . 166
5.4.1 Results overview . . . 166
5.4.2 Implications . . . 168
5.5 RO4: Design and implementation of system for ASR systems benchmarking 169
5.5.1 Results overview . . . 169
5.6 RO5: Using a curated dataset to benchmark ASR systems for Polish . . . 170
5.6.1 Results overview . . . 170
5.6.2 Observations . . . 186
5.6.3 Implications . . . 186
5.6.4 Methodological gaps in ASR benchmarking addressed in this study . 187
5.6.5 Limitations and future work . . . 188
5.7 **RO6: Organization of competition for the ASR community** . . . 190

6 Conclusion 191

6.1 Main Research Questions and Answers . . . 191
6.2 Contributions and achievements . . . 194

6.3 Future Directions 195

6.4 Research Impact 195

7 Appendix 198

7.1 ASR speech datasets survey 198

7.1.1 Attributes of speech datasets catalog 198

7.1.2 Attributes of ASR benchmarks survey 203

7.1.3 Freely available speech datasets for Polish ASR 206

7.1.4 Commercially available speech datasets for Polish ASR 207

7.1.5 Dataset subsets cards 207

7.1.6 Commercial ASR systems pricing 223

7.1.7 Freely available ASR models sizes 223

7.1.8 Call for participation in 2024 Polish ASR challenge 224

List of Figures

2.1 Internal and external issues identified in the ML evaluation practices. Source: [72] 18

2.2 ML tasks and learning problems universe. Source: [72] 19

2.3 Examples of WER sliced into groups A, B, and C, with the width of the bars reflecting relative sizes of those groups. Source: [2] 43

2.4 ASR speech dataset lifecycle 46

3.1 Overall research framework 57

3.2 Process of analysis of curated datasets. 76

3.3 ASR evaluation process 82

3.4 ASR evaluation process data flow 82

3.5 BIGOS data management framework 94

4.1 Normalized cumulative size of Polish ASR speech datasets 98

4.2 ASR benchmark results — POLSL dataset. Source: [99]. 119

4.3 ASR benchmark results — BOR dataset scenario 1, year 2018. Source: [99] 120

4.4 ASR benchmark results — PolEval year 2019. Source: [59] 121

4.5 ASR benchmark results — DiaBiz corpus, year 2022. Source: [110] 122

4.6 ASR benchmark results — Whisper, MLS corpus, year 2022. Source: [117] . 122

4.7 ASR benchmark results — Whisper, CommonVoice corpus, year 2022. Source: [117] 122

4.8 ASR benchmark results — Whisper, VoxPopuli corpus, year 2022. Source: [117] 123

4.9 ASR benchmark results — Whisper, FLEURS corpus, year 2022. Source: [117] 123

4.10 ASR evaluation results — SpokesBiz corpus, year 2023. Source: [111] 123

4.11 ASR benchmark results — Accuracy of medical terms recognition, Kuligowska et. al, year 2023. Source: [68] 125

4.12 ASR benchmark results — Recognition errors classification, Kuligowska et. al, year 2023. Source: [68] 125

4.13 ASR benchmark results — Medical terms recognition, Zielonka et. al, year 2023. Source: [153] 125

4.14 Mean WER per system for all BIGOS dataset subsets. 129

4.15 Mean WER for all PELCRA dataset subsets. 129

4.16 Box plot of WER for all systems per specific subset of BIGOS dataset. . . . 131

4.17 Box plot of WER for all systems per specific subset of PELCRA dataset. . . 132

4.18 Comparison of WER for the most accurate free and paid ASR systems. BIGOS dataset. 134

4.19 Comparison of WER for the least accurate free and paid ASR systems. BIGOS dataset. 135

4.20 Comparison of WER for the most accurate free and paid ASR systems. PELCRA dataset. 136

4.21 Comparison of WER for the least accurate free and paid ASR systems. PELCRA dataset. 137

4.22 WER for freely available systems for various model sizes. BIGOS dataset. . 138

4.23 WER for freely available systems for various model sizes. PELCRA dataset. 139

4.24 Average WER in function of audio duration. BIGOS dataset. 140

4.25 Mean WER in function of audio duration. PELCRA dataset. Best paid and free systems. The size of the point corresponds to the number of samples. . 141

4.26 Mean WER in function of speech rate for top systems. BIGOS dataset. . . 142

4.27 Mean WER in function of speech rate for top systems. PELCRA dataset. . 142

4.28 Standard deviation in WER across speaker age groups. PELCRA dataset. . 144

4.29 Impact of normalization on error rates on BIGOS dataset. 145

4.30 Impact of normalization on error rates on PELCRA dataset. 146

4.31 Management framework extension to incorporate results from PolEval open challenge. 150

5.1 WER scores of top systems in Polish ASR benchmarks. 167
5.2 Number of datasets and vocabulary domains in Polish ASR benchmarks. . . 188
5.3 Number of evaluated models in Polish ASR benchmarks. 189
5.4 Number of dataset-system-model combinations in Polish ASR benchmarks . 189

List of Tables

2.1 ASR use scenarios overview. Inspired by work of Aksenova et al.[2] 28

2.2 Types of sources, recipients, and modes for various ASR use scenarios. Inspired by work of Aksenova et al.[2] 28

2.3 Vertical aspects of ASR challenges. Inspired by: [2] 29

2.4 Practical challenges of ASR evaluation process 31

2.5 Metrics used for ASR evaluation 32

2.6 Differences between WER, MER and WIL values for different input/output combinations. Source: [90] 36

2.7 Evaluation results analysis dimensions 44

2.8 Stages and methods of speech data management 46

2.9 Tools for ASR datasets management 53

2.10 Tools for ASR evaluation 55

3.1 Attributes of datasets and their relevance to ASR evaluation 64

3.2 Sample of PELCRA SNUV references and ASR outputs 70

3.3 Overview of factors enhancing specific dataset utility for ASR evaluation purposes. 71

3.4 Overview of factors decreasing datasets' utility for ASR evaluation purposes. 71

3.5 Meta-data and partitioning of source datasets — BIGOS dataset 72

3.6 Meta-data and partitioning of source datasets — PELCRA dataset 73

3.7 Meta-data and partitioning of source datasets 73

3.8 Meta-data and partitioning of source datasets 74

3.9 Attributes in the BIGOS utterance data object 74

3.10 Metrics used for analysis of datasets contents. 76

3.11 Design considerations for ASR evaluation system 81

3.12 Methods of normalizing references and hypotheses 84

3.13 Evaluation scenarios and their analysis dimensions 87

3.14 Relation between research question and evaluation scenarios. 87

3.15 Whisper model types. Source: Whisper model card. 89

3.16 ASR systems evaluated in the study. 91

3.17 Evaluated ASR systems usage cost and license type. 92

4.1 Polish ASR datasets survey summary 97

4.2 Summary of audio dataset availability and characteristics by year 98

4.3 Institutions contributing speech datasets for Polish. 100

4.4 Data catalogs and platforms hosting ASR speech datasets for Polish 101

4.5 Audio devices for all available datasets 102

4.6 Audio devices for publicly available datasets 102

4.7 Audio devices for commercially available datasets 103

4.8 Distribution of sampling rate for publicly reported ASR speech datasets for Polish. 103

4.9 Distribution of speech types for publicly reported ASR speech datasets for Polish. 104

4.10 Speaker and recordings meta-data availability in available speech datasets . 104

4.11 Summary statistics of curated datasets 105

4.12 BIGOS dataset subset license and language coverage. 105

4.13 PELCRA for BIGOS dataset subset license and language coverage. 106

4.14 BIGOS dataset subset domains and speech types. 106

4.15 PELCRA for BIGOS dataset subset domains and speech types. 107

4.16 PELCRA for BIGOS dataset subset domains and speech types. 107

4.17 PELCRA for BIGOS dataset subset domains and speech types. 108

4.18 Audio content size metrics for BIGOS dataset 108

4.19 Audio content size metrics for PELCRA dataset 109

4.20 Text content size metrics for BIGOS dataset 109

4.21 Text content size metrics for PELCRA dataset 110

4.22 Text content features for BIGOS dataset 110

4.23 Text content features for PELCRA dataset 111

4.24 Audio content features for BIGOS dataset 111

4.25 Audio content features for PELCRA dataset 112

4.26 Average duration of audio recordings and utterances — BIGOS dataset. . . 112

4.27 Average duration of audio recordings and utterances — PELCRA dataset. . 113

4.28 Coverage of speaker meta-data — BIGOS dataset 113

4.29 Coverage of speaker meta-data — PELCRA dataset 114

4.30 Publication date and number of downloads of BIGOS datasets as of June 6th 2024. 114

4.31 Overview of sections providing relevant results to research questions RQ7-RQ13 115

4.32 Overview of ASR use-cases covered in Polish ASR benchmarks to date. . . . 116

4.33 Public domain ASR benchmarks 2018-2023. 116

4.34 Overview of domains, speech types, audio sources and recording devices. . . 117

4.35 Datasets size and number of domains, recordings, and speakers. 117

4.36 Acoustic conditions, annotations, and speaker meta-data across Polish ASR benchmarks 118

4.37 Overview of metrics employed in Polish ASR systems benchmarks. 118

4.38 Publicly reported evaluations of ASR models for Polish language. 119

4.39 Number of reported independent evaluations and benchmarks per system. . 119

4.40 ASR systems supporting Polish not yet evaluated in the public domain. . . 120

4.41 Types of ASR systems evaluated in public domain ASR benchmarks 2018-2023. 121

4.42 ASR benchmarks performed in this study. 126

4.43 ASR systems evaluation scenarios overview 126

4.44 Evaluation details for BIGOS dataset 127

4.45 Evaluation details for PELCRA dataset 127

4.46 WER statistics – BIGOS dataset 128

4.47 WER statistics – PELCRA dataset 130

4.48 WER statistics for all ASR systems and specific subsets of BIGOS dataset. 131

4.49 WER statistics for all ASR systems and specific subsets of PELCRA set. . . 132

4.50 WER statistics for free and paid ASR systems on BIGOS dataset. 133
4.51 Best and worse systems for BIGOS dataset. 133
4.52 WER statistics for the most accurate free and commercial systems. BIGOS dataset. 133
4.53 WER statistics for the most accurate free and commercial systems. BIGOS dataset. 134
4.54 WER statistics for free and paid ASR systems evaluated on PELCRA dataset.135
4.55 Best and worst accurate systems for PELCRA dataset. 135
4.56 The most accurate systems WER statistics. PELCRA dataset. 136
4.57 The least accurate systems WER statistics. PELCRA dataset. 136
4.58 Average WER for free systems with information about model size. 137
4.59 Average WER for free systems with information about model size. 138
4.60 Mean WER for specific audio duration ranges. BIGOS dataset. Best paid and free systems. 140
4.61 Mean WER for specific audio duration ranges for top paid and free systems. PELCRA dataset. 141
4.62 Number of samples with speaker gender information. 143
4.63 Values and differences in mean WER scores per speaker gender. 144
4.64 Number of samples with speaker gender information. 145
4.65 Values and differences in average WER scores per speaker gender for 689 samples from PELCRA dataset per gender. 146
4.66 Mean WER across systems and age ranges. PELCRA dataset. 147
4.67 Standard Dev. and maximum difference in WER across age groups. PELCRA dataset. 148
4.68 Reduction of error rates caused by normalization of references and hypothesis for BIGOS dataset. 149
4.69 Reduction of error rates caused by normalization of references and hypothesis for PELCRA dataset. 149

5.1 Benchmark dataset design requirements validation 165
5.2 Evaluated models, datasets, and their combinations. 187

7.1 Publicly and freely available speech datasets for Polish. 206

7.2 Commercially available speech datasets for Polish. 207

7.3 Dataset size per split — Clarin Mobile. 207

7.4 Dataset features per split — Clarin Mobile. 207

7.5 Dataset size per split — Clarin Studio. 208

7.6 Dataset features per split — Clarin Studio. 208

7.7 Dataset size per split — MLS. 208

7.8 Dataset features per split — MLS. 209

7.9 Dataset size per split — Munich AI Labs Librivox. 209

7.10 Dataset features per split — Munich AI Labs Librivox. 209

7.11 Dataset size per split — Common Voice. 210

7.12 Dataset features per split — Common Voice. 210

7.13 Dataset size per split — AZON read. 210

7.14 Dataset features per split — AZON read. 211

7.15 Dataset size per split — AZON spontaneous. 211

7.16 Dataset features per split — AZON spontaneous. 211

7.17 Dataset size per split — PWR Maleset. 212

7.18 Dataset features per split — PWR Maleset. 212

7.19 Dataset size per split — PWR Shortwords. 212

7.20 Dataset features per split — PWR Shortwords 213

7.21 Dataset size per split — PWR Very Important Utterances. 213

7.22 Dataset features per split — PWR Very Important Utterances. 213

7.23 Dataset size per split — Google FLEURS. 214

7.24 Dataset features per split — Google FLEURS. 214

7.25 Dataset size per split — Minds-14. 214

7.26 Dataset features per split — Minds-14. 215

7.27 Dataset size per split — PolEval 22. 215

7.28 Dataset features per split — PolEval 22. 215

7.29 Dataset size per split — Spokes Mix Emo. 216

7.30 Dataset features per split — Spokes Mix Emo. 216

7.31 Dataset size per split — Spokes Mix Luz. 216

7.32 Dataset features per split — Spokes Mix Luz. 217

7.33 Dataset size per split — Spokes Mix Parl. 217

7.34 Dataset features per split — Spokes Mix Parl. 217

7.35 Dataset size per split — Spokes Biz Bio. 218

7.36 Dataset features per split — Spokes Biz Bio. 218

7.37 Dataset size per split — Spokes Biz Interviews. 218

7.38 Dataset features per split — Spokes Biz Interviews. 219

7.39 Dataset size per split — Spokes Biz Luz. 219

7.40 Dataset features per split — Spokes Biz Luz. 219

7.41 Dataset size per split — Spokes Biz Podcasts. 220

7.42 Dataset features per split — Spokes Biz Podcasts. 220

7.43 Dataset size per split — Spokes Biz Presentations 220

7.44 Dataset features per split — Spokes Biz Presentations 221

7.45 Dataset size per split — Spokes Biz Various 1. 221

7.46 Dataset features per split — Spokes Biz Various 1. 221

7.47 Dataset size per split — Spokes Biz Various 2. 222

7.48 Dataset features per split — Spokes Biz Various 2. 222

7.49 Dataset size per split — Spokes Biz Interviews. 222

7.50 Dataset features per split — Spokes Biz Interviews. 223

7.51 Commercial ASR services pricing 223

7.52 Size of freely available ASR models 224

Glossary

Machine Learning task An abstract problem statement, defined either in natural language or formally. Tasks vary in granularity, creating a hierarchy, such as “dog vs. cat classification” to “image classification.” They frame contributions in Machine Learning field and are instantiated by Learning problems for evaluation. Examples include MNIST, CIFAR-10, and ImageNet for the “image classification” task.. 24

AMU ASR Leaderboard Publicly accessible leaderboard presenting results of ASR systems benchmarking supporting Polish on BIGOS datasets. 56, 90, 149

ASR Automatic Speech Recognition (ASR) is a technology that enables machines to process speech input and translate it into text. Also known as Speech Recognition or Speech-to-Text (STT). 1, 10, 17, 24, 44

Audio encoding Manner in which digital audio signal is encoded for storage or transmission. The most popular audio encodings for speech and ASR applications are lossless encodings of *PCM* or *FLAC* and lossy encodings of *Opus* and *Speex*. 29

AZON Repository of open data from Wrocław University of Technology (Atlas Zasobów Otwartej Nauki).[1]. 102

Benchmark A learning problem serving as an indicator of progress on a ML task. Benchmarks often include a leaderboard and open competition. For example, within the ILSVRC competition (ImageNet Large Scale Visual Recognition Challenge) increasing accuracy on ImageNet benchmark dataset reflects advancements in image classification task.. 2

[1] AZON project

Benchmark dataset A benchmark dataset is a curated, widely accepted reference used to evaluate and compare algorithms, models, or systems in a specific domain. It provides a consistent basis for comparison and objective assessment.[58, 9, 27, 96] Benchmark datasets with specific metrics are referred to as learning problems and represent more abstract tasks.[72]. 2, 11, 24, 63, 66, 69, 85, 160

benchmark saturation Phenomenon, which occurs when a learning problem becomes too easy for current ML models, leading to a plateau in performance. This can happen due to various reasons, such as overfitting to test data, advancements in technology or not challenging enough dataset or evaluation metric.. 20, 24, 62

BIGOS BIGOS stands for Benchmark Intended Grouping of Open Speech). BIGOS is a set of curated datasets intended to facilitate benchmarking of ASR systems. Currently, BIGOS is focused on the Polish language. [2]. x, 56, 63, 69, 70, 90, 94

CER Character Error Rate. xxiii, 54, 84, 169

Common Voice Large-scale, multilingual speech dataset collected with crowdsourcing via Mozilla Corporation. 2, 102, 123, 163, 171–173, 177

data curation Broad set of data management techniques, such as acquisition, formatting, documentation, enrichment, annotation or quality verification, aimed at improving the practical utility of datasets. 11

FLEURS Few-shot Learning Evaluation of Universal Representations of Speech. 123

Forced alignment Method of analyzing and synchronizing the speech content with its transcription to achieve temporal alignment. 53

Gated datasets Access to datasets on Hugging Face which allows authors to control dataset usage by requiring users to request access, providing their username and email. Authors can approve requests manually or automatically and may ask for additional information.[3]https://huggingface.co/docs/hub/en/datasets-gatedHugging Face's documentation on gated datasets. 77, 162, 163

[2] BIGOS in Polish means "cabbage stew". The name is inspired by the work of Google on SpeechStew[12]

[3] ,

GitHub Web-based platform for version control and collaborative software development using Git. Supports public and private repositories, and features like pull requests, issue tracking, and project wikis. [4]. 54, 194

GMM Gaussian Mixture Model. 121

Hallucinations In ASR systems, hallucinations are outputs that do not match any spoken input. They can result from background noise, poor audio quality, or ASR model limitations, leading to incorrect transcriptions and reduced system reliability.. 175, 178

Hugging Face Datasets Python library by Hugging Face designed for efficient handling and processing of large datasets. Offers simple access to a wide range of datasets, tools for dataset loading, transformation, and evaluation. Supports various data formats and integration with machine learning workflows.[5]. 53, 55, 77, 81, 162, 169, 170

Hugging Face Hub Platform for hosting, sharing, and discovering machine learning models and datasets. Provides tools for collaborative development, discovering resources, version control, and deployment of models and datasets, enhancing accessibility and community engagement.[6]. 77, 162, 163, 194

JIWER Python library for calculating evaluation metrics for ASR systems such as WER based on minimum-edit distance between one or more reference and hypothesis sentences. 54, 170

Kaldi framework Toolkit for speech recognition written in C++, licensed under the Apache License v2.0 intended for speech recognition researchers. 121

Learning problem A learning problem consists of a dataset of (input, output) pairs and an evaluation metric to score solutions (functions from input to output). It is fully defined by these components without needing external semantics or data; for example,

[4] GitHub

[5] HF datasets

[6] HF Hub

the ILSVRC-2012 dataset (ImageNet) with top-1 accuracy as the metric.[72]. xix, 24

LibriVox Python library for music and audio analysis. Provides the building blocks necessary to create music information retrieval systems. 101

Librosa Python library for music and audio analysis.. 70

Machine Learning Development of algorithms and statistical models that perform specific tasks without explicit instructions. These algorithms and models learn from and make predictions or decisions based on data. xix, 3, 17, 20, 24

MER Match Error Rate (MER) is calculated as the number of errors (insertions, deletions, and substitutions) divided by the number of words in the hypothesis, addressing the issue of unbounded WER in cases of high insertion errors.. 35, 54, 84, 169

ML evaluation set A subset of data used to assess the performance of machine learning models. It is separate from the training data and is utilized to provide an unbiased evaluation of a model's accuracy, generalization, and effectiveness on unseen data.. 31

MLS Multilingual LibriSpeech. 2, 123, 172, 175, 177

Natural Language Processing (NLP) Research field lying at the intersection of computer science, artificial intelligence, and linguistics that focuses on making human communication, such as speech and text, understandable to computers. Involves a variety of tasks (speech or language generation or understanding), techniques (parsing, stemming, tokenization, etc.), and applications (translation, question-answering, summarization, etc.). 9

NeMo toolkit NLP development toolkit provided by NVidia. 53–55

Pandas Open-source Python library for tabular and time-series text data analysis and manipulation. 52, 53, 70, 169, 170

PELCRA Research group from the University of Łódź. PELCRA stands for Polish and English Language Corpora for Research and Applications. xxiii

PELCRA for BIGOS Openly accessible speech dataset in BIGOS format curated from datasets developed by the PELCRA group.[111, 112, 109]. 56, 90

Polish ASR speech data catalog Structured information about existing Polish ASR speech datasets. Includes information about availability, license, size, content characteristics etc. Available on GitHub and Hugging Face and described in the article in PSICL journal[53]. 2, 61, 62, 70, 105

Polish ASR speech datasets survey Survey on available speech datasets for Polish ASR development. 125

RTF Metric for determining how long a system takes to process a given length of input signal compared to the duration of the signal itself. For real-time systems must be lower than 1. 25

Sampling rate Number of sound samples per time unit. Usually expressed in kiloHertz (kHz), which means 1,000 times per second. 29

SDE Speech Data Explorer. Tool. 53, 55, 169, 170

Semantic Word Error Rate (SWER) An ASR evaluation metric that extends traditional WER by incorporating semantic weights. Proposed by Somnath Roy[125], SWER assigns higher weights to errors involving key semantic words or named entities, reflecting their impact on transcript meaning. It uses NLP techniques to calculate semantic similarity and includes CER for spelled-out entities, providing a nuanced measure of ASR accuracy that aligns with human judgment.. 37

SemDist Semantic Distance (SemDist) is a metric proposed by Kim et al.[56] that uses advanced language models like BERT to measure the semantic similarity between a reference transcription and ASR output. Unlike WER, SemDist evaluates the semantic correctness of the outputs, identifying deviations that alter the conveyed message. It uses token-level embeddings and similarity measures, such as cosine distance, to provide a quantitative measure of semantic accuracy.. 37

SER Sentence Error Rate. 54, 84, 169

SOX SOX (Sound eXchange) is a cross-platform command line utility that converts various formats of computer audio files into other formats. It can also apply various effects to these sound files and play and record audio files on most platforms. 70

Speech dataset A collection of digital recordings of speech together with their annotations, meta data, and documentation.. 2, 10, 11, 24, 26

test data leakage Leakage occurs when a model accesses unauthorized information during training, leading to inflated performance metrics and compromised evaluation integrity.. 18, 171, 172

Text to Speech (TTS) Technology that converts written text into spoken words. This enables machines to read text in a human-like voice. Also known as speech synthesis.. 47, 165

Transformers Open-source Python library by Hugging Face offering state-of-the-art pretrained NLP models for tasks such as text classification, translation, and question answering. It simplifies model integration, fine-tuning, and deployment.[7]. 55, 169

WAV RIFF WAVE audio format. 83

WER Word Error Rate (WER) is a metric for evaluating ASR performance. Calculated as number of word-level edit operations (deletions, insertions, substitutions) required to match the two texts divided by the total number of words in the reference text. xxi–xxiv, 1, 7, 18, 19, 23–25, 31, 33, 37, 42, 54, 84, 89, 90, 127, 129, 143, 159, 166, 169

WIL WIL stands for Word Information Lost. It used as the metric to assess the effeciency of conveying information encoded as text e.g. accuracy of ASR systems. WIL provides a simple approximation of the proportion of lost word information in a sequence of words. Defined as 1 - WIP, where *WIP* stands for *Word Information Preserved.* [90]. 54, 169

WRR Word Recognition Rate (WRR) is a metric for evaluating ASR performance, calculated as 100% minus the WER. It measures the proportion of correctly recognized words in the ASR output.. 54, 169

[7] HF Transformers

Chapter 1

Introduction

ASR systems process human speech signals into the corresponding textual transcriptions. Recent progress in machine learning technology, powerful computational resources, and abundance of data have significantly advanced ASR technology, resulting in a notable improvement in the accuracy of speech-to-text conversion. In 2017, Microsoft announced that the precision of its ASR system for English is on par with manual transcription when evaluated using the *Switchboard* corpus[49]. The quality in terms of the WER metric (word error rate) was below 5%. In 2022, the Whisper ASR system achieved an average *WER* of 5% for multiple test datasets and high-resource languages [117]. ASR technology is now widely used in various applications such as virtual assistants, meeting aids, voice search, smart home controls, and transcription tools. The increasing global demand for ASR solutions has made it a focal point of research aimed at improving speech recognition performance. Companies that develop ASR systems are constantly working to reduce error rates for new applications, domains, and languages to enhance the user experience and increase market adoption.

The Polish language is spoken by more than 50 million people around the world and is the sixth most spoken language in the European Union. The number of commercial and freely available voice technology solutions and applications for Polish is steadily growing [95]. In July 2023, more than 50 speech data resources were available for training and evaluating ASR systems [53]. New language resources are being introduced thanks to global initiatives like Mozilla Common Voice [3] or Multilingual Librispeech [115] and local projects, e.g. DiaBiz [112], Spokes[111] etc.

So far, no research has been conducted to survey, validate, or improve the usefulness of existing Polish ASR speech datasets. There are also no widely adopted Speech datasets for performing Benchmarks of ASR systems for the Polish language. International studies typically use popular multilingual datasets, for example Common Voice[3] or MLS[115], while Polish studies mostly use locally created datasets[111, 110, 65]. Although many datasets are available under permissive licenses, they are often used exclusively for specific studies due to interoperability concerns. Same practical obstacles may also contribute to the restricted use of all accessible speech datasets for Polish ASR benchmarks. This restriction limits usefulness for researchers and the public, who rely on benchmarks and leaderboards to track progress and identify suitable models. [91, 34]

There is an ongoing debate within the international ASR community about the establishment of a standardized evaluation methodology. [2, 137] Adopting standard methodology and datasets enables comparing results between different studies. However, to make the results relevant to various usage scenarios, a variety of representative datasets are needed. Examples from other fields of ML [96] show that it is important for the community to take advantage of existing and easily contributing new datasets for benchmarking purposes [10, 24, 145, 126]. To monitor technological advances over time, it is preferable to perform systematic benchmarking instead of one-time assessments. However, since there are many applications and variables of ASR that impact its effectiveness, establishing a common standard of evaluation is not trivial.[2]

The objective of this thesis is to increase the practical utility of available speech datasets for the evaluation of Polish ASR systems. The proposed framework consists of a set of methods for speech data management and ASR evaluation. The key contributions include:

1. Survey of Polish ASR speech datasets and curation of Polish ASR speech data catalog

2. Survey of Polish ASR benchmarks

3. Curation of Benchmark dataset from publicly available sources

4. Development of framework for ASR systems benchmarking

The thesis also discusses the strengths and limitations of existing speech datasets and outlines potential research directions to further improve ASR data management and benchmarking practices for Polish ASR.

1.1 Problem background

1.1.1 The role of datasets in the training and evaluation of machine learning systems

Datasets are essential in the development of Machine Learning (machine learning) because they convey the signal used during the training, testing and validation of ML models. These datasets encode useful information, allowing algorithms to recognize patterns within the input data. Relevant information can be obtained from the original source or annotated via a dedicated process, typically by trained humans. ML datasets must be diverse and representative of target usage to ensure the accurate performance of models on new data during operation.

In 2020 Andrew Ng, the co-creator of the Google Brain project, introduced the term "Data-Centric AI" to the public discourse. He noted that currently 90% academic work follows the "Model-Centric" paradigm, which assumes that data are fixed and that quality improvement is achieved through changes in architecture and the model training process. According to the *"Data-Centric AI"* paradigm, the model architecture and training process remain constant, and the quality improvement of the model operation is achieved by increasing the quality and size of the data used for training and testing the system. He notes that developing datasets that are not only widely applied but also actively maintained by the community is a challenge. In practice, the data preparation process is often treated as a one-time effort. The limited adoption of standards for documentation or quality assurance methods adds to the challenge[36, 8, 53, 45, 116]. As a result, the benchmark results of the ML systems from academic conferences can present a distorted picture of the state of technology development [72]. This can result from the relative simplicity of the task represented by the test set, e.g., the *librispeech* speech corpus that contains only records of spoken speech in a quiet environment)[100]. Another factor contributing to the reduced reliability of the benchmark results is inaccuracies in the data labels. For example, recent research from December 2022 indicates that in the 10 most popular test sets used in the benchmarks, an average of 3.3% of the data had incorrect labels[93].

1.1.2 The role of speech datasets in the training and evaluation of ASR systems

The ASR community is highly dependent on extensive training datasets that accurately represent the speech and acoustic patterns of the target population, as well as the operating conditions of the ASR systems. The construction of datasets of this kind is a difficult undertaking that requires specialized infrastructure, meticulous planning, nimble recruitment operations, and resource-intensive data quality control [42, 3, 115, 52]. Moreover, to conduct responsible and informative testing of ASR systems [2], one needs access to evaluation datasets that are free of errors, contain abundant metadata, and are up-to-date. This necessity makes the management of ASR speech datasets even more complex and demanding.

Another challenge facing the ASR community is the discovery of relevant datasets that already exist. Currently, there is no centralized repository dedicated to ASR speech datasets, either multilingual or for the Polish language. As a result, researchers and industry practitioners have to rely on information dispersed among many sources and may struggle to accurately determine the number of available datasets and their characteristics, such as size, recording devices, utterance domain, audio and transcription quality, and others. Ideally, a comprehensive data catalog should include download links to dataset samples, allowing seamless and in-depth inspection of the datasets of interest, in addition to the aforementioned metadata descriptors.

Speech datasets commonly include distinct sets for training, validation, and testing. Validation sets assist in fine-tuning model parameters, while test sets gauge the final model's performance, offering an impartial evaluation of its functionality in real-world scenarios. It is worth highlighting that speech datasets shall encompass various languages, dialects, accents, speech styles, and noise environments in order for ASR systems to be robust. This diversity guarantees that the ASR system can handle a wide range of speech variations and operate effectively in different settings and demographics of speakers.

In addition, training of ASR models is heavily based on extensive and varied speech datasets. These datasets encode a wide spectrum of phonetic, linguistic, and acoustic attributes essential for precise speech recognition. datasets featuring conversational speech, ambient noise, and authentic speech patterns (e.g., pauses and interruptions) enable developers to efficiently handle real-world use cases like voice-activated assistants or IVR (Interactive Voice Response) systems. Finally, speech datasets also serve a role in the examination and mitigation of biases in ASR systems. datasets comprising a diverse array of voices and speech attributes can help to recognize and minimize bias related to accents, dialects, age, gender, and more. [1]

New methods and resources are actively developed for the evaluation of ASR systems in academia and technology companies such as Google, Apple, Amazon, Meta, Appen [75] or Rev[21, 22]. However, details on specific methods or confidential datasets used to create commercial products are not disclosed because of the confidentiality nature. For-profit entities contribute predominantly to the curation of novel datasets from publicly accessible sources, e.g. [115, 19]. The relevant findings are presented at conferences related to speech technologies such as *Interspeech* or *Language Resources and Evaluation (LREC)*[1], as well as workshops such as *NeurIPS workshop on Evaluation and Benchmarks.* [2]

1.1.3 Challenges in ASR speech dataset management

ASR practitioners managing speech datasets face numerous practical challenges.

Data identification

Identifying the right data for the task is often difficult. The information is spread across numerous data repositories and publications. Furthermore, there is no widely established standard for documenting and evaluating the potential application of speech datasets. Often without manual inspection of the content of the datasets, it is not feasible to determine their quality or suitability for a specific task.

Data formatting

Although some data may be freely available and easily accessible, the diversity of audio and text file formats, data quality issues, and limited documentation can require significant data wrangling efforts before a dataset can be used effectively.

Legal and licensing concerns

Legal and licensing limitations may apply to the use of speech data, particularly when using data from public sources or third parties.

Data privacy and ethics

[1] LREC

[2] NeurIPS

Managing datasets that contain sensitive or personal data requires strict adherence to privacy laws and ethical guidelines, including obtaining consent from participants and anonymizing data where possible.

Language evolution and terminology

Language is constantly changing, with new vocabulary, expressions, and meanings frequently evolving. It is an ongoing challenge to ensure that speech datasets remain up-to-date with these linguistic shifts.

Data bias

Speech datasets can unintentionally exhibit biases toward specific demographic groups (such as age, gender, accent, and dialect), resulting in disparities in the performance of ASR systems for different user groups. [1]

Audio data quality

The accuracy of ASR may decrease due to background noise or poor audio quality. Therefore, it is crucial to manage these factors during data collection. If background noise or distorted speech are essential for the ecological validity of the ASR application under study, relevant metadata and documentation must be included to ensure an accurate interpretation of the evaluation results.

Data annotation quality

Annotation can be time-consuming and susceptible to human errors, particularly when dealing with large datasets, complex domains, and diverse annotation teams.

Managing versions of datasets

It is essential to maintain version control and provide users with accurate and up-to-date datasets as they are modified and improved over time. Effective dataset management practices are necessary for this purpose.

Data storage and retrieval

The size of high-fidelity audio files presents difficulties in storage and distribution, particularly with large datasets.

Striking a balance between size and manageability

Although larger datasets can improve ASR performance, they also present difficulties in terms of computational resources and training duration. Therefore, determining the optimal balance between the size of the dataset and the ease of management is a critical issue.

1.1.4 Challenges in ASR evaluation

Common challenges

These are the challenges faced in the ASR evaluation process.

Lack of ground truth

There may not be definitive ground-truth transcription for the audio data being analyzed, for example, in the case of multiple spelling conventions.

Domain-specific challenges

ASR systems may perform differently depending on the domain or context. For example, a system trained on news broadcasts may not perform as well on telephone conversations. Hence, a careful selection of appropriate evaluation datasets that represent the target domain is required. For example, significant discrepancies have been reported in a recent comparison of the accuracy of ASR systems for medical terminology in Polish. [68] and [153].

Metric selection

Different metrics are used in the scientific literature, the most popular being WER (Word Error Rate). Depending on the ASR application, the appropriate evaluation metric and method should be used.

Annotation consistency

The annotation of the evaluation data must be consistent and unbiased between multiple annotators. This requires the use of standardized annotation protocols and thorough training of the annotators.

Limited resources

Evaluation of ASR requires significant resources including data storage, computing and cloud usage costs, human expertise, and time for results analysis.

Conflicts of interest

Commercial sources often showcase and explain ASR solutions through company reports, testimonials, or white papers. These providers typically strive to highlight the strengths of their products. As a result, there is a need for independent comparative research on existing ASR systems, focusing on evaluating their performance, scalability, and accessibility to provide practical benefits for particular applications or domains.

Challenges in the industrial settings

Additional factors must be taken into account when creating an ASR system within industrial environments. To ensure and continuously monitor the quality of technology, products or services, companies conduct continuous research, implementations, and tests with the aim of improving product features and eliminating defects and their causes. To test the quality of a solution based on machine learning algorithms under conditions that match actual use, it is necessary to prepare and continuously update test data that are representative of the specific requirements of the offered solution, for example, language of target user group, device, and domain. Moreover, ASR systems must be tested to determine the impact of disturbances and modifications of the acoustic signal, such as:

- Variable characteristics of sound processing in a given type or specific model of device,
- Distance and position of the user relative to the device.
- The presence of discontinuities and additive noise in the speech signal.

Ideally, ASR testing should also verify the robustness to speech variations resulting from individual user characteristics such as gender, accent, age, language proficiency, ethnic background, emotional or health condition, articulation quality, and so on.

To check whether the quality requirements of an ASR-based product or service are met, it is necessary to perform a series of tests on a representative sample for real-use conditions. In practice, obtaining representative test data before deploying a service/product to the market is a significant challenge and requires substantial investments in preparing the appropriate environment, scenarios, and processes to acquire and control data quality. This is because numerous companies do not possess sufficient resources and know-how to record new statements under controlled conditions or transcribe and annotate existing recordings.

The requirements and characteristics of real-world usage data evolve rapidly. The more quality criteria are considered, the more extensive resources are required to design, create, curate, and validate ASR evaluation datasets. Companies developing ASR commercially require dedicated processes and systems to ensure the quality and availability of data for the continuously changing product requirements. The coherent methodology includes data typologies, data standards, annotation protocols, operating procedures, and systems for data collection and annotation.

1.1.5 State of the ASR speech datasets and ASR evaluation for Polish

In recent years, the field of NLP has experienced a surge in benchmarks designed to evaluate the most widely available systems in a wide range of datasets [145, 144]. The most advanced research on the methodology for evaluating ASR systems and requirements for data used for this purpose relates to the English language and, to a much lesser extent, selected European languages, such as German. In addition, there has been a growing interest in ASR benchmarks in the international community [139, 34, 2, 1, 26].

In Poland, the growing interest in data for AI development and benchmarks for the Natural Language Processing (NLP) (Natural Language Processing) area is evidenced by the PolEval competitions [59] organized annually and the *KLEJ* initiative (Comprehensive List of Language Evaluations) [126] and *LEPISZCZE* [4].

The first benchmark for Polish ASR systems was conducted in 2018. Three commercial ASR systems were evaluated on a set of recordings representing domain and acoustic conditions of security officer training. [99] In 2019, the first open competition was organized under the PolEval initiative [59]. Six community-provided systems were evaluated using datasets created by recordings of the Polish Parliament. The next benchmark in 2022 compared the accuracy of 3 commercial ASR systems using recordings from the customer support domain [112]. The most recent benchmarks focused on the accuracy of medical terms recognition accuracy.[153, 68]+

The major challenges of Polish ASR benchmarks include:

- limited utilization of publicly available speech datasets
- limited reproducibility due to lack of access to evaluation datasets
- lack of independent quality verification of test sets used in evaluations
- limited number of evaluated systems

1.2 Research aim

The primary aim of this thesis was to design and implement a data management framework to increase the utility of the available Polish speech datasets for the evaluation of ASR systems.

The initial stage involved creating a taxonomy and organizing metadata on existing speech datasets using publicly accessible information. The subsequent stage covered the quantitative evaluation of the characteristics of the datasets to determine their usefulness for the ASR evaluation. The selected datasets were then consolidated, refined and made openly accessible. The final stage was the development of an evaluation system and the use of curated Speech dataset to compare various ASR systems for the Polish language.

1.3 Research hypothesis

The hypothesis advanced in this thesis is the following:

The creation of an extensive data management framework will make it possible to reliably and objectively evaluate the ASR systems available for Polish.

1.4 Research objectives and questions

This section presents the main research objectives (RO) and research questions (RQ).

RO1: Survey of ASR speech datasets for Polish The first objective was to survey existing ASR speech datasets for Polish. The research questions addressed were:

- **RQ 1:** How to systematically categorize Polish ASR speech datasets using public information?
- **RQ 2:** What is the current state of Polish ASR speech datasets?
- **RQ 3:** How can the survey findings be shared for community feedback?

RO2: Design and curation of the speech dataset for Polish The second objective was to curate the dataset to evaluate ASR systems for Polish. The research questions considered were:

- **RQ 4**: What factors are crucial in designing and curating dataset for benchmarking purposes?

- **RQ 5**: What are data curation steps required to create Benchmark dataset from publicly available speech datasets?
- **RQ 6**: Which public Polish speech datasets can be used as benchmarks?
- **RQ 7**: How can the curated dataset be shared with the community?

RO3: Survey of ASR benchmarks for Polish Next goal was to categorize and review Polish ASR benchmarks with respect to datasets, systems, tasks, domains and evaluation metrics. The specific questions research included:

- **RQ 8:** How to categorize Polish ASR benchmarks using public information?
- **RQ 9:** What methods, datasets, and ASR systems have been used in Polish ASR benchmarks?
- **RQ 10:** Which Polish ASR systems have not been evaluated?
- **RQ 11:** Which benchmarks evaluated commercial and free systems?
- **RQ 12:** Which ASR system performs best?
- **RQ 13:** What are the main conclusions from the ASR benchmarks?
- **RQ 14:** How to share the survey results with the community?

RO4: Design and implementation of system for ASR systems benchmarking The following objective was the development of a system enabling the evaluation and comparison of ASR systems. The research was focused on the following aspects:

- **RQ 15:** What tools and systems exist for ASR benchmarking?
- **RQ 16:** What challenges arise in evaluating multiple ASR systems, and what strategies can address them?
- **RQ 17:** How can the system be extended to new ASR systems, datasets, languages, metrics, and normalization methods?

RO5: Using a curated dataset to benchmark ASR systems for Polish RO5 goal was to use the self-curated Speech dataset (RO3) and the evaluation system (RO4) to compare ASR systems for Polish. The specific research questions included:

- **RQ 18:** What is the ASR accuracy for different datasets?
- **RQ 19:** What is the accuracy gap between commercial and free systems?
- **RQ 20:** Does ASR accuracy vary with speech features?
- **RQ 21:** Is there an accuracy difference by age or gender?
- **RQ 22:** How to share evaluation results with the community?

RO6: Organization of an open competition for the ASR community The goal was to organize a public contest for ASR practitioners to compare their solutions with the latest advances.

- **RQ 22:** What programs can organize the Polish ASR community challenge?
- **RQ 23:** How to compare community solutions with state-of-the-art ASR systems?

1.5 Research scope

1. **Curation of Polish ASR speech data catalog** Publicly available information about Polish speech datasets was manually annotated with a dedicated taxonomy. The resulting Polish ASR speech data catalog was used to select datasets for further curation. The practical utility of the catalog was evaluated through a user survey.
2. **Curation of benchmark datasets from publicly available speech datasets** The datasets were selected from the speech data catalog according to the ASR evaluation criteria. They underwent automatic refinement, including standardizing audio and metadata formats, and were organized into training, validation, and test sets. Erroneous samples were removed.
3. **Analysis of curated datasets contents and preparation of dashboard for dataset features inspection** Detailed analysis of the curated datasets was performed. To inspect and explore the characteristics of these datasets a dedicated dashboard was created. This tool allowed for a comprehensive inspection of the attributes of the dataset and facilitated better understanding of the data.

4. **Survey of Polish ASR benchmarks** A comprehensive survey was conducted to identify existing benchmarks for Polish ASR systems. The survey involved analyzing the available benchmarks, their methodologies and the datasets they used. Insights were derived to highlight the gaps and areas for improvement in current Polish ASR benchmarks.

5. **Implementing system for ASR evaluation** Developed a robust system to evaluate ASR systems. This system included tools for automatic and manual assessment of ASR output, incorporating various evaluation metrics such as WER (Word Error Rate), CER (Character Error Rate), and others. The system was designed to be scalable and adaptable for continuous benchmarking.

6. **Benchmarking ASR systems for the Polish language** The curated datasets were used to evaluate and compare the performance of ASR systems for the Polish language. In total, 25 models were evaluated. The results were made available to the community through the ASR leaderboard.

7. **Publication of Polish ASR leaderboard** A publicly accessible ASR leaderboard was developed, enabling comparison of the performance of the ASR system. Interactive dashboards were included to allow users to explore the results in detail and compare different systems based on various criteria.

8. **Organization of open ASR challenge** The curated datasets were used to organize an open challenge for the Polish ASR community. This challenge aimed to engage the community in improving ASR technology for Polish and to benchmark new systems against the curated datasets.

1.6 Limitations

This section lists the limitations of the research conducted.

1. **Language specificity:** The research is confined to the Polish language, a language with distinct linguistic attributes. Its findings may not extend to ASR systems for languages with divergent phonetic or grammatical structures.

2. **Datasets selection:** This study is based on a selection of publicly accessible Polish speech datasets intended for ASR. The limited scope of datasets might influence the applicability of the research to broader speech data contexts and corpus linguistic research.

3. **Data curation constraints:** Collecting new speech or annotations is beyond this work's scope. Manual annotation was used to inspect existing data and validate automatic curation methods. No new recordings or annotations were added.

4. **Technological focus:** The study focused on ASR technology, particularly speech-to-text accuracy. Metrics like latency, real-time factor, voice biometrics, and downstream task evaluation were not considered.

5. **Resource availability**: Research on the accuracy of commercial ASR systems and large ASR models was limited by funding and computational resources.

6. **Temporal constraints:** The study covers speech datasets available up to December 2023 and ASR systems up to March 2024.

7. **Demographic and use case coverage:** The research does not fully represent all segments of the Polish-speaking population, including unique dialects or speech variances.

8. **Methodological boundaries:**, Evaluation results are based on selected automatic metrics. The linguistic and acoustic analysis was limited to selected aspects.

9. **Commercial and academic solutions:** The analysis included various commercial and free ASR systems for Polish, though not all solutions are covered due to the rapidly evolving landscape.

1.7 Methodology adopted

The methodology adopted in the research consisted of several steps listed below.

Survey of Polish ASR speech datasets The method consisted of a review of publicly accessible information to catalog Polish ASR speech datasets Specific activities include:

- Literature review and identification of existing speech datasets.

- Development of a taxonomy classification framework. identify and

- Cataloging of speech datasets according to the framework.

- Developing a publicly accessible digital repository and dashboard.

Curation of datasets for Polish ASR systems evaluation The method utilized publicly available sources to curate diverse datasets for Polish ASR development . Specific activities include:

- Selection of speech datasets based on the curated data catalog.

- Data unification, normalization, and formatting.

- Developing a publicly accessible digital repository and dashboard.

Evaluation of ASR Systems for Polish The method used curated datasets to compare ASR systems in various scenarios. Specific activities include:

- Selecting evaluation metrics

- Evaluating ASR systems using recordings from curated datasets

- Analyzing performance, highlighting strengths and weaknesses

- Developing a public dashboard with results

Organization of Polish ASR challenge

Curated datasets were used to organize open competition to allow the comparison of state-of-the-art ASR systems with community-developed systems. Specific activities include:

- Selecting a competition platform.

- Establish participation and evaluation guidelines.

1.8 Contributions

Below are the major contributions of this work to the Polish ASR field:

1. Creation of the largest Polish ASR speech data catalog, documenting 53 datasets with 65 attributes.

2. Development of a metadata schema for cataloging ASR speech datasets.
3. Analysis of the current state of the Polish ASR datasets and the proposal of future research directions.
4. Distribution of two datasets curated from 24 publicly available datasets.
5. Performing and sharing the analysis of the content of curated datasets.
6. Performing the survey and creating the catalog of Polish ASR benchmarks.
7. Development of an extensible system for ASR evaluation.
8. Comprehensive evaluation of Polish ASR systems involving 7 systems, 25 models and 24 datasets
9. Development of a publicly accessible ASR leaderboard with interactive dashboards.
10. Improvement of reproducibility and guidance for future ASR advancements by providing public access to data catalogs, curated datasets, evaluation tools, and dashboards.
11. Organization of an open challenge for the ASR community using curated datasets.

Chapter 2

Literature Review

2.1 Introduction

This section presents literature relevant to the following topics:

- Challenges in benchmarking of Machine Learning and ASR systems.
- Challenges, methods and tools for the management of ASR speech datasets.
- ASR speech datasets and benchmarks for the Polish language.

Based on the review, relevant datasets, methods and tools required to create research artifacts and achieve research objectives were selected.

2.2 Benchmarking of Machine Learning Systems

2.2.1 Challenges in ML benchmarking

Liao et al. provides a comprehensive overview of challenges and systemic issues in benchmarking practices in various subfields of machine learning (ML) [72]. In the meta-review, the authors studied more than 107 articles that describe benchmarks from subfields such as computer vision, natural language processing, recommender systems, and reinforcement learning. The major conclusion is that the inconsistency in evaluation standards and methodologies has led to claimed advances in machine learning that do not withstand thorough examination or do not possess the broad applicability initially assumed.

The authors introduced concepts of internal and external validity of ML evaluations. Internal validity concerns the "correctness and fairness of evaluations in the context of a

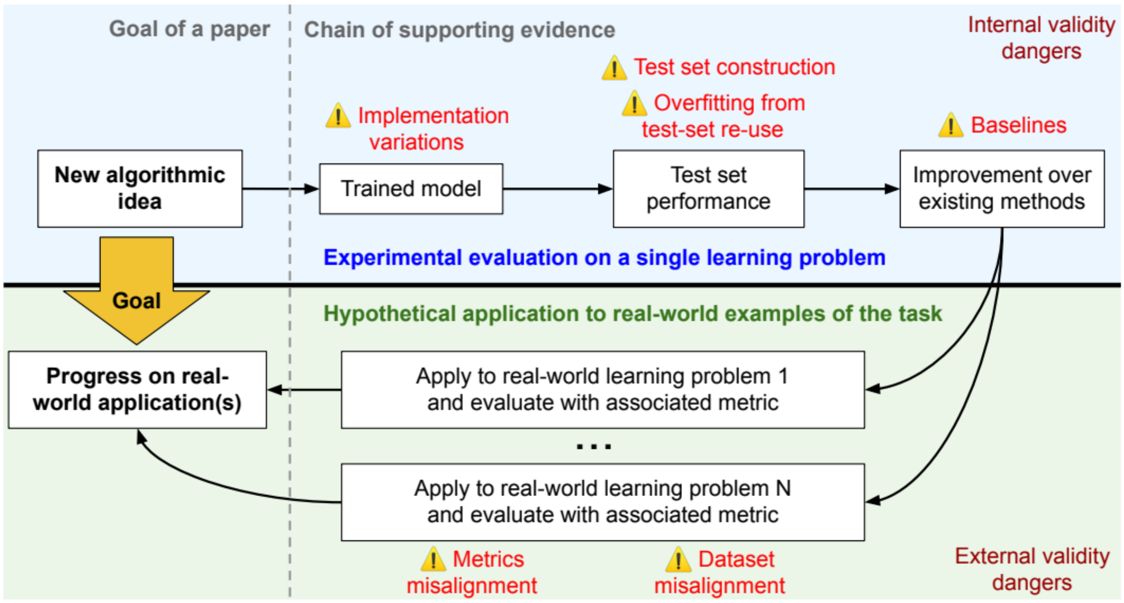

Goal of a paper
Chain of supporting evidence
Internal validity dangers
Test set construction
Overfitting from test-set re-use
Implementation variations
Baselines
New algorithmic idea
Trained model
Test set performance
Improvement over existing methods
Experimental evaluation on a single learning problem
Goal
Hypothetical application to real-world examples of the task
Progress on real-world application(s)
Apply to real-world learning problem 1 and evaluate with associated metric
...
Apply to real-world learning problem N and evaluate with associated metric
Metrics misalignment
Dataset misalignment
External validity dangers

Figure 2.1: Internal and external issues identified in the ML evaluation practices. Source: [72]

specific learning problem".[72]. Internal validity is negatively affected by incorrect baseline comparisons, errors in the construction of the test set, and overfitting due to test data leakage. External validity, on the other hand, refers to the 'applicability and generalizability of the evaluation findings in different learning problems, tasks, or real-world scenarios'[72]. In case of misalignment of metrics and dataset with respect to the real-world scenario, the benchmark result may not accurately reflect the progress or performance of the ML application under the target conditions. Failures of both types are common and contribute to a misleading representation of progress within the ML field. Figure 2.1 presents specific issues of internal and external validity throughout the ML lifecycle. The authors also propose a useful distinction between terms that are often used interchangeably in the ML benchmarking context: *learning problems* and *tasks*. A *learning problem* comprises a dataset of input and output pairs and an associated evaluation metric to score the proposed solutions (functions that correspond to the input space). The example is the Librispeech dataset with WER as a metric to score ASR systems. A task is described in a more general manner, either in the everyday language or formally. There is no fixed definition of a task, and the goal is not to set specific task definitions. Tasks can be found at different levels of detail, for example, from '*dog vs. cat classification*' to '*animal classification*' to '*image classification*', which naturally gives rise to a hierarchy (see Figure 2.2). For the purpose

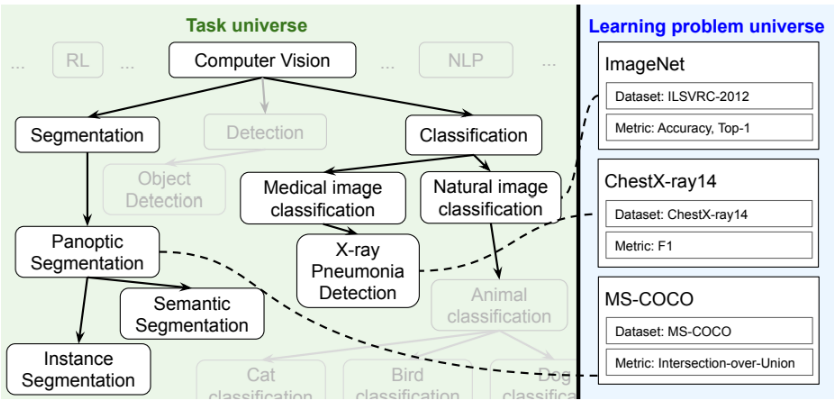


Figure 2.2: ML tasks and learning problems universe. Source: [72]

of evaluation, tasks are usually instantiated by learning problems. Given the above definitions, a "*benchmark is a learning problem framed as an indicator of progress on some task*" [72]. Benchmarks typically include a ranking system, contest, or other framework that defines the current state-of-the-art. Enhancing WER performance on the English Librispeech dataset can be seen as an improvement in *ASR task*, but only within the specific scope and use case determined by the dataset.

The recommendations to improve the robustness and reliability of ML benchmarks include:

1. adoption of more rigorous experimental designs
2. improved documentation standards
3. sharing of research artifacts, enabling replication and inspection
4. development of benchmarks that more accurately reflect real-world conditions.

2.2.2 Examples of methods for curating ML benchmarking datasets

Introduction

Evaluation of ML solutions can be challenging. Factors such as the specific learning problem, the task at hand, the context of the application, and the objectives of the study must be taken into account for the benchmark to be useful. In addition, evaluation datasets are available from various sources, but their formatting, documentation, or access methods are often inconsistent. As a result, choosing and organizing the evaluation process can be an additional burden for ML professionals and data scientists. Therefore, accessible, curated, and maintained public benchmark resources are essential to identify the strengths and weaknesses of different ML methodologies. The curation involves several processes to ensure the utility of the datasets for benchmarking purposes. This section presents examples of such curation processes and selected methods based on examples of popular benchmarks from various ML subfields.

Examples of datasets curated for benchmarking purposes

Penn Machine Learning Benchmark (PMLB) alpha 2017 [96] is a curated collection of 165 datasets from a wide range of sources covering real-world, simulated and toy problems. The datasets were standardized with numerically encoded categorical features. Instances with fewer than 10 examples per class were removed to maintain reasonable learning scenarios. The curated datasets were then made available via a Python interface to simplify retrieval and working with the data. The authors performed a comparison of meta-features of datasets and found that they lacked the diversity to properly benchmark ML algorithms. The study also identified datasets for which the corresponding benchmarks matched or exceeded human baselines or achieved a plateau in performance, resulting in a so-calledbenchmark saturation. The study also identified more challenging datasets, offering a range of difficulties to test Machine Learning methods. The original 2017 article was presented as an ongoing project and is still being developed.

Penn Machine Learning Benchmark (PMLB) v1.0 2020[121] The updated version of the PMLB benchmarking suite was released in 2020 [1]. The original collection that covered classification tasks has been expanded to include regression tasks. Each dataset has been enhanced with a standardized metadata file that contains information about its original source, purpose description, related publications, keywords, and details about individual features and their coding schemes. The structured metadata format simplified the validation process, leading to improved data accuracy and easier addition of new datasets by the community. The user experience has been enhanced with a new contribution guide and an improved website interface that allows browsing, sorting, filtering, and searching for datasets. Support for the R library was also added. Pandas-profiling reports for each

[1] https://epistasislab.github.io/pmlb/

dataset were added that cover feature correlations and identification of duplicates and missing values, allowing users to make informed decisions regarding necessary modifications prior to using a specific dataset.

GLUE 2019 The GLUE (General Language Understanding Evaluation) benchmark[2] is the collection of tools and assembly of existing datasets for nine NLP tasks, such as question answering, sentiment analysis, and textual entailment. *GLUE* includes test data that were never made public and a hand-crafted diagnostic dataset for detailed linguistic analysis. Manually annotated examples *serve as a tool for error analysis, qualitative model comparison, and the development of adversarial examples* [145]. The benchmark focus is not to reflect overall performance or generalization in downstream applications, but rather to understand the performance of general versus specialized models and their capabilities and limitations in handling complex linguistic phenomena.

SUPERGLUE 2020 *SuperGLUE* [144] builds on its forerunner, the *GLUE* benchmark, by incorporating a range of more challenging language comprehension tasks. *SuperGLUE* was developed in response to the realization that performance on the GLUE benchmark exceeded that of non-specialist humans. New tasks were collected by issuing an open invitation for task suggestions within the NLP community. The tasks were selected based on their level of challenge for existing NLP methods and covered a variety of formats, such as coreference resolution and question answering. The datasets were derived from preexisting data to guarantee availability and consistency. The tasks must have available public training data, have an automatic performance measure that correlates well with human evaluation, and should not require specialized knowledge beyond standard English proficiency. Human performance benchmarks were established for all tasks, ensuring ample scope for enhancing model performance. The benchmark was launched with a modular toolkit that facilitates model training, testing, and assessment. This toolkit was based on commonly used frameworks such as PyTorch and includes conventional models like BERT for initial evaluations. The leaderboard[3] was structured to promote fair competition and meaningful comparisons of models. The guidelines for submissions are explicit on data usage and the tasks are designed to reduce overfitting and enhance the interpretability of model performance across a range of NLP tasks.

[2] https://gluebenchmark.com/

[3] super.gluebenchmark.com

MMLU 2021[4] The Massive Multitask Language Understanding (MMLU) benchmark is designed to assess text models across a broad spectrum of fields and complexity levels. *MMLU* covers 15,908 questions from 57 topics. The questions were manually collected by graduate and undergraduate students from openly accessible online resources. The few-shot development (training) set has 5 questions for each subject, the validation set has 1,540 questions, and the test set has 14,079 questions. Each subject has questions of different difficulty levels, from elementary to high school, college, and professional. This enables one to gauge the depth of knowledge of a model and its capacity to deal with increasingly difficult content. Baseline results from both non-specialized human test-takers and experts are available. This comparison offers a context for assessing the performance of language models in relation to human abilities. The *MMLU* is designed for zero-shot and few-shot settings to evaluate the ability of models to generalize and apply knowledge without extensive fine-tuning, as in many real-world scenarios.

BIG-Bench 2022[5] BIG-bench[135], which stands for Beyond the Imitation Game, is a benchmark for language models, comprising 204 tasks put forward by 450 authors from 132 different institutions. The tasks are varied and cover a wide range of topics, including linguistics, childhood development, mathematics, common sense reasoning, biology, physics, social bias, software development, and more. BIG-bench's emphasis is on tasks that are thought to exceed the abilities of current language models. The tasks come in various formats, such as multiple choice and text-complete questions. The curation process was carried out transparently and cooperatively. Contributions were collected through *GitHub* pull requests and then subjected to a peer review process. This approach guaranteed a broad spectrum of tasks and viewpoints. Expert human raters were employed to complete all tasks, establishing a reference point to evaluate the performance of the language models. *BIG-bench* was created with the intention of facilitating the ongoing contributions of tasks and evaluations, ensuring its continued relevance.

SUPERB 2021[6] *SUPERB (Speech Processing Universal PERformance Benchmark)* [139] is a toolkit and leaderboard to benchmark the performance of a shared model in a wide range of speech processing tasks with minimal architecture changes and labeled

[4] https://huggingface.co/datasets/cais/mmlu

[5] https://huggingface.co/datasets/bigbench

[6] https://arxiv.org/abs/2105.01051

data. Multiple speech processing is included, for example, phoneme recognition, automatic speech recognition, keyword spotting, speaker identification, speaker verification, speaker diarization, intent classification, slot filling, and emotion recognition. For the dataset to be included in the benchmark, it must adhere to the conventional protocols accepted by the speech community, be publicly accessible, and allow universal participation. datasets considered to be the standard benchmarks for various tasks are included, e.g.

- *LibriSpeech*: Used for phoneme recognition and automatic speech recognition tasks.
- *Speech Commands V1.0:* Utilized for keyword spotting to detect predefined words.
- *VOXCELEB1*: Employed for speaker identification and verification tasks.
- *Fluent Speech Commands:* Used for intent classification.
- *IEMOCAP*: Chosen for emotion recognition tasks.

Each task has specific metrics for evaluation, such as the WER for speech recognition, the accuracy for keyword spotting and speaker identification, and the diarization error rate (DER) for speaker diarization. The benchmark goal is to encourage the development of models that can perform well on diverse speech processing tasks with minimal specific tuning for each task.

ASR-GLUE 2022 ASR-GLUE [29] is a benchmark to study the effect of ASR error on NLU tasks in terms of noise intensity, error type and speaker variants. Six NLU tasks that are prevalent in speech-based scenarios are included: sentiment analysis, paraphrase detection, and natural language inference. Data instances were manually selected from existing NLU task datasets. The selection criteria excluded samples with non-standard words or overly long sentences to ensure clarity and quality in speech-to-text conversion. Six native speakers recorded the selected test samples in different noise environments. This was done to simulate real-world speech variations and introduce controlled ASR errors. The recordings were converted to text using an ASR system trained for this purpose. For tasks that require labeled data, the dataset maintained the original labels of the source datasets, ensuring that the impact of ASR errors could be assessed against known outcomes. The dataset is maintained by *Tencent AI Lab*, is publicly available, and open to community contributions.[7]

[7] ASR GLUE audio

ESB 2022[8] The *End-to-End Speech Benchmark (ESB)* [34] aims to evaluate ASR systems in various domains, eliminating the need for domain-specific adjustments. *ESB* consists of a range of speech datasets from various domains, including audiobooks, political speeches, educational talks, among others. Data instances are sourced from existing datasets such as *LibriSpeech, Common Voice, VoxPopuli, TED-LIUM, GigaSpeech, SPGISpeech, Earnings-22,* and *AMI.* The source datasets of ESB are freely available and accessible datasets to encourage broad participation and usage in the speech research community. Transcription artifacts, such as punctuation and casing, which are usually normalized in many ASR systems, are preserved in this benchmark to enhance the complexity and realism of speech recognition tasks. The diagnostic dataset with manually verified transcriptions is used for the public leaderboard available on the Hugging Face platform. [9]

2.3 Benchmarking of Automatic Speech Recognition Systems

This section presents a relevant work on the problem of evaluation of ASR systems. Popular methods, metrics, taxonomies, and analysis frameworks are discussed, along with known challenges and design considerations.

2.3.1 Introduction

The evaluation process involves a numerical measurement of the usefulness of the output generated automatically for a given Machine Learning task. In case of ASR, typically aSpeech dataset and WER metric are used to represent Machine Learning task as a specific Learning problem [72]. For example, the English ASR task can be assessed as a learning problem consisting of Librispeech Speech dataset and the metric WER[100]. The task of automatic recognition of Polish customer support conversations can be defined as the learning problem using the DiaBiz corpus and WER metric [112, 110]. The task of recognizing clean English speech defined using the Librispeech dataset reached the stage of benchmark saturation[148]. Furthermore, ASR systems can show on-par performance with humans on one set of Benchmark datasets and subpar accuracy across other set of use cases. As reported by Likhomenanko et al. “No single validation or test set from public

[8] https://huggingface.co/datasets/esb/datasets
[9] Open ASR Leaderboard

datasets is adequate to gauge transferability to other public datasets or to real-world audio data" [73]. ASR systems based on an end-to-end architecture could even generate incoherent output when tested on speech from a domain that was not present in the training data [55]. Furthermore, the error rates of contemporary ASR systems evaluated on popular datasets can be lower than those achieved by trained humans [152]. Given the limited transferability of the evaluation results between learning problems and datasets, Aksenova et al. [2] suggest that the ultimate objective of the ideal ASR benchmark should be to verify the capacity of the ASR system to generalize in a wide range of use cases. Methods for comparing systems or ASR technologies can be classified as subjective or objective [15]. Subjective methods involve humans in the evaluation process and are best suited to assess the impact of ASR recognition error and root cause [101, 58, 28] or validate the quality of the evaluation data [148]. Their drawback is the inconsistency in quality assessment by human subjects and the cost of applying at scale. Objective methods offer the advantage of generating reproducible results because they do not require human involvement. Their key benefit is automation, with the resulting lower cost and faster execution. However, effectively evaluating the practical usability of ASR output in the context of the target application remains a challenge due to the complexity of the processes involved [104, 137]. To decide which system offers the best performance, relying solely on accuracy metrics such as WER may not be enough. Additional metrics to be considered include latency (real-time factor RTF [127] or precision in the downstream task [129].

2.3.2 Overview of ASR benchmark design considerations

The following aspects have impact on the utility of ASR benchmark:

- scope of evaluated ASR systems,
- diversity of datasets and use scenarios,
- reliability of datasets,
- diversity of analysis dimensions,
- availability of evaluation results,
- reproducibility of evaluation results.

ASR systems, with their wide range of applications and tasks, should ideally be resilient to different types of speech input variation. For instance, an ASR system that generates automatic captions for video meetings should be capable of recognizing words from diverse semantic fields, adjusting to the meeting's subject. The characteristics of speech can also differ across various contexts: for instance, the style of speech used for dictating text messages is different from that of a group discussion, where participants might occasionally interrupt each other. Therefore, the benchmark can cover many '*horizontal*' and '*vertical*' challenges [2]. Horizontal challenges refer to ASR use cases, while vertical challenges refer to diversity of subjects, encoding formats, etc. The authors argue that "the more horizontal and vertical areas are covered by a benchmark, the more representative it will be, and hence it is more appropriate to measure ASR progress".[2] These challenges and related aspects are discussed in more detail in the following subsections.

2.3.3 ASR use scenarios

Ideally, the benchmark for ASR systems covers many ASR use cases. The best way to represent various usage scenarios is the creation of a comprehensive Speech dataset, either by merging existing datasets [73, 12] or by collecting new data to fill the gaps. Aksenova et al. [2] proposed a taxonomy of ASR use cases based on their experience developing an ASR-based customer-facing product at Google. The overview of the challenges and differences in the use cases can be found in tables 2.1 and 2.2, respectively.

Text dictation function is to enable the input of text into a digital device without manual typing. Typically, it involves relatively slow speech from a single speaker. As the user consciously interacts with a device, the speech is adjusted to maximize the chance of correct understanding [18]. Typical applications include general purpose dictation on desktop / mobile / portable devices, medical records transcription [78, 87], legal proceedings transcription[41, 23], language learning with computer-aided pronunciation feedback[82, 119] and speech-to-speech translation [134].

Voice search and control allow individuals to retrieve information or perform tasks through verbal commands. Speech patterns have *human-to-device* interaction characteristics and often contain specific nouns required to perform the task, for example, navigate to a location of interest or play a song on a streaming service. Another example is interactive voice response (IVR) applications, where individuals contacting customer service engage with a voice-operated chatbot. This chatbot can either assist in collecting data before transferring the call or be capable of addressing the problems on its own. [86]

Voicemails, oration, and audiobooks scenarios include using the ASR system to provide transcription for voicemail messages [48, 5], parliamentary speeches [65, 66, 143, 35, 107, 57, 62, 76, 51, 67, 133], and audiobooks [115, 100]. In these scenarios' speech typically originate from a single speaker. Spontaneity artifacts such as hesitations, fillers, back-channel speech, disfluencies, false starts, and corrections are present [37, 84]. In case of *audiobooks* the *human-to-human speech* features are less prevalent[50].

Conversations and meetings scenario typically involves transcribing spontaneous speech among several participants within a single audio recording. As with *voicemails, oration and audiobooks,* this type of speech is considered *human-to-human* speech. The presence of noise, overlapping, and distant speech adds to the challenge of recognizing spontaneous speech [54]. Practical applications include the transcription of video meetings [150] and customer-agent conversations [112, 113].

Podcasts, movies and TV podcasts scenario involve transcribing interviews or motion pictures to make them more accessible [17, 109, 111]. Such applications require ASR systems to be robust to non-speech audio like music and special effects. The challenge in recognizing multi-speaker *human-to-human* speech arises from the existence of fillers, overlapped dialogue, and interruptions. It is important to differentiate between movie subtitling and TV closed captioning. Subtitling is considered an '*offline*' task because the entire audio is accessible to the ASR system at the time of recognition, allowing for multiple iterations, including human post-editing. In contrast, closed captioning involves real-time processing of the audio stream under strict latency restrictions. Subtitles frequently include non-verbal cues that aid in understanding for those with hearing impairments, and they are designed for readability. Close captions are typically displayed in capital letters with fewer restrictions. [2]

2.3.4 Technical challenges

ASR applications also vary in other aspects such as the semantic content of the input speech (for instance, a lecture on physics vs. a phone conversation to arrange a medical

Aspect	**Description**	**Challenge**
Dictation	Slow speech, awareness of the interaction with the device, domain-specific jargon.	Application-specific vocabulary
Voice Search and Control	Short device queries, including proper nouns and specific tokens.	Application-specific vocabulary
Voicemails, Oration, Audiobooks	Spontaneous elements such as fillers, hesitations, and disfluencies.	Speech Variation
Conversations and Meetings	Spontaneous speech with challenges in transcribing overlapping speech.	Speech Variation
Podcasts, Movies, TV	Requires robustness to non-speech audio, with distinctions between subtitling and closed captioning.	Acoustic Environment

Table 2.1: ASR use scenarios overview. Inspired by work of Aksenova et al.[2]

Scenario	**Source and recipient**	**Interaction type**	**No of sources**	**Speech type**
Dictation	Human-to-device	monolog	Single	Spontaneous
Voice Search and Control	Human-to-device	dialog	Single	Spontaneous
Voicemails, Oration, Audiobooks	Human-to-human	monolog	Single	Read, Spontaneous
Conversations and Meetings	Human-to-human	dialog	Multiple	Spontaneous
Podcasts, Movies, TV	Human-to-human	dialog	Multiple	Spontaneous

Table 2.2: Types of sources, recipients, and modes for various ASR use scenarios. Inspired by work of Aksenova et al.[2]

appointment), the audio encoding format and the sample rate, among other factors. Ideally, a benchmark should take into account as many of these elements as possible. The aspects are summarized below and in Table 2.3

Terminology ASR systems, given their wide application across various fields, must have the ability to recognize a wide range of unique words. Ideally, datasets used for benchmarking encompass terms and phrases from a multitude of fields, such as medical terminology or historical phrases. ASR systems should also be adept at recognizing neologisms, despite the inherent challenge posed by their rapidly evolving nature and trending status. Special attention during measurements should be given to loanwords, as they often involve atypical grapheme-to-phoneme correspondences.

Speed The ideal benchmark should include samples with a variety of speech rates to assess the impact on recognition accuracy [132]. This is particularly relevant for paid services. In these services, users may be tempted to intentionally speed up recordings. They might also eliminate easily identifiable silence segments to reduce costs. Consequently, this can result in unnatural pitch shifts or sentence boundaries.

Acoustic Environment The context in which the audio input was recorded, be it a real-life or phone conversation, video call, or dictation, can significantly influence ASR performance. Ideally, datasets should be designed to gauge the robustness of an ASR system against background noise and other environmental variables [130]. The audio may contain content not intended for transcription, for example, background music.

Encoding Formats Finally, ASR accuracy can be affected by Audio encoding or Sampling rate [25, 108, 106]. Ideally, the datasets and the evaluation process should take these phenomena into account.

Aspect	**Description**	**Challenge**
Terminology and Phrases	Diversity of vocabulary across domains, including neologisms and domain-specific terminology.	Vocabulary
Speech Speed	Challenges presented by varying speech rates, requiring diverse samples.	Speech Variation
Acoustic Environment	Impact of recording environment and background noise on performance.	Noise
Encoding Formats	Effects of audio encodings and sample rates on recognition quality.	Diversity

Table 2.3: Vertical aspects of ASR challenges. Inspired by: [2]

Practical considerations

Aksenova et al. list the practical considerations when designing the ideal ASR benchmark.[2] The summary can be found in the following paragraphs and Table 2.4

Transcription Conventions Maintaining consistency in speech transcriptions can be a challenging task. Numerous issues arise in practical scenarios. There are multiple standards and conventions for speech transcription [128]. Non-speech events like hesitations or fillers can be transcribed in multiple ways. The same is true for proper names and slang, which often occur in various spellings and pronunciations. Therefore, the choice of a specific transcription convention should always consider the downstream task. For example, voice control, message dictation, or podcast transcriptions can exclude repetitions, disfluencies, and filler words. However, for subtitling or dialog systems, information about conversational clues can be highly relevant. To create or curate high-quality datasets, appropriate transcription conventions should be used according to established best practices [128], especially in multilingual [7] or medical settings [79]. Detecting and rectifying transcription errors is also important [148, 122]. Transcriptions can be expressed in a '*spoken domain*' or a '*written domain*' form. In former numbers are expressed as words, e.g., '*twenty-two'*, while in latter numbers are expressed as numbers '*22*'. Real-world ASR applications for readability or downstream use (e.g., for a natural language understanding system) can benefit from the fully-formatted, written-domain transcripts[97, 2]

Representativeness The primary consideration for representativeness is the similarity of test recordings to real-world audio signal and utterance domain characteristics. For example, when evaluating an ASR system's proficiency in processing speech amidst background noise, the noise level in the test sets should not exceed the threshold at which humans would find it challenging to accurately transcribe the audio.[2] Another factor accounting for the representation of reality for noise robustness evaluation is the impact of the Lombard effect [141]. datasets with artificially introducing noise do not account for it and therefore are less realistic than recordings collected under noisy conditions [106, 77, 58]. The secondary consideration is the set size (number of recordings). It should be large enough to ensure the proper predictive power of the error estimation metrics. Ideally, the benchmarking process should cover validation if performance differences are statistically significant.

Practical aspects	**Challenges**
Transcription format	Different conventions across datasets: ‘spoken-domain’ – *three thirty* ‘written-domain’ – *3:30*
Representativeness of test set	Sample must represent real-world case. Impact of noise level (Lombard effect). Sample size must be large enough to represent target group.

Table 2.4: Practical challenges of ASR evaluation process

2.3.5 Performance metrics

Overview

Within a practical deployment scenario, determining the ’*best*’ system often depends on a comprehensive analysis of various metrics, not just the average WER in all subsets of the ML evaluation set. A system may have a lower WER, but if it has a considerably higher latency, it might be less suitable for deployment, even if it has a lower WER score. Latency is typically defined as the average time delay from the completion of each spoken word to when it is outputted by the ASR system. The total latency encompasses all processes from the initiation of microphone activity to the final display of results, inclusive of network overhead and potential post-processing such as capitalization, punctuation, etc. A ’pure’ ASR latency metric disregards these factors and concentrates on the processing duration of the recognition engine. However, latency in relation to voice assistant commands might take into account the delay prior to the successful recognition of a command, which can occasionally occur before the completion of the spoken phrase. [2] Finally, given the limitations of WER, making an informed decision about the performance of downstream tasks requires additional human inspection or the use of more advanced metrics. The overview of the most popular metrics examples or categories can be found in Table 2.5.

Precision and Recall

Precision measures the proportion of correctly recognized words out of all words that the ASR system recognized, while recall measures the proportion of correctly recognized words out of all correct words in the reference.

$Precision = \frac{TruePositives}{TruePositives+FalsePositives}$

$Recall = \frac{TruePositives}{TruePositives+FalseNegatives}$

Metric	**Definition**	**Challenges**
Precision	Proportion of correctly recognized words out of all words that the ASR system recognized	Ignores context and syntax (word order). Over-penalizes ommission errors.
Recall	Proportion of correctly recognized words out of all correct words in the reference	Ignores context and syntax (word order). Ignores insertion errors.
F1-score	Harmonic mean of precision and recall.	Ignores context and syntax (word order).
Word Error Rate (WER)	Number of word-level operations (substitutions, insertions, and deletions) required to transform the ASR output into the reference transcript, divided by the total number of words in the reference.	Depends on transcription conventions. Does not reflect functional accuracy. Ignores semantics.
Character Error Rate (CER)	Number of character-level operations (substitutions, insertions, and deletions) required to transform the ASR output into the reference transcript, divided by the total number of characters in the reference.	Less intuitive for languages using words as base semantic unit. Ignores semantics. Does not reflect functional accuracy.
Sentence Error Rate (SER)	Number of sentences containing error divided by the total number of sentences considered during evaluation.	Over-penalization of small errors. Ignores semantics. Does not reflect functional accuracy.
Vocabulary recall	Number of correctly recognized domain-specific keywords.	Over-sensitive to non-canonical spelling of named-entities.
Real Time Factor (RTF)	Time required to process the audio divided by the audio length.	In case of cloud systems sensitive to network latency.
Latency	Time from the start/end of the recognition process to the availability to the downstream task.	Application specific definition of latency.
Semantic distance	Distance between a reference and hypothesis pair in a sentence-level embedding space.	Requires a pre-trained encoder.
Human annotation derived	Assessment of error severity, type, root cause etc. context-breaking, normalization issue, incorrect reference transcript.	Resource intensive.

Table 2.5: Metrics used for ASR evaluation

Both precision and recall provide bounds between 0 and 1 (0% to 100%), where 1 indicates perfect recognition. The balance between these two metrics is crucial, especially in applications like medical transcription, where both missing a critical term (low recall) and recognizing noncritical terms (low precision) have different implications for usability.

Precision and recall offer a nuanced view of ASR performance by highlighting the trade-offs between missing important words and incorrectly adding unneeded words. However, these metrics alone do not capture the overall efficiency of the system unless combined into a composite metric such as the F1 score, which is defined as the harmonic mean of precision and recall.

$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

An ASR system designed for medical transcription that achieves high precision, but low recall, might be preferred in scenarios where false positives (e.g., incorrect drug names) are more dangerous than false negatives (missed but not critical terms). However, this preference could reverse in a different application context, for example, meeting transcriptions, where the omission of specific words could be detrimental to understand the conversation.

Word Error Rate (WER)

WER is calculated as the sum of substitutions, insertions, and deletions divided by the total number of words in the reference text. $WER = (S+D+I)/N = (S+D+I)/(H+S+D)$ where:

- S – number of substitution errors
- D – number of deletion errors
- I – number of insertion errors
- H – number of hits (correctly recognized words).
- N – number of words in the reference transcription. Sum of hits, substitutions, and deletions.

WER can also be expressed as a percentage. $WER = (S + D + I)/N * 100$

WER metric is sensitive to the format of the reference transcription. The '*spoken-domain*' format has informal text that represents exactly what was pronounced by the speaker. The '*written-domain*' follows specific text formatting rules. For example, if the reference is ‘Set the timer for 00:02:30’ and the ASR outputs 'set an alarm for two minutes and thirty seconds’, the differences in format, punctuation, and capitalization are penalized by the WER metric, despite the fact that the ASR recognition result is correct.[137] Normalization of errors by the total number of words in the reference transcription can also lead to situations where WER exceeds 1 or 100%. This happens when there are more errors in the insertion than there are words in the reference. This characteristic of WER complicates its interpretation, making it challenging to compare systems with very high error rates. It is also nontrivial to understand the impact of different types of transcription errors on the final score, as each type of error (insertion, substitution, deletion) contributes equally to the error rates level.

The primary advantage of WER is its straightforward representation of the accuracy of the ASR with a single number. WER is relatively easy to calculate and commonly used, making it a standard metric in the field. The major drawback of WER is the over-penalization of insertion errors or sensitivity to text normalization differences[137]. The WER also does not reflect the varied significance of errors in different application contexts[148, 58]. In case of ASR for virtual assistants, high WER for recognition of phone numbers used by voice dialing app or places of interest used by navigation application is more detrimental than for less critical applications e.g., rarely used advanced camera settings. Furthermore, two ASR systems can have identical average WERs, but the first one may consistently not recognize numbers, while the second one has more evenly distributed errors across various token types. Although their average WERs are the same, the first system might be unsuitable for tasks requiring precise number recognition, demonstrating how the average WER might not fully capture the practical usability of ASR systems in specific contexts.

It is worth to note that the symmetrical metrics to WER called *Word Accuracy Rate (WAR)* or *Word Recognition Rate (WRR)* were also used in the pass[81, 131]. WAR and WRR metrics are calculated by subtracting the value of WER from 1. WAR $= 1 - -$WER WRR $= 1 - -$WER

Local WER and Variability Measures

Local WER refers to the WER calculated for specific recordings or recording segments, such as individual sentences or turns in a conversation. This approach helps identify variability in ASR performance in different types of content or speaking styles. Variability can also be quantified using statistical measures like standard deviation and percentiles to describe the distribution of WER scores between these segments. The standard deviation of the WER scores across segments provides insight into the consistency of the performance of an ASR system. The lower standard deviation indicates more reliable ASR behavior across different speech inputs. Percentiles, particularly higher percentiles like the 95th, show the upper bound of errors, highlighting the worst-case scenarios. Segmenting WER calculations allows for a more detailed analysis of where an ASR system performs well or poorly, enabling targeted improvements. However, this method requires a larger amount of annotated data to ensure statistical significance and could obscure overall performance trends if not interpreted carefully. If an ASR system shows a low average WER but a high 95th percentile WER, it indicates that while most of the system's outputs are correct, there are occasional but significant accuracy issues. This variability may be acceptable in general consumer applications but unacceptable in high-stakes environments such as emergency services or industrial device control.

Match Error Rate (MER) and Word Information Lost (WIL)

According to Morris et al. the more relevant performance metric for ASR than WER is the ratio of information conveyed correctly. The authors proposed two new metrics: MER (match error rate) and WIL (word information lost). MER is the proportion of I/O word matches that are errors. [90]

$MER = (S + D + I)/(H + S + D + I)$ where:

- S – number of substitution errors
- D – number of deletion errors
- I – number of insertion errors
- H – number of hits (correctly recognized words).

WIL is a simple approximation to the proportion of lost word information [90] $WIL = 1 - -WIP$

$WIP = \frac{H}{N_1} \cdot \frac{H}{N_2}$ where H is the number of hits (correctly recognized tokens), N1 is the length of the reference in tokens, and N2 is the length of the ASR hypothesis.

Both MER and WIL were designed to fall in the range from 0 to 1, where 0 indicates perfect performance and 1 indicates complete information loss. These designs address the drawback of WER, which is the lack of upper bound in case of a high number of insertion errors. The difference between the values of WER, MER, and WIL for various scenarios are presented in Table 2.6. Table explanation:

Reference	Output	H	S	D	I	%WER	%MER	%WIL
X	X	1	0	0	0	0	0	0
Xiii	XXYY	1	0	0	3	300	75	75
XYX	XZd	1	1	1	0	67	67	83
X	Y	0	1	0	0	100	100	100
Xi	YZ	0	1	0	1	200	100	100

Table 2.6: Differences between WER, MER and WIL values for different input/output combinations. Source: [90]

- Reference – Words to be recognized (ground truth). XYZ represent arbitrary words. Symbol i represents the insertion error.
- Output – Word(s) recognized by the ASR system. Symbol d represents deletion error
- H: Hits – Number of correctly recognized words.
- S: Substitutions – Number of words that were incorrectly changed.
- D: Deletions – Number of words that were omitted in the ASR output
- I: Insertions – Number of words that were added in the ASR output.
- %WER: Word Error Rate – The percentage of word-level errors in the ASR output compared to the reference.
- %MER: Match Error Rate – The percentage of word-level errors in the ASR output compared to the length of the ASR output.
- %WIL: Word Information Lost – The percentage of information lost due to errors.

Semantic-WER (SWER) Semantic Word Error Rate (SWER) is the evaluation metric to assess the usability of ASR in downstream tasks such as understanding spoken language and information retrieval proposed by Somnath Roy [125]. $SWER$ extends the traditional WER by incorporating semantic weights into the evaluation process. $SWER$ assigns weights to errors based on their semantic significance, differentiating between types of errors (substitutions, deletions, insertions) and their impact on the meaning of the transcript. For instance, errors involving named entities or key semantic words are weighted more heavily than other errors. This approach attempts to provide a more nuanced understanding of the accuracy of the ASR, reflecting how well the ASR output preserves the intent and meaning of the spoken input. $SWER$ integrates the concept of semantic distance using NLP techniques to calculate the similarity between words in the reference and hypothesis. It uses character error rate (CER) for specific cases like spelled-out entities, enhancing the metric's sensitivity to errors that affect understanding significantly. The calculation involves complex weighting schemes that consider both the type of word affected by an error and its semantic role in the discourse. This approach allows $SWER$ to better align with human judgements of transcript quality, particularly in tasks where semantic accuracy is crucial.

The primary advantage of $SWER$ is its ability to reflect the qualitative impact of ASR errors on the usability of transcripts in specific applications, which traditional WER cannot adequately capture. This makes $SWER$ particularly valuable for applications where the precise understanding of the content is more critical than the sheer precision of the transcription. However, the complexity of $SWER$, including the need for semantic analysis and custom weight settings for different applications, can be a drawback. It requires more computational resources and deeper linguistic analysis, which may not be feasible in all settings.

Semantic Distance (SemDist)

Semantic Distance (SemDist) proposed by Kim et al. utilizes advanced language models such as BERT, RoBERTa, and XLM to calculate the semantic similarity between a reference transcription and the ASR hypothesis. Unlike WER, which assesses the syntactic correctness by counting each error equally, *SemDist* assesses the semantic correctness of the outputs. This approach helps identify whether deviations from the reference significantly alter the conveyed message or the user's intent. Calculating *SemDist* involves extracting token-level embeddings from the speech content and comparing them using similarity metrics, which requires careful selection of the language model and embedding strategy. *SemDist* reflects a more nuanced understanding of linguistic variations, recognizing that not all deviations from a transcript are equally detrimental to understanding. This metric often uses similarity measures, such as cosine distance, to evaluate how closely the meanings of two sets of embeddings align, providing a quantitative measure of semantic accuracy. The main advantage of *SemDist* is its ability to correlate more effectively with user satisfaction and functional correctness in scenarios where semantic comprehension is critical. It is particularly useful in evaluating systems where the precise conveyance of information is more important than verbal accuracy. However, the complexity of its computation and the need for advanced NLP tools can be a limitation, especially in resource-constrained environments or languages with fewer computational resources. *SemDist* was shown to align closely with user judgments, effectively distinguishing between semantically critical and non-critical errors in more than 100,000 user-annotated ASR outputs. [56]

Stokke's Semantic Distance Metric

Stokke developed a semantic evaluation metric using Norwegian BERT models to compute the semantic distance between words in a transcript.[136] Similarly as *SemDist* metric measures the cosine distance between word embeddings, reflecting the semantic similarity perceived by humans. The challenge lies in accurately representing words with embedded words and ensuring that the distance calculation meaningfully corresponds to the semantic differences perceived by users. Using cosine similarity, Stokke's metric quantifies semantic similarity in a way that aims to mirror human perception. The cosine distance measures the cosine of the angle between two vectors, with a smaller angle (and a higher cosine value) indicating greater similarity. This approach provides a direct assessment of the semantic coherence between the reference and the ASR output.

The primary advantage of Stokke's semantic metric is its focus on semantic content, which can offer more detailed insights into ASR performance, especially in terms of user perception and usability. However, this method's reliance on high-quality, language-specific

BERT models can be a limitation, particularly for languages with limited NLP resources. Furthermore, the interpretation of cosine distances as measures of semantic similarity may not always align perfectly with human judgments, necessitating calibration and validation against human evaluations. In Stokke's study, the semantic metric was compared with both WER and human judgments using an online survey, showing that it is more closely aligned with human perceptions of semantic accuracy than WER. This validation underscores the potential of semantic metrics to provide more meaningful evaluations of ASR systems, especially in understanding the impact of ASR errors on user experience and task performance.

Manual Error Analysis – German language Wirth and Peinl performed a detailed error analysis for ASR for the German language [148]. The goal was to understand the qualitative aspects of model performance, which the WER metric alone is not able to provide. The analysis included root-cause analysis and categorization of errors based on their nature and impact. The error categories covered negligible, minor (non-context-breaking), major (context-breaking) errors, as well as errors related to proper names, homophones, flawed ground truth, and invalid audio. The authors provided examples of errors with significant semantic implications, such as incorrectly recognizing "waffeln" (waffles) as "waffen" (weapons), which completely changes the meaning of the sentence.

Manual Error Analysis – Korean language

Another recently introduced evaluation process and metric is *KEBAP (Korean Error Explainable Benchmark dataset for ASR and Post-processing)*[58]. The objective is to improve the accuracy and readability assessment of ASR systems. *KEBAP* framework differentiates speech- and text-level errors in ASR output. As a result, it provides a detailed analysis of ASR system performance in various noisy environments and speaker characteristics. The speech-level category includes 37 noise types and speaker characteristics, while the text-level category lists 13 types of textual errors. This granular categorization helps in pinpointing specific vulnerabilities of ASR systems, moving beyond the traditional metrics like WER and CER, which do not fully capture the nuances of ASR performance in real-world scenarios. Multiple types of error allow one to examine the variance in system response under different conditions. The authors also assessed how varying types of noise and textual errors correlate, providing insights into the robustness of ASR systems under different speaking conditions. The new method revealed the complexity of evaluating ASR systems, as traditional metrics might not fully reflect the user experience or the functional correctness of the transcriptions. The *KEBAP* framework enables better diagnostics and targeted improvements in ASR systems, but at the cost of complexity and extensive data annotation. Effective categorization requires a deep understanding of the types and sources of potential errors, which could be resource-intensive.

Real-Time Factor

RTF is calculated by dividing the time taken to process an input signal by the duration of that signal. [20]. $RTF = \frac{f(d)}{d}$, where f(d) is the time needed to process an input of duration d. The RTF is interpreted as follows:

- RTF $=$ 1: The system processes the input in real time. This means that it takes exactly as long to process the input as the input lasts. For example, a one-hour audio file is processed in one hour.

- RTF $>$ 1: Processing takes longer than the input duration. For example, an RTF of 2 means that a one-hour input takes two hours to process.

- RTF $<$ 1: The system processes the input faster than in real time. An RTF of 0.5 means that a one-hour input is processed in 30 minutes.

For streaming ASR systems, it is necessary to maintain an *RTF* less than one. Similarly to WER and latency, *RTF* samples create a distribution. Understanding its shape is crucial to determine the worst-case scenario. In practical settings, there is a trade-off between ASR system speed and quality. A broader space for hypotheses slows the search process, but enhances the chances of identifying the right hypothesis. In contrast, a smaller search space facilitates rapid decoding, but typically leads to higher WER. It is common practice to present an *RTF* vs. *WER* curve that illustrates all potential operating points, allowing for a balanced trade-off. [2]

Hallucination

Modern ASR models can *hallucinate* transcriptions i.e. provide random outputs for speech out-of-domain pr audio without any speech present [55, 71]. The susceptibility of an ASR system to these *hallucinations* can be assessed by testing it on datasets from domains not exposed during the training phase. Moreover, the use of *reject sets* is feasible, which incorporate various types of audio that should not be transcribed: such *reject sets*, for instance, might include various noises (e.g. *AudioSet*) [37], silence, speech in other languages, and so on [118, 88, 31]. A related topic is adversarial attacks, when a particular message is 'hidden' in audio in a way humans cannot hear, but may deceive ASR systems into transcribing in an unexpected way. [151, 47]

Summary

This subsection provides a summary of the different metrics used to evaluate the performance of the ASR system. The field of ASR evaluation is rapidly evolving. Traditional metrics such as WER might not align closely with human evaluations or performance in actual applications. Semantic-based methods are gaining popularity to overcome the limitations of commonly used metrics such as WER, CER, and MER. A similar shift has been observed in the Machine Translation (MT) field, where the BLEU metric, which is based on lexical information, has been replaced by the COMET metric. Similarly to the MT field, gaining insight into the root causes of ASR quality issues requires even more costly, annotation-based evaluation. Such a fine-grained quality assessment can accelerate the development of new ASR technologies that take advantage of phonetic and linguistic features more effectively. [98]

2.3.6 Evaluation results analysis

As proposed by Aksenova et al. '*the ideal benchmark for ASR systems would cover as many horizontals and verticals as possible and would involve various kinds of metrics beyond just WER'.*[2], The authors advocate for an analysis utilizing available demographic information about speakers, for example, measuring differences in recognition performance for different accents, age ranges, and gender. Since their call for action, numerous studies on measuring ASR bias along demographic characteristics were conducted for English [1, 39, 11, 74, 80, 92, 32], Portuguese [69], and Dutch [30]. The survey of accented speech evaluation was conducted by Hinsvark et al. [44]

Linguistic variations often corresponding to demographic characteristics include: • phonetic differences, how vowel realizations for specific accent • phonological differences, e.g., differences in phonemes across dialects of a language • lexical differences, e.g., region-specific terminology • voice quality differences, e.g., pitch differences correlated with parameters such as gender [111] and age [94]

The authors advocated following the example of Common Voice with respect to collecting and documenting speaker demographics metadata. They also recommend to practitioners to validate whether the sociolinguistic profile of the data used for training and evaluation matches the target user base. Linguists and creators of ASR systems should work together to determine and represent the most important linguistic aspects from a practical standpoint. Even if a large array of demographic metadata is available, organizing and interpreting the valuation results can be challenging. Therefore, the authors propose a "*metric-independent population-weighted visualization framework designed to evaluate ASR systems based on demographic metadata*". The core idea involves computing evaluation metrics for specific *slices* based on metalinguistic parameters. These *sliced metrics* help identify performance disparities between groups. Given that evaluation sets often do not perfectly represent the target user base, the authors recommend adjusting slice metrics using real-world population statistics to reflect actual demographic distributions, such as reweighting scores for male and female slices if the dataset predominantly contains male recordings.

To simplify the analysis, they suggest dividing all speakers into distinct groups based on particular linguistic or demographic characteristics. For example, the population can be divided into three distinct groups: group A (65% of the population), group B (30%) and group C (5%). The two subplots in Figure 2.3 show evaluations of two ASR models for these groups, where the WER scores are indicated by bar heights, and the bar widths show the size of the group. In the real test dataset, group A constitutes 80% of the data, while groups B and C each constitute 10%. Encoding the actual population distribution using the bar widths provides an intuitive understanding of how ASR systems handle linguistic diversity across the target group. Adjusted weights can also be used to calculate single-

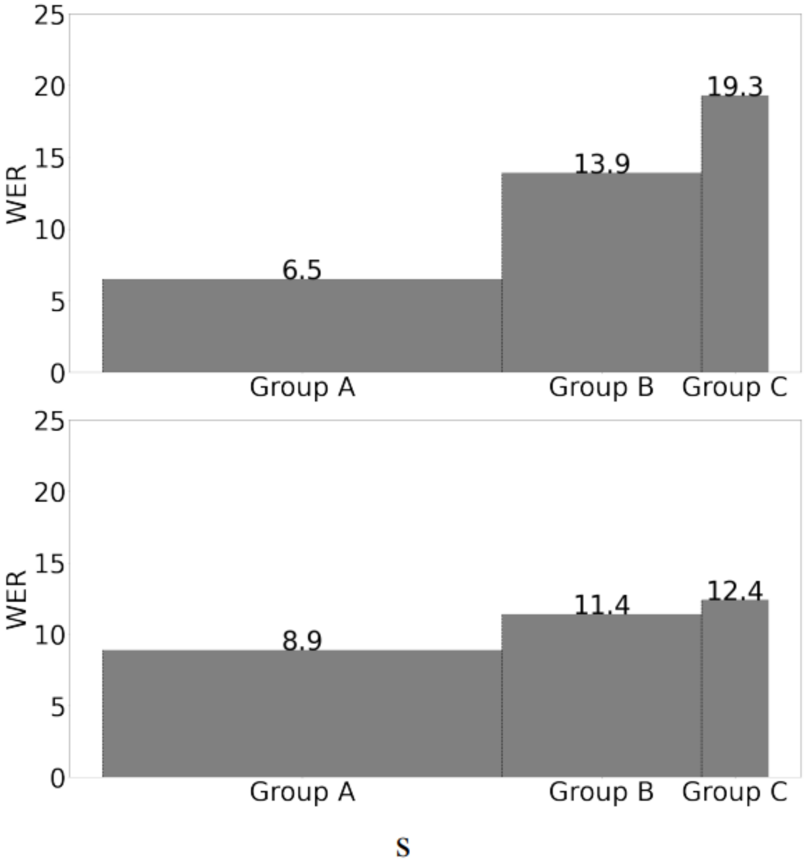


Figure 2.3: Examples of WER sliced into groups A, B, and C, with the width of the bars reflecting relative sizes of those groups. Source: [2]

weighted performance scores, which are more representative of actual ASR performance given the demographics of the target group. Authors argument that also the disparity of the ASR performance across various groups should be calculated. The systems with WER scores for 3 groups shown in Figure 2.3 have the same average WER, but the system shown in subplot 2 is clearly more consistent. The system in the bottom subplot shows 3.5 absolute points difference between the worst and best groups, while the top one has 12.8 absolute points. Slicing can be based on just a single parameter or on parameter intersections. Naturally, the more groups are intersected, the more challenging it is to fill every bucket with enough samples to obtain solid statistics and to control for all other variables not considered. The authors proposed a non-exhaustive list of representative and mutually exclusive slices to be considered when benchmarking ASR systems. The categories are presented in Table 2.7

Analysis dimension	**Scope**
Regional language variation	Impact of phonological, lexical and syntactic differences for various dialects.
Sociolects	Impact of phonological, lexical and syntactic differences for various social groups.
L2 background	Impact of L1 language characteristics on L2 speech recognition.
Gender, age, and pitch	Impact of voice pitch, speed, pronunciation clarity.

Table 2.7: Evaluation results analysis dimensions

2.4 ASR speech datasets management methods and tools

2.4.1 Introduction

Speech datasets are the cornerstone of the research and development of ASR (Automatic Speech Recognition) systems. These datasets typically comprise digital speech recordings, annotations, metadata, and documentation. The purpose of speech datasets includes training the ASR system or its various components (such as language models, acoustic models, and post-processing modules) and facilitating quality evaluation.[147]

Typically, datasets are segmented into at least two subsets: one for training and the other for testing. The training set is used to fine-tune the model parameters of the system. In contrast, the testing set evaluates the system's performance after training. Additionally, a validation split is often provided as a third partition or derived from the training set for hyperparameter tuning during training. [105]

Maintaining a clear separation between training and testing data is vital to ensure unbiased evaluation of system performance and prevent leakage of test data[72]. This separation helps to validate the model's ability to generalize to new, unseen data, which is a crucial factor in the development of effective ASR systems and the main objective of the evaluation process [2]

The success of developing or evaluating an ASR system depends on the dataset that accurately reflect the essential characteristics needed for the task. It is critical because irrelevant speech and language variations within the dataset can negatively affect the accuracy of recognition or the ecological validity of evaluations.[72] For example, when creating a system designed to recognize isolated words on mobile phones [147], it is important to use a corpus consisting of mobile speech recordings of the words to be recognized. In contrast, a general (large-vocabulary) ASR system requires training on a diverse set of speech samples to effectively cover a wide range of real-life speech scenarios. [3, 43, 13, 33, 115] The larger the volume of utterances with varied characteristics (e.g. recording conditions, speaker gender, age, and accents), the more robust the resulting ASR system [75, 114]. To effectively select data for training or evaluation, detailed and precise metadata are required. This includes transcription and annotation guidelines, as well as information about a speaker, such as native language, dialect, age, gender, and level of education. Documentation of recording conditions, noise sources, and recording equipment further increases the practical utility of the dataset for evaluation purposes.

2.4.2 ASR speech dataset lifecycle

The ASR speech dataset lifecycle typically involves four main stages:

1. **Collection stage**: Acquisition of recordings in digital format.
2. **Annotation stage**: Transcription of speech data, annotating duration of speech segments, or tagging non-speech events such as background noise or laughter.
3. **Curation stage**: Preparation of the dataset for release, for example, quality validation, normalization, and documentation.
4. **Release stage**: Making metadata and dataset available to the target audience

Iterative refinement of the dataset prior and post-release is achieved by adding feedback loops to the process. The typical feedback loops during the dataset production are as follows:

1. Quality issues are identified during annotation and triggers additional data collection.
2. Quality issues are detected during the curation or after a dataset release:
3. Quality issues detected during the release stage:
   (a) If issues concern dataset content, an additional annotation is performed.
   (b) If issues concern structure, documentation, or legal aspects, additional curation is performed.

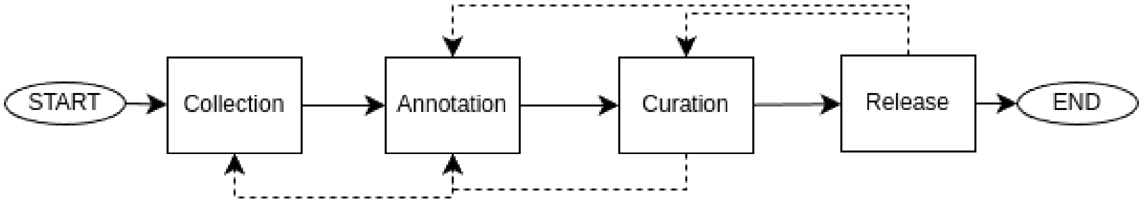


Figure 2.4: ASR speech dataset lifecycle

2.4.3 Overview of the ASR dataset management methods

Stage	**Methods**
Collection	Controlled recording sessions Crowdsourcing speech collection Public and archival sources Speech generation
Annotation	Manual annotation Automated annotation Human in a loop annotation
Curation	Validation and quality control Data augmentation Standardization and formatting Metadata and documentation Legal and ethical verification
Release	Proprietary or public repositories Data user engagement

Table 2.8: Stages and methods of speech data management

Collection stage:

- **Controlled speech recordings:** Speech recordings are obtained in regulated settings by speakers recruited to capture particular dialects, accents, or styles. This method is expensive and less scalable, but quality can be strictly controlled.

- **Crowdsourcing speech recordings:** Speech recordings are collected from a wide range of participants representing different demographic and linguistic backgrounds, increasing the diversity of the dataset. As both the recordings and quality control are often performed by volunteers, the quality may be worse than in controlled settings.

- **Public and archival sources:** Existing speech databases and public domain audio are used, including oratory speeches, broadcasts, or recordings of historical events. This method can be time- and cost-effective; however, the quality of the speech data strongly depends on the source.

- **Speech generation:** Synthetic speech data is created using Text to Speech (TTS) (Text-To-Speech) engines to increase the volume of the dataset. This method is cost-effective and time-effective, but the quality of the resulting recordings depends on the quality of the TTS engine and the vocabulary domain.

Annotation stage

- **Manual annotation:** Linguists or trained annotators accurately transcribe spoken words and may annotate specific linguistic or acoustic characteristics.
- **Automated annotation:** Initial transcriptions and annotations are generated by ML-based engines. Transcription can be generated from speech using an ASR system, or existing transcriptions can be forced-aligned to the specific speech fragments.
- **Human-in-a-loop annotation:** Automatically enriched data are reviewed and corrected by human annotators.

Curation stage:

- **Validation and quality control:** Validation processes are implemented to check the accuracy of transcriptions and annotations, involving manual reviews or automated checks.
- **Data augmentation:** The dataset is enhanced by altering existing recordings to improve the robustness of ASR models.[103]
- **Standardization and formatting:** Data is formatted consistently to ensure compatibility with common ASR tools, involving standard file formats and data structures. [34]
- **Metadata and documentation:** Detailed metadata about speakers and recording conditions, along with comprehensive documentation, are provided to facilitate the use of the dataset.[36, 8, 149, 116]
- **Legal and ethical verification:** Addresses copyright, privacy, and ethical concerns, ensuring data collection and usage comply with laws and guidelines.[102]

Release stage:

- **Proprietary or public repositories**: datasets designed for open access are typically indexed in public catalogs[53] or published on open data platforms like *Hugging Face datasets*[70]. datasets intended for commercial use are listed in public catalogs [16], while those created for proprietary purposes are shared by internal means of the organization that owns the data.
- **Data users engagement**: Feedback from dataset users is collected to iteratively refine and improve dataset quality and relevance.

2.4.4 Challenges related to ASR speech datasets management

Numerous challenges can arise when curating and using ML datasets. This section summarizes the most frequently occurring issues when using or curating open and commercial ASR speech datasets.

Open source datasets

- **Quality and diversity of data**: Open source datasets are often collected from multiple sources, which means that the quality and diversity of the data can vary significantly. Some datasets may be carefully annotated and diverse, while others may be noisy, biased, or incomplete. [93]
- **Lack of standardization**: Open-source datasets are typically not standardized, which means that they can use different data formats, annotation schemes, and evaluation metrics. This can make it difficult to compare results across different datasets or integrate data from multiple sources.
- **Data size**: ASR models require large amounts of training data to achieve high accuracy, and many open source datasets may not be large enough to train high-performance models for some applications.
- **Legal and ethical considerations**: Open-source datasets may contain sensitive or personal information, and their use may be subject to legal or ethical restrictions. Users must comply with relevant laws and ethical guidelines when using these datasets.

- **Data bias**: Open source datasets may be biased towards certain demographic groups or speech patterns, which can impact the performance of ASR models resulting from the availability of subjects, funding, or existing speech material collected for other purposes. Users must be aware of the potential bias in the data and take steps to address it.

- **Data preprocessing**: The audio data in open source datasets may require significant pre-processing before they can be used for ASR. This includes tasks such as audio normalization, noise reduction, and audio segmentation.

Commercial datasets: Using commercial ASR datasets also presents several challenges, including the following:

- **Cost**: Commercial datasets can be expensive to acquire, especially for smaller research teams or organizations with limited budgets.

- **Data bias**: Commercial datasets may also be biased towards certain demographics or speech patterns.

- **Quality and diversity of data**: Commercial datasets can be accurate and diverse, while others may be incomplete or noisy. Careful inspection of dataset samples is required prior to purchase.

- **Legal and ethical considerations**: Commercial datasets may be subject to legal restrictions. A detailed inspection of the terms and conditions is required prior to purchasing.

- **Compatibility**: Commercial datasets come in different data formats and annotation schemes, making them difficult to integrate with other datasets without custom pre-processing, e.g., audio normalization, transcription normalization, audio segmentation, etc.

2.5 ASR speech datasets and benchmarks for Polish

2.5.1 ASR speech datasets for Polish

The number of speech datasets available for Polish ASR continues to grow. The major projects include initiatives such as Mozilla Common Voice [3], FAIR's Multilingual Librispeech [115], VoxPopuli [146], EU [14] and Polish Parliament[65]. The most recent Polish ASR speech survey in 2019 identified eight datasets and 226 hours of transcribed speech [142]. This amount may not be sufficient to train a robust transformer-based ASR system. However, more than 200 hours of speech is reported to be enough to fine-tune a model for a resource rich language such as English through cross-language transfer learning [85], as well as to train the ASR system based on HMM [138, 142] or based on WFTS [89], or from scratch [65]. Although Polish speech datasets are available publicly, they often remain underutilized for ASR evaluation purposes. Publicly reported ASR research experiments for Polish typically used datasets created by local institutions involved in research [154, 60, 61, 65]. A detailed survey and analysis were performed as part of this study. The methodology of the surveys is explained in Sections 3.4 and 3.2. The results are presented in Sections 4.3.2 and 4.1.

2.5.2 ASR speech benchmarks for Polish

This section presents ASR benchmarks for the Polish language reported in the public domain to-date:

- BOR (*BOR POLSL PS 18*) [99]
- PolEval 19 ASR challenge (PolEval PJATK 19) [59]
- DiaBiz commercial ASR systems benchmark [112]
- Medical PG [153]
- Medical PŚ [68]

Detailed survey and analysis of ASR benchmarks for Polish was performed for the purpose of this thesis. Details about methodology can be found in Section 3.4, while results in Section 4.3.

Benchmark	**Year**	**Models evaluated**	**Best model**	**Best model WER**	**Observations**
BOR POLSL PS 18	2018	ARM, Skrybot, Google	Google	clean-50% noisy-90%	Tested systems are not accurate enough to be used in training of government agents.
PolEval PJATK 19	2019	GOLEM, ARM-1, SGMM2, tri2a, clarin-pl-studio, clarin-pl-sejm	GOLEM	11.80%	– All systems, except ARM-1, based on Kaldi – All systems, except for clarin-pl, using GMM models – Fixed systems were the only using in-domain data
DiaBiz CLARIN Voicelab 22	2022	Azure, Google, Voicelab	Azure	10.50%	Azure achieved the best results (10.51 WER for both channels), followed by Voicelab's ASR (11.51 WER). Google's Polish ASR performed worse on the DiaBiz dataset (20.84 WER). Azure outperforms other ASRs in 8 of 9 domains. Voicelab's results are slightly better for telecommunications customer support dialogs.
SpokesBiz CLARIN 23	2023	Whisper (large)	Whisper	20%	Whisper accuracy varies significantly from official evaluations on CommonVoice and FLEURS. Recording quality and vocabulary domain greatly affect WER (15.2% – 26%).
Medical UW SOVVA PS 23	2023	Azure, Google, Techmo	Google	14%	All three ASR systems showed over 86% accuracy, with only a 1.7% difference between the best and worst results.
Medical PG 23	2023	Azure, Google, Whisper (large-v2)	Azure	56%	Tested models are not suitable for voice-filling medical records, case descriptions, or treatment prescriptions due to high error rates (WER 56%, CER 16%).

2.6 Overview of tools for dataset management and ASR evaluation

2.6.1 ASR speech datasets management tools

This section describes the most frequently used general or ASR-specific data management tools accessible under open licenses.

- **pandas** [83] is an open-source Python library that provides high-performance data manipulation and analysis tools. The objective of Pandas is to simplify the handling of data structures such as SQL tables, Excel, or text files, spanning from tabular data with different types of columns to time series and labeled matrices. The library gas two core data structures, the Series for one-dimensional data, and the DataFrame for two-dimensional data. *Pandas* excels in various data operations, such as managing missing data, modifying the size of data structures, aligning data based on labels, and grouping data for analysis. It also simplifies the conversion of heterogeneous data forms into DataFrame objects, provides easy data slicing, indexing, concatenation, reshaping, and data fields renaming. Available under BSD 3-Clause license.

- **The Hugging Face datasets** [70] is a Python library designed to simplify data handling in ML projects. Its main benefit is the extensive support for public datasets in different formats and languages, which allows users to load the dataset with just one line of code. The library is also compatible with popular ML frameworks like *Numpy*, *Pandas*, *PyTorch*, *TensorFlow*, and *JAX. datasets* library facilitate efficient data preparation thanks to standardized data pre-processing tools that can handle datasets in various file formats. Furthermore, it simplifies the sharing of new datasets using the *HF datasets hub* [10]. Advanced library functionalities include:

  - handling large datasets beyond RAM capacity through memory-mapping,
  - smart caching to avoid redundant processing,
  - compatibility with different data types, including audio and image
  - streaming mode for efficient use of disk space and immediate data iteration.

[10] https://huggingface.co/datasets

- **Speech Data Explorer (SDE)** [6] is a tool for the exploration and analysis of speech datasets.[11] SDE was created by the NVIDIA team responsible for the development of the ASR system and the NLP framework NeMo toolkit.[12] Researchers used SDE to investigate errors and fine-tune the process of constructing a speech dataset using the Forced alignment technique. The main features of SDE are:
  - calculating dataset statistics e.g., number of recordings, alphabet, vocabulary, duration-based histograms
  - dataset exploration with interactive data-tables for filtering and sorting
  - audio data inspection tools e.g., waveforms, spectrograms, audio playback
  - transcriptions and hypotheses analysis tools e.g., ASR accuracy metrics, alignments
  - audio signal measurements e.g., encoding, amplitude, spectrum

Summary information on tools for the management of ASR speech datasets is provided in Table 2.9.

Tool	**Language**	**Features**	**License**
Pandas	Python	support for wide range of data formats and types, comprehensive tools for data manipulation and analysis	BSD 3-Clause
Hugging Face Datasets	Python	dataloaders for public datasets, datasets hub, handling large datasets through memory-mapping, smart caching and streaming, multimedia data types support	Apache 2.0
SDE	Python	dataset statistics, dataset exploration tools, audio data inspection tools, transcriptions and hypotheses analysis tools, audio signal measurements	Apache 2.0

Table 2.9: Tools for ASR datasets management

2.6.2 ASR evaluation tools

This section outlines the most commonly used tools for the evaluation of ASR systems, which are available under permissive open-source licenses.

[11] SDE User Guide

[12] NVIDIA NeMo ASR toolkit

- **sclite**: Developed by the National Institute of Standards and Technologies (NIST), written in C, this tool uses the WER as its primary metric. Its features include speaker-level statistics, identification of commonly misrecognized words, and the ability to count hits, insertions, deletions, and substitutions. It also provides alignment capabilities. The software is available on *GitHub* and falls under NIST's software license.

- **jiwer**: A product of Jitsi, implemented in Python, JIWER calculates WER, along with Character Error Rate (CER), Match Error Rate (MER) and Word Information Lost (WIL) It supports aligning hypothesis and reference, as well as native support for text normalization transformations. The library is hosted on GitHub and released under the Apache 2.0 license.

- **asr-evaluation**: Created by Ben Lambert and also in Python, this tool measures WER, the word recognition rate WRR, and the sentence error rate SER). It can handle simple normalization, removal of empty utterances, and calculation of the WER relative to the reference length. In addition, it generates confusion tables. Available on GitHub, *asr-evaluation* is licensed under Apache 2.0.

- **fstalign**: Developed by Rev and written in Python/C++, *fstalign* assesses WER and supports multiple input formats such as CTM, NLP, FST, and CSV. It natively supports text normalization and synonym handling and provides detailed error analysis based on metadata (WER tags) in NLP format. This tool is available on *GitHub* under the Apache 2.0 license.

- **evaluate**: From Hugging Face and built with Python/C++, this tool focuses on WER and is integrated with the Hugging Face *datasets* and *transformers*[13] libraries, enhancing its utility for users in the Hugging Face ecosystem. It can be found on *GitHub*, with an Apache 2.0 license.

- **asr-evaluator** ASR evaluation tool from the NVIDIA's NeMo toolkit toolkit, with the following features:

  - On-the-fly data augmentation for ASR robustness evaluation.

[13] https://huggingface.co/docs/transformers/index

- Analysis of insertion, deletion, and substitution error rates.
- Reliability assessment across metadata available, e.g. gender, audio length, etc.

Detail information on tools for the evaluation of ASR systems is provided in Table 2.10.

Tool	Author	Language	Metric(s)	Features	License
sclite	NIST	C	WER	Speaker-level statistics. Frequent recognition errors. Hits, insertions, deletions, substitutions. Alignments.	NIST software
jiwer	Jitsi	Python	WER, CER, MER, WIL, WIP	Alignments. Supports text normalization. Python and CLI interface.	Apache 2.0
asr-evaluation	Ben Lambert	Python	WER, WRR, SER	Remove empty references. Lowercase text. WER by reference length. Confusion tables.	Apache 2.0
fstalign	Rev	Python, C++	WER	Supports CTM, NLP, FST, CSV formats. Native text normalization and synonym handling. WER-based error analysis in NLP format.	Apache 2.0
evaluate	Hugging Face	Python/C++	WER	Integration with Hugging Face Datasets and Transformers libraries.	Apache 2.0
asr-evaluator	NVidia	Python	WER, CER	Integration with NeMo toolkit ASR models and SDE tool.	Apache 2.0

Table 2.10: Tools for ASR evaluation

Chapter 3

Methodology

3.1 Overview

The study aimed to develop a framework and resources to benchmark Polish ASR systems based on publicly available datasets. This section outlines the methods used to design, create, and publish research artifacts. Initially, a survey of existing speech data was created. The selected datasets were then consolidated intoBIGOS and PELCRA for BIGOS datasets. The next step was the evaluation of commercial and freely available ASR systems for Polish. The results were shared as publicly available AMU ASR Leaderboard. Finally, the curated dataset was published, inviting the Polish ASR community to participate in an open challenge. The research framework is shown in Fig. 3.1.

3.2 RO1: Survey of ASR speech datasets for Polish

3.2.1 Research objectives and questions

The first research objective was to survey the available Polish ASR speech datasets. The following research questions (RQ) were considered:

- **RQ 1:** How to systematically categorize Polish ASR speech datasets using public information?
- **RQ 2:** What is the current state of Polish ASR speech datasets?
- **RQ 3:** How can the survey findings be shared for community feedback?

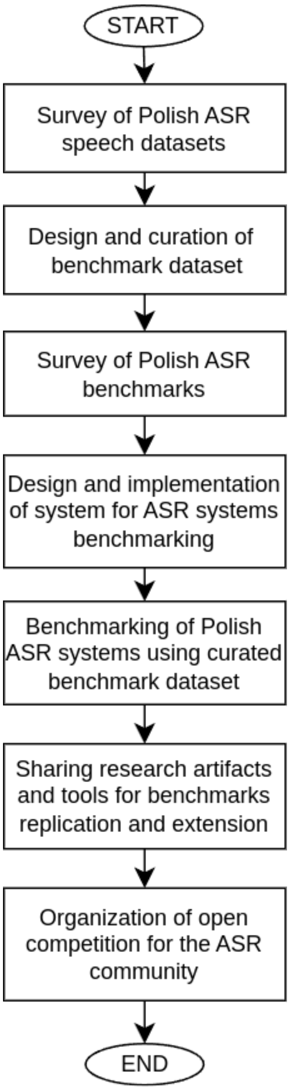

START
Survey of Polish ASR speech datasets
Design and curation of benchmark dataset
Survey of Polish ASR benchmarks
Design and implementation of system for ASR systems benchmarking
Benchmarking of Polish ASR systems using curated benchmark dataset
Sharing research artifacts and tools for benchmarks replication and extension
Organization of open competition for the ASR community
END

Figure 3.1: Overall research framework

3.2.2 Research methodology

The research method comprised of:

- Literature search to identify existing Polish ASR speech datasets
- Development of a taxonomy covering key dataset features
- Cataloging of speech datasets according to the taxonomy framework.
- Developing a publicly accessible digital repository with a data catalog and survey results.

Overview of methodology

A keyword-based literature review process [124] was adopted to identify and document relevant datasets. The information about the datasets was analyzed and manually annotated. The scope of the annotation was refined iteratively. The final methodology of the survey consisted of the following steps:

1. Conduct a keyword search in relevant sources.
2. Perform manual analysis and annotation of the available documentation.
3. If there are multiple sources of documentation, cross-check the information from all sources to ensure consistency and accuracy.
4. If the dataset is downloadable, validate the documentation and analyze the dataset content for further information.
5. Analyze collected metadata to derive insights about the state of Polish ASR speech datasets
6. Make the catalog and extracted insights publicly available

The initial search and annotation process took place between March and May 2022. The cross-check and validation were concluded in August 2022. New datasets were included in the catalog and survey in February and July 2023.[1]

Information sources

Language data repositories

Repositories can be thought of as *libraries for linguistic data and tools.* Several global organizations aim to facilitate the efficient distribution of language resources, including Polish ASR datasets. For instance, the US-based LDC Consortium manages the LDC Data Catalog, while the European Language Resources Association (ELRA) supports initiatives such as the META-SHARE repository. International cross-institutional initiatives, such as CLARIN ERIC and its Virtual Language Observatory, also contribute to this effort.

In Poland, cross-institutional language data repositories include Dspace[2] by CLARIN-PL and Open Science Resource Atlas 2.0 (AZON). Single-institution-based repositories,

[1] Polish ASR speech data catalog changelog

[2] `https://clarin-pl.eu/dspace/`

such as the University of Łódź (PELCRA research group)[3], also play a role. Furthermore, two Polish speech datasets suitable for ASR-related usage can be found at *Hamburger Zentrum für Sprachkorpora (HZSK)*[4] in Germany.

Other sources Additional sources of information on ASR datasets come from a wide variety of channels. These include, but are not limited to:

- Individual authors' web pages[5]
- Open challenges such as PolEval[6]
- Various reports, publications, conference proceedings, and technical papers were discovered through targeted keyword searches such as "*Polish*", "*ASR*", "*speech corpus*", "*dataset*" and others, as well as general web searches.

Selected datasets, such as PELCRA EMI and CLARIN PL Parliament, are also available on popular data sharing platforms such as Kaggle[7] and Hugging Face[8], thanks to individual contributions from researchers and enthusiasts.

ASR datasets taxonomy

The final dataset card taxonomy consists of 66 attributes. The selected attributes correspond to the metadata fields commonly encountered during the survey, state-of-the-art recommendations for comprehensive benchmarking of ASR processes [2], and the authors' experience in managing ASR speech datasets. The taxonomy covers useful aspects of the ASR dataset's lifecycle, such as the creator, funding institution, license, publication date, quality assurance process, and more. It also includes content characteristics such as audio file format, number of recorded speakers, metadata distributions, etc. The complete taxonomy is presented in the Appendix 7.1.1.

Data annotation and catalog curation

The data annotation and catalog curation process was carried out using the spread sheeting tool, with taxonomy metadata fields defined as columns.

[3] `http://pelcra.pl/new/tools_and_resources`

[4] `https://www.slm.uni-hamburg.de/hzsk.html`

[5] `https://www.ii.pwr.edu.pl/~sas/ASR/`

[6] `http://poleval.pl/`

[7] `https://www.kaggle.com/`

[8] https://huggingface.co/

The manual annotation process consisted of the following steps:

1. For each identified Polish ASR speech dataset, examine the provided documentation.
2. Assess the feasibility of downloading the dataset.
3. If the dataset can be downloaded, review any references.
4. Fill in the metadata entries in the spreadsheet according to the information examined.
5. If a specific attribute of the dataset is not mentioned, insert *"no info"* into the corresponding cell.
6. If the attribute of the reported dataset is not included in the taxonomy, add a new column to the spreadsheet and fill in the values for the datasets already analyzed.
7. If a discrepancy is observed between the metadata provided by the resource link and the information in the published article, highlight the conflict using color for further verification.
8. If a dataset is downloadable:
    (a) Examine any embedded documentation, such as README files.
    (b) Manually inspect the dataset's content and update or supplement the relevant metadata, such as the number of recordings and the audio file encoding format.
9. If a contact point to the author or publisher is provided:
    (a) Request information on missing metadata attributes.
    (b) (in case the dataset is not downloadable) Request a sample of the dataset.

Developing a publicly accessible digital repository and dashboard

The last step was the development of publicly accessible repositories for catalog and survey results. Scientific journals [53] and platforms (GitHub[9], Hugging Face[10]), which are popular among the Polish NLP community, were used to distribute research results.

[9] Polish ASR speech data survey – GitHub

[10] Polish ASR speech data survey – Hugging Face

3.3 RO2: Design and curation of ASR benchmark dataset for Polish

3.3.1 Research objectives and questions

The goal was to establish a comprehensive dataset to evaluate Polish ASR systems. Taking into account the suggestions of Aksenova et al. (2021) [2], the main objective was to build a dataset that covers a wide range of usage scenarios and demographic categories. Given that many publicly available datasets are used for ASR benchmarking purposes [117, 3, 115], the secondary objective was to streamline the laborious task associated with the processing of datasets of different origin.

The following research questions were addressed:

- **RQ 4**: What factors are crucial in designing and curating an ASR benchmark dataset?
- **RQ 5**: What steps are needed to curate a benchmark dataset from public resources?
- **RQ 6**: Which public Polish speech datasets can be used for benchmarks?
- **RQ 7**: How can the benchmark dataset be shared for ASR community feedback?

3.3.2 Research methodology

The method consists of using the information collected in Polish ASR speech data catalog to compile and organize a diverse dataset in a standardized format that meets quality control criteria and is easily accessible to the ASR community.

Specific activities include:

- Selection of speech datasets based on the information collected in RO1.
- Data cleaning, normalization, and formatting for consistent evaluation.
- Developing a publicly accessible digital repository and dashboard with analysis results.

Design considerations

Below are the main requirements for designing a benchmark dataset.

- **Task appropriate:** Relevant and practical for the intended ASR task, with clearly outlined limitations in covering typical, edge-case scenarios.

- **Accessible:** Available online under a license that allows the free use and creation of derivative works.

- **Discoverable:** Easy to find and acquire (without time-consuming registration or other access barriers).

- **Diverse and challenging:** Containing various examples to test the adaptability of the model, as well as complex cases to encourage community participation and minimize the risk of benchmark saturation.

- **Annotated**: The metadata of speakers and recordings contains information that allows nuanced analysis and interpretation of the results.

- **Optimally sized:** Large enough to be representative, but manageable to download and explore.

- **Clean yet realistic:** Free of major errors, but noisy enough to represent the complexity of the real world.

- **Well-documented:** Provided with documentation that is understandable to users without technical skills.

- **Well-explained:** Provided with evaluation baselines and how-to-use script examples.

Leveraging speech data catalog for sourcing open datasets

Creating a comprehensive benchmark dataset for Polish ASR systems from publicly available datasets required careful consideration of the attributes of the source datasets. Polish ASR speech dataset catalog was used to extract the information required for selection. This section describes this process in detail.

The attributes available in Polish ASR speech data catalog were classified into three categories, depending on their level of relevance and impact on the utility of the dataset for the evaluation of ASR:

- **Mandatory** — attributes representing criteria that are essential for the dataset to be considered usable for ASR evaluation, e.g., license, data availability, transcriptions availability. These attributes are typically sourced from the original documentation and it may not be feasible to determine them correctly otherwise.

- **Optional** — attributes which enhance the utility of datasets but are not strictly mandatory, e.g., information about transcription protocol, quality assurance practices, recording sampling rate. These attributes are also derived from the original documentation. If the author does not provide information on a specific attribute, it might not be possible to curate the missing information, either automatically or manually. The exception is the audio format and sampling rate, which can be determined automatically but only under the assumption that the format of the distributed audio files is the same as the original recordings.

- **Task-specific** — attributes that determine the suitability of the dataset for a specific evaluation task, for example, measurement of accuracy bias by speaker gender, age, nativity, and accent; device bias, read vs. conversational speech bias, etc. These attributes are sourced from the original documentation or dataset contents (metadata). If not available, selected characteristics can be automatically recovered at varying quality levels. The cost of manual curation on a large scale can be prohibitive. For example, the age or gender of the speaker can be determined based on audio characteristics with acceptable precision for the subsequent sociodemographic analysis of the results of the ASR evaluation. However, determining the recording device or speaker nativity of a short speech sample may not be feasible.

The detailed list of the attributes considered is presented in Table 3.1

The BIGOSBenchmark dataset was created from the datasets that met the following criteria:

- The dataset was downloadable.

- The license allowed free use for non-commercial purposes.

- The transcriptions were available and aligned with the recordings.

- The sampling rate of the audio recordings was at least 8 kHz.

Attribute	**Type**	**Info source**
Availability	Mandatory	Documentation
License	Mandatory	Documentation
Sampling Rate $>$ 8 kHz	Mandatory	Documentation
Bits Per Sample $>$ 16 bits	Mandatory	Documentation
Test/Dev/Train Split	Optional	Documentation
Standard Audio Encoding Format	Optional	Documentation
Transcription Availability	Mandatory	Documentation
Lossless Original Recording	Optional	Documentation
UTF8 Text Encoding	Optional	Content inspection
Text Normalization	Optional	Content inspection
Transcription Accuracy	Optional	Content inspection
Annotation Accuracy	Optional	Content inspection
Domain Specific Vocabulary	Optional	Documentation
Type of Speech	Optional	Documentation
Annotation Audio Segmentation	Optional	Documentation
Annotation Speaker Gender	Task specific	Documentation
Annotation Speaker Age	Task specific	Documentation
Annotation Speaker Nativity	Task specific	Documentation
Annotation Speaker Accent	Task specific	Documentation
Annotation Audio Device	Task specific	Documentation
Annotation Acoustic Environment	Task specific	Documentation
Annotation Utterance Domain	Task specific	Documentation
Annotation Part-of-Speech	Task specific	Documentation
Annotation Named-Entities	Task specific	Documentation

Table 3.1: Attributes of datasets and their relevance to ASR evaluation

- The audio was encoded using at least 16 bits per sample.

The following is an overview of 24 datasets that met the mandatory criteria and were selected for curation.

- **The Common Voice dataset** *(mozilla-common_voice_15-23)* [11] is an open source multilingual resource developed by Mozilla foundation. [3]. This project aims to democratize voice technology by providing a wide-ranging and freely available dataset that covering wide range of languages and accents. Contributors from around the globe donate their voices, reading out pre-defined sentences or validating the accuracy of other contributions. Common Voice is recognized as the most comprehensive and diverse voice dataset available, spanning more than 60 languages and representing many underrepresented groups. datasets are released every three months under a Creative Commons 0 (CC-0) license.
- **The Multilingual LibriSpeech (MLS) dataset** *(fair-mls-20)* [12] is a large multilingual corpus created for speech research by Facebook AI Research (FAIR)[115]. This dataset is derived from LibriVox audiobooks and covers eight languages, including approximately 44,000 hours of English and a total of around 6,000 hours for other languages. The Polish speech data include 137 hours of read speech from 25 books, recorded by 16 speakers. Transcriptions in the test sets were evaluated by humans.
- **The Clarin Studio dataset** *(clarin-pjatk-studio-15)*[13] is provided by CLARIN-PL, a CLARIN subsection devoted to the Polish language. This corpus includes 13,802 short utterances, which add up to about 56 hours, spread over 554 audio sessions by 317 speakers. Each session contains between 20 and 31 audio files. All utterances were recorded in a studio, guaranteeing clear audio files free from background noise and other environmental factors.
- **The Clarin Mobile dataset** *(clarin-pjatk-mobile-15)*[14] is a Polish speech corpus of read speech recorded on the phone. It includes many speakers, each reading several dozen different sentences, and a list of words containing rare phonemes. It is designed for the analysis of modern Polish pronunciation in a telephony environment.

[11] Common Voice webpage
[12] MLS dataset webpage
[13] Clarin Studio dataset CLARIN-PL DSpace repository
[14] Clarin Mobile dataset CLARIN-PL Dspace repository

- **The Jerzy Sas PWR datasets** (Politechnika Wrocławska) According to the documentation available online[15] speech samples were collected using a variety of microphones and in a relatively noise-free acoustic conditions. Three set of recordings were downloaded and curated:
  - Male speaker speech set *(pwr-maleset-unk)* – single male speaker recordings used to build the acoustic model for experiments.
  - Utterances containing short words (*pwr-shortwords-unk*) – recordings containing single-phoneme conjunctions and prepositions that are likely to be falsely recognized.
  - *Spoken commands as very important utterances (VIUs) (pwr-viu-unk)*– the set of editor control commands that can be interleaved with the domain-specific utterances.
- **The M-AI Labs Speech corpus** *(mailabs-19)*[16], similarly to the MLS corpus, was created from LibriVox audiobooks. This corpus covers nine languages and was created by the European company M-AI Labs with the goal of empowering (European) companies to leverage AI & ML while retaining control and expertise. The M-AILABS Speech dataset is provided free of charge and is intended to be used as training data for speech recognition and speech synthesis. Training data consist of nearly a thousand hours of audio for all languages, including 53.5 hours for Polish.
- **The AZON Read and Spontaneous Speech datasets** *(pwr-azon_ spont-20, pwr-azon_ read-20)*[17] is a collection of recordings of academic staff, mainly in the physical chemistry domain. The corpus is divided into two parts: supervised, where the speaker reads the provided text, and unsupervised spontaneous recordings, such as live-recorded interviews and conference presentations by scientific staff. The dataset contains recordings of 27 and 23 speakers, totaling up to 5 and 2 hours of transcribed speech, respectively. The AZON database is available under a CC-BY-SA license.
- **Google FLEURS** *(google-fleurs-22)* is a parallel speech Benchmark dataset in 102 languages built on top of the FLoRes-101 machine translation benchmark, with ap-

[15] Jerzy Sas webpage

[16] Munich AI Labs speech dataset webpage

[17] AZON speech dataset

proximately 12 hours of supervised speech per language. [19] It is hosted on the Hugging Face platform [18], and is available under a CC-BY license.

- **PolyAI Minds14** (*polyai-minds14-21*) is a dataset designed to train and evaluate intent recognition systems using spoken data. It includes 14 intentions obtained from a commercial e-banking system, along with spoken samples in 14 different language variations.[38] It is hosted on the Hugging Face platform[19], and is available under the CC-BY license.

- **PolEval 22 Diabiz sample** (*ul-diabiz_poleval-22)* is a dataset provided for the punctuation restoration task in the 2022 PolEval competition. It is a subset of *DiaBiz* [20] dialog corpus of phone-based customer-agent interactions created by the PELCRA group of the University of Łódź and the *VoiceLab*[21] company. It contains more than 4,000 conversations, totaling nearly 410 hours. The recordings were provided by the five call center agents and 191 participants as customers. The dataset covers nine high-demand business domains for conversational analytics and automation solutions. The data is available under CC-BY-SA-NC-ND. The creator has authorized sharing a curated version of the corpus for the purpose of open challenge organization.

- **SpokesMix** [22] is a freely available time-aligned corpus of conversational Polish developed by the PELCRA group of the University of Łódź. Each corpus consists of speech recordings (in WAV format) and word-by-word transcriptions, which also include some non-speech events. The transcriptions are complemented with words, phone annotations, PDF transcripts, and video content (if available). The corpus is available under a CC-BY license [109]. The following subsets have been made available for download by the authors:

  - `PELCRA_EMO` (*ul-spokes_ mix_ emo-18*) is a subcorpus of focused interviews of people reflecting on their emotions. It contains speech from the 80 speakers and has a total size of 28 hours. It is available under CC-BY

  - PELCRA_LUZ| (*ul-spokes_ mix_ luz-18*) represents a subcorpus consisting of

[18] FLEURS dataset on Hugging Face

[19] Minds14 dataset on Hugging Face

[20] Diabiz corpus webpage

[21] VoiceLab webpage

[22] SpokesMix webpage

open interviews. It encompasses conversational speech that involves 42 speakers and spans a total duration of 20 hours.

  - PELCRA_PARL| *(ul-spokes_mix_parl-18)* is a subset created from examples of spoken parliamentary content. It covers a total duration of 14 hours and includes recordings of oratory speeches by 241 different speakers.

- **SpokesBiz** [23] is a freely available time-aligned corpus of conversational Polish developed within the CLARIN-BIZ project, which currently comprises more than 650 hours of recordings from nearly 600 speakers [111]. The transcribed recordings were diarized and manually annotated for punctuation and casing. The corpus is divided into multiple subsets:

  - CBIZ_BIO *(ul-spokes_biz_bio-23)* – Biographical interviews covering childhood, current job and family situation, and future plans, with an informal tone.
  - CBIZ_INT *(ul-spokes_biz_int-23)* – Job interviews for potential babysitters.
  - CBIZ_LUZ *(ul-spokes_biz_luz-23)* – Unrestricted conversations among friends and families, characterized by their free and natural flow.
  - CBIZ_POD *ul-spokes_biz_pod-23* – Internet podcasts focusing on board games, nature photography, society, traveling, and international affairs.
  - CBIZ_PRES *(ul-spokes_biz_pres-23)* – Student presentations on a broad range of topics including culture, literature, parenting, and gender roles.
  - CBIZ_VC & CBIZ_VC2 *(ul-spokes_biz_vc-23 & ul-spokes_biz_vc2-23)* - Thematic discussions on topics of society and lifestyle.
  - CBIZ_WYW *(ul-spokes_biz_wyw-23)* – Interviews with a fixed set of questions on personal preferences and experiences. The SpokesBiz corpus is available under the CC-BY-NC-ND license. The authors consent for the distribution of a curated version of the data specifically to organize the open challenge. The curated datasets originating from the PELCRA catalog were distributed as separate artifact on Hugging Face platform.[24]

[23] SpokesBiz webpage

[24] PELCRA for BIGOS dataset

Summary statistics on the size of the dataset and the characteristics of the content can be found in Section 4.2.3. Detailed attributes of the datasets sourced can be found in the appendix 7.1.5.

Manual analysis of datasets

Prior to the creation of scripts for automated pre-processing, the original datasets were subjected to a manual quality check. The goal was to identify issues that could potentially hinder the effectiveness of the datasets for the evaluation of the ASR system. Datasets that were found to significantly undermine the reliability of the evaluation without additional manual content curation were removed from the process. Below are examples of two datasets that met the mandatory requirements specified in the catalog but were excluded from further processing due to distinctive transcription formats.

- The *Spelling and Numbers Voice (SNUV)*[25] dataset from University of Łódź is available under CC-BY license. It contains more than 220 hours of recordings of Polish speakers reading numbers and spelling words. A written representation of the recordings is provided with the original sound files. The transcription represents how each letter was pronounced by the speaker, e.g. spelled out word "*pstrąg*" is transcribed as "*py sy ty ry ą gy*". Many ASR systems transcribe spelled words as a list of letters, e.g. *"p st r ą g".* Initial experiments revealed high number of false negatives resulting from mismatched text normalization standards, rather than incorrectly recognized speech (examples are provided in Table 3.2).
- The *CLARIN Cyfry* dataset of the Polish Japanese Academy of Technology contains only transcriptions of numeric expressions. This leads to high error rates, despite the system correctly recognizing non-numeric terms.

During initial evaluation experiments, it was discovered that both issues led to inflated error rates. Therefore, both datasets were excluded from the further curation of BIGOS Benchmark dataset.

Manual dataset inspection prior to the curation process unveiled several pragmatic challenges inherent in the content and structural composition of the data set. Diverse

[25] SNUV

Table 3.2: Sample of PELCRA SNUV references and ASR outputs

Reference	**Hyp. Whisper**	**Hyp. Google**	**Hyp. Azure**
py ly ą sy	PEU-LE-ON-SE	p l o s	Py ly s.
py (o kreskowane) źi ny i e ji	P. Okreskowane. ZI. N. I. E. J.	p okreskowane zi e j	Py. Okres kodowany zi. My. I. E. Ji.
(czterysta osiemdziesiąt pięć)	485	485	485.

factors were qualitatively evaluated. Tables 3.3 and 3.4 provide a comprehensive overview of the factors intrinsic to the specific source dataset. Tables 3.3 and 3.4 outline positive or negative influences on the utility of the dataset for evaluation purposes, respectively. After curating to a format that allows rigorous analysis, those factors were confirmed using quantitative analysis, the results of which are presented in section 4.2.3.

Automatic curation process

The initial stage of automated curation involved acquiring the accessible datasets. Whenever feasible, URLs for web-hosted datasets were included in the configuration files to enable automatic batch download. Manual download was required for the Common Voice dataset, as it required consenting to the custom license terms.[26] URLs for automatic and manual downloads of datasets were retrieved from Polish ASR speech data catalog. Subsequently, the downloaded data were extracted and transformed into the target format (see Table 3.9) using bash and Python scriptsBIGOS. Text data were pre-processed using the Pandas Python library and regular expressions. The preprocessing of the audio data was performed with the SOX command-line utility and Librosa Python library.

The scope of the automatic curation is as follows:

- Dataset level:
  - Creating train/dev/test splits if not available in the original dataset.
  - Assigning standard IDs to speakers and files.
- Audio files:
  - validation of audio file availability,

[26] CommonVoice on Hugging Face

Dataset	**Positive utility factors**
pjatk-clarin_mobile-15, clarin-pjatk-studio-15	Simple format, Transcription quality, Low noise environment
fair-mls-20	Speakers pool and meta, Large vocabulary
mailabs-librivox_corpus-19	Large vocabulary
mozilla-common_voice_15-23	Speakers pool, Speakers meta availability.
pwr-azon_read-21, polyai-minds14-21	Speakers pool, Domain representative
pwr-azon_spontaneous-21	Spontaneous speech, Domain terminology
pwr-male_sample-unk, pwr-short_words-unk	Simple format
pwr-vui-unk	Simple format, Speech commands
pelcra-snuv-12	Simple format, Large number of speakers, Large size
pjatk-clarin_cyfry-16	Audio quality, Numerals rich
pelcra-spokes_mix, pelcra-spokes_biz	Speaker metadata availability, Large size, Manual transcriptions, Realistic audio quality.

Table 3.3: Overview of factors enhancing specific dataset utility for ASR evaluation purposes.

Dataset	**Negative utility factors**
pjatk-clarin_mobile-15, pjatk-clarin_studio-15	Lack of speakers meta-data
pjatk-clarin_cyfry-16	Only numerals are transcribed
fair-mls-20	Archaic language
mailabs-corpus_librivox-19	Only 2 speakers
mozilla-comm-voice-22	Gaps in meta-data coverage
pwr-azon-read-21	Complex structure
pwr-azon-spontaneous-21, polyai-minds14-21	Audio quality
pwr-male-sample-unk	Only one speaker, Lack of speakers meta-data
pwr-short-words-unk	Non UTF text encoding
pwr-vui-unk	Limited vocabulary
pelcra-snuv-12	Spelling transcription format
pelcra-spokes_mix, pelcra-spokes_biz	None

Table 3.4: Overview of factors decreasing datasets' utility for ASR evaluation purposes.

- unification of audio format to WAV 16 bits/16 kHz,
- normalization of audio amplitude to -3 dBFS,
- splitting long audio files into shorter segments based on time-alignment annotations,

- Text files (transcripts and meta-data):
  - conversion of source data text encoding to UTF8,
  - extraction of original transcription,
  - removal of redundant characters,
  - extraction and unification of available metadata,
  - generation of metadata from text and audio content.

The automatic curation steps were adjusted to the specific format and content of the dataset. For datasets delivered without partitions, a pseudo-random deterministic division into train, dev, and test splits was applied. Tables 3.5 and 3.6 present partitioning of splits in original and curated datasets in BIGOS and PELCRA for BIGOS datasets, respectively. Examples of metadata derived from text and audio analysis can be found in 3.9.

Subset	**Original part.**	**BIGOS splits**	**Entity for split**
google-fleurs-22	train, test, dev	original splits preserved	N/A
polyai-minds14-21	none	pseudorandom	audio file id
pjatk-clarin_mobile-15	none	pseudorandom	session (speaker id)
pjatk-clarin_studio-15	none	pseudorandom	session (speaker id)
pwr-azon_read-20	none	pseudorandom	session (speaker id)
pwr-azon_spont-20	none	pseudorandom	session (speaker id)
fair-mls-20	train, test, dev	original splits preserved	N/A
mozilla-cv15-23	train, test, dev	original splits preserved	N/A
mailabs-corpus_librivox-19	none	pseudorandom	audio file id
pwr-maleset-unk	none	pseudorandom	audio file id
pwr-shortwords-unk	none	pseudorandom	audio file id
pwr-viu-unk	none	pseudorandom	audio file id

Table 3.5: Meta-data and partitioning of source datasets — BIGOS dataset

Subset	**Original part.**	**BIGOS splits**	**Entity for split**
ul-diabiz_poleval-22	train, test, dev	original splits preserved	N/A
ul-spokes_biz_bio-23	none	pseudorandom	recording id
ul-spokes_biz_int-23	none	pseudorandom	recording id
ul-spokes_biz_luz-23	none	pseudorandom	recording id
ul-spokes_biz_pod-23	none	pseudorandom	recording id
ul-spokes_biz_pres-23	none	pseudorandom	recording id
ul-spokes_biz_vc-23	none	pseudorandom	recording id
ul-spokes_biz_vc2-23	none	pseudorandom	recording id
ul-spokes_biz_wyw-23	none	pseudorandom	recording id
ul-spokes_mix_emo-18	none	pseudorandom	recording id
ul-spokes_mix_luz-18	none	pseudorandom	recording id
ul-spokes_mix_parl-18	none	pseudorandom	recording id

Table 3.6: Meta-data and partitioning of source datasets — PELCRA dataset

Dataset-specific transcription conventions were retained. Metadata relevant to the evaluation, such as the age or gender of the speaker, were extracted and standardized whenever available. To maintain consistency, the metadata standardization conventions of the Common Voice format were adopted whenever possible due to its widespread usage and diversity. The speaker metadata available in specific source datasets is presented in Tables 3.7 and 3.8.

Subset	**Speaker ID**	**Speaker gender**	**Age info**
google-fleurs-22	no	no	no
polyai-minds14-21	no	no	no
pjatk-clarin_mobile-15	yes	no	no
pjatk-clarin_studio-15	yes	no	no
pwr-azon_read-20	yes	yes	no
pwr-azon_spont-20	yes	yes	no
fair-mls-20	yes	no	no
mozilla-cv15-23	yes	yes	yes
mailabs-corpus_librivox-19	yes	yes	no
pwr-maleset-unk	no	yes	no
pwr-shortwords-unk	no	yes	no
pwr-viu-unk	no	yes	no

Table 3.7: Meta-data and partitioning of source datasets

Table 3.9 presents the *utterance data object* resulting from the curation process.

Subset	Speaker ID	Speaker gender	Age info
ul-diabiz_poleval-22	yes	no	no
ul-spokes_biz_bio-23	yes	yes	yes
ul-spokes_biz_int-23	yes	yes	yes
ul-spokes_biz_luz-23	yes	yes	yes
ul-spokes_biz_pod-23	yes	yes	yes
ul-spokes_biz_pres-23	yes	yes	yes
ul-spokes_biz_vc-23	yes	yes	yes
ul-spokes_biz_vc2-23	yes	yes	yes
ul-spokes_biz_wyw-23	yes	yes	yes
ul-spokes_mix_emo-18	yes	yes	yes
ul-spokes_mix_luz-18	yes	yes	yes
ul-spokes_mix_parl-18	yes	yes	yes

Table 3.8: Meta-data and partitioning of source datasets

Table 3.9: Attributes in the BIGOS utterance data object

Field name	Description
audioname	Standardized unique identifier for each audio recording in the dataset.
split	Indicates the dataset split the recording belongs (e.g., train, test, validation).
dataset	Source dataset identifier.
ref_orig	The original transcript associated with the audio recording.
ref_spoken	Transcription in the spoken domain format.
ref_written	Transcription in the written domain format.
audio	Object for storing audio data in HF datasets format.
sampling_rate	The sampling rate of the audio recording in the dataset. Can be the same as the original or adjusted for standardization.
samplingrate_orig	The original sampling rate of the audio recording.
speaker_id	A unique identifier of the speaker in the recording.
audiopath_bigos	The relative path to the audio file from distributed data archive.
audiopath_local	The absolute path to the extracted audio file, typically in the default HF datasets cache directory.
audio_duration_samples	Recording duration in samples.
audio_duration_seconds	Recording duration in seconds.
speaker_gender	Information about the speaker's gender in the CommonVoice format. If not available, it is indicated as N/A (Not Available).
speaker_age	Information about the speaker's age in CommonVoice format. If not available, it is indicated as N/A (Not Available).
speech_rate_words	Speech rate expressed in words per second.
speech_rate_chars	Speech rate expressed in characters per second.
utterance_length_words	Length of the utterance in words.
utterance_length_chars	Length of the utterance in characters.

3.3.3 Dataset analysis process

The methodology considers the availability of datasets in two forms: public and non-public. In the split test, the public version includes only audio recordings with hidden corresponding references, essential for facilitating an open competition where participants can formulate hypotheses without access to the actual answers. The procedure begins by accessing the publicly available dataset from Hugging Face. Should the dataset feature masked elements, it verifies the existence of a nonpublic version. The absence of such a version results in the termination of the process without further action on the masked elements. Conversely, if a non-public version exists, its masked elements are integrated into the dataset for comprehensive feature analysis. The results from this analysis are subsequently synthesized into reports and visualizations that shed light on the dataset's attributes and significant discoveries. The analysis phase is completed with these reports and visualizations. process is presented in Figure 3.2.

Dataset metrics

Quantitative assessment of the dataset was performed using various metrics. For audio content, these metrics included the number of speakers, the total duration of the audio in hours, and the total number of speech recordings. Text data were analyzed on the basis of the total and unique counts of utterances, words, and characters. Additionally, the analysis assessed metadata coverage, detailing the percentage of recordings that included metadata on the speaker's sex and age. The speech rate metrics included the number of words and characters per second. The linguistic structure of the recordings was evaluated by calculating the average number of words and characters per utterance and the average recording duration in seconds. The metrics used for data analysis are presented in Table 3.10 The metrics were calculated for the BIGOS and PELCRA datasets, as well as their individual subsets and splits. The aggregate results for the curated datasets are presented in Section 4.2.3, while the metrics per split for individual subsets are presented in Appendix 7.1.5.

3.3.4 Dataset release

The final step in curation was releasing the dataset, which involved:

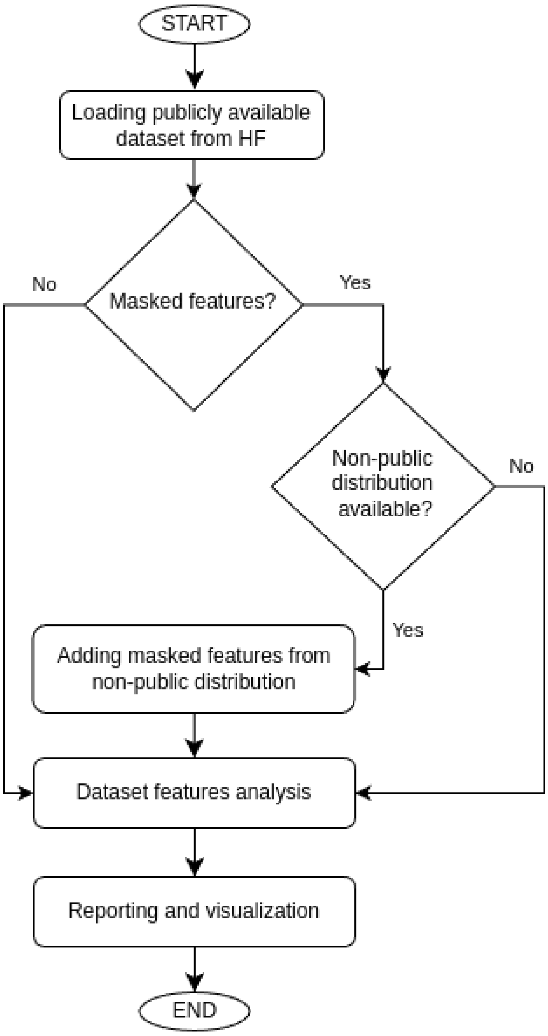

START
Loading publicly available dataset from HF
Masked features?
No
Yes
Non-public distribution available?
No
Yes
Adding masked features from non-public distribution
Dataset features analysis
Reporting and visualization
END

Figure 3.2: Process of analysis of curated datasets.

Metric	**Definition**
Speakers	Number of individual speakers represented in the dataset.
Audio [h]	Total duration of audio material, expressed in hours.
Recordings	Number of individual speech recordings.
Utterances	Total number of speech transcriptions.
Words	Total number of words.
Characters	Total number of characters.
Unique utterances	Number of distinct utterances.
Unique words	Number of distinct words.
Unique characters	Number of distinct characters.
Meta coverage – gender [%]	Percentage of recordings with speaker gender metadata.
Meta coverage – age [%]	Percentage of recordings with speaker age metadata.
Speech rate [words per second]	Average number of words spoken per second.
Speech rate [characters per second]	Average number of characters spoken per second.
Words per utterance	Average number of words per utterance.
Characters per utterance	Average number of characters per utterance.
Average recording duration [s]	Statistics of audio duration, expressed in seconds.

Table 3.10: Metrics used for analysis of datasets contents.

- Converting to Hugging Face Datasets format
- Masking references of test split
- Uploading public and secret datasets to Hugging Face Hub
- Creating and uploading Hugging Face Datasets build script
- Referencing the original licenses and authors in the README
- Setting Gated datasets to acknowledge the original licenses

3.4 RO3: Survey of ASR benchmarks for Polish

3.4.1 Research objectives and questions

The objective was to determine the current status of the benchmarks of the ASR systems for the Polish language, specifically to find the answer to the following questions:

- **RQ 8:** How to categorize Polish ASR benchmarks using public information?
- **RQ 9:** What methods, datasets, and ASR systems have been used in Polish ASR benchmarks?
- **RQ 10:** Which Polish ASR systems have not been evaluated?
- **RQ 11:** Which benchmarks evaluated commercial and free systems?
- **RQ 12:** Which ASR system performs best?
- **RQ 13:** What are the main conclusions from the ASR benchmarks?
- **RQ 14:** How to share the survey results with the community?

3.4.2 Research methodology

The study involved the identification of benchmarks through a review of the literature and manual annotation of key aspects such as the datasets used, the evaluated systems, the tasks, the domains, the evaluation metrics, etc. This iterative process led to the development of a taxonomy consisting of 40 attributes that facilitated a quantitative comparison of the benchmarks implemented so far. The resulting catalog and analysis results were shared with the community[27].

The research method comprises:

- Literature search to identify existing Polish ASR benchmarks
- Development of a taxonomy covering key dataset features
- Cataloging ASR benchmarks datasets according to the taxonomy framework.
- Developing a publicly accessible digital repository with a data catalog and survey results.

Literature review

Most of the ASR benchmarks for Polish were already identified during the survey of Polish ASR speech datasets (see Section 3.2 for a description of the methodology and Section 4.1 for results). In February 2024, an additional keyword-based search [124] was performed to identify publicly reported benchmarks after the last update to the speech data survey. As a result, the survey covers benchmarks reported in years 2018-2023.

Development of the taxonomy

The original function of the taxonomy was to unify diverse and unstructured information extracted from the review of the literature. Once the information was standardized, it was possible to analyze and compare benchmarks in terms of the scope of the evaluation, the characteristics of the datasets, and the methodological aspects.

The final taxonomy included 40 attributes and is presented in Appendix 7.1.2. Key attributes include:

- **Benchmark**: Benchmark codename.
- **Catalog update information**: Information when the catalog entry was last updated.
- **Publication reference**: URL to the publication detailing the benchmark.
- **Temporal information**: The year the benchmark was created.

[27] Polish ASR benchmarks catalog

- **Evaluation focus:** Details on the evaluated systems and models, including the model with the best performance and its average Word Error Rate (WER).
- **Benchmark outcomes and limitations**: Overview of the main conclusions and methodological limitations.
- **Evaluated systems:** Information on whether commercial, freely available and community-provided systems were evaluated.
- **Replicability**: Information about whether resources for benchmark replication were made available.
- **Dataset accessibility**: Details on the availability of the evaluation dataset.
- **Evaluation methodology:** Information on the frequency of the benchmark and the type of evaluation (automatic or human).
- **Metrics used:** List of lexical, language model-based, and annotation-based metrics used in the benchmark.
- **Scope of benchmark:** Description of ASR use cases, sociodemographic analyzes, types of speech, acoustic conditions, etc. used in the benchmark.
- **Data collection details**: Details on recording devices, vocabulary domains, audio sources, and available annotations.
- **Quantitative metrics**: Details on the number of datasets, vocabulary domains, recordings, speakers, system and model variants evaluated, etc.

Analysis

Collected data were used to check aspects relevant to the design of the new ASR benchmarking system:

1. What Polish ASR systems remained unevaluated (both commercial and open-source)?
2. What benchmarking methodologies have been used so far (metrics, datasets)?
3. How does the number of evaluated systems change over time?

4. Does an ASR system exist that is superior in multiple benchmarks?
5. Are different conclusions drawn from ASR benchmarks for similar use cases?

The results of the analysis are presented in Section 4.3

3.5 RO4: Design and implementation of a system for ASR systems benchmarking

3.5.1 Research objectives and questions

The goal was to design and implement the system that allows comparing the performance of ASR systems. The system was developed to incorporate the recommended functionalities and considerations outlined in Section 2.3.2. The research questions addressed are as follows:

- **RQ 15:** What tools and systems exist for ASR benchmarking?
- **RQ 16:** What challenges arise in evaluating multiple ASR systems, and what strategies can address them?
- **RQ 17:** How can the system be extended to new ASR systems, datasets, languages, metrics, and normalization methods?

3.5.2 Research methodology

System design considerations

The primary design goal was to simplify the evaluation of new ASR systems using new datasets. In addition, the benchmarking system was designed to accommodate new metrics and analytical dimensions in the future. Whenever possible, established tools and platforms were used. The overview of major design considerations is presented in the table 2.4.

Overview of the evaluation process

The process consists of 7 steps, as depicted in Figure 3.3

1. **Configurations loading:** Involves loading the necessary configurations for the evaluation process. Input data include common, user-specific, and evaluation run-specific

Table 3.11: Design considerations for ASR evaluation system

Aspect	**Considerations**
Metrics	Support for well-established metrics.
Extensibility	Straightforward integration of new datasets, normalization methods, metrics, and new ASR systems.
Availability	Publicly accessible and intuitive presentation of results.
Comprehensiveness	Performance analysis across scenarios, system parameters, and user groups.
Analysis of bias	Analysis of system performance in various scenarios and user groups should be feasible.

configurations. The output of this step is a set of runtime parameters of the system that will guide the ASR system during the evaluation.

2. **ASR systems initialization:** In this step, ASR systems are prepared to be tested using the system runtime parameters obtained from the previous step. The output comprises initialized models and a cache to retrieve existing hypotheses and save newly generated hypotheses.

3. **Evaluation dataset loading:** In this step, the benchmark dataset is loaded into the system. The input is the names of the datasets, and the output is the dataset objects, as handled by the Hugging Face Datasets library.

4. **ASR hypotheses generation:** This step generates or retrieves ASR hypotheses, which are the predicted transcriptions produced by ASR systems. The input of this function is the list of audio files and the output is the ASR hypotheses.

5. **Metrics calculation:** Performance metrics for ASR outputs are calculated in this step. The inputs include the ASR hypotheses along with the reference transcriptions, and the output consists of evaluation metrics.

6. **Results analysis:** The performance of ASR systems is analyzed. The input is the evaluation metrics generated in the previous step, and the output is an analytical report detailing the performance of the ASR systems for various analytic dimensions.

7. **Results visualization:** Finally, the results of the analysis are visually interpretable. The analytical report serves as input, and the output is a series of graphs, graphs,

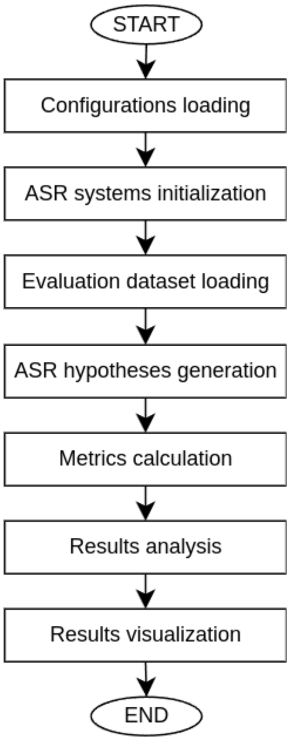


Figure 3.3: ASR evaluation process

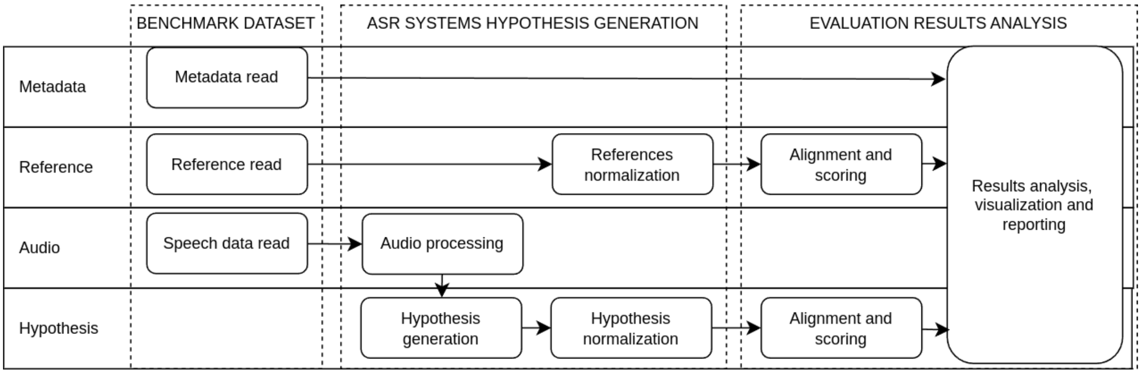


Figure 3.4: ASR evaluation process data flow

or other visual representations that convey the findings of the evaluation of the ASR system.

The data flow of specific data types (metadata, references, audio files and ASR hypotheses) is presented in Figure 3.4

ASR systems integration

Object-oriented design was used to separate common elements and those specific to the ASR system. The common elements include:

- Savings and reading from hypotheses cache

- Error handling and logging

The ASR system functionalities include:

- Local model initialization
- Configuring web clients (model type, language etc.)
- Handling API requests
- Parsing ASR system outputs

Audio processing:

The major function of audio processing model was to ensure that audio file is available and not corrupted. No further processing was needed, as the audio files were already standardized to a 16 bits/16 kHz WAV format, which is compatible with all ASR systems evaluated.

References and transcripts normalization

False recognition errors may arise due to inconsistencies in normalization of references and ASR system output. [140]

The following automatic post-processing operations were applied to the reference transcripts and the output from the evaluated ASR systems:

- Elimination of unnecessary white spaces
- Conversion of all characters to lowercase
- Removal of all punctuation symbols
- Replacement of words using lexicon
- Removal of all special purpose words (tags)
- Combination of all the above.

Normalization methods are presented in table 3.12

Table 3.12: Methods of normalizing references and hypotheses

Normalization method	**Scope**
blanks removal	Elimination of redundant white spaces.
lowercasing	Conversion of all characters to lowercase.
punctuation removal	Removal of punctuation symbols.
lexicon-based	Removal of specific words e.g. fillers "um", "mhm" etc. Unification of spelling e.g. Kissindżer -> Kissinger
tags removal	Removal of tags e.g. 'trunc' in PELCRA dataset.

Alignment and scoring

ASR systems predictions were evaluated against target transcriptions using the following metrics:

- **Sentence Error Rate** (SER), which calculates the proportion of sentences that are not perfectly recognized, i.e., sentences that contain at least one error.

- **Word Error Rate** (WER), which is defined as the minimum number of operations (substitutions, insertions, and deletions) required to transform the system output into the reference transcript, divided by the total number of words in the reference. The result is expressed as a percentage. A lower WER indicates a more accurate system. The WER value can be greater than 100%.

- **Match Error Rate** (MER), which calculates the ratio of the total number of errors (substitutions, insertions, and deletions) to the total number of words in the reference and hypothesis (system output) transcripts. Unlike WER, which is normalized by the number of words in the reference, MER is normalized by the total number of words in both the reference and hypothesis. This makes the MER potentially less sensitive to the insertion of incorrect words by the ASR system, offering a different perspective on the accuracy of the system. MER value is equal to or less than 100%.

- **Character Error Rate** (CER), which calculates the minimum number of character-level operations (substitutions, insertions, and deletions) needed to change the system output to the reference transcript, divided by the total number of characters in the reference.

3.6 RO5: Use of curated dataset for benchmarking ASR systems for Polish

3.6.1 Research objectives and questions

Developed Benchmark dataset was used to compare the available ASR systems for the Polish language. This included evaluating performance in different scenarios and metrics, to develop a better understanding of the strengths and weaknesses of current ASR technologies.

- **RQ 18:** What is the ASR accuracy for different datasets?
- **RQ 19:** What is the accuracy gap between commercial and free systems?
- **RQ 20:** Does ASR accuracy vary with speech features?
- **RQ 21:** Is there an accuracy difference by age or gender?
- **RQ 22:** How to share evaluation results with the community?

3.6.2 Research methodology

Overview

The curated benchmark dataset and the developed evaluation system, were used to compare the precision of the ASR systems available for Polish. The systems were compared across 4 major evaluation scenarios:

1. general and per dataset accuracy
2. commercial vs. free systems accuracy
3. impact of speech variations on accuracy
4. accuracy across sociodemographic groups

Evaluation Scenarios

The first evaluation goal is focused on practical application, that is, the accuracy for various datasets. This comparison enables practitioners to choose the ASR system that offers the best generalizability or the best accuracy for a specific application (device, speech type, etc.).

The evaluation involves four recognition tasks (use cases) outlined by Aksenova. et al. [2]:

1. human-human dialogue (transcription of meetings or interviews)

2. human-human monologue (transcription of lectures and presentations)

3. human-machine dialogue (voice-commands for device control)

4. human-machine monologue (dictation).

The second objective is also motivated by practical aspects and deals with choosing between a commercial or freely accessible ASR system. The purpose of the evaluation is to compare performance of free vs. paid systems to assess whether there is a trade-off between cost and quality.

The third objective is to examine the connections between recognition accuracy and various variations in spoken language, such as the duration of utterances, the speaking rates of the speaker, or interruptions in informal and spontaneous speech. The aim is to identify which of the factors studied poses the most significant challenges to Polish ASR technology.

The fourth objective is to study the recognition accuracy for different sociodemographic groups, e.g. age groups and genders.

Table 3.13 presents specific scenarios, metrics, and analysis dimensions. Table 3.14 presents the relation between the evaluation scenarios and the research questions.

Evaluated ASR systems

Seven types of ASR system were evaluated: Google STT, Azure STT, Whisper, AssemblyAI, NeMo, MMS and Wav2Vec. All systems, except Azure STT, offer different variants of the models that support the Polish language. In total, 22 combinations of pairs of system-model were compared. The evaluated systems and models are presented in Table 3.16.

Eval. scenario ID	Analysis dimension	Description
ES1	systems	Accuracy per system-model variant across all subsets.
ES2	dataset	Accuracy per subset across all variants of the system model.
ES3	system types	Accuracy per system type (free or commercial) across all subsets. WER per subset for the best performing free and commercial systems.
ES4	model size	Accuracy across all subsets for systems with known model size.
ES5	audio duration	Accuracy in function of audio duration for the most accurate free and paid systems.
ES6	speaking rate	Accuracy in function of speech rate for the most accurate free and paid systems.
ES7	speaker age group	Accuracy for speaker gender group for all systems.
ES8	speaker gender	Accuracy for speaker age group for all systems.

Table 3.13: Evaluation scenarios and their analysis dimensions

Identifier	Research question	Eval. scenario ID
RQ 18	What is the ASR accuracy for different datasets?	ES1. ES2
RQ 19	What is the accuracy gap between commercial and free systems?	ES3, ES4
RQ 20	Does ASR accuracy vary with speech features?	ES5, ES6
RQ 21	Is there an accuracy difference by age or gender?	ES7, ES8

Table 3.14: Relation between research question and evaluation scenarios.

- **Google Cloud Speech-to-Text** [28] supports more than 125 languages and variants. Google's service offers several useful features, such as noise cancelation, support for streaming, automatic punctuation, and the capability to recognize specific phrases or words when provided with context (e.g., specialized vocabulary or formats for spoken numbers, addresses, years, currencies, etc.). For selected languages, it also provides domain-specific models, multichannel audio support, and filtering of profanity content. Two generations of service are available: v1[29] and v2 [30]. For Polish, multiple model variants are available and were evaluated: *v1_ default*, *v1_ latest_ long*, *v1_ latest_ short*, *v1_ command_ and_ search*, *v2_ long* and *v2_ short*.

- **Microsoft's Azure Speech Service** [31] as of May 2023 supports more than 100 languages and variants. In addition to standard transcription, the Azure Speech Service supports continuous real-time speech recognition and provides robust noise reduction capabilities. It allows users to apply custom models to improve the accuracy of domain-specific terminology. Additional services include text search or analytics on transcribed content, as well as speaker diarization. The *latest default* model for Polish (dated for January 2023) was used, as no specialized model types support this language.

- **Whisper** [32] is an ASR system developed by the OpenAI company. It is trained on a large amount of weakly supervised multilingual and multitask data collected from the Internet [117]. According to the literature, Whisper is capable of handling different languages, dialects, and accents, demonstrating strong performance in diverse applications when evaluated on well-known benchmark datasets, e.g. Common Voice [117]. Whisper is available via a web API or as a pre-trained model for local use. Five versions of models of varying sizes are available for free download as shown in Table 3.15. The large model is available in 3 versions.[33] For this benchmark, the commercial model available via API and eight locally run models were used.

- **NVIDIA NeMo** is the ASR system based on the *quartznet* model, which con-

[28] https://cloud.google.com/speech-to-text

[29] https://cloud.google.com/speech-to-text/docs/speech-to-text-requests?hl=en

[30] https://cloud.google.com/speech-to-text/v2/docs?hl=en

[31] https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text

[32] https://github.com/openai/whisper/tree/main

[33] Whisper ModelCard

Size	**Parameters**	**English-only model**	**Multilingual model**
tiny	39 M	Yes	Yes
base	74 M	Yes	Yes
small	244 M	Yes	Yes
medium	769 M	Yes	Yes
large	1550 M	No	Yes

Table 3.15: Whisper model types. Source: Whisper model card.

sists of 79 layers and has a total of 18.9 million parameters. [64] Two models supporting the Polish language are available: *stt_ pl_ fastconformer_ hybrid_ large_ pc*, *stt_ pl_ quartznet15x5* and *stt_ multilingual_ fastconformer_ hybrid_ large_ pc*. The English version was trained on $\tilde{3}$,000 hours of public English data. Polish models were fine-tuned from English to Polish on the *Mozilla Common Voice (MCV)* dataset. [3] The authors report on 14 % WER on the *dev set* from the Polish MCV dataset. All models are available for free use under a CC-BY-NC license.

- **MMS**: Facebook AI's massive multilingual pre-trained model for speech ("MMS"). It was pre-trained on about 500,000 hours of speech data in more than 1,400 languages[114]. MMS system supports over 1000 languages and other speech processing tasks such as *Text-to-Speech (TTS)* generation and *Speech Language Identification (LID)* [34]. The MMS system is available for free[35] under the CC-BY-NC 4.0 license. The following versions of the fine-tuned model of ASR are available:

  - *1b-fl102* – 1 billion parameter model fine-tuned on *FLEURS* dataset [19]
  - *1b-l1107* – 1 billion parameter model fine-tuned *MMS-lab* [114] dataset.
  - *1b-all* – 1 billion parameter model fine-tuned on *MMS-lab, FLEURS, CommonVoice, MLS* and*VoxPopuli* datasets. [3, 114, 115, 146]

- **Wav2Vec** is the automated speech recognition (ASR) system created by Facebook AI. It employs self-supervision to learn from unlabeled training data. Upon its launch in 2020, wav2vec2 exceeded the top semi-supervised approach with only a fraction of labeled training data [46]. Two models fine-tuned for Polish are available on the Hugging Face platform: *xls-r-1b-polish* and *large_ xlsr-53-polish.*

[34] https://huggingface.co/spaces/mms-meta/MMS

[35] https://huggingface.co/facebook/mms-1b-all

- **Assembly AI**[36] provides an advanced automatic speech recognition service supporting multiple languages. Key features include real-time transcription, automatic punctuation, and robust noise cancellation. The service supports domain-specific vocabulary through custom models, filtering of sensitive content and integration with various platforms via a web API. The system is designed to handle diverse accents and dialects, ensuring high accuracy across different use cases. According to the authors, their system "leverages a diverse training dataset comprising unsupervised (12.5M hours), supervised (188k hours), and pseudo-labeled (1.6M hours) data across four languages"[118]. It is also reported that the *Universal-1* model achieves comparative WER scores to larger and more computationally expensive models, such as Whisper large and Canary-1B.[118]. The amount of training data for Polish is not reported.

Table 3.16 presents the evaluated system model. The table 3.17 presents the details about the usage cost and the license. Pricing of commercial ASR systems and sizes of freely available model are available in the Appendix sections 7.1.6 and 7.1.7, respectively.

Evaluation dataset

The test splits from the BIGOS and PELCRA for BIGOS datasets were used for benchmarking. The average length of the audio recording and the size of the test split subsets differ between subsets (see Section 4.2.3). To ensure a fair comparison of the system's performance across subsets, the amount of speech for each subset should be similar. Due to computational and financial limitations, the duration of speech per subset was capped at 20 minutes. If a test subset exceeded 20 minutes, a random selection of recordings was used from that subset.

Sharing results using Polish ASR leaderboard

The benchmark results were made available to the community throughAMU ASR Leaderboard on the Hugging Face [37] platform. Relevant research artifacts (evaluation results, datasets, surveys) were also shared as the *Hugging Face collection.*[38]

[36] Assembly AI

[37] AMU ASR Leaderboard

[38] AMU BIGOS collection

Shortname	System	Model
assembly_best	assembly_ai	best
assembly_nano	assembly_ai	nano
azure_latest	azure	latest
google_cmd_search	google	command_and_search
google_default	google	default
google_long	google	latest_long
google_short	google	latest_short
google_v2_long	google_v2	long
google_v2_short	google_v2	short
mms_all	mms	1b-all
mms_102	mms	1b-fl102
mms_1107	mms	1b-l1107
nemo_multilang	nemo	stt_multilingual_fastconformer_hybrid_large_pc
nemo_pl_confromer	nemo	stt_pl_fastconformer_hybrid_large_pc
nemo_pl_quartznet	nemo	stt_pl_quartznet15x5
w2v-53-pl	wav2vec2	large-xlsr-53-polish
w2v-1b-pl	wav2vec2	xls-r-1b-polish
whisper_cloud	whisper_cloud	whisper-1
whisper_base	whisper_local	base
whisper_large_v1	whisper_local	large-v1
whisper_large_v2	whisper_local	large-v2
whisper_large_v3	whisper_local	large-v3
whisper_medium	whisper_local	medium
whisper_small	whisper_local	small
whisper_tiny	whisper_local	tiny

Table 3.16: ASR systems evaluated in the study.

Shortname	**Usage cost**	**License**
assembly_best	commercial	Proprietary
assembly_nano	commercial	Proprietary
azure_latest	commercial	Proprietary
google_cmd_search	commercial	Proprietary
google_default	commercial	Proprietary
google_long	commercial	Proprietary
google_short	commercial	Proprietary
google_v2_long	commercial	Proprietary
google_v2_short	commercial	Proprietary
mms_all	free	CC-BY-NC
mms_102	free	CC-BY-NC
mms_1107	free	CC-BY-NC
nemo_multilang	free	CC-BY
nemo_pl_confromer	free	CC-BY
nemo_pl_quartznet	free	CC-BY
w2v-53-pl	free	Apache
w2v-1b-pl	free	Apache
whisper_cloud	commercial	Proprietary
whisper_base	free	MIT
whisper_large_v1	free	MIT
whisper_large_v2	free	MIT
whisper_large_v3	free	MIT
whisper_medium	free	MIT
whisper_small	free	MIT
whisper_tiny	free	MIT

Table 3.17: Evaluated ASR systems usage cost and license type.

3.7 RO6: Organization of competition for the ASR community

3.7.1 Research objectives and questions

The objective was to allow professionals, academics, and companies to compare solutions with the latest advances in the field of ASR. The secondary purpose was to encourage the adoption of the curated benchmark dataset among the Polish and global ASR communities.

- **RQ 22:** What programs can organize the Polish ASR community challenge?
- **RQ 23:** How to compare community solutions with state-of-the-art ASR systems?

3.7.2 Research methodology

The first stage was selecting a suitable program and platform. The second step involved preparing the dataset and describing the contest participants. Consent from dataset authors and rights owners was obtained for curation and redistribution. The final step was designing a solution to integrate community-provided ASR results with the open ASR leaderboard for commercial and free systems.

3.8 Summary

3.8.1 Overview of the data management framework

The research artifacts presented in this chapter can be combined into the data management framework presented in Figure 3.5. The framework combines three processes as follows:

1. **Survey of Polish ASR speech datasets and benchmarks:**
   (a) Curation of information available in public domain
   (b) Development of taxonomies and tools for catalog management
   (c) Sharing survey results and catalogs of datasets and benchmarks
2. **Curation of benchmark dataset for Polish ASR systems**
   (a) Curation of speech datasets content obtained from public domain
   (b) Development of tool chain for speech datasets curation

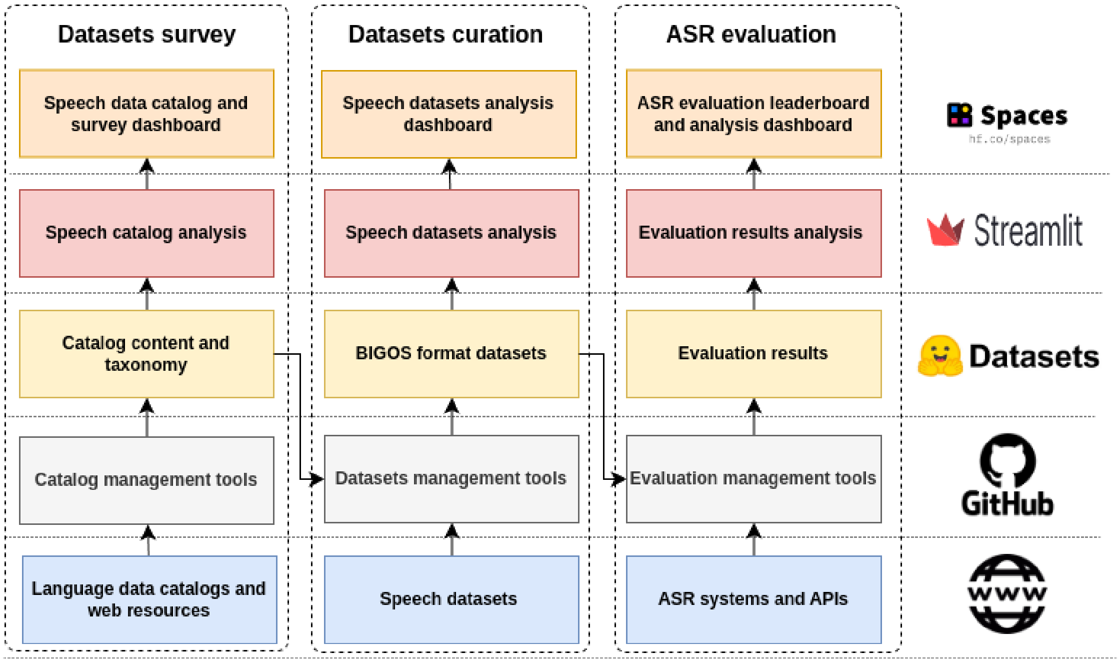

Datasets survey
Datasets curation
ASR evaluation
Speech data catalog and survey dashboard
Speech datasets analysis dashboard
ASR evaluation leaderboard and analysis dashboard
Spaces
hf.co/spaces
Speech catalog analysis
Speech datasets analysis
Evaluation results analysis
Streamlit
Catalog content and taxonomy
BIGOS format datasets
Evaluation results
Datasets
Catalog management tools
Datasets management tools
Evaluation management tools
GitHub
Language data catalogs and web resources
Speech datasets
ASR systems and APIs
www

Figure 3.5: BIGOS data management framework

(c) Sharing curated benchmark dataset and its documentation

3. **Evaluation of Polish ASR systems**:

   (a) Development of tool chain for evaluation and analysis

   (b) Evaluating ASR systems using curated dataset

   (c) Sharing publicly available leaderboard with benchmark results.

Individual actions within the survey, curation and evaluation processes can be conceptually organized according to their functional purpose as follows:

1. **Acquisition:**

   (a) **Language data catalogs and web resources:** Accessing various language data catalogs and Web resources to obtain information about publicly available datasets

   (b) **Speech datasets:** Collecting publicly available speech datasets for further processing.

   (c) **ASR systems and API:** Obtaining ASR hypotheses for evaluation.

2. **Curation:**

   (a) **Catalog management tools:** Managing and organizing the speech data catalog, e.g. validating if content complies with the taxonomy.
   (b) **Datasets management tools:** Managing speech datasets, including format standardization, version control, partitioning, etc.
   (c) **Evaluation management tools:** Managing the evaluation of ASR systems, including storing results and metadata.

3. **Data transfer:**
   (a) **Catalog content and taxonomy:** Delivery of speech data catalog and taxonomy for further analysis.
   (b) **BIGOS format datasets:** Delivery of standardized datasets in BIGOS format for processing and analysis.
   (c) **Evaluation results:** Delivery of ASR evaluations results for further analysis.

4. **Data analytics:**
   (a) **Speech catalog analysis:** Extract insights from the speech catalog.
   (b) **Speech datasets analysis:** Analysis of speech datasets to assess quality and characteristics.
   (c) **Evaluation results analysis:** Assessing ASR systems performance based on set of metrics and scenarios.

5. **Results application :**
   (a) **Datasets survey and dashboard:** Public interface to speech datasets catalog and derived insights.
   (b) **Datasets curation dashboard:** Public dashboard with insights on BIGOS composition and quality.
   (c) **ASR leaderboard and analysis dashboard:** A system for assessing the performance of ASR systems and creating a ranking system using evaluation outcomes.

Chapter 4

Results

4.1 RO1: Survey of ASR speech datasets for Polish

4.1.1 Introduction

This section presents results of the ASR speech datasets survey. The objective was to provide answers to the following research questions:

- **RQ 1:** How to identify and systematically categorize Polish ASR speech datasets using publicly available information?
- **RQ 2:** What is the current state of the ASR speech data?
- **RQ 3:** How can the survey findings be shared and available for feedback from the ASR community?

The answer to RQ 1 was provided in Section 3.2 on the survey methodology. The following sections 4.1.2 and 4.1.4 answer RQ 2 and RQ 3, respectively.

4.1.2 ASR speech datasets survey results overview

The investigation has cataloged 53 distinct datasets, authored by 16 unique entities. Among these, 44 datasets, representing 83%, are accessible through public domain resources or through commercial entities. The total volume of speech exceeds 27 thousand hours, with more than 95% (approximately 25.9k hours) readily available. The corpus of transcribed speech is approximately 6000 hours, with more than 80% accessible. More than 1600 hours of transcribed speech are available at no cost. Close to 3200 hours of transcribed speech data can be purchased from commercial sources. Table 4.1 shows the metrics regarding the availability of speech data for the development of ASR systems for the Polish language.

Metric	**Value**
Catalog last update date	2023-12-19
Unique Polish speech datasets producers	16
Identified datasets reported in the public domain	53
Datasets available to the public (free and paid)	44
Fraction of reported datasets available to the public [%]	83
Speech data reported in the public domain [hours]	27099.1
Speech data available total [hours]	25926.1
Available vs reported speech data ratio [%]	95.67
Transcribed speech data reported in the public domain [h]	5986.1
Transcribed speech data available total [h]	4813.1
Transcribed speech data available free of charge [h]	1641.1
Transcribed speech data available commercially [h]	3172
Available vs reported transcribed speech data ratio [%]	80.4

Table 4.1: Polish ASR datasets survey summary

4.1.3 ASR speech data survey results

This section contains answers to specific research questions related to RQ2 about the general state of the ASR speech datasets for Polish.

What is the oldest publicly reported Polish ASR speech dataset? The *Corpora* dataset, created in 1997 by Stefan Grocholewski [40].

How many ASR speech datasets have been reported publicly for Polish? 53 datasets have been reported publicly between 1997 and 2023. The number of datasets created in specific years and the aggregated statistics are presented in Table 4.2.

What are the largest publicly reported ASR speech datasets for Polish? The *Mobile speech dataset of scripted monolog*[1] produced by the company *Shaip*[2], followed by the *JURISDIC* dataset [23] and *Diabiz* [112]. The datasets contain 1482, 855, and 410 hours of transcribed speech, respectively.

What is the total size of publicly reported ASR speech datasets for Polish? The survey identified 5986 hours of transcribed speech data created between 1997 and July 2023. The evolution of availability of speech data for ASR development across years

[1] `https://www.shaip.com/offerings/speech-data-catalog/polish-dataset/`

[2] https://www.shaip.com/offerings/speech-data-catalog/

Year	Datasets	Trans. speech [h]	Recordings	Speakers
1997	1	6	365	45
1998	1	16	no-info	60
2002	2	50	no-info	198
2005	2	252	no-info	600
2007	2	11	500	505
2008	1	855	no-info	1000
2010	1	78	no-info	1000
2011	2	no-info	no-info	no-info
2012	2	240	no-info	220
2014	1	205	no-info	781
2015	4	136	no-info	557
2016	2	10	488	25
2018	8	180	163	1047
2019	3	335	29	200
2020	6	302	504	3285
2021	6	1895	770	3064
2022	4	464.1	3955	342
2023	1	650	925	590
no info	4	301	939	353

Table 4.2: Summary of audio dataset availability and characteristics by year

is presented in Figure 4.1

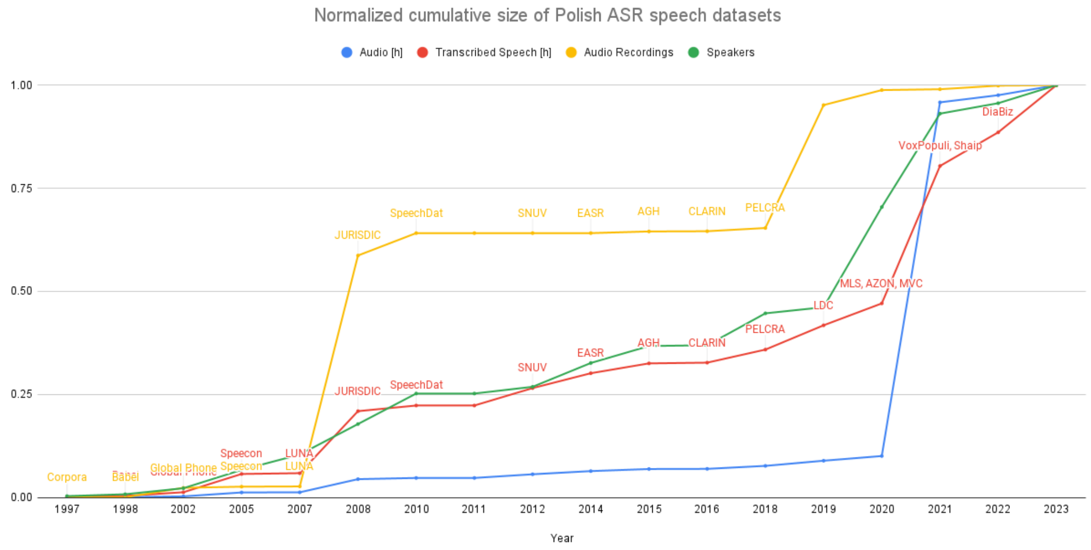


Figure 4.1: Normalized cumulative size of Polish ASR speech datasets

What is the total size of the datasets available for free use? How does it compare with the amount reported in scientific publications so far? According to the data collected, 1,641 hours of transcribed speech and 27,099 hours of speech in total are available for free use. It should be noted that the previous survey of Polish ASR speech datasets [142] in 2019 reported only 223 hours of transcribed speech.

How do the available transcribed speech data for Polish compare with English? The three largest ASR speech datasets for English are MLS [115], People's Speech [33], and Gigaspeech [13]. These contain, respectively, 32,000, 30,000 and 10,000 hours of transcribed speech. The total estimated size for English may exceed 100,000 hours. For Polish, the total amount of public domain transcribed speech data is two orders of magnitude smaller, at approximately 1400 hours.

What is the total size of the Polish ASR speech datasets available from commercial providers? The total size is 3171 hours of transcribed speech.

How do commercially offered datasets compare with public domain datasets? The amount of transcribed Polish speech in commercially offered ASR speech datasets is nearly two and a half times higher than in datasets available under a free-of-charge license.

What is the amount of speech material transcribed contributed by specific institutions involved in the production of ASR speech data? According to the survey data, the largest amount of speech transcribed for ASR was created by the Shaip company, followed by the PELCRA research group (Polish and English Language Corpora for Research and Applications) of the University of Łódź (UL). The third largest contribution comes from Adam Mickiewicz University. Details are presented in Table 4.3.

Which organizations provided the most substantial amounts of open-access ASR speech databases for Polish? Table 4.3 lists the institutions and the size of freely available ASR speech datasets contributed by them. It should be noted that more than half of the freely available speech datasets for Polish were created at the University of Łódź PELCRA group. One third of the contributed transcribed speech material originates from institutions outside of Poland: FAIR, Mozilla Common Voice, and M-AILABS.

What Polish ASR speech corpora are available in the public domain under permissive licensing? There are 31 datasets available in total under Creative Commons or proprietary licenses, allowing free use for noncommercial purposes. Table 7.1 in Appendix 7.1.3 presents the full names and sizes of all available datasets divided into categories according to the specific license type.

What Polish ASR speech corpora are available from commercial providers?

Publisher	Datasets	Tran. speech [h]	Recordings	Speakers
UL	15	1415	1154	1824
PJATK	8	211	1209	1241
WUST	5	15	1395	56
ELRA	4	494	no info	1540
Appen	3	396	no info	1452
HZSK	3	20	no info	15
Shaip	2	1751	no info	2582
LDC	2	284	no info	200
FAIR	2	248	no info	298
AGH	2	67	no info	557
PUT	2	15	365	45
AMU	1	855	no info	1000
Mozilla Foundation	1	148	no info	3062
M-AILABS	1	54	no info	no info
Google	1	12.1	3937	no info
PolyAI	1	1	578	no info

Table 4.3: Institutions contributing speech datasets for Polish.

There are 12 datasets available from commercial providers such as ELRA, LDC, Appen, Shaip, and CLARIN-PL. Table 7.2 in Section 7.1.4 presents the complete list of datasets and their respective producers.

What language data repository offers the largest selection of Polish ASR speech datasets? Table 4.4 shows the number and size of datasets available in various language data repositories. The largest selection is available in the catalog of University of Łódź PELCRA group. The next most extensive collection (9 datasets) is located in the DSpace catalog offered by the Polish CLARIN consortium.

What is the availability of transcribed speech data for specific recording devices? Most of the recordings (1775 hours) originate from mobile devices (Table 4.5). However, these recordings are exclusively accessible through commercial entities (Table 4.7). The second most common type (1370 hours) and the most prevalent in datasets that are freely available (1034 hours), are recordings collected on various types of device. This situation is typical for community-driven initiatives such as *LibriVox* or *CommonVoice*. Recordings made with headsets constitute the third highest volume (705 hours), with recordings from studio-quality microphones coming next (473 hours). There is also a significant amount of data collected using landline telephones, mostly from commercial offering (534 hours).

Repository	Datasets	Trans. speech [h]	Recordings	Speakers
PELCRA	12	963	1136	1482
DSpace CLARIN PL	9	651	698	1083
ELRA	4	494	no-info	1540
Appen Pre-Labelled Datasets	3	396	no-info	1452
Author's homepage (WUST)	3	8	939	no-info
Shaip data catalog	2	1751	no-info	2582
LDC	2	284	no-info	200
Github	2	248	no-info	298
HZSK	2	20	no-info	15
Hugging Face Data Catalog	2	13.1	4515	no-info
AZON WUST	2	7	456	56
Common Voice	1	148	no-info	3062
Coqui Free Corpora Catalog	1	54	no-info	no-info

Table 4.4: Data catalogs and platforms hosting ASR speech datasets for Polish

What is the availability of transcribed speech data for various sampling frequencies? More than 50% of the datasets (22) contain recordings collected with a sampling rate of 16 kHz. More than a quarter of the datasets have a sampling rate of 48 kHz. Information about the sampling rate is missing for 17 datasets. Six legacy datasets with a sampling frequency of 8 kHz were collected. Details are presented in Table 4.8.

What is the availability of transcribed speech data for various types of speech? Read speech constitutes the most documented transcribed speech (56% or 3,362 hours) and almost half of all reported datasets (25). This prominence of read speech can be traced back to two main factors. First, numerous corpora draw on existing read speech, such as those found in LibriVox audiobooks. Second, the collection of read speech is deemed the most efficient and scalable method, as it entails less manual oversight for quality control and post-processing (transcription, diarization, segmentation) compared to spontaneous speech or dialogues. Validation of whether audio recordings match the original prompts is more feasible on a large scale and can also be performed by volunteers,

Table 4.5: Audio devices for all available datasets

Rec. device	Datasets	Trans. speech [h]	Recordings	Speakers	Percent of total
mobile phone	2	1775	no info	2402	29.65
various	11	1370.1	5632	4958	22.89
no info	21	1105	1549	2394	18.46
headset	3	705	no info	1191	11.78
landline phone	7	558	518	1842	9.32
studio mic	8	473	939	1080	7.9
lavalier mic	1	no info	no info	5	0

Table 4.6: Audio devices for publicly available datasets

Rec. device	Datasets	Trans. speech [h]	Recordings	Speakers	Percent of total
various	8	1034.1	5632	3868	63.01
headset	1	220	no info	210	13.41
no info	16	199	1184	1279	12.13
studio mic	5	175	939	282	10.66
landline phone	1	13	no info	no info	0.79

as in the case of Common Voice project[3]. In contrast, the transcription and annotation of conversational speech often require specialized training and tools.

Conversational speech either collected in controlled environments (e.g., interviews) or sourced from existing resources (e.g., radio shows, podcasts) constitutes nearly 20% of all data.

A notable amount of available data (275 hours) is classified as public speech (8 datasets and nearly 5% of total). This category covers academic lectures captured in corpora AZON and recordings of politicians from the Polish or European Parliament.

Three datasets lack available information on the type of speech or the process used for its collection. These constitute only 1% of the material and generally refer to older corpora. Lastly, the *JURISDIC* dataset, which includes a diverse mix of speech types collected through both controlled and uncontrolled processes, makes up 14.88% or 855 hours of total documented data. Details can be found in Table 4.9.

What meta-data is available in publicly available datasets? The more detailed the speaker-level metadata, the more analytical dimensions are available to understand the factors that influence ASR performance [2]. Metadata is also required for corpus linguistic analyses, such as comparing differences in the average fundamental frequency of voice

[3] common voice validation guidelines

Rec. device	**Datasets**	**Trans. speech [h]**	**Recordings**	**Speakers**	**Percent of total [%]**
mobile phone	2	1775	no info	2402	55.96
landline phone	5	534	18	1342	16.83
studio mic	3	298	no info	798	9.39
headset	1	280	no info	200	8.83
various	1	269	no info	533	8.48
no info	1	16	no info	60	0.5

Table 4.7: Audio devices for commercially available datasets

Sampling rate	**Datasets**	**Trans. speech [h]**	**Recordings**	**Speakers**	**Percent of total [%]**
16000	22	3340.1	5571	5335	55.8
48000	2	1630	no info	5111	27.23
8000	6	535	596	1342	8.94
no info	17	252	1076	1813	4.21
22050	1	220	no info	210	3.68
44100	5	9	1395	61	0.15

Table 4.8: Distribution of sampling rate for publicly reported ASR speech datasets for Polish.

among different demographics.[111] The survey showed that half of the publicly available recordings are provided with metadata about the speaker's age and native language, and more than a third with information about the speaker's age. Approximately a third of the recordings also contain information about the speaker's accent or region. It should be noted that most of the recordings with rich speaker-level annotations originate from the recently created corpus (2023), *SpokesBiz* [4] [111]. Details can be found in Table 4.10.

4.1.4 Survey availability

The survey results, along with the searchable data catalog, have been made accessible through various public platforms such as:

1. GitHub repository where users can report an issue or request the registration of a new dataset.[5].

2. Dedicated website.[6].

[4] http://docs.pelcra.pl/doku.php?id=spokesbiz

[5] Polish ASR Speech Data Survey repo

[6] Polish ASR Speech Data Survey homepage

Speech type	Datasets	Trans. speech [h]	Recordings	Speakers	Percent of total [%]
read	25	3362.1	5942	9080	56.17
conversational	13	1184	1558	1717	19.78
various	4	1134	48	1684	18.94
public speech	8	275	725	1286	4.59
no info	3	31	365	105	0.52

Table 4.9: Distribution of speech types for publicly reported ASR speech datasets for Polish.

Table 4.10: Speaker and recordings meta-data availability in available speech datasets

	No of datasets		Trans. speech [h]		Coverage [%]	
	Free	Paid	Free	Paid	Free	Paid
Age info	3	4	798	578	48.63	18.22
Gender info	7	7	1008.1	2739	61.43	86.35
Nativity info	5	5	993.1	595	60.51	18.76
Time alignement annotation	2	3	682	452	41.56	14.25
Accent or region info	1	0	650	0	39.61	0

3. Google Sheet document and TSV files for independent analysis and automatic processing.[7].

4. Interactive catalog and up-to-date survey results on Hugging Face.[8]

Each platform is equipped with search functionality that enables the location of specific information within the dataset catalog. The initial version of the catalog was made public in January The survey results were published in March 2024 as a scientific article in the Poznan Studies in Contemporary Linguistics (PSICL)[9]. The catalog is updated regularly.

4.2 RO2: Design and curation of ASR benchmark dataset for Polish

4.2.1 Introduction

This section presents structured information on the dataset curated in this study. Dataset features presented in this section were derived from the original documentation through

[7] Polish ASR speech data catalog - Google Sheet

[8] AMU Polish ASR Survey

[9] PSICL article

the catalog and through the automatic analysis of the contents of the curated dataset. The curation and analysis processes are described in the Methodology chapter Section 3.3 Table 4.11 shows the total hours, recordings, and speakers in curated datasets.

Dataset	**Trans. speech [h]**	**Recordings**	**Speakers**
BIGOS	293	111272	3945
PELCRA	529	283258	972
Total	822	394530	4917

Table 4.11: Summary statistics of curated datasets

4.2.2 Datasets features derived from the documentation

This subsection discusses the characteristics derived from the initial documentation, organized based on the taxonomy outlined in Section3.3. The relevant characteristics of the curated datasets were collected from the Polish ASR speech data catalog.

Licensing and language coverage

Tables 4.12 and 4.13 provide licensing and dataset type information.

Dataset	**Codename**	**License**	**Languages**
Clarin Studio	pjatk-clarin_studio-15	CC-BY	monolingual
Clarin Mobile	pjatk-clarin_mobile-15	CC-BY	monolingual
Munich AI Labs LibriVox	mailabs-corpus_librivox-19	Proprietary	multilingual
Mozilla Common Voice	mozilla-common_voice_15-23	CC-0	multilingual
Multilingual Librispeech	fair-mls-20	CC-BY	multilingual
Azon Read	pwr-azon_read-20	CC-BY-SA	monolingual
Azon Spotaneous	pwr-azon_spont-20	CC-BY-SA	monolingual
PWR Male Set	pwr-maleset-unk	Public domain	monolingual
PWR Short Words	pwr-shortwords-unk	Public domain	monolingual
PWR Very Important Utter.	pwr-viu-unk	Public domain	monolingual
Google FLEURS	google-fleurs-22	CC-BY	multilingual
PolyAI Minds14	polyai-minds14-21	CC-BY	multilingual

Table 4.12: BIGOS dataset subset license and language coverage.

Dataset	Codename	License	Languages
DiaBiz ASR PolEval 22	ul-diabiz_poleval-22	Public domain	monolingual
SpokesBiz CBIZ_BIO	ul-spokes_biz_bio-23	CC-BY-NC-ND	monolingual
SpokesBiz CBIZ_INT	ul-spokes_biz_int-23	CC-BY-NC-ND	monolingual
SpokesBiz CBIZ_LUZ	ul-spokes_biz_luz-23	CC-BY-NC-ND	monolingual
SpokesBiz CBIZ_POD	ul-spokes_biz_pod-23	CC-BY-NC-ND	monolingual
SpokesBiz CBIZ_PRES	ul-spokes_biz_pres-23	CC-BY-NC-ND	monolingual
SpokesBiz CBIZ_VC	ul-spokes_biz_vc-23	CC-BY-NC-ND	monolingual
SpokesBiz CBIZ_VC2	ul-spokes_biz_vc2-23	CC-BY-NC-ND	monolingual
SpokesBiz CBIZ_WYW	ul-spokes_biz_wyw-23	CC-BY-NC-ND	monolingual
SpokesMix PELCRA_EMO	ul-spokes_mix_emo-18	CC-BY	monolingual
SpokesMix PELCRA_LUZ	ul-spokes_mix_luz-18	CC-BY	monolingual
SpokesMix PELCRA_PARL	ul-spokes_mix_parl-18	CC-BY	monolingual

Table 4.13: PELCRA for BIGOS dataset subset license and language coverage.

Domains, speech and interaction types.

Tables 4.14 and 4.15 outline domains, speech types, and interaction types for the BIGOS and PELCRA datasets.

Codename	Domain	Speech type	Interaction type
pjatk-clarin_studio-15	open domain	read	monolog
pjatk-clarin_mobile-15	open domain	read	monolog
mailabs-corpus_librivox-19	audiobook	read	monolog
mozilla-common_voice_15-23	open domain	read	monolog
fair-mls-20	audiobook	read	monolog
pwr-azon_read-20	scientific	read	monolog
pwr-azon_spont-20	scientific	spontaneous	monolog
pwr-maleset-unk	commands	read	monolog
pwr-shortwords-unk	commands	read	monolog
pwr-viu-unk	commands	read	monolog
google-fleurs-22	wikipedia	read	monolog
polyai-minds14-21	banking	read	monolog

Table 4.14: BIGOS dataset subset domains and speech types.

Codename	Domain	Speech type	Interaction type
ul-diabiz_poleval-22	customer service	spontaneous	dialog
ul-spokes_biz_bio-23	open domain	spontaneous	dialog
ul-spokes_biz_int-23	open domain	spontaneous	dialog
ul-spokes_biz_luz-23	open domain	spontaneous	dialog
ul-spokes_biz_pod-23	open domain	spontaneous	dialog
ul-spokes_biz_pres-23	open domain	spontaneous	dialog
ul-spokes_biz_vc-23	open domain	spontaneous	dialog
ul-spokes_biz_vc2-23	open domain	spontaneous	dialog
ul-spokes_biz_wyw-23	open domain	spontaneous	dialog
ul-spokes_mix_emo-18	open domain	spontaneous	dialog
ul-spokes_mix_luz-18	open domain	spontaneous	dialog
ul-spokes_mix_parl-18	open domain	spontaneous	monolog

Table 4.15: PELCRA for BIGOS dataset subset domains and speech types.

Sources, acoustic environments and devices

Details regarding speech sources, acoustic environments, and audio devices for BIGOs and PELCRA are available in Tables 4.16 and 4.17.

Codename	Speech source	Acoustic env.	Audio device
pjatk-clarin_studio-15	volunteers	quiet	studio mic
pjatk-clarin_mobile-15	volunteers	quiet	mobile phone
mailabs-corpus_librivox-19	volunteers	quiet	various
mozilla-common_voice_15-23	crowd	various	various
fair-mls-20	volunteers	various	various
pwr-azon_read-20	volunteers	quiet	studio mic
pwr-azon_spont-20	public speakers	mixed	lavalier
pwr-maleset-unk	volunteers	quiet	studio mic
pwr-shortwords-unk	volunteers	quiet	studio mic
pwr-viu-unk	volunteers	quiet	studio mic
google-fleurs-22	volunteers	quiet	mobile phone
polyai-minds14-21	crowd	quiet	mobile phone

Table 4.16: PELCRA for BIGOS dataset subset domains and speech types.

Codename	Speech source	Acoustic environment	Audio device
ul-diabiz_poleval-22	volunteers	quiet	telephone
ul-spokes_biz_bio-23	volunteers	quiet	lavalier mic
ul-spokes_biz_int-23	volunteers	quiet	lavalier mic
ul-spokes_biz_luz-23	volunteers	quiet	lavalier mic
ul-spokes_biz_pod-23	public speakers	quiet	various
ul-spokes_biz_pres-23	public speakers	quiet	various
ul-spokes_biz_vc-23	volunteers	quiet	lavalier mic
ul-spokes_biz_vc2-23	volunteers	quiet	lavalier mic
ul-spokes_biz_wyw-23	volunteers	quiet	lavalier mic
ul-spokes_mix_emo-18	volunteers	quiet	lavalier mic
ul-spokes_mix_luz-18	volunteers	quiet	lavalier mic

Table 4.17: PELCRA for BIGOS dataset subset domains and speech types.

4.2.3 Datasets features derived from the analysis of datasets contents

The content of the curated datasets was analyzed following the process described in Section 3.3.3

Size of the audio content

Tables 4.18 and 4.19 show the size metrics for BIGOS and PELCRA, including total hours of speech, number of samples, and unique speakers.

Subset	Transcribed audio[h]	Samples	Speakers
fair-mls-20	107.86	26072	24
google-fleurs-22	12.07	3937	3
mailabs-corpus_librivox-19	32.14	14862	2
mozilla-common_voice_15-23	53	36910	2920
pjatk-clarin_mobile-15	12.48	3495	117
pjatk-clarin_studio-15	56.43	13810	553
polyai-minds14-21	3.07	562	3
pwr-azon_read-20	5.72	2788	29
pwr-azon_spont-20	2.14	456	27
pwr-maleset-unk	6.38	4738	3
pwr-shortwords-unk	1.43	939	3
pwr-viu-unk	1.04	2703	3
total	293.76	111272	3945

Table 4.18: Audio content size metrics for BIGOS dataset

Subset	Transcribed audio[h]	Samples	Speakers
ul-diabiz_poleval-22	9.83	8950	170
ul-spokes_biz_bio-23	137.98	54917	158
ul-spokes_biz_int-23	2.25	1109	9
ul-spokes_biz_luz-23	74.27	41966	158
ul-spokes_biz_pod-23	55	22807	113
ul-spokes_biz_pres-23	32.25	17174	55
ul-spokes_biz_vc-23	52.07	45272	78
ul-spokes_biz_vc2-23	81.04	25802	84
ul-spokes_biz_wyw-23	28.21	11357	38
ul-spokes_mix_emo-18	25.61	24329	40
ul-spokes_mix_luz-18	18.74	20919	21
ul-spokes_mix_parl-18	12.27	8656	48
total	529.52	283258	972

Table 4.19: Audio content size metrics for PELCRA dataset

Size of the text content

Tables 4.20 and 4.21 show the size metrics for BIGOS and PELCRA, including the total number of samples, words, and characters.

Subset	Samples	Words	Characters
fair-mls-20	26072	886046	5639669
google-fleurs-22	3937	72641	509844
mailabs-corpus_librivox-19	14862	252479	1650672
mozilla-common_voice_15-23	36910	305333	2136502
pjatk-clarin_mobile-15	3495	91142	620158
pjatk-clarin_studio-15	13810	582840	2339971
polyai-minds14-21	562	10160	64431
pwr-azon_read-20	2788	27767	237161
pwr-azon_spont-20	456	17254	112521
pwr-maleset-unk	4738	39305	270386
pwr-shortwords-unk	939	9003	61752
pwr-viu-unk	2703	4776	30951
total	111272	2298746	13674018

Table 4.20: Text content size metrics for BIGOS dataset

Subset	Samples	Words	Characters
ul-diabiz_poleval-22	8950	105206	585481
ul-spokes_biz_bio-23	54917	1278269	7694395
ul-spokes_biz_int-23	1109	23123	141643
ul-spokes_biz_luz-23	41966	786593	4490695
ul-spokes_biz_pod-23	22807	605852	3650700
ul-spokes_biz_pres-23	17174	251841	1642817
ul-spokes_biz_vc-23	45272	568780	3348648
ul-spokes_biz_vc2-23	25802	755885	4526688
ul-spokes_biz_wyw-23	11357	259517	1552980
ul-spokes_mix_emo-18	24329	252380	1379695
ul-spokes_mix_luz-18	20919	204587	1132428
ul-spokes_mix_parl-18	8656	100992	669210
total	283258	5193025	30815380

Table 4.21: Text content size metrics for PELCRA dataset

Unique utterances, vocabulary, and alphabet size

The counts of unique utterances, vocabulary (words), and alphabet size (characters) are shown in Tables 4.22 and 4.23.

Subset	Unique utt.	Unique words	Unique chars
fair-mls-20	26069	89464	37
google-fleurs-22	1919	13826	71
mailabs-corpus_librivox-19	14796	51144	77
mozilla-common_voice_15-23	36853	66815	87
pjatk-clarin_mobile-15	3487	26424	35
pjatk-clarin_studio-15	13525	57853	39
polyai-minds14-21	550	1636	69
pwr-azon_read-20	1517	7628	32
pwr-azon_spont-20	456	5004	32
pwr-maleset-unk	4006	12970	62
pwr-shortwords-unk	668	3649	54
pwr-viu-unk	13	18	27

Table 4.22: Text content features for BIGOS dataset

Subset	Unique utt.	Unique words	Unique chars
ul-diabiz_poleval-22	8760	13716	72
ul-spokes_biz_bio-23	54096	108163	113
ul-spokes_biz_int-23	1100	5195	68
ul-spokes_biz_luz-23	41600	87990	105
ul-spokes_biz_pod-23	22753	69735	101
ul-spokes_biz_pres-23	17155	47352	100
ul-spokes_biz_vc-23	44647	63913	96
ul-spokes_biz_vc2-23	25567	79725	114
ul-spokes_biz_wyw-23	11192	39147	94
ul-spokes_mix_emo-18	20798	15485	67
ul-spokes_mix_luz-18	19526	20101	83
ul-spokes_mix_parl-18	8502	15338	78

Table 4.23: Text content features for PELCRA dataset

Speech rates

Speech rates derived from the analysis of audio and text content can be found in Tables 4.24 and 4.25.

Subset	Words per second	Chars per second
fair-mls-20	2.28	12.24
google-fleurs-22	1.67	10.06
mailabs-corpus_librivox-19	2.18	12.08
mozilla-common_voice_15-23	1.6	9.6
pjatk-clarin_mobile-15	2.03	11.77
pjatk-clarin_studio-15	2.87	8.65
polyai-minds14-21	0.92	4.91
pwr-azon_read-20	1.35	10.17
pwr-azon_spont-20	2.24	12.36
pwr-maleset-unk	1.71	10.05
pwr-shortwords-unk	1.76	10.32
pwr-viu-unk	1.27	6.98

Table 4.24: Audio content features for BIGOS dataset

Subset	**Words per second**	**Chars per second**
ul-diabiz_poleval-22	2.97	13.56
ul-spokes_biz_bio-23	2.57	12.92
ul-spokes_biz_int-23	2.85	14.62
ul-spokes_biz_luz-23	2.94	13.85
ul-spokes_biz_pod-23	3.06	15.38
ul-spokes_biz_pres-23	2.17	11.98
ul-spokes_biz_vc-23	3.03	14.83
ul-spokes_biz_vc2-23	2.59	12.93
ul-spokes_biz_wyw-23	2.56	12.74
ul-spokes_mix_emo-18	2.74	12.23
ul-spokes_mix_luz-18	3.03	13.75
ul-spokes_mix_parl-18	2.29	12.86

Table 4.25: Audio content features for PELCRA dataset

Average utterance durations

Tables 4.26 and 4.27 provide data on audio and utterance durations.

Subset	**Avg. dur. [s]**	**Avg. len. [words]**	**Avg. len. [chars]**
fair-mls-20	14.89	33.98	216.31
google-fleurs-22	11.04	18.45	129.5
mailabs-corpus_librivox-19	7.79	16.99	111.07
mozilla-common_voice_15-23	5.17	8.27	57.88
pjatk-clarin_mobile-15	12.86	26.08	177.44
pjatk-clarin_studio-15	14.71	42.2	169.44
polyai-minds14-21	19.65	18.08	114.65
pwr-azon_read-20	7.38	9.96	85.06
pwr-azon_spont-20	16.9	37.84	246.76
pwr-maleset-unk	4.85	8.3	57.07
pwr-shortwords-unk	5.44	9.59	65.76
pwr-viu-unk	1.39	1.77	11.45

Table 4.26: Average duration of audio recordings and utterances — BIGOS dataset.

Subset	Avg. dur. [s]	Avg. len. [words]	Avg. len. [chars]
ul-diabiz_poleval-22	3.96	11.75	65.42
ul-spokes_biz_bio-23	9.04	23.28	140.11
ul-spokes_biz_int-23	7.31	20.85	127.72
ul-spokes_biz_luz-23	6.37	18.74	107.01
ul-spokes_biz_pod-23	8.68	26.56	160.07
ul-spokes_biz_pres-23	6.76	14.66	95.66
ul-spokes_biz_vc-23	4.14	12.56	73.97
ul-spokes_biz_vc2-23	11.31	29.3	175.44
ul-spokes_biz_wyw-23	8.94	22.85	136.74
ul-spokes_mix_emo-18	3.79	10.37	56.71
ul-spokes_mix_luz-18	3.22	9.78	54.13
ul-spokes_mix_parl-18	5.1	11.67	77.31

Table 4.27: Average duration of audio recordings and utterances — PELCRA dataset.

Meta-data coverage

Tables 4.28 and 4.29 illustrate the levels of availability for speaker meta-data in the BIGOS and PELCRA datasets, respectively.

Subset	Gender coverage [%]	Age coverage [%]
fair-mls-20	N/A	N/A
google-fleurs-22	100.0	N/A
mailabs-corpus_librivox-19	100.0	N/A
mozilla-common_voice_15-23	63.92	63.95
pjatk-clarin_mobile-15	N/A	N/A
pjatk-clarin_studio-15	N/A	N/A
polyai-minds14-21	N/A	N/A
pwr-azon_read-20	100.0	N/A
pwr-azon_spont-20	100.0	N/A
pwr-maleset-unk	100.0	N/A
pwr-shortwords-unk	100.0	N/A

Table 4.28: Coverage of speaker meta-data — BIGOS dataset

Subset	Gender coverage [%]	Age coverage [%]
ul-diabiz_poleval-22	N/A	N/A
ul-spokes_biz_bio-23	100.0	100.0
ul-spokes_biz_int-23	100.0	100.0
ul-spokes_biz_luz-23	100.0	100.0
ul-spokes_biz_pod-23	100.0	100.0
ul-spokes_biz_pres-23	100.0	100.0
ul-spokes_biz_vc-23	100.0	100.0
ul-spokes_biz_vc2-23	100.0	100.0
ul-spokes_biz_wyw-23	100.0	100.0
ul-spokes_mix_emo-18	100.0	100.0
ul-spokes_mix_luz-18	100.0	100.0

Table 4.29: Coverage of speaker meta-data — PELCRA dataset

4.2.4 Availability of curated datasets

The following methods were used to share the dataset and gather feedback from the community:

1. **Accessibility, discoverability, and tracking**: The curated BIGOS [10] and PELCRA [11] datasets were uploaded to the Hugging Face datasets hub, ensuring discoverability and long-term accessibility with easy access control. It also enables tracking of number of downloads. Table 4.30 shows publication dates and download counts.

2. **Open licensing**: The datasets were shared under open licenses: BIGOS under CC-BY-SA and PELCRA under CC-BY-NC-ND.

3. **Public dashboards**: Dataset content analysis results were made available on a public dashboard,[12] providing insights into the dataset contents and more informed interpretation of evaluation results.

4. **Feedback mechanisms**: The Hugging Face platform allows direct feedback from users of datasets.

Dataset	Publication date	Downloads
BIGOS V2	November 2023	4,886
PELCRA for BIGOS	December 2023	977

Table 4.30: Publication date and number of downloads of BIGOS datasets as of June 6th 2024.

[10] BIGOS dataset on HF datasets
[11] PELCRA for BIGOS dataset on HF hub
[12] AMU BIGOS dataset dashboard

4.3 RO3: Survey of ASR benchmarks for Polish

4.3.1 Introduction

Research questions concerning the survey of ASR benchmarks include:

- **RQ 8:** How to identify and systematically categorize Polish ASR benchmarks using publicly available information?

- **RQ 9:** What methods, datasets and ASR systems have been considered in the Polish ASR benchmarks so far?

- **RQ 10:** What automatic speech recognition (ASR) systems supporting Polish have not yet been evaluated?

- **RQ 11:** Which ASR benchmarks have evaluated commercial and freely available systems?

- **RQ 12:** What ASR system is ranked as the best performing one?

- **RQ 13:** How are the main conclusions derived from the ASR benchmarks so far?

- **RQ 14:** How the results of the survey of the Polish ASR benchmark be shared with the community?

The survey methodology presented in Section 3.4 constitutes the answer to research question 7. The following sections provide answers to the remaining research questions, as outlined in Table 4.31.

Research question ID	**Subsection with relevant results**
RQ8	3.4
RQ9	4.3.2
RQ10	4.3.2
RQ11	4.3.2
RQ12, RQ13	4.3.2
RQ14	4.3.2

Table 4.31: Overview of sections providing relevant results to research questions RQ7-RQ13

4.3.2 Results

Polish ASR systems benchmarks overview

As of February 2024, six benchmarks of Polish ASR systems were reported in the public domain.

Benchmark	Use cases
BOR POLSL PS 18	Voice Control
PolEval PJATK 19	Oration
DiaBiz CLARIN Voicelab 22	Conversations
SpokesBiz CLARIN 23	Conversations, Meetings, Orations
Medical UW SOVVA PS 23	Dictation
Medical PG 23	Dictation

Table 4.32: Overview of ASR use-cases covered in Polish ASR benchmarks to date.

Table 4.33 displays the benchmarks reported by year, the number of systems, the metrics, and the datasets used.

Benchmark	Year	Systems	Datasets	Metrics automatic	Metrics manual
BOR POLSL PS 18	2018	3	1	3	0
PolEval PJATK 19	2019	6	1	1	0
DiaBiz CLARIN Voicelab 22	2022	3	7	3	0
Medical PG 23	2023	3	1	6	0
Medical UW PS 23	2023	3	1	5	3
SpokesBiz CLARIN 23	2023	1	8	3	0

Table 4.33: Public domain ASR benchmarks 2018-2023.

Datasets and domains

Metrics

Systems Table 4.38 presents the benchmarks conducted from 2018 to 2023, along with the count and details of the ASR systems evaluated. In total 19 systems or system-model combinations were benchmarked.

Table 4.39 shows the number of independent evaluations of various ASR systems for the Polish language that have been publicly reported.

Polish ASR systems lacking public benchmark evaluations.

Table 4.40 presents ASR systems that support the Polish language (as of February 2024) and have not been publicly evaluated before this research.

Benchmark	Domain	Speech types	Audio sources	Recording devices
BOR POLSL PS 18	government training	read	field recordings	lavalier microphone
PolEval PJATK 19	parliamentary speech	read	field recordings	venue microphone
DiaBiz CLARIN 22	customer support	spontaneous	phone conversations	phone
SpokesBiz CLARIN 23	various	spontaneous	podcasts, interviews	various
Medical UW PS 23	medical terms	read	field recordings	lavalier microphone
Medical PG 23	medical terms	read	field recordings	lavalier microphone

Table 4.34: Overview of domains, speech types, audio sources and recording devices.

Benchmark	Audio [hours]	Domains	Recordings	Speakers
BOR POLSL PS 18	1	1	140	18
PolEval PJATK 19	1	1	29	29
DiaBiz CLARIN 22	41	7	400	151
SpokesBiz CLARIN 23	52	7	79	79
Medical UW PS 23	1	1	1000	no info
Medical PG 23	1	1	1200	10

Table 4.35: Datasets size and number of domains, recordings, and speakers.

Types of ASR systems included in benchmarks

Best performing ASR systems and major conclusions across benchmarks

BOR POLSL PS 18

- Tested systems are insufficient for BOR officer training applications characterized by the noisy acoustic conditions and emotional, rapid speech. The tested systems are more suited to recognize speech dictated by a single speaker in quiet acoustic conditions.

- Correct recognition of entire speech segments is rare. Many recognitions are completely incorrect or contain similar-sounding words. The best upper-bound recognition rate for selected commands was achieved with the Google Cloud Speech-to-text system (90% and 60% correctly recognized words in clean and noisy conditions, respectively.) The average recognition rate in all test cases and scenarios was not reported. The examplary results for clean and noisy conditions are presented in Fig-

Benchmark	**Acoustic conditions**	**Recordings annotations**	**Speaker meta-data**
BOR POLSL PS 18	mixed	none	none
PolEval PJATK 19	mixed	none	none
DiaBiz CLARIN 22	mixed	timestamped diarizations, non-speech events	age, gender, education
SpokesBiz CLARIN 23	mixed	timestamped diarizations, non-speech events	age, gender, education
Medical UW PS 23	clean	none	age, gender, region

Table 4.36: Acoustic conditions, annotations, and speaker meta-data across Polish ASR benchmarks

Benchmark	**Automatic evaluation**	**Human evaluation**	**Lexicon based metrics**	**Annotation based metrics**
BOR POLSL PS 18	yes	no	SRR, WRR	none
PolEval PJATK 19	yes	no	WER	none
DiaBiz CLARIN Voicelab 22	yes	no	WER	none
SpokesBiz CLARIN 23	yes	no	WER, MER, WIL	none
Medical UW PS 23	yes	yes	Accuracy, WER, LED, JWS	Error types (misrecognition, quality, word boundary)
Medical PG 23	yes	no	WER, MER, WIL, CER, LED, Jaccard distance	none

Table 4.37: Overview of metrics employed in Polish ASR systems benchmarks.

ure 4.2 and 4.3, respectively. Results are presented for individual commands in test set (denoted as *K1, K2 etc.*)

Benchmark	Evaluated systems	Models evaluated
BOR POLSL PS 18	ARM, Skrybot, Google	3
PolEval PJATK 19	GOLEM, ARM-1, SGMM2, tri2a, clarin-pl-studio, clarin-pl-sejm	6
DiaBiz CLARIN Voicelab 22	Azure, Google, Voicelab	3
SpokesBiz CLARIN 23	Whisper (large)	1
Medical UW PS 23	Azure, Google, Techmo	3
Medical PG 23	Azure, Google, Whisper (large-v2)	3
Total		19

Table 4.38: Publicly reported evaluations of ASR models for Polish language.

System	Benchmarks system
azure_latest	3
google_default	4
skrybot_default	1
voicelab_default	1
arm_default	2
techmo_default	1
clarin_studio_kaldi_default	1
clarin_pl_sejm_default	1
golem_default	1
sgmm2_default	1
tri2a_default	1
whisper_local_large-v2	2
Total	19

Table 4.39: Number of reported independent evaluations and benchmarks per system.

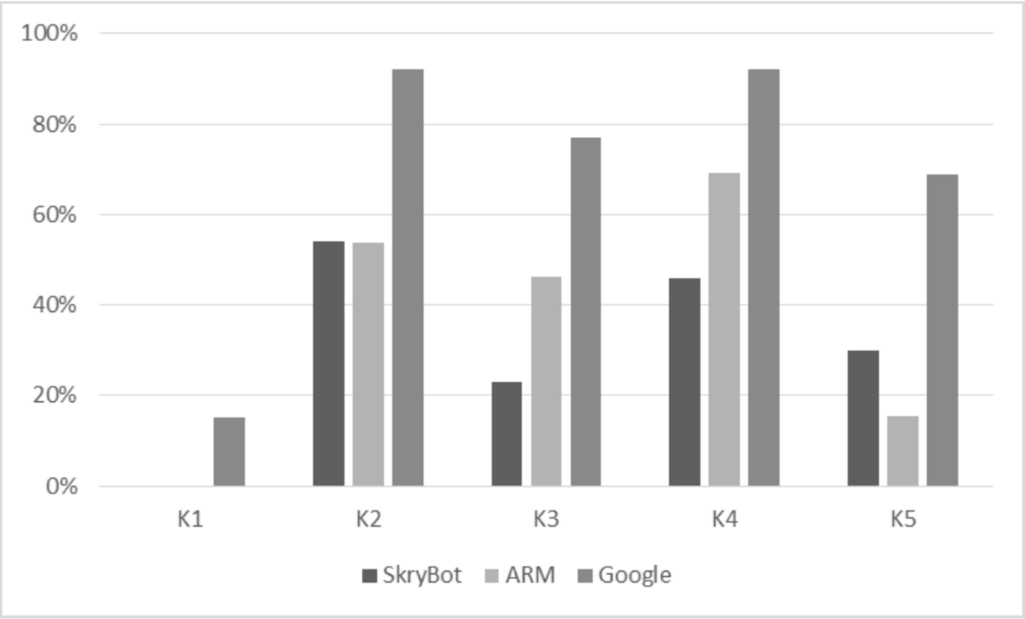


Figure 4.2: ASR benchmark results — POLSL dataset. Source: [99].

System	Model	Type	License
google_v2	long	commercial	Proprietary
google_v2	short	commercial	Proprietary
google	latest_long	commercial	Proprietary
google	latest_short	commercial	Proprietary
google	command_and_search	commercial	Proprietary
whisper_cloud	whisper-1	commercial	Proprietary
assembly_ai	best	commercial	Proprietary
assembly_ai	nano	commercial	Proprietary
notta.ai	default	commercial	Proprietary
mms	1b-all	free	CC-BY-NC
mms	1b-fl102	free	CC-BY-NC
mms	1b-l1107	free	CC-BY-NC
nemo	stt_pl_fastconformer_hybrid_large_pc	free	CC-BY
nemo	nemo_stt_multilingual_fastconformer...	free	CC-BY
nemo	stt_pl_quartznet15x5	free	CC-BY
whisper_local	tiny	free	MIT
whisper_local	base	free	MIT
whisper_local	small	free	MIT
whisper_local	medium	free	MIT
whisper_local	large-v1	free	MIT
whisper_local	large-v3	free	MIT
wav2vec	xls-r-1b-polish	free	Apache
wav2vec	large_xlsr-53-polish	free	Apache

Table 4.40: ASR systems supporting Polish not yet evaluated in the public domain.

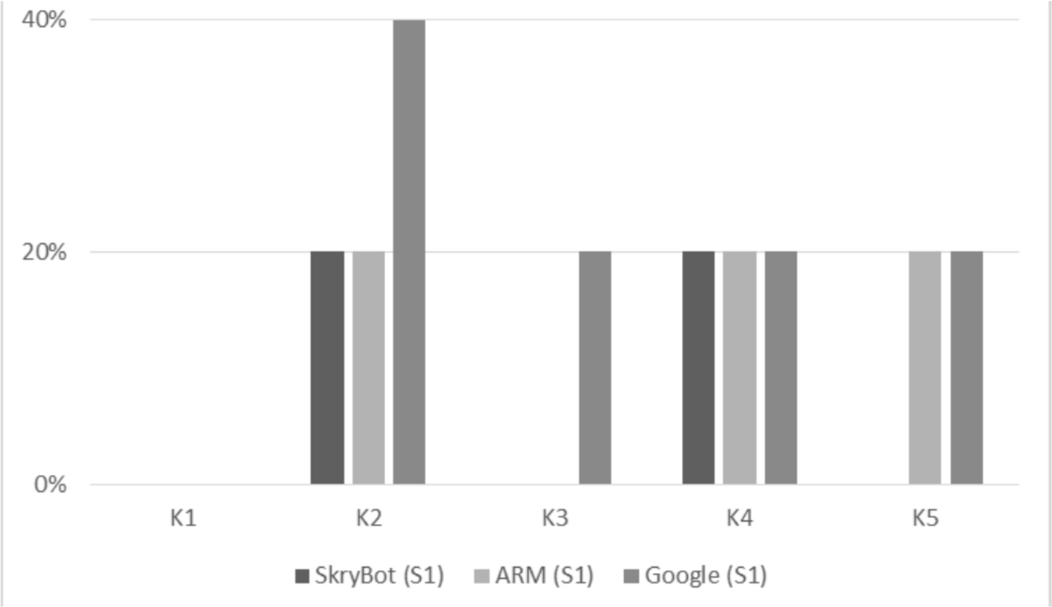
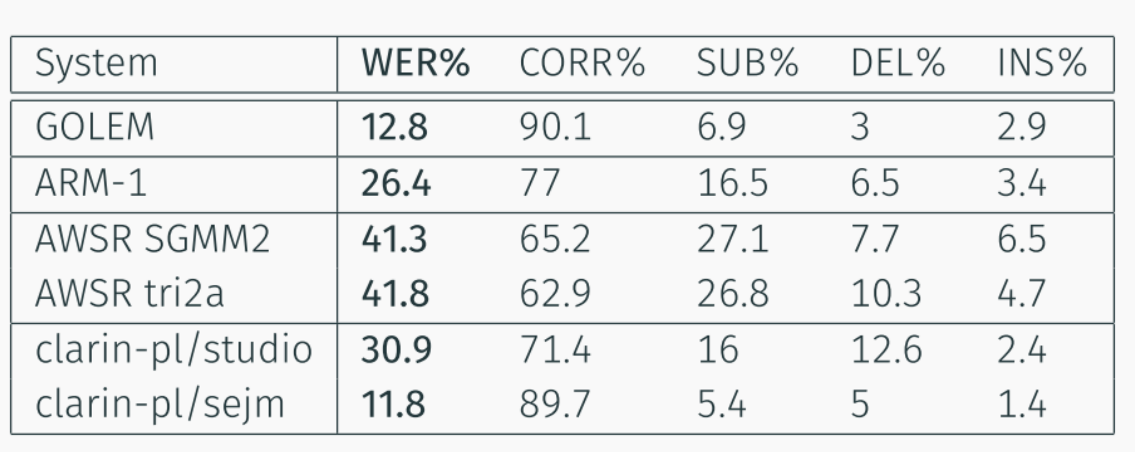


Figure 4.3: ASR benchmark results — BOR dataset scenario 1, year 2018. Source: [99]

PolEval PJATK 19

Benchmark	Year	System types
BOR POLSL PS 18	2018	Commercial
PolEval PJATK 19	2019	Community provided
DiaBiz CLARIN 22	2022	Commercial
SpokesBiz CLARIN 23	2023	Commercial
Medical UW PS 23	2023	Commercial
Medical PG 23	2023	Commercial + Public domain

Table 4.41: Types of ASR systems evaluated in public domain ASR benchmarks 2018-2023.



System	**WER%**	CORR%	SUB%	DEL%	INS%
GOLEM	**12.8**	90.1	6.9	3	2.9
ARM-1	**26.4**	77	16.5	6.5	3.4
AWSR SGMM2	**41.3**	65.2	27.1	7.7	6.5
AWSR tri2a	**41.8**	62.9	26.8	10.3	4.7
clarin-pl/studio	**30.9**	71.4	16	12.6	2.4
clarin-pl/sejm	**11.8**	89.7	5.4	5	1.4

Figure 4.4: ASR benchmark results — PolEval year 2019. Source: [59]

- All systems, except for ARM-1, were based on Kaldi framework framework.

- All systems, except for *clarin-pl/sejm* and *clarin-pl/studio*, used GMM models.

- The best systems in the *fixed* competition (limited training data available) was *GOLEM* with WER of 12.8%. The best system in the *open* competition was *clarin-pl/sejm* with WER of 11.8 %. Full results from the original report are presented in Figure 4.4

DiaBiz CLARIN Voicelab 22

- Microsoft's Azure service achieved the best WER (10.51%) for both channels.

- Voicelab's ASR had an overall WER of 11.51 %, close to Azure's performance.

- Google's ASR service for Polish had a worse WER of 20.84% on the DiaBiz dataset.

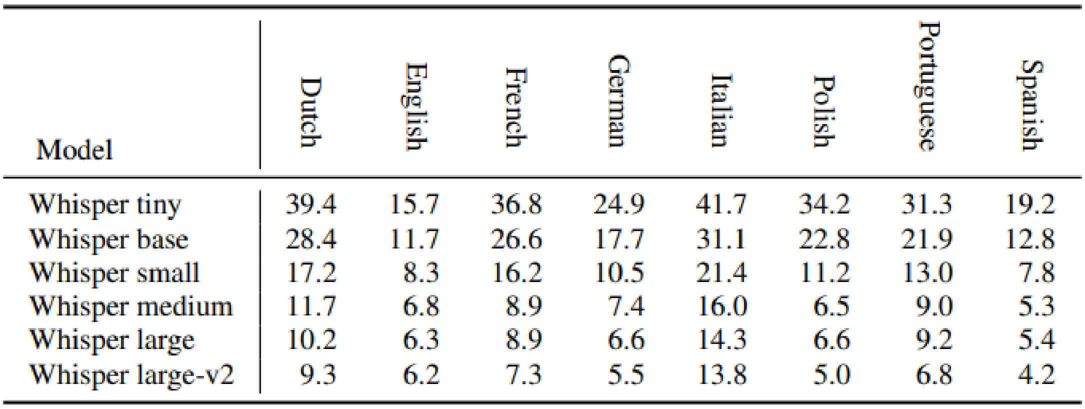
- Azure outperformed other ASR systems in 8 out of 9 domains.

- Voicelab's ASR was slightly better for telecommunications customer support dialogues.

Vendor	Total WER	Client WER	Agent WER
Microsoft Azure	10.51*	13.9 *	8.89*
Voicelab	11.51	14.86	9.92
Google	20.84	24.95	18.89

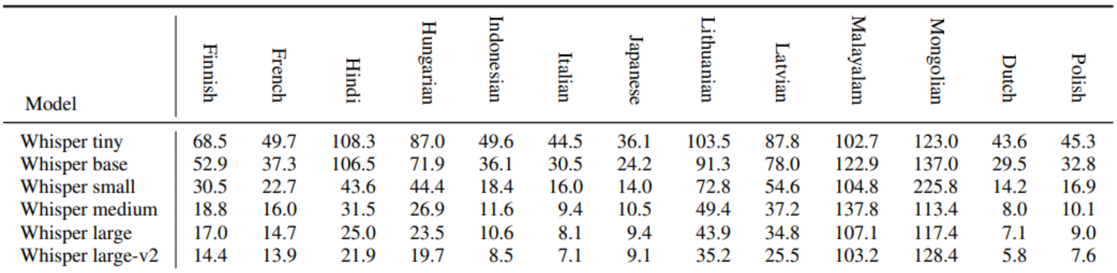
Figure 4.5: ASR benchmark results — DiaBiz corpus, year 2022. Source: [110]

Whisper 22



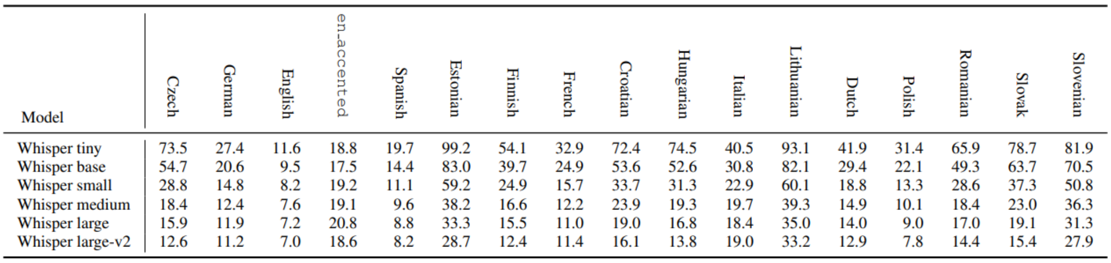
Model	Dutch	English	French	German	Italian	Polish	Portuguese	Spanish
Whisper tiny	39.4	15.7	36.8	24.9	41.7	34.2	31.3	19.2
Whisper base	28.4	11.7	26.6	17.7	31.1	22.8	21.9	12.8
Whisper small	17.2	8.3	16.2	10.5	21.4	11.2	13.0	7.8
Whisper medium	11.7	6.8	8.9	7.4	16.0	6.5	9.0	5.3
Whisper large	10.2	6.3	8.9	6.6	14.3	6.6	9.2	5.4
Whisper large-v2	9.3	6.2	7.3	5.5	13.8	5.0	6.8	4.2

Figure 4.6: ASR benchmark results — Whisper, MLS corpus, year 2022. Source: [117]



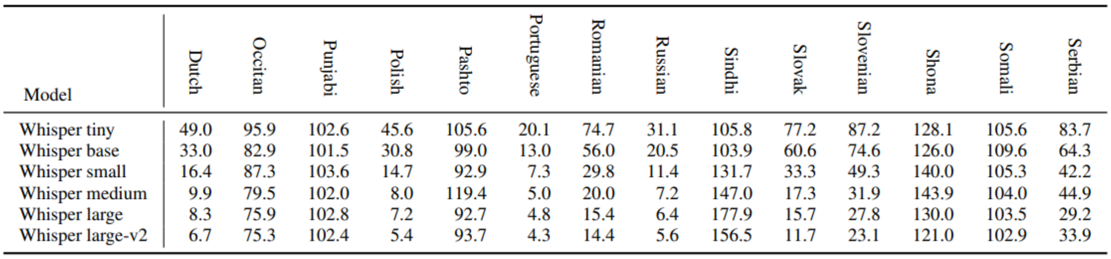
Model	Finnish	French	Hindi	Hungarian	Indonesian	Italian	Japanese	Lithuanian	Latvian	Malayalam	Mongolian	Dutch	Polish
Whisper tiny	68.5	49.7	108.3	87.0	49.6	44.5	36.1	103.5	87.8	102.7	123.0	43.6	45.3
Whisper base	52.9	37.3	106.5	71.9	36.1	30.5	24.2	91.3	78.0	122.9	137.0	29.5	32.8
Whisper small	30.5	22.7	43.6	44.4	18.4	16.0	14.0	72.8	54.6	104.8	225.8	14.2	16.9
Whisper medium	18.8	16.0	31.5	26.9	11.6	9.4	10.5	49.4	37.2	137.8	113.4	8.0	10.1
Whisper large	17.0	14.7	25.0	23.5	10.6	8.1	9.4	43.9	34.8	107.1	117.4	7.1	9.0
Whisper large-v2	14.4	13.9	21.9	19.7	8.5	7.1	9.1	35.2	25.5	103.2	128.4	5.8	7.6

Figure 4.7: ASR benchmark results — Whisper, CommonVoice corpus, year 2022. Source: [117]



Model	Czech	German	English	en_accented	Spanish	Estonian	Finnish	French	Croatian	Hungarian	Italian	Lithuanian	Dutch	Polish	Romanian	Slovak	Slovenian
Whisper tiny	73.5	27.4	11.6	18.8	19.7	99.2	54.1	32.9	72.4	74.5	40.5	93.1	41.9	31.4	65.9	78.7	81.9
Whisper base	54.7	20.6	9.5	17.5	14.4	83.0	39.7	24.9	53.6	52.6	30.8	82.1	29.4	22.1	49.3	63.7	70.5
Whisper small	28.8	14.8	8.2	19.2	11.1	59.2	24.9	15.7	33.7	31.3	22.9	60.1	18.8	13.3	28.6	37.3	50.8
Whisper medium	18.4	12.4	7.6	19.1	9.6	38.2	16.6	12.2	23.9	19.3	19.7	39.3	14.9	10.1	18.4	23.0	36.3
Whisper large	15.9	11.9	7.2	20.8	8.8	33.3	15.5	11.0	19.0	16.8	18.4	35.0	14.0	9.0	17.0	19.1	31.3
Whisper large-v2	12.6	11.2	7.0	18.6	8.2	28.7	12.4	11.4	16.1	13.8	19.0	33.2	12.9	7.8	14.4	15.4	27.9

Figure 4.8: ASR benchmark results — Whisper, VoxPopuli corpus, year 2022. Source: [117]

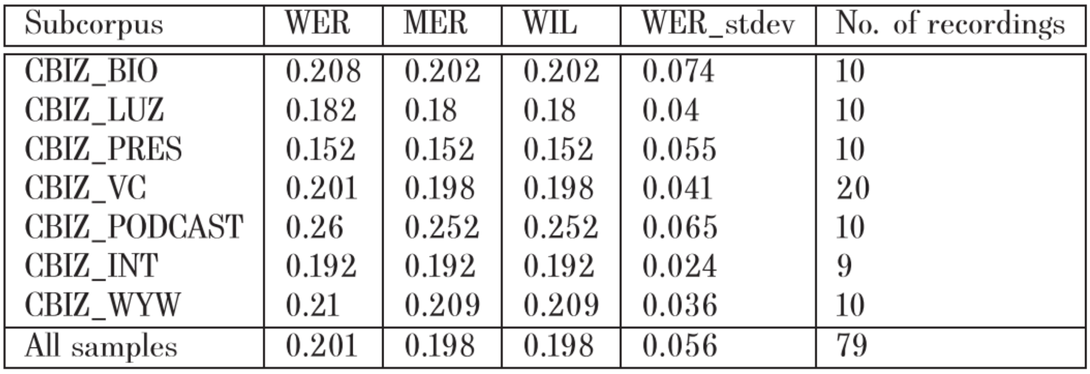


Model	Dutch	Occitan	Punjabi	Polish	Pashto	Portuguese	Romanian	Russian	Sindhi	Slovak	Slovenian	Shona	Somali	Serbian
Whisper tiny	49.0	95.9	102.6	45.6	105.6	20.1	74.7	31.1	105.8	77.2	87.2	128.1	105.6	83.7
Whisper base	33.0	82.9	101.5	30.8	99.0	13.0	56.0	20.5	103.9	60.6	74.6	126.0	109.6	64.3
Whisper small	16.4	87.3	103.6	14.7	92.9	7.3	29.8	11.4	131.7	33.3	49.3	140.0	105.3	42.2
Whisper medium	9.9	79.5	102.0	8.0	119.4	5.0	20.0	7.2	147.0	17.3	31.9	143.9	104.0	44.9
Whisper large	8.3	75.9	102.8	7.2	92.7	4.8	15.4	6.4	177.9	15.7	27.8	130.0	103.5	29.2
Whisper large-v2	6.7	75.3	102.4	5.4	93.7	4.3	14.4	5.6	156.5	11.7	23.1	121.0	102.9	33.9

Figure 4.9: ASR benchmark results — Whisper, FLEURS corpus, year 2022. Source: [117]

SpokesBiz CLARIN 23

- The WER on all samples is 20.1%.
- WER ranges between 15.2% and 26%.
- The quality of the recording and the vocabulary domain greatly affect WER.
- Whisper V2 accuracy differs from official evaluations reported by Radford et al. [117] on Common Voice (7.6%), VoxPopuli (7.8%), MLS (5%) and FLEURS (5.4%) datasets.



Subcorpus	WER	MER	WIL	WER_stdev	No. of recordings
CBIZ_BIO	0.208	0.202	0.202	0.074	10
CBIZ_LUZ	0.182	0.18	0.18	0.04	10
CBIZ_PRES	0.152	0.152	0.152	0.055	10
CBIZ_VC	0.201	0.198	0.198	0.041	20
CBIZ_PODCAST	0.26	0.252	0.252	0.065	10
CBIZ_INT	0.192	0.192	0.192	0.024	9
CBIZ_WYW	0.21	0.209	0.209	0.036	10
All samples	0.201	0.198	0.198	0.056	79

Figure 4.10: ASR evaluation results — SpokesBiz corpus, year 2023. Source: [111]

Medical UW PS 23

- Accuracy of all three tested ASR systems was greater than 86%, with a difference of only 1.7% between the best and worst result. Google ASR was best, followed by Techmo ASR and Microsoft ASR,
- The best, average and worst result remained the same for remaining metrics (WER, Levenshtein editing distance algorithm and Jaro-Winkler similarity), indicating strong correlation among metrics.
- Manual errors classification was performed. Three types of errors were considered:
  - *Misrecognitions* – instances where the ASR system did not recognize the recorded expression, either completely or with an altered meaning.
  - *Quality Problems* – situations where the ASR system identified the recorded expression with slight inaccuracies that did not alter the overall meaning of the phrase.
  - *Word Boundaries* – instances where the ASR system accurately detected the spoken phrase, yet failed to correctly identify initial or final segments of some words.
- Manual inspection highlighted differences in types of errors of individual systems:
  - Techmo ASR had the lowest percentage of *Misrecognitions* (61,7%) as compared to

  Microsoft ASR (71,2%) and Google ASR (73,2%).

  - Microsoft ASR had the lowest percentage of *Quality Problems* (14,8%) as compared to Google ASR (23,6%) and Techmo ASR (26,6%)
  - Google ASR had the lowest percentage of *Word Boundaries* (3,2%) as compared to Microsoft ASR and Techmo ASR (14,0%) for Microsoft ASR (11,7%).

Medical PG 23 According to the authors, the results obtained "show that the highest efficiency for most cases was obtained by Azure speech-to-text. However, none of the tested models is ready for voice filling medical records, describing cases, or prescribing treatment,

	Google ASR	Microsoft ASR	Techmo ASR
Accuracy	88.1%	86.4%	87.5%

Figure 4.11: ASR benchmark results — Accuracy of medical terms recognition, Kuligowska et. al, year 2023. Source: [68]

	Misrecognitions	Quality Problems	Word Boundaries
Google ASR	73.2%	23.6%	3.2%
Microsoft ASR	71.2%	14.8%	14.0%
Techmo ASR	61.7%	26.6%	11.7%

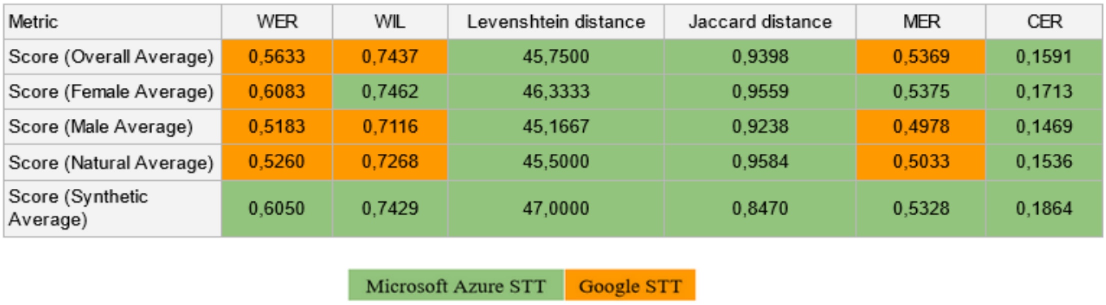
Figure 4.12: ASR benchmark results — Recognition errors classification, Kuligowska et. al, year 2023. Source: [68]

because the number of errors made when converting speech to text is too high."[153] Specifically, the average WER for the best ASR system (Google STT) for all speech types (male, female, synthetic) exceeds 56%.

Benchmarks survey results sharing

The survey results were made publicly available on GitHub [13] and Hugging Face[14] platforms along Polish ASR speech datasets survey, as well as publicly accessible spreadsheet [15].

[13] Polish ASR survey GitHub
[14] Polish ASR survey HuggingFace
[15] Polish ASR benchmarks catalog



Metric	WER	WIL	Levenshtein distance	Jaccard distance	MER	CER
Score (Overall Average)	🟧 0,5633	🟧 0,7437	🟩 45,7500	🟩 0,9398	🟧 0,5369	🟩 0,1591
Score (Female Average)	🟧 0,6083	🟩 0,7462	🟩 46,3333	🟩 0,9559	🟩 0,5375	🟩 0,1713
Score (Male Average)	🟧 0,5183	🟧 0,7116	🟩 45,1667	🟩 0,9238	🟧 0,4978	🟩 0,1469
Score (Natural Average)	🟧 0,5260	🟧 0,7268	🟩 45,5000	🟩 0,9584	🟧 0,5033	🟩 0,1536
Score (Synthetic Average)	🟩 0,6050	🟩 0,7429	🟩 47,0000	🟩 0,8470	🟩 0,5328	🟩 0,1864

🟩 Microsoft Azure STT 🟧 Google STT

Figure 4.13: ASR benchmark results — Medical terms recognition, Zielonka et. al, year 2023. Source: [153]

4.4 RO5: Use of curated dataset for benchmarking ASR systems for Polish

4.4.1 Introduction

The benchmark of ASR systems supporting Polish was performed on BIGOS and PELCRA datasets. The curation process and datasets content were described in detail in Sections 3.3 and 4.2.3, respectively. For simplicity, in this section the subsets of test splits from the BIGOS and PELCRA4BIGOS datasets used for evaluation are referred to as BIGOS and PELCRA, respectively. Table 4.42 outlines the types and number of evaluated systems.

Benchmark name	**System types**	**Number of systems**
BIGOS	Commercial + Public domain	25
PELCRA FOR BIGOS	Commercial + Public domain	25

Table 4.42: ASR benchmarks performed in this study.

The evaluated systems are described in Section 3.6.2. All available ASR systems for Polish, which have not been evaluated to date (4.40.), except *notta.ai*[16], were included in this benchmark.

Table 4.43 outlines the evaluation scenarios and the relevant sections that present the results. The relation between specific evaluation scenarios and research questions was described in Section 3.6.2.

Eval scenario	**Scenario codename**	**Relevant section**
ES1	accuracy_across_systems	4.4.3
ES2	accuracy_per_dataset	4.4.3
ES3	accuracy_per_system_type	4.4.3
ES4	accuracy_per_model_size	4.4.3
ES5	accuracy_per_audio_duration	4.4.3
ES6	accuracy_per_speaking_rate	4.4.3
ES7	accuracy_per_speaker_age_group	4.4.3
ES8	accuracy_per_speaker_gender	4.4.3

Table 4.43: ASR systems evaluation scenarios overview

4.4.2 Evaluation setups

Tables 4.44 and 4.45 show an overview of the evaluation parameters for the BIGOS and PELCRA datasets, respectively.

[16] notta.ai/transcribe-pl

Attribute	Value
Evaluation date	March 2024
Number of evaluated system-model variants	25
Dataset	pl-asr-bigos-v2
Split	test
Text reference type	original
Normalization steps	all
Number of dataset subsets	12
Number of evaluated system-model-subset combinations	300
Number of unique speakers in dataset	83
Number of unique recordings used for evaluation	1993
Total size of the evaluation dataset	4.85 hours
Total number of test cases (audio-hypothesis pairs)	49726

Table 4.44: Evaluation details for BIGOS dataset

Attribute	Value
Evaluation date	March 2024
Number of evaluated system-model variants	25
Dataset	pl-asr-pelcra-for-bigos
Split	test
Text reference type	original
Normalization steps	all
Number of evaluated subsets	12
Number of evaluated system-model-subsets combinations	300
Number of unique speakers	22
Number of unique recordings used for evaluation	2345
Total size of the dataset	4.71 hours
Total number of test cases (audio-hypothesis pairs)	56354

Table 4.45: Evaluation details for PELCRA dataset

4.4.3 Evaluation scenarios

This section shows the results of the evaluation scenarios described in 3.6.2. For each scenario, the results are presented independently for the BIGOS and PELCRA datasets.

ES1 — Accuracy across systems

BIGOS dataset Table 4.46 shows the median, mean, standard deviation, minimum, and maximum values of WER. These scores are for the variants of the system model evaluated on a subset of recordings from the BIGOS dataset. The results are sorted from the best to the worst median WER score.

Figure 4.14 shows the box plot of WER scores (without outliers) of 25 systems evaluated on BIGOS dataset. The x-axis lists the different ASR systems evaluated. The y-axis

	Word Error Rate (WER) [%]				
Model	**Median**	**Mean**	**Std.**	**Min.**	**Max.**
whisper_large_v3	5.53	8.38	7.5	3.79	29.17
whisper_cloud	5.97	10.05	8.53	4.17	31.85
assembly_best	6.25	8.47	6.15	4.46	25.2
whisper_large_v2	6.61	9.58	8.68	5.03	35.32
whisper_large_v1	7.84	10.5	8.75	4.11	36.01
whisper_medium	9.03	11.52	8.62	5.94	36.61
google_long	10.26	11.12	7.74	0	26.28
google_v2_long	10.38	11.24	7.78	0.28	27.37
w2v-1b-pl	10.38	13.74	8.12	7.43	32.93
google_short	11.09	12.46	7.02	0	28.27
mms_all	11.93	16.9	11.29	7.15	42.46
nemo_multilang	12.75	15.03	10.65	2.1	42.56
nemo_pl_confromer	12.75	15.03	10.65	2.1	42.56
nemo_pl_quartznet	12.75	15.03	10.65	2.1	42.56
google_v2_short	14.98	17.8	13.71	0	52.73
google_cmd_search	15.19	14.52	6.69	0.28	22.97
azure_latest	15.38	18.85	13.2	1.14	39.87
whisper_small	15.92	20.52	16.26	8.11	68.65
google_default	16.51	15.53	7.29	0.28	25.22
mms_1107	19.16	22.25	10.47	11.29	47.52
w2v-53-pl	20.7	25.53	16.77	6.81	69.35
mms_102	20.83	22.56	10.35	11.69	45.83
whisper_base	31.79	35.6	17.86	21.49	85.81
assembly_nano	32.83	52.53	66.27	14.54	260.86
whisper_tiny	44.4	54.28	28.92	33.71	139.29

Table 4.46: WER statistics – BIGOS dataset

represents the WER scores in percentages. A lower WER score signifies better performance.

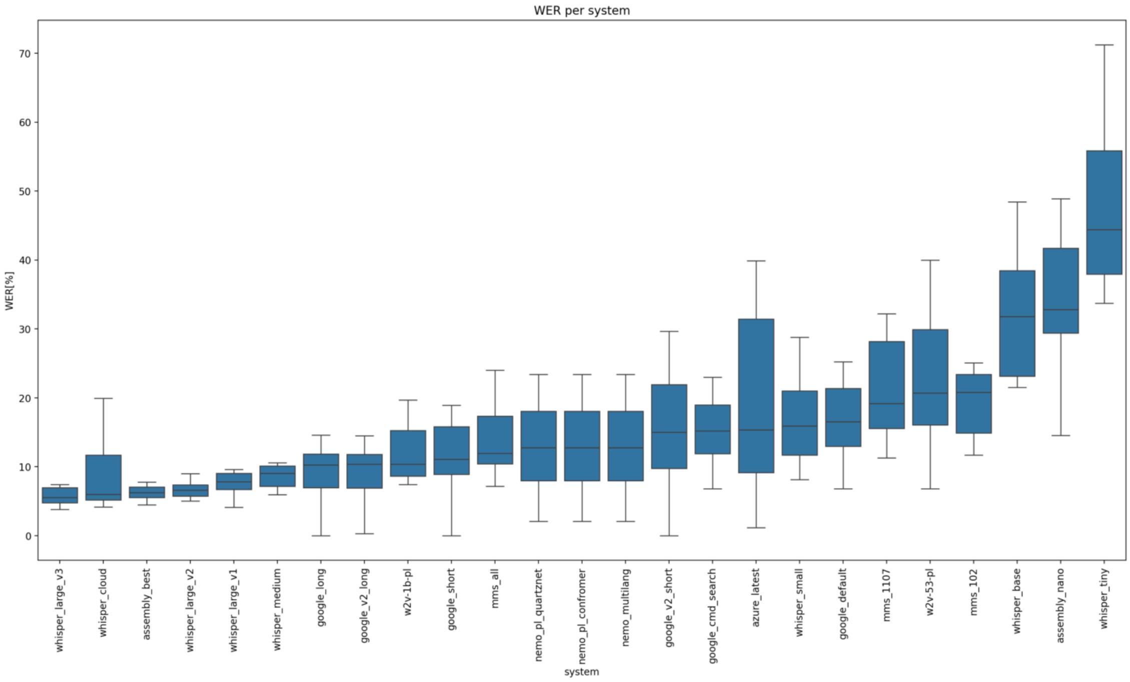


Figure 4.14: Mean WER per system for all BIGOS dataset subsets.

PELCRA dataset Table 4.47 presents the WER scores statistics derived from evaluation on subset of recordings from the PELCRA dataset test split. The results are ordered from the lowest to the highest median WER score. Figure 4.15 illustrates the box plot of WER scores (excluding outliers) for 25 systems assessed on the PELCRA dataset.

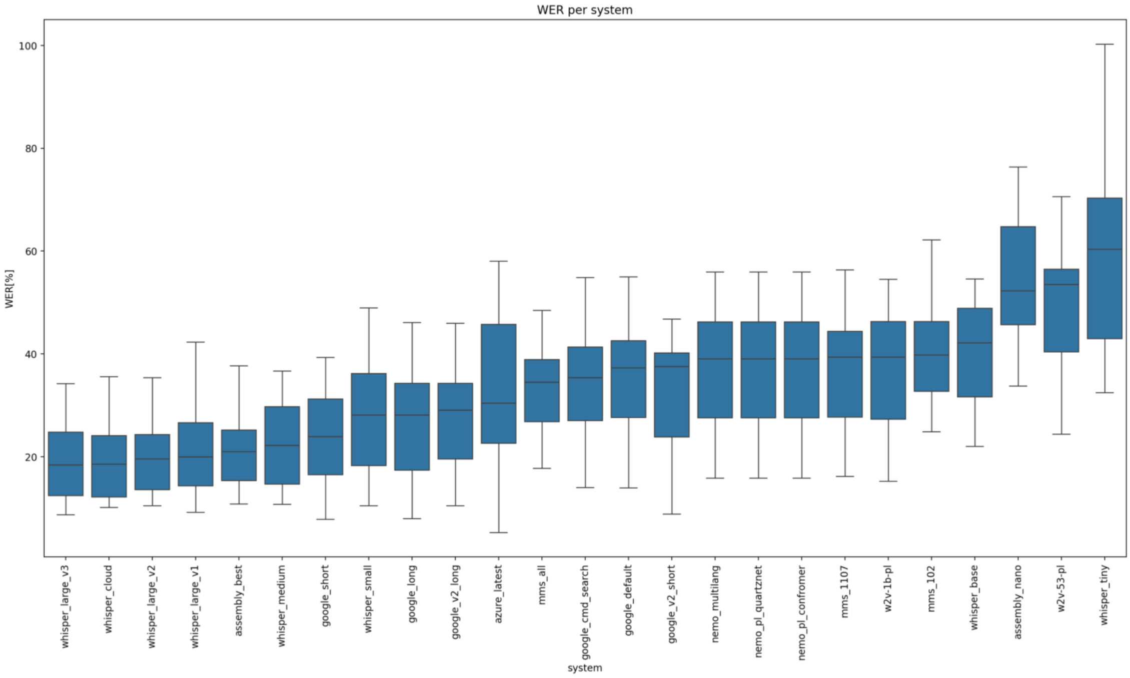


Figure 4.15: Mean WER for all PELCRA dataset subsets.

	Word Error Rate (WER) [%]				
Model	**Median**	**Mean**	**Std.**	**Min.**	**Max.**
whisper_large_v3	18.45	19.45	8.02	8.74	34.25
whisper_cloud	18.54	19.64	8.31	10.12	35.6
whisper_large_v2	19.55	20.57	8.36	10.49	35.4
whisper_large_v1	20.02	21.94	10.1	9.21	42.29
assembly_best	21.01	21.53	8.02	10.8	37.72
whisper_medium	22.2	22.66	9.09	10.76	36.65
google_short	23.91	24.58	9.95	7.82	39.3
whisper_small	28.12	28.97	12.42	10.49	48.97
google_long	28.16	26.86	10.91	7.95	46.1
google_v2_long	29.06	27.84	10.5	10.49	45.99
azure_latest	30.42	33.51	16.77	5.27	58.08
mms_all	34.48	34.03	9.28	17.74	48.51
google_cmd_search	35.42	35.21	11.81	13.98	54.89
google_default	37.26	36.08	11.76	13.92	55.01
google_v2_short	37.58	32.68	11.55	8.86	46.77
nemo_multilang	39.02	38.02	13.16	15.81	55.93
nemo_pl_quartznet	39.02	38.02	13.16	15.81	55.93
nemo_pl_confromer	39.02	38.02	13.16	15.81	55.93
mms_1107	39.36	37.06	12.75	16.17	56.37
w2v-1b-pl	39.39	37.49	12.79	15.23	54.5
mms_102	39.78	41.04	12.11	24.84	62.16
whisper_base	42.17	45.65	20.28	21.99	89.14
assembly_nano	52.3	54.93	13.63	33.74	76.4
w2v-53-pl	53.52	50.28	13.62	24.42	70.58
whisper_tiny	60.38	61.86	25.12	32.46	114.1

Table 4.47: WER statistics – PELCRA dataset

ES2 — Accuracy per subset

BIGOS dataset Table 4.48 presents the statistics of the WER scores for each subset of the BIGOS dataset in different combinations of systems-models. Figure 4.16 presents the WER scores box plots for individual subsets from BIGOS dataset derived from evaluations across all system-model combinations.

	Word Error Rate (WER) [%]				
Subset	Median	Mean	Std.	Min.	Max.
pwr-shortwords-unk	7.77	11.05	10.11	4	46.06
pwr-maleset-unk	9	12.07	10.14	3.79	43.79
mozilla-common_voice_15-23	10.09	14.89	12.81	2.1	55.77
mailabs-corpus_librivox-19	10.28	13.77	10.05	4.9	44.06
pwr-viu-unk	10.54	25.96	52.07	0	260.86
pjatk-clarin_studio-15	10.93	12.79	6.42	5.86	34.96
fair-mls-20	11.29	14.76	12.02	4.42	52.73
google-fleurs-22	14.52	17.73	12.11	5.33	56.06
pjatk-clarin_mobile-15	14.57	16.93	10.53	5.57	52.96
pwr-azon_read-20	16.98	17.4	9.44	5.31	38.47
pwr-azon_spont-20	24.03	25.32	7.3	16.4	44.74
polyai-minds14-21	36.61	42.43	26.47	13.59	139.29

Table 4.48: WER statistics for all ASR systems and specific subsets of BIGOS dataset.

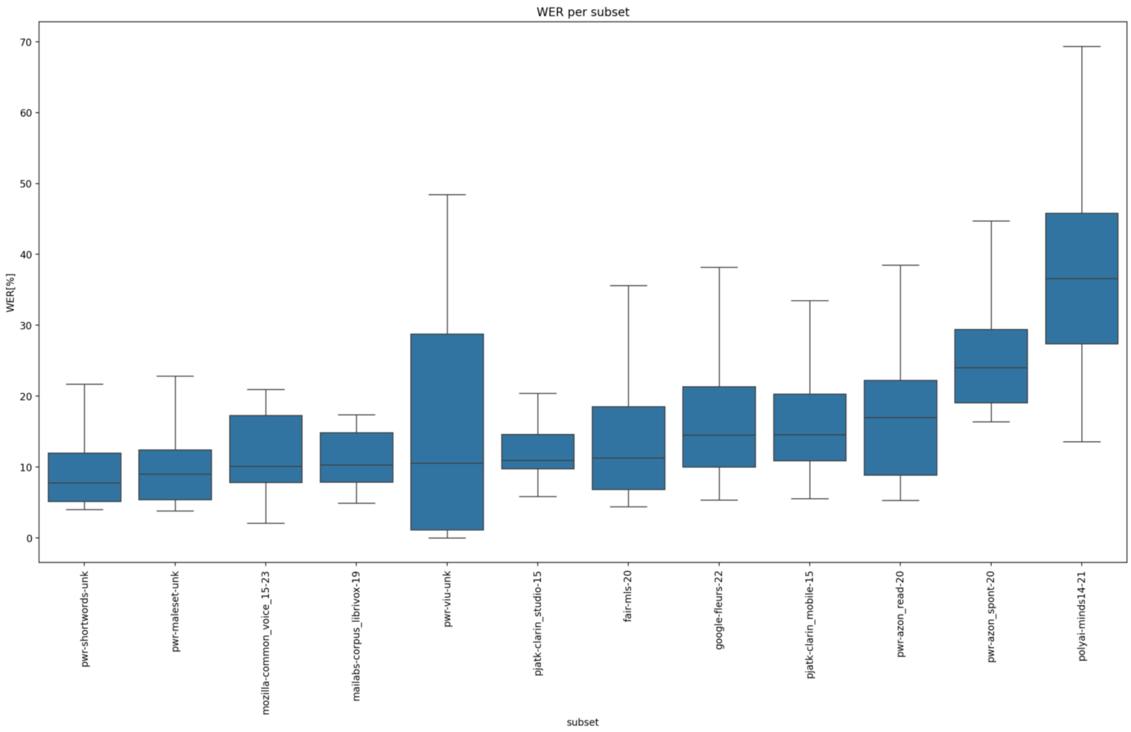


Figure 4.16: Box plot of WER for all systems per specific subset of BIGOS dataset.

PELCRA dataset Table 4.49 shows the statistics of WER scores for various subsets of the PELCRA dataset, evaluated across different system-model combinations. Figure 4.17 displays the box plots of WER scores for each subset of the PELCRA dataset, based on evaluations from all system-model combinations.

	Word Error Rate (WER) [%]				
Subset	**Median**	**Mean**	**Std.**	**Min.**	**Max.**
ul-spokes_mix_emo-18	13.92	16.07	11.55	5.27	62.46
ul-spokes_mix_parl-18	23.39	25.1	12.76	13.33	72.62
ul-spokes_biz_int-23	25.24	24.26	9.48	12.1	49.4
ul-spokes_biz_vc-23	27.39	24.99	10.6	11.75	58.35
ul-spokes_biz_pres-23	27.7	27.27	12.06	10.54	53.22
ul-spokes_biz_vc2-23	36.5	35.63	9.81	22.46	62.59
ul-spokes_biz_pod-23	37.22	36.09	11.15	20.25	59.47
ul-spokes_biz_luz-23	38.15	36.59	11.47	19.67	61.29
ul-spokes_biz_wyw-23	39.78	36.9	13.57	17.14	68.61
ul-spokes_mix_luz-18	43.33	44.24	14.77	24.15	76.4
ul-diabiz_poleval-22	48.97	50.58	17.43	30.04	100.27
ul-spokes_biz_bio-23	51.52	49.3	19.02	27.89	114.1

Table 4.49: WER statistics for all ASR systems and specific subsets of PELCRA set.

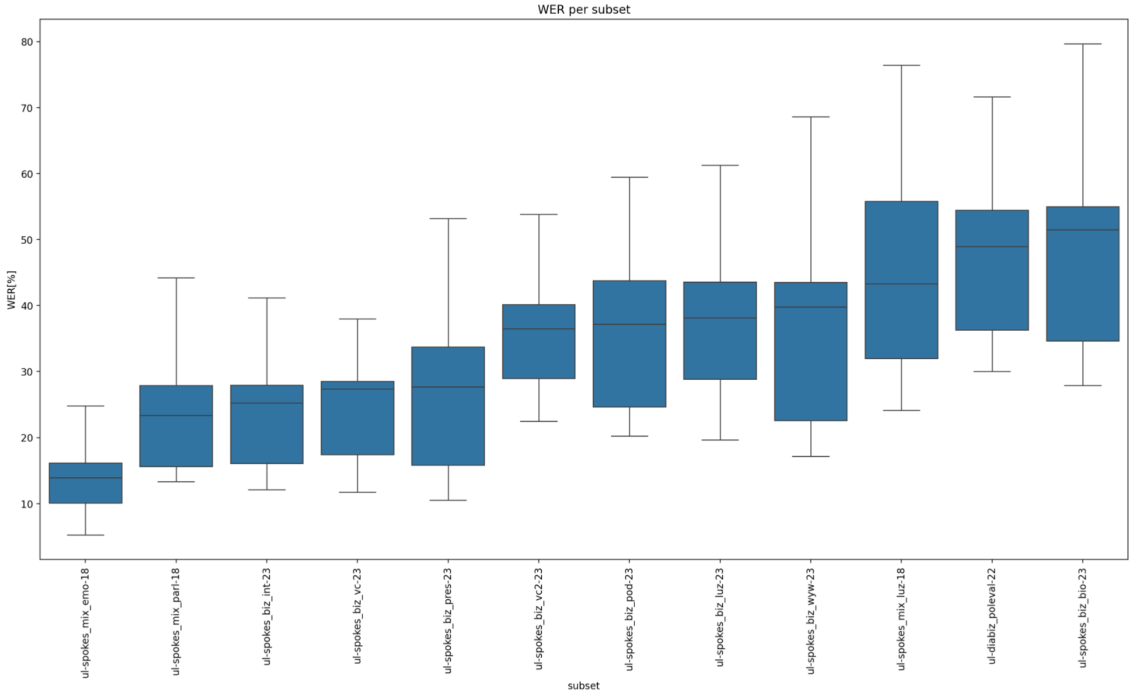


Figure 4.17: Box plot of WER for all systems per specific subset of PELCRA dataset.

ES3 — Accuracy per system type

This section shows the results of comparing the accuracy of ASR systems based on whether they are available freely or commercially.

BIGOS dataset Tables 4.50 shows the WER statistics calculated for all subsets and systems of type *free* and *commercial* for BIGOS.

Table 4.51 presents the systems recognized for having the highest and lowest average

	Word Error Rate (WER) [%]				
Type	Median	Mean	Std.	Min.	Max.
commercial	12.96	17.26	24.98	0	260.86
free	14.57	19.76	17.36	2.1	139.29

Table 4.50: WER statistics for free and paid ASR systems on BIGOS dataset.

WER across all subsets of the BIGOS dataset.

Type	Best system	Worse system
Free	whisper_large_v3	whisper_tiny
Commercial	whisper_cloud	assembly_nano

Table 4.51: Best and worse systems for BIGOS dataset.

Table 4.52 presents the WER statistics for the most accurate systems of each type obtained for the BIGOS dataset. WER statistics for the least accurate commercial and

	Word Error Rate (WER) [%]				
System	Median	Mean	Std.	Min.	Max.
Best free	5.53	8.38	7.5	3.79	29.17
Best commercial	6.25	8.47	6.15	4.46	25.2

Table 4.52: WER statistics for the most accurate free and commercial systems. BIGOS dataset.

free systems for the BIGOS dataset are presented in table 4.52.

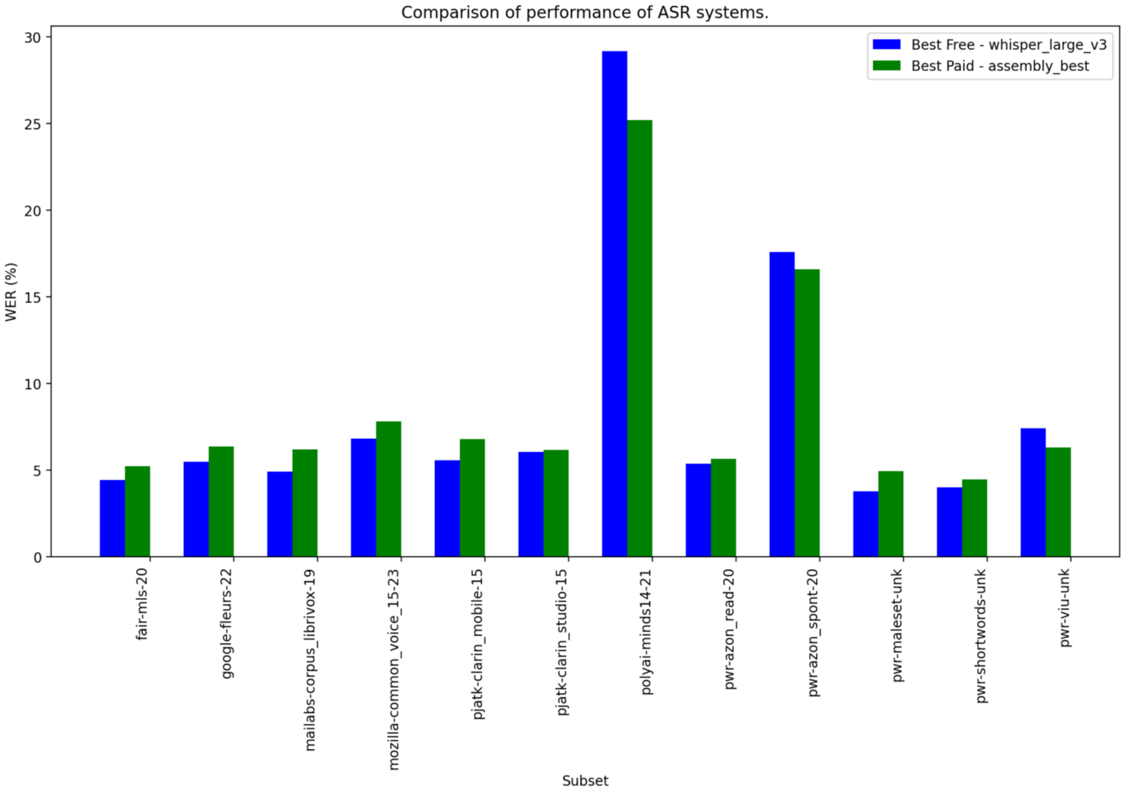


Figure 4.18: Comparison of WER for the most accurate free and paid ASR systems. BIGOS dataset.

Figure 4.18 illustrate a comparison of accuracy between the top-performing systems across various subsets of the BIGOS dataset.

Table 4.53 illustrates the WER statistics for the lowest performing systems in the BIGOS dataset. A visual comparison can be found in Figure 4.19.

	Word Error Rate (WER) [%]				
System	**Median**	**Mean**	**Std.**	**Min.**	**Max.**
Worst commercial	32.83	52.53	66.27	14.54	260.86
Worst free	44.4	54.28	28.92	33.71	139.29

Table 4.53: WER statistics for the most accurate free and commercial systems. BIGOS dataset.

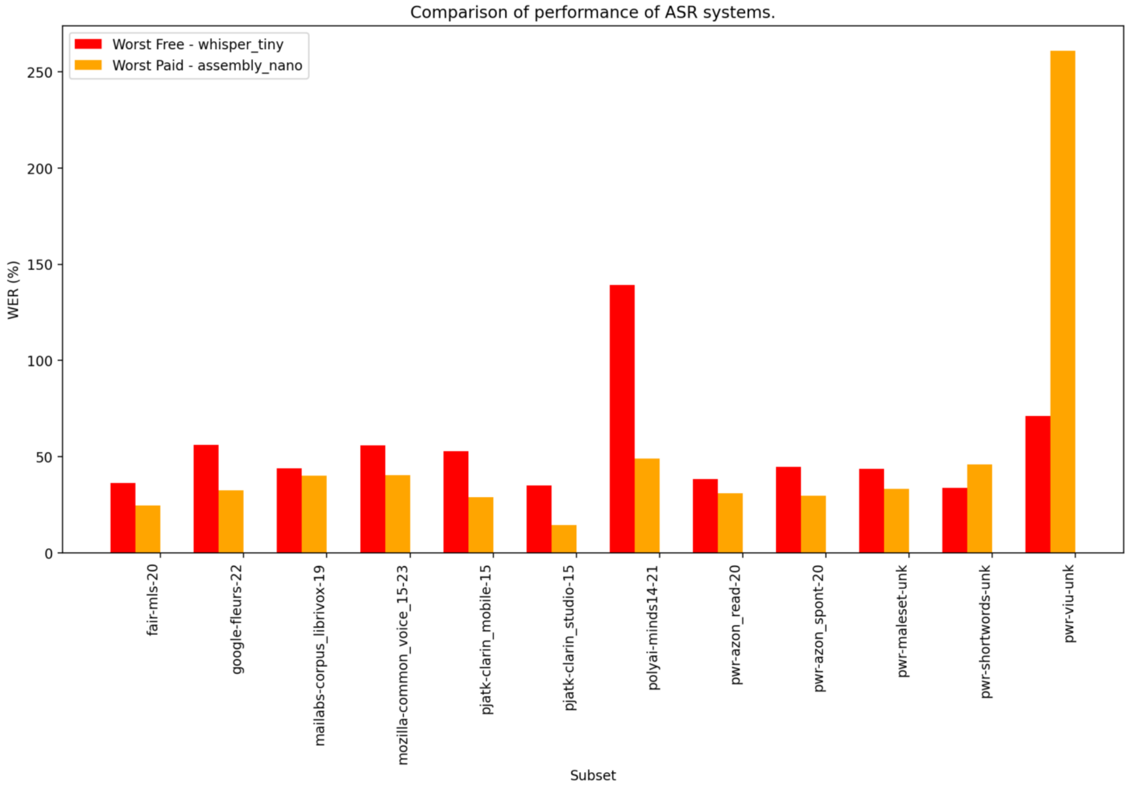


Figure 4.19: Comparison of WER for the least accurate free and paid ASR systems. BIGOS dataset.

PELCRA dataset Table 4.54 presents the WER statistics for the PELCRA dataset, comparing average, median, minimum and maximum WER for free and commercial ASR systems.

	Word Error Rate (WER) [%]				
System	**Median**	**Mean**	**Std.**	**Min.**	**Max.**
Commercial	29.88	31.29	14.69	5.27	76.4
Free	33.58	35.67	17.36	8.74	114.1

Table 4.54: WER statistics for free and paid ASR systems evaluated on PELCRA dataset.

Table 4.55 shows the names of the systems identified as having the best and worst overall accuracy.

Type	**Best accuracy**	**Worst accuracy**
Free	whisper_large_v3	whisper_tiny
Commercial	whisper_cloud	assembly_nano

Table 4.55: Best and worst accurate systems for PELCRA dataset.

Table 4.56 shows the WER statistics for the top performing systems assessed using the

PELCRA dataset.

	Word Error Rate (WER) [%]				
System	**Median**	**Mean**	**Std.**	**Min.**	**Max.**
Best free	18.45	19.45	8.02	8.74	34.25
Best commercial	18.54	19.64	8.31	10.12	35.6

Table 4.56: The most accurate systems WER statistics. PELCRA dataset.

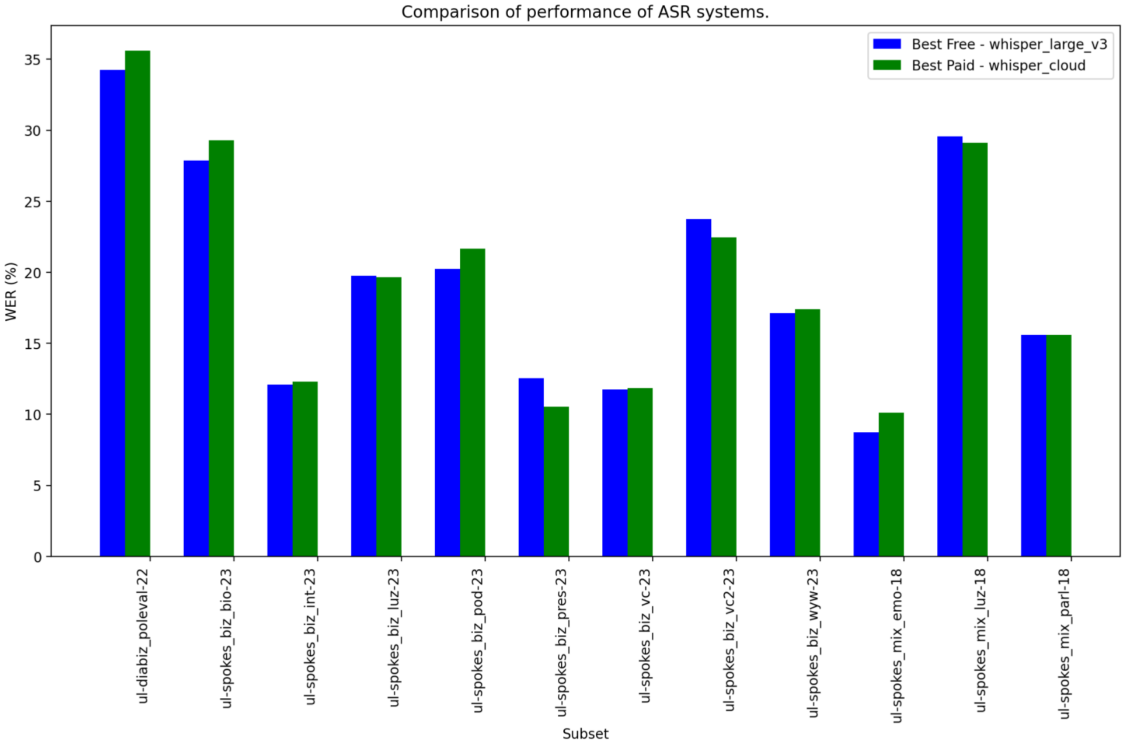


Figure 4.20: Comparison of WER for the most accurate free and paid ASR systems. PELCRA dataset.

The performance analysis of the least efficient systems is shown in Table 4.57. A graphical comparison across different subsets of the PELCRA dataset is illustrated in Figure 4.21.

	Word Error Rate (WER) [%]				
System	**Median**	**Mean**	**Std.**	**Min.**	**Max.**
Worst commercial	52.3	54.93	13.63	33.74	76.4
Worst free	60.38	61.86	25.12	32.46	114.1

Table 4.57: The least accurate systems WER statistics. PELCRA dataset.

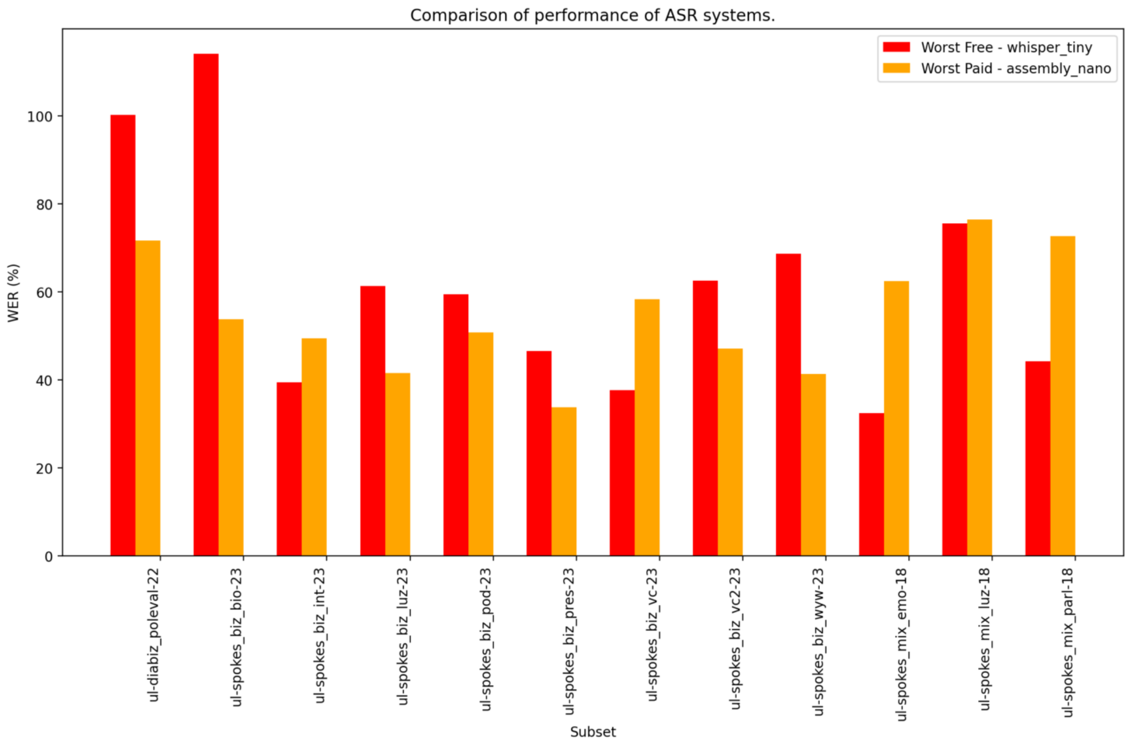


Figure 4.21: Comparison of WER for the least accurate free and paid ASR systems. PELCRA dataset.

ES4 — Accuracy per model size

BIGOS dataset Table 4.58 contains information on the accuracy of model size recognition, represented as average WER. Figure 4.22 illustrates the relation between model

System	Parameters [M]	Avg. WER [%]
whisper_large_v3	1550	8.38
whisper_large_v2	1550	9.58
whisper_large_v1	1550	10.5
whisper_medium	769	11.52
w2v-1b-pl	1000	13.74
nemo_pl_quartznet	19	15.03
nemo_multilang	114	15.03
nemo_pl_confromer	118	15.03
mms_all	1000	16.9
whisper_small	244	20.52
mms_1107	1000	22.25
mms_102	1000	22.56
w2v-53-pl	300	25.54
whisper_base	74	35.6
whisper_tiny	39	54.28

Table 4.58: Average WER for free systems with information about model size.

size and its performance on the BIGOS dataset. The X-axis denotes the model size in millions of parameters, while the Y-axis represents the average WER across all subsets of the BIGOS dataset.

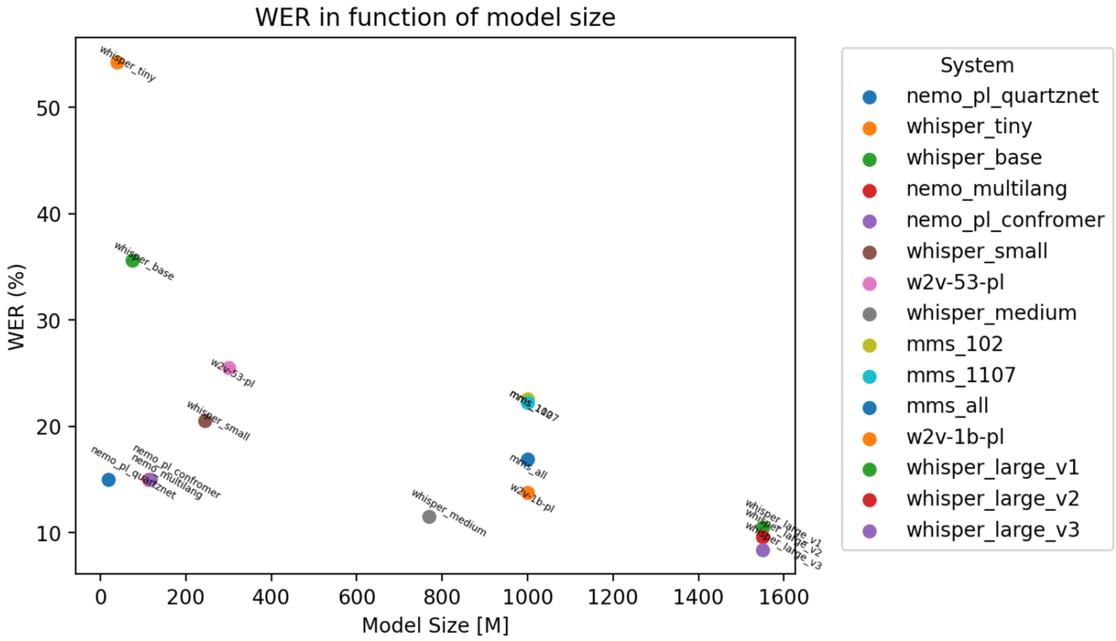


Figure 4.22: WER for freely available systems for various model sizes. BIGOS dataset.

PELCRA dataset Table 4.59 shows the average WER for the BIGOS dataset for freely available systems, with information about the size of the model. Figure 4.23 shows the relationship between the size of the model and its performance on the PELCRA dataset.

System	Parameters [M]	Avg. WER [%]
whisper_large_v3	1550	19.45
whisper_large_v2	1550	20.57
whisper_large_v1	1550	21.94
whisper_medium	769	22.66
whisper_small	244	28.97
mms_all	1000	34.03
mms_1107	1000	37.06
w2v-1b-pl	1000	37.49
nemo_pl_quartznet	19	38.02
nemo_multilang	114	38.02
nemo_pl_confromer	118	38.02
mms_102	1000	41.04
whisper_base	74	45.65
w2v-53-pl	300	50.28
whisper_tiny	39	61.86

Table 4.59: Average WER for free systems with information about model size.

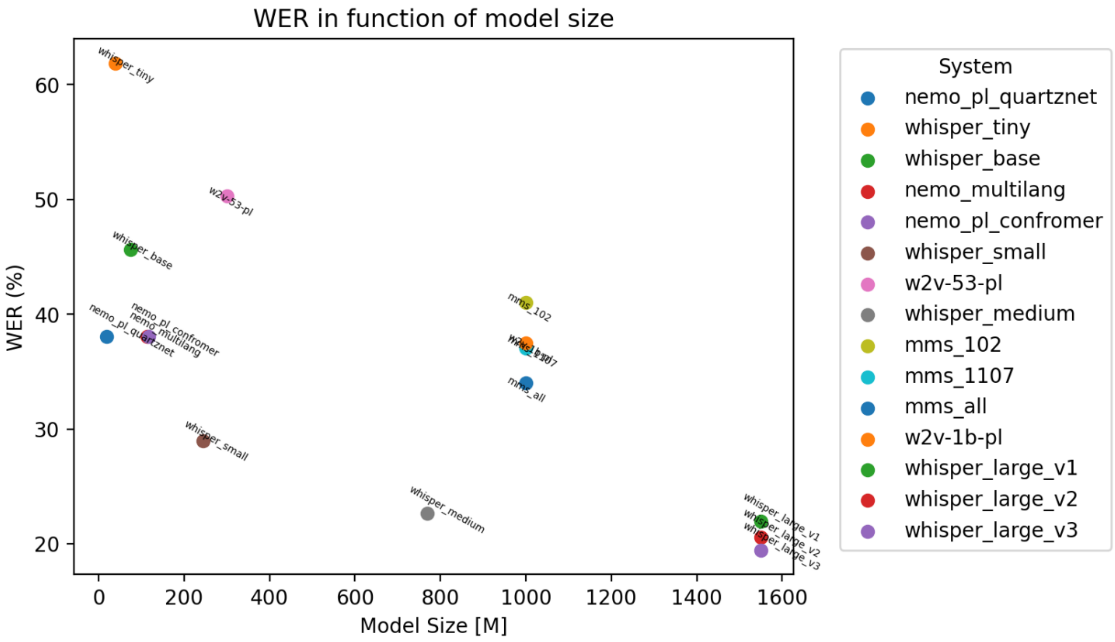


Figure 4.23: WER for freely available systems for various model sizes. PELCRA dataset.

ES5 — Accuracy per audio duration

This subsection presents experimental results on how speech recording duration affects speech recognition precision.

BIGOS dataset Table 4.60 shows the average WER for various audio duration ranges for the most accurate free and commercial systems. For each audio duration range, the number of samples across all BIGOS dataset subsets is provided. The relationship between WER and audio duration is also depicted in figure 4.24. The x-axis represents the duration of the audio samples in seconds. The y-axis represents the mean WER across all samples falling into specific audio duration range. The point size indicates the quantity of samples for each specific audio duration.

		Word Error Rate (WER) [%]	
Min. duration [s]	No of samples	Assembly Best	Whisper Large V3
1	284	8.22	8.92
2	155	6.54	5.6
3	338	6.1	3.86
4	402	5.09	5.28
5	748	7.32	6.51
10	1,148	6.36	5.43
15	584	6	5.33
20	266	9.01	9.43
30	38	21.08	29.43
40	8	184.78	202.16
50	12	93.11	91.23
60	2	57.14	100

Table 4.60: Mean WER for specific audio duration ranges. BIGOS dataset. Best paid and free systems.

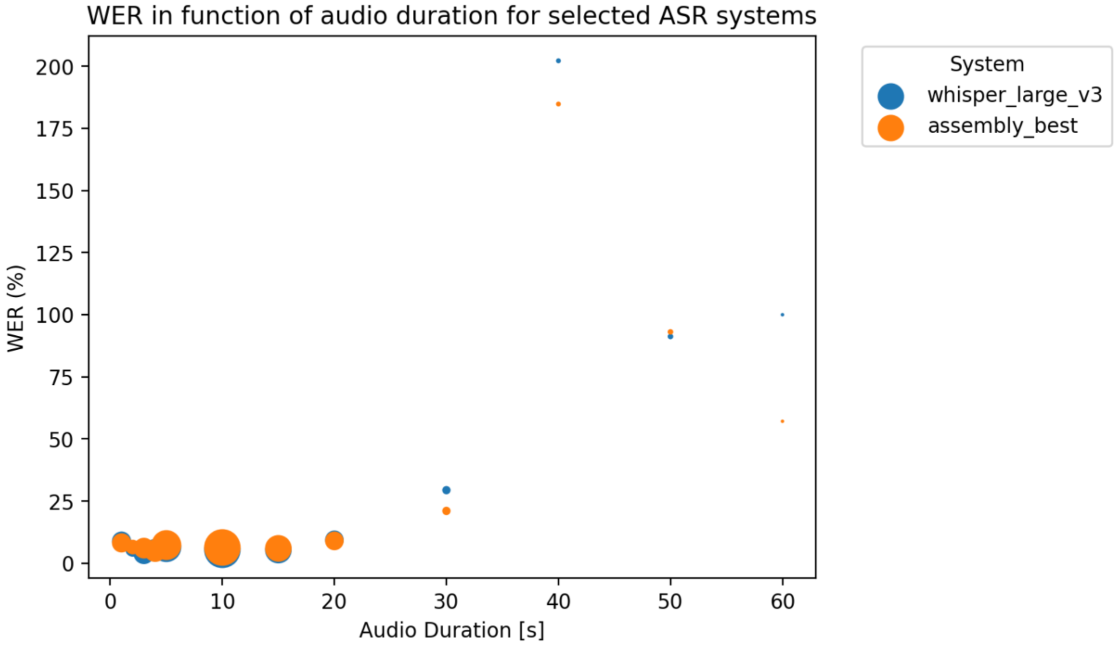


Figure 4.24: Average WER in function of audio duration. BIGOS dataset.

PELCRA dataset Table 4.61 presents the mean WER for different audio duration intervals for the most performing free and commercial systems. The count of samples used to compute the mean WER for a specific duration range is provided in the second column. The relationship between mean WER and audio duration is also illustrated in Figure 4.25.

		Word Error Rate (WER) [%]	
Min. duration [s]	**No of samples**	**Whisper Cloud**	**Whisper Large V3**
1	498	60.74	51.14
2	565	37.48	35.31
3	551	27.35	26.15
4	570	31.16	25.33
5	1086	24.38	23.85
10	712	20.98	20.58
15	320	20.29	21.15
20	206	19.93	21.13
30	106	20.08	20.3
40	22	38.33	37.09
50	16	16.67	16.28
60	20	25.81	28.07

Table 4.61: Mean WER for specific audio duration ranges for top paid and free systems. PELCRA dataset.

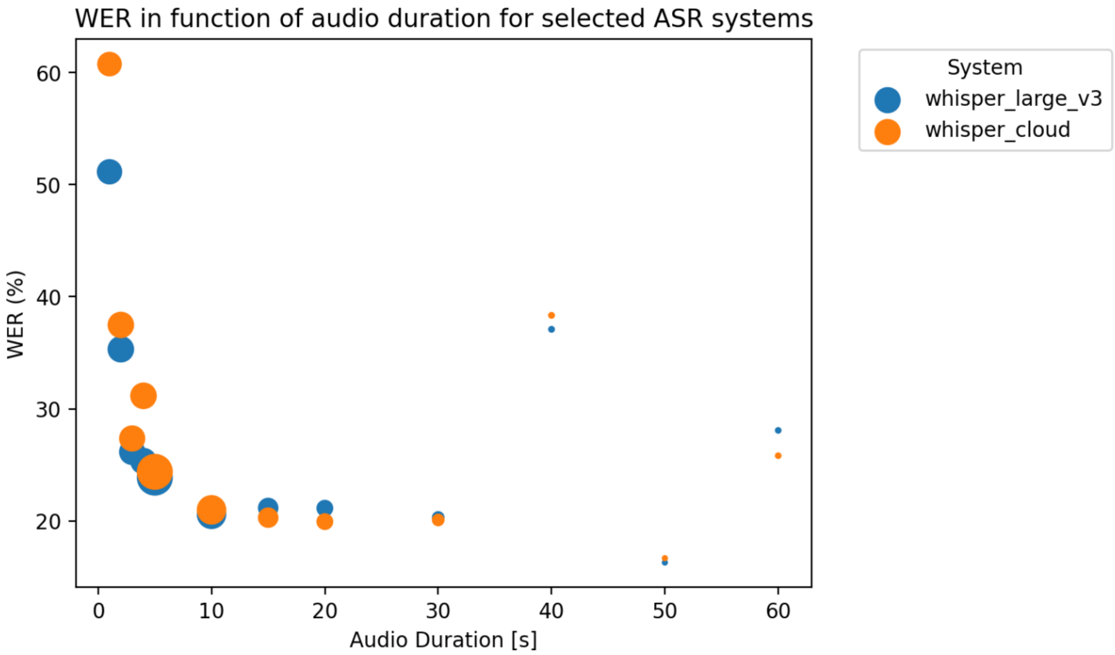


Figure 4.25: Mean WER in function of audio duration. PELCRA dataset. Best paid and free systems. The size of the point corresponds to the number of samples.

ES6 — Accuracy per speaking rate

BIGOS dataset

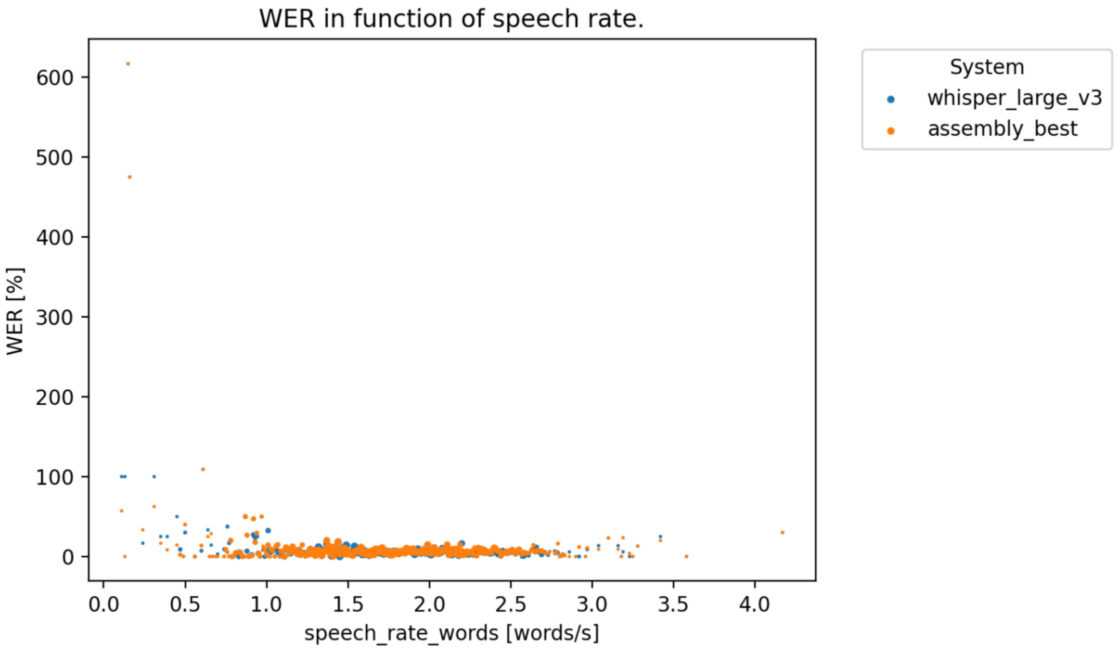


Figure 4.26: Mean WER in function of speech rate for top systems. BIGOS dataset.

PELCRA dataset

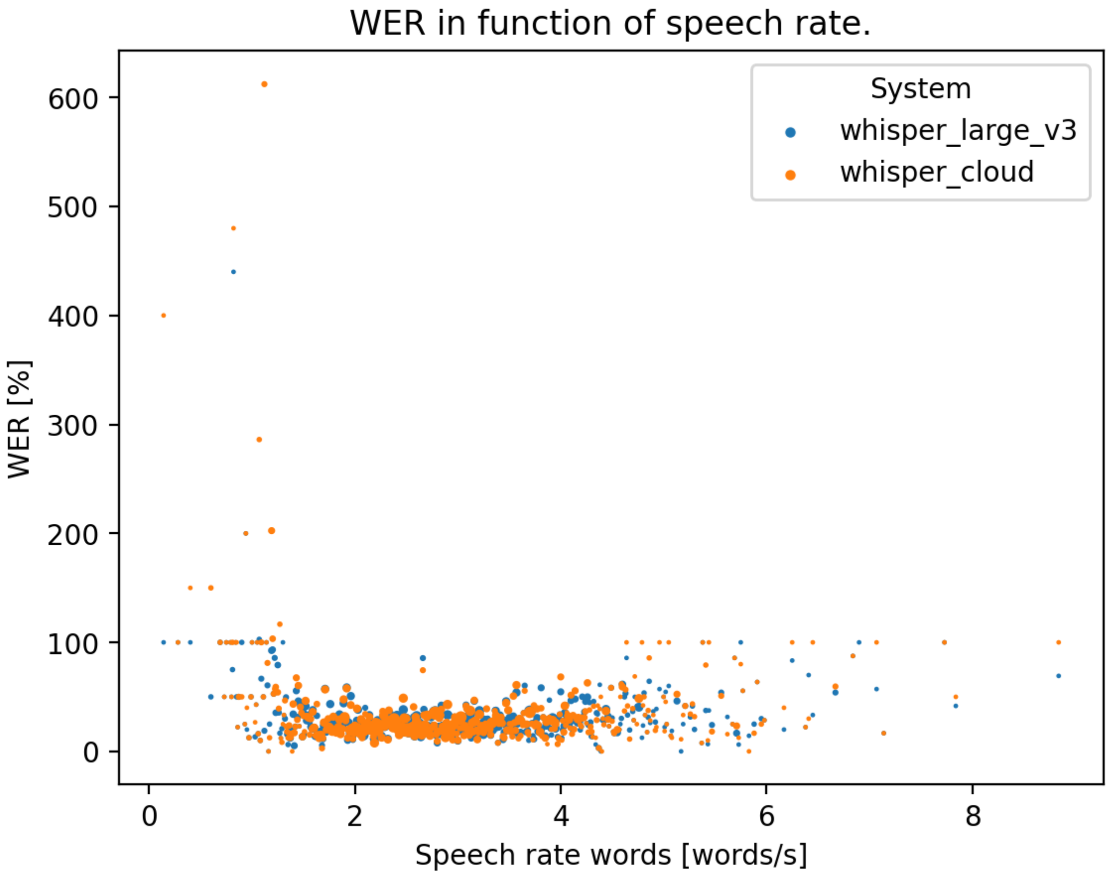


Figure 4.27: Mean WER in function of speech rate for top systems. PELCRA dataset.

ES7 — Accuracy per speaker gender

BIGOS dataset Table 4.63 shows the average WER for female and male speakers and the difference between them. A negative difference means male speakers have lower WER, indicating a bias toward males. A positive difference means female speakers have lower WER. Values close to zero indicate no gender bias. The BIGOS subset used for evaluation in this study contained 442 samples from female speakers and 758 from male speakers (Table 4.62). Equal number of samples from both groups were used to calculate mean WER scores.

Gender	**No of samples**
female	442
male	758
total	1200

Table 4.62: Number of samples with speaker gender information.

PELCRA dataset Table 4.65 presents the average WER for both female and male speakers, along with the difference between them. A negative difference indicates that male speakers have a lower WER, suggesting a bias towards males. Conversely, a positive difference indicates that female speakers have a lower WER. Values near zero suggest no gender bias. The BIGOS subset used for evaluation in this study included 908 samples from female speakers and 689 from male speakers (Table 4.64). An equal number of samples from both groups were used to compute the mean WER scores.

ES8 — Accuracy per speaker age group

Table 4.66 shows the average WER for different age categories. The variation in precision among specific groups is evaluated using the standard deviation and the range between the lowest and highest WER for all groups (table 4.67)

PELCRA dataset

Figure 4.28 shows the standard deviation in mean WER between age groups for PELCRA dataset.

	Word Error Rate (WER) [%]		
System	Females	Males	Diff. [p.p.]
azure_latest	23.06	7.87	-15.19
google_short	20.68	7.11	-13.57
google_v2_short	20.27	7.91	-12.36
google_default	20.25	8.99	-11.26
google_cmd_search	19.9	8.69	-11.21
w2v-53-pl	25.68	18.93	-6.75
google_v2_long	11.18	5.26	-5.92
google_long	11.1	5.31	-5.79
nemo_pl_confromer	15.89	13.56	-2.33
whisper_large_v2	7.27	6.17	-1.1
whisper_medium	9.72	8.64	-1.08
whisper_large_v1	8.67	7.6	-1.07
nemo_pl_quartznet	15.89	14.97	-0.92
whisper_large_v3	6.32	5.5	-0.82
assembly_best	7.07	6.27	-0.8
nemo_multilang	15.89	15.36	-0.53
w2v-1b-pl	11.4	13.22	1.82
mms_1107	20.14	23	2.86
whisper_cloud	6.74	10.3	3.56
mms_all	13.4	18.91	5.51
whisper_small	15.96	22.85	6.89
whisper_base	29.23	37.89	8.66
mms_102	19.49	29.44	9.95
whisper_tiny	44.59	59.53	14.94
assembly_nano	36	92.93	56.93
median	15.89	10.3	-0.92
average	17.43	18.25	0.82
std	9.33	19.87	13.95

Table 4.63: Values and differences in mean WER scores per speaker gender.

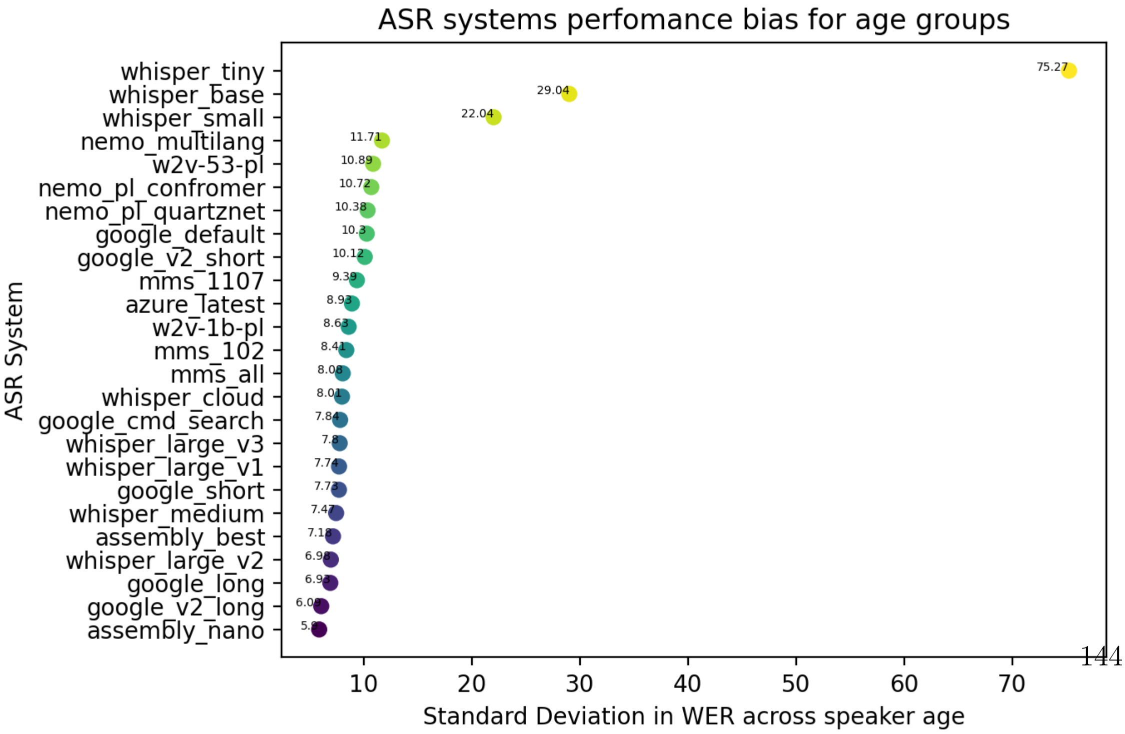


Figure 4.28: Standard deviation in WER across speaker age groups. PELCRA dataset.

Gender	No of samples
female	908
male	689
total	1597

Table 4.64: Number of samples with speaker gender information.

4.4.4 Reference and ASR Transcripts Normalization

Table 4.68 shows the reduction of various errors resulting from the application of specific normalization procedures as well as all the methods together.

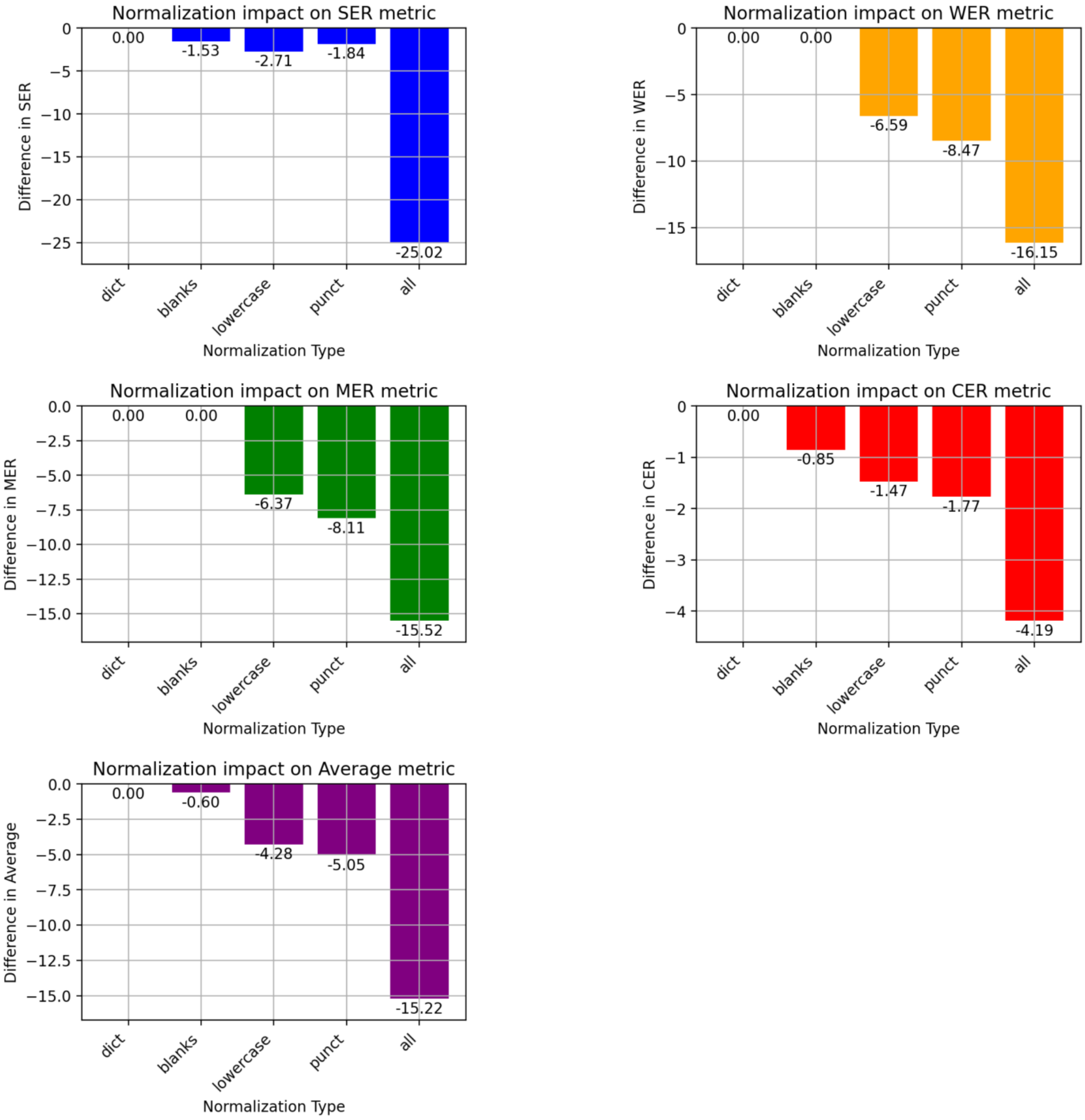
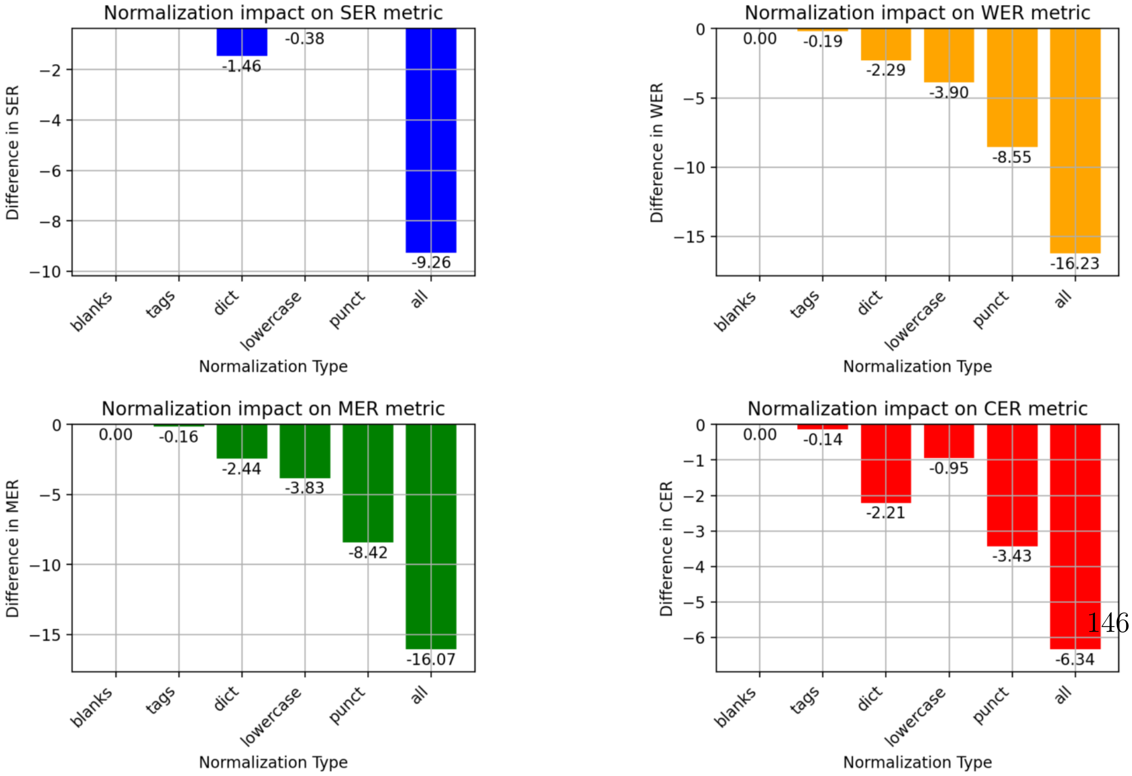


Figure 4.29: Impact of normalization on error rates on BIGOS dataset.

	Word Error Rate (WER) [%]		
System	**Females**	**Males**	**Diff. [p.p.]**
assembly_nano	79.59	53.36	-26.23
whisper_large_v2	40.25	22.9	-17.35
whisper_base	88.66	74.59	-14.07
mms_102	55.89	46.37	-9.52
nemo_multilang	53.48	44.81	-8.67
whisper_large_v1	35.16	26.78	-8.38
nemo_pl_quartznet	51.76	44.03	-7.73
nemo_pl_confromer	52.2	44.49	-7.71
whisper_cloud	30.6	23.6	-7
w2v-1b-pl	48.97	42.81	-6.16
whisper_medium	34.86	28.72	-6.14
whisper_small	43.18	37.53	-5.65
mms_all	45.57	40.87	-4.7
assembly_best	31.59	26.99	-4.6
whisper_large_v3	29.47	25.02	-4.45
mms_1107	48.82	44.46	-4.36
w2v-53-pl	59.98	56.12	-3.86
google_short	34.1	30.68	-3.42
google_cmd_search	44.44	41.7	-2.74
google_default	43.77	42.72	-1.05
google_v2_short	35.64	39.62	3.98
google_v2_long	31.42	36.62	5.2
google_long	30.62	36.6	5.98
azure_latest	33.95	44.5	10.55
whisper_tiny	91.17	104.56	13.39
median	43.77	41.7	-4.7
average	47.01	42.42	-4.59
std	17.38	17.31	8.31

Table 4.65: Values and differences in average WER scores per speaker gender for 689 samples from PELCRA dataset per gender.



	Word Error Rate (WER) [%]				
System	**20s**	**30s**	**40s**	**60s**	**70s**
assembly_nano	62.88	66.14	78.59	70.49	70.46
google_v2_long	33.84	36.31	39.62	23.64	30.68
google_long	38.91	37.93	39.33	24.23	28.64
whisper_large_v2	30.01	27.73	31.39	15.49	32.88
assembly_best	28.37	30.24	38.3	20.94	37.66
whisper_medium	28.73	37.38	36.84	19.16	32.74
google_short	31.8	32.13	46.57	25.47	33.74
whisper_large_v1	32.65	26.38	38.59	18.24	32.77
whisper_large_v3	29.13	26.69	38.05	16.87	32.13
google_cmd_search	39.71	46.87	53.46	32.68	40.94
whisper_cloud	24.32	26.5	32.41	14.13	34.5
mms_all	39.45	48.35	55.13	34.22	42.68
mms_102	50.05	53.28	62.99	43.78	63.16
w2v-1b-pl	49.6	50.59	60.09	36.41	45.07
azure_latest	44.67	44.92	43.38	29.78	26.6
mms_1107	52.41	50.58	61.26	35.3	47.75
google_v2_short	36.64	42.1	53.69	28.22	30.95
google_default	47.32	51.8	58.04	31.94	39.17
nemo_pl_quartznet	44.83	50.18	62.07	33.55	51.13
nemo_pl_confromer	44.68	52.52	62.65	34.39	54.62
w2v-53-pl	57.59	62.35	69.96	41.41	50.99
nemo_multilang	53.25	60.79	64.07	33.88	52.67
whisper_small	36.7	47.93	82.85	24.1	41.72
whisper_base	55.58	53.59	68.69	120.52	51.49
whisper_tiny	75.62	228.49	94.38	41.07	165.15
median	39.71	47.93	55.13	31.94	40.94
average	42.75	51.67	54.9	34	46.81
std	12.43	38.63	16.52	21.53	27.07

Table 4.66: Mean WER across systems and age ranges. PELCRA dataset.

4.4.5 Evaluation results sharing

The scripts utilized to produce ASR hypotheses and evaluation outcomes have been shared via the GitHub repository[17]. The resulting data artifacts, including evaluation results, data analysis scripts, and the generated diagrams and tables, have been made accessible on the Hugging Face platform. [18]

[17] BIGOS tools

[18] PL ASR Leaderboard

	Word Error Rate (WER)	
System	**Std. [p.p.]**	**Range [p.p.]**
assembly_nano	5.9	15.71
google_v2_long	6.09	15.98
google_long	6.93	15.1
whisper_large_v2	6.98	17.39
assembly_best	7.18	17.36
whisper_medium	7.47	18.22
google_short	7.73	21.1
whisper_large_v1	7.74	20.35
whisper_large_v3	7.8	21.18
google_cmd_search	7.84	20.78
whisper_cloud	8.01	20.37
mms_all	8.08	20.91
mms_102	8.41	19.38
w2v-1b-pl	8.63	23.68
azure_latest	8.93	18.32
mms_1107	9.39	25.96
google_v2_short	10.12	25.47
google_default	10.3	26.1
nemo_pl_quartznet	10.38	28.52
nemo_pl_confromer	10.72	28.26
w2v-53-pl	10.89	28.55
nemo_multilang	11.71	30.19
whisper_small	22.04	58.75
whisper_base	29.04	69.03
whisper_tiny	75.27	187.42
median	8.41	21.1
average	12.54	31.76
std	14.01	34.76

Table 4.67: Standard Dev. and maximum difference in WER across age groups. PELCRA dataset.

4.5 RO6: Organization of open competition for the ASR community

4.5.1 Introduction

RQ 22: What programs can organize the Polish ASR community challenge? **RQ 23:** How to compare community solutions with state-of-the-art ASR systems?

Method	SER	WER	MER	CER	Average
blanks	-1.53	0	0	-0.85	-0.6
lowercase	-2.71	-6.37	-6.59	-1.47	-4.28
punct	-1.84	-8.11	-8.47	-1.77	-5.05
all	-25.02	-15.52	-16.15	-4.19	-15.22

Table 4.68: Reduction of error rates caused by normalization of references and hypothesis for BIGOS dataset.

Method	SER	WER	MER	CER	Average
blanks	-0.34	0	0	0	-0.08
tags	-0.35	-0.16	-0.19	-0.14	-0.21
dict	-1.46	-2.44	-2.29	-2.21	-2.1
lowercase	-0.38	-3.83	-3.9	-0.95	-2.26
punct	-0.36	-8.42	-8.55	-3.43	-5.19
all	-9.26	-16.07	-16.23	-6.34	-11.98

Table 4.69: Reduction of error rates caused by normalization of references and hypothesis for PELCRA dataset.

4.5.2 Program selection and task creation

PolEval was chosen to organize open challenge for the Polish ASR community. An online call for participation was issued on June 3rd 2024. [19] The details of the participation call can be found in Appendix 7.1.8. The results of the competition were not available at the time of writing this thesis.

4.5.3 Comparison of community ASR solutions with other systems for Polish

The evaluation of publicly accessible ASR systems against community-developed solutions is facilitated by the AMU ASR Leaderboard. After the conclusion of the open PolEval challenge, the datasets containing the participants' submitted ASR hypotheses will be formatted and integrated into the ASR leaderboard as illustrated in Figure 4.31.

[19] PolEval 2024 – ASR challenge beta

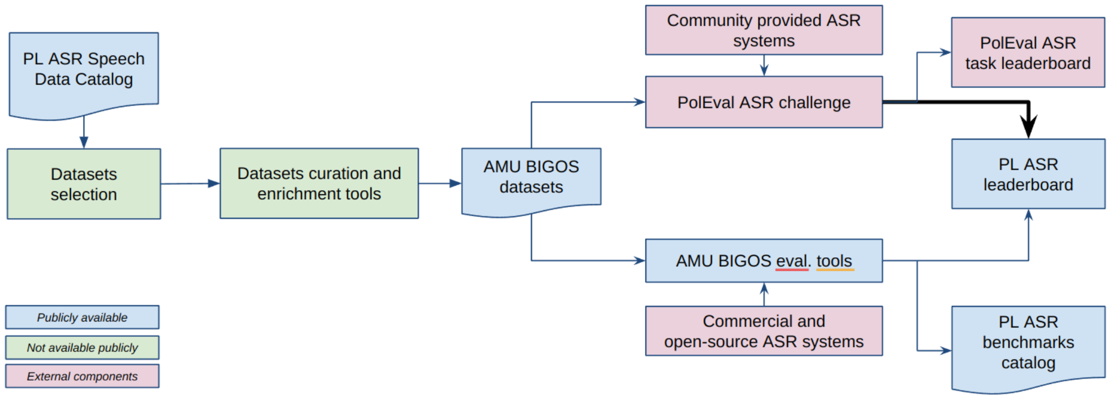

PL ASR Speech Data Catalog
Datasets selection
Datasets curation and enrichment tools
AMU BIGOS datasets
Community provided ASR systems
PolEval ASR challenge
PolEval ASR task leaderboard
PL ASR leaderboard
AMU BIGOS eval. tools
Commercial and open-source ASR systems
PL ASR benchmarks catalog
Publicly available
Not available publicly
External components

Figure 4.31: Management framework extension to incorporate results from PolEval open challenge.

Chapter 5

Discussion

5.1 Overview

The dissertation addresas lack of centralized information and benchmarking tools for Polish ASR systems. By surveying speech datasets, curating a benchmark dataset, and developing an evaluation system, this research aimed to improve accessibility, usability, and standardization of Polish ASR resources. The major contributions are:

1. **Catalog of Polish ASR Datasets**: Created a metadatsociodemographictent ASR dataset descriptions. Comsplitting the dataset,R datasets into a centralized repository, aiding researchers and industry professionals. The standardized framework reduces time and effort to locate datasets, accelerating ASR development.

2. **Curated Benchmark Dataset**: Compiled a diverse benchmark dataset for various use cases and demographics, released on a popular ML dataset platform. This resource sets a new standard for Polish ASR benchmarking, enabling result comparisons across studies.

3. **Evaluation System Development**: Developed an evaluation framework for Polish ASR systems, enabling outcome replication. The extendable design supports new systems, metrics, datasets, and languages, aiding ASR development.

4. **Identification of Research Gaps**: Highlighted gaps in existing datasets and benchmarks, guiding future research and development in the Polish ASR community.

5.2 RO1: Survey of ASR speech datasets for Polish

5.2.1 Results overview

RQ1: How to identify and systematically categorize Polish ASR speech datasets using publicly available information?

To systematically identify and categorize Polish ASR speech datasets using publicly available information, the approach involves surveying of the literature, data repositories, and web sources to find relevant datasets. A taxonomy with a rich set of attributes was created to describe and categorize each dataset. This involved both automatic metadata extraction and manual annotation, supported by community contributions. The final step was publishing and regularly updating the catalog and adding search and filter functionalities to facilitate dataset discovery.

A detailed description of the methodology was provided in Section 3.2 An overview of the process developed in this work is provided below.

1. **Survey of existing sources:**

    - **Literature review:** Review of previous studies, reports, and publications on Polish ASR datasets.
    - **Exploration of data repositories:** Exploration of repositories such as CLARIN, ELRA, and Hugging Face for speech datasets.
    - **Engagement in academic and industry collaborations:** Engagement with institutions, industry, and government organizations to find proprietary or unpublished datasets.
    - **Conducting web searches:** Conducting web searches to find datasets in conference papers, articles, and reports.
    - **Access to institutional repositories:** Access to universities' repositories for relevant datasets.
    - **Community contributions:** Gathering contributions from platforms like GitHub for shared datasets and tools related to ASR.

2. **Definition of attributes for categorization:**

- **Dataset characteristics:** Naming, description, sizing (hours of audio), and detailing transcriptions of datasets.
- **Source information:** Documentation of author(s), funding institution, and publication date.
- **Licensing and accessibility:** Identification of license type and accessibility (public, restricted, commercial).
- **Technical specifications:** Specification of audio format, sampling rate, and recording environment (studio, mobile, etc.).
- **Content and quality:** Detailing the number of speakers, speaker demographics (gender, age), and quality control processes.

3. **Collection of metadata:**
    - **Automatic extraction:** Use of web scraping tools and APIs to gather metadata from online sources.
    - **Manual annotation:** Manual annotation of datasets based on detailed examination of dataset documentation and publications.
    - **Community feedback:** Allowance for community contributions and updates to information about datasets to ensure completeness and accuracy.

4. **Catalog creation:**
    - **Development of a centralized repository:** Development of a centralized repository or catalog that hosts the metadata of all identified datasets.
    - **Implementation of search and filter functionality:** Implementation of search and filter functionalities to allow users to easily find datasets that meet their specific criteria.
    - **Ensuring regular updates:** Regular updates to the catalog with new datasets and revised metadata as more information becomes available

RQ2: What is the current state of the ASR speech data?

The survey results are presented in Section 4.1.2 The overview is presented below.

- **Oldest dataset:** The oldest publicly reported Polish ASR speech dataset is the \emph{Corpora} dataset, created in 1997 by Stefan Grocholewski.

- **Total number of datasets:** 53 datasets have been reported publicly for Polish ASR between 1997 and 2023, with a detailed breakdown provided by year.

- **Largest datasets:** The largest datasets are the \emph{Mobile Speech Dataset of Scripted Monolog} (1482 hours), \emph{JURISDIC} (855 hours), and \emph{Diabiz} (410 hours).

- **Total transcribed speech data:** The survey identified 5986 hours of transcribed speech data created between 1997 and July 2023.

- **Freely available data:** There are 1641 hours of transcribed speech and 27,099 hours of speech available for free use. This is a significant increase from the 223 hours reported in 2019.

- **Comparison with English data:** The total amount of public domain transcribed speech data for Polish is approximately 1400 hours, which is significantly smaller compared to over 100,000 hours for English.

- **Commercial data:** There are 3171 hours of transcribed speech available from commercial providers, which is nearly two and a half times higher than freely available datasets.

- **Top contributors:** The largest contributions of transcribed speech come from the Shaip company, University of Łódź (PELCRA group), and Adam Mickiewicz University.

- **Open-access data:** 31 datasets are available under permissive licenses, predominantly created by the University of Łódź PELCRA group and institutions outside Poland like FAIR and Mozilla.

- **Commercial providers:** 12 datasets are available from providers like ELRA, LDC, Appen, Shaip, and CLARIN-PL.

- **Largest repository:** The University of Łódź PELCRA group offers the largest selection of datasets, followed by the Polish CLARIN consortium's DSpace catalog.

- **Recording devices:** Most recordings are from mobile devices (1775 hours), followed by various devices (1370 hours), headsets (705 hours), and studio-quality microphones (473 hours).
- **Sampling rates:** Over 50% of datasets use a 16 kHz sampling rate, with other rates including 48 kHz and 8 kHz.
- **Types of speech:** Read speech is the most documented (56%), followed by conversational speech (20%) and public speech (5%).
- **Metadata availability:** Metadata such as speaker's age, gender, native language, and accent/region is available for many datasets, with recent datasets like \emph{SpokesBiz} providing rich annotations.

RQ3: How can the survey findings be shared and available for feedback from the ASR community?

The detailed answer was provided in Section 4.1.4 Below is the overview of actions taken.

1. Created a GitHub repository and Hugging Face space page to host the dataset catalog, metadata, and documentation, as well as provide a feedback form and discussion forum. GitHub allows for community feedback and contributions through issues and pull requests. Hugging Face additionally provides search and filter functionalities.
2. Published findings in peer-reviewed journal [53].

5.2.2 Observations from community feedback

Users of the survey and catalog can provide their opinions on the usability and their personal perspectives on the current availability of ASR speech data. As of writing the work, feedback has been received from six ASR professionals. The main findings are summarized below.

Professional Affiliations and Roles

- **Affiliations**: Respondents were from academia (3), a combination of academia and industry (1), non-profit or government organizations (1), and a combination of industry and non-profit or government organizations (1).

- **Roles**: The roles included Machine Learning Engineering (2), Data Management and Operations (2), and Applied Research (2).

Professional experience and estimates regarding data availability

- **Years of experience**: Respondents had 4-10 years (3), more than 10 years (2), and 0-3 years (1) of ASR-relevant professional experience.

- **Estimates before getting familiar with the catalog**:

  – **Total transcribed speech material**: The estimated amount ranged from 300 to 50,000 hours, with a mean of approximately 20,217 hours.

  – **Number of available datasets**: Estimates ranged from 4 to 80 datasets, with a mean of approximately 27 datasets.

Perception of dataset availability

- **Public-domain datasets**: The availability was mostly rated as low (4), with some rating it as medium (2).

- **Commercial datasets**: The availability was generally hard to assess (3), with one rating each for medium and low availability.

Impact of the catalog

- **Change in opinion on dataset availability**: After familiarizing themselves with the catalog, two respondents reported a moderate change in their opinion, while two noted a significant change.

- **Quality of the catalog**: The catalog was rated highly, with a mean score of 4.25 out of 5.

- **Practical usefulness of the catalog**: The practical usefulness of the catalog received a mean score of 4.75 out of 5, indicating high usefulness.

Summary The survey among ASR professionals shows the catalog positively impacts the perceived availability and accessibility of Polish ASR datasets. Initial estimates varied, but familiarity with the catalog led to more informed assessments. The catalog was highly rated for quality and usefulness, emphasizing its value in Polish ASR research and development. Respondents appreciated the clear format and detailed information. This feedback underscores the catalog's role in supporting academia and industry. However, the small sample size is not representative of the broader ASR community.

5.2.3 Implications

1. **Enhanced research efficiency:** Researchers can find suitable datasets more easily, accelerating ASR development and experimentation. Identifying dataset gaps can guide future research. Broader access to datasets lead to higher-quality publications and increased citations for dataset authors.

2. **Collaborative efforts:** The open catalog allows for updates and improvements through collaboration. Researchers and institutions can contribute new datasets and report issues, fostering a collaborative environment and keeping the catalog *up-to-date.*

3. **Benchmarking and evaluation:** As shows in this thesis, the catalog can serve as a foundation for creating a robust evaluation suite for Polish ASR systems, aligning with established practices in other ML fields and the international research community.

4. **Improved ASR systems and product development:** Access to diverse datasets can enhance the robustness and accuracy of commercial ASR systems and lead to more versatile and accurate ASR products.

5. **Strategic Investments:** Companies can potentially identify areas needing investment, such as data collection for underrepresented environments or speech types. Public-private partnerships can address dataset gaps, especially in high-demand areas like mobile speech recognition and non-native demographics.

5.2.4 Limitations

1. **Information accuracy:** The metadata for the catalog are mainly based on online information from language repositories and scientific articles. Despite efforts to cross-check the information with the authors, some metadata may be missing or inaccurate due to errors in the original sources.

2. **Difficulties in collecting feedback and tracking adoption:** Collective curation and verification is a proven method to address errors in catalog contents and share practical issues with the use of datasets. However, to engage the community, proactive promotion is required.

5.2.5 Future research directions

1. **Impact assessment**: Collect more feedback from the ASR community to measure the impact of the catalog and gather suggestions for improvement. Several methods can be used for the assessment, categorized as follows:

   (a) *Direct-subjective* – This method uses a feedback form among representatives of the ASR community to evaluate the perceived availability and usability of the Polish ASR speech dataset before and after reviewing the catalog. The form quantifies the difference between estimated and actual counts of available datasets, reflecting the 'accessibility gap' addressed by our research. A form[1] was created for this purpose and will remain open for ongoing feedback. The initial results presented above are encouraging, although the number of responses is currently too small to draw conclusions.

   (b) *Indirect-objective* – This method measures the number of datasets downloaded from language repositories before and after catalog publication. It requires log data from external parties, making it infeasible for the authors. Although objective, other factors may influence the download numbers, and downloading a dataset does not imply its use for ASR development.

   (c) *Direct-objective* – This method assumes that each dataset use is tracked in a central repository, allowing direct measurement of usage frequency. Currently not feasible from technical reasons.

2. **Catalog enrichment**:

   (a) Adding attributes from automatic analysis, for example, audio signal statistics, token, and phone distributions.

[1] https://forms.gle/tp9bWeJNDqa696do7

(b) Adding attributes from manual analysis, for example, transcription accuracy, speaker metadata, and speech and audio characteristics. While feasible, large-scale manual assessment needs efficient, low-effort auditing processes and a simple yet comprehensive error taxonomy. For example, datasets having transcriptions with WER below 2% are reliable for evaluation, while those above 5% should be used with caution to avoid performance degradation in training or biased evaluation results. Such an inspection is especially valuable for automatically created datasets like MLS, which often lack manual quality control. Recent work on multilingual parallel text corpora [63] has shown systematic errors from unsupervised methods.

(c) Adding information about unstructured audio and text resources, which are often used as the building material for ASR corpora, e.g., LibriVox, open subtitles YouTube, parliamentary speech text corpus, etc. along with manuals and tools for use. This could further reduce the time and effort required to find the appropriate resources to train and evaluate ASR models.

3. **Usability improvements**: Developing APIs and tools for easier integration of the catalog into various platforms and enhancing its usability for different end-user needs.

(a) Developing dedicated web API endpoints to extract relevant information from the catalog. The catalog is currently available for download in TSV format, which facilitates automatic processing such as batch download and analysis, but has limited functionalities with respect to version control.

(b) The information in the catalog could be fed back to public catalogs and language repositories used by the ASR and ML communities. Tools for automatic or semi-automatic generation of data cards in supported formats would be required to achieve that.

4. **Catalog promotion**:

(a) Organizing workshops and webinars to present findings.

(b) Sharing updates and soliciting feedback through mailing lists and newsletters.

(c) Using Twitter, LinkedIn, and ResearchGate to share findings, updates.

(d) Partnering with initiatives like Common Voice or CLARIN, for broader dissemination.

5.3 RO2: Design and curation of ASR benchmark dataset for Polish

5.3.1 Results overview

RQ4: What factors are crucial in designing and curating a new ASR benchmark dataset?

A curated Benchmark dataset for Polish ASR systems is intended to have the following features:

- **Task-appropriate:** Relevant and practical for the intended ASR task.
- **Accessible:** Available online under a license that allows the free use and creation of derivative works.
- **Discoverable:** Easy to find and acquire (without time-consuming registration or other access barriers).
- **Diverse and challenging:** Containing various examples to test the adaptability of the model, as well as complex cases to encourage community participation and minimize the risk of benchmark saturation.
- **Annotated**: With metadata about speakers and recordings, allowing nuanced analysis and interpretation of the results.
- **Optimally sized:** Large enough to be representative, but manageable to download and explore.
- **Clean yet realistic:** Free of major errors, but noisy enough to represent the complexity of the real world.
- **Well-documented:** Provided with documentation that is understandable to users without technical skills.

- **Well-explained:** Provided with evaluation baselines and how-to-use script examples.

RQ5: What steps are necessary to curate a benchmark dataset from publicly available resources?

The preliminary step is selection of datasets meeting mandatory criteria:

- Datasets are available online under a license allowing free use for non-commercial purposes.
- Transcriptions are aligned with the recordings.
- Recording sampling rate is at least 8 kHz.
- Audio files are encoded using at least 16 bits per sample.

The following is an overview of the curation process applied to selected datasets.

1. **Dataset structure curation:**
    - Downloading and manually inspecting format and contents
    - Creating train/dev/test splits if not available
    - Assigning standard IDs to speakers and files
2. **Audio file curation:**
    - Removal of invalid audio files
    - Unifying audio format to WAV 16 bits/16 kHz
    - Normalizing audio amplitude to -3 dBFS
    - Splitting long audio files into shorter segments based on time-alignment annotations
3. **Text files (transcripts and metadata) curation:**
    - Converting text encoding to UTF8
    - Extracting original transcription and removing redundant characters
    - Extracting and unifying metadata contents

- Generating metadata from text and audio content
- Saving in the standard tabular format

4. **Dataset distribution**

- Converting to Hugging Face Datasets format
- Uploading to Hugging Face Hub
- Creating Hugging Face Datasets build script
- Referencing the original licenses and authors in the README
- Setting Gated datasets to acknowledge the original licenses

RQ6: What publicly available Polish speech datasets can be used to curate benchmark datasets? The BIGOS V2 dataset comprises diverse speech datasets across various domains, speech types, and interaction types. Key datasets include Clarin Studio and Clarin Mobile (CC-BY, monolingual), and Munich AI Labs LibriVox (proprietary, multilingual). Other notable datasets are Mozilla Common Voice (CC-0), Multilingual Librispeech (CC-BY), and Google FLEURS (CC-BY), all multilingual. Speech types range from read speech and audiobooks to spontaneous speech (PWR AZON). Interaction types are mainly monologs, recorded in quiet environments using various audio devices (Common Voice).

The PELCRA for BIGOS dataset focuses on spontaneous dialogues in domains like customer service, open domain, and public speeches. It includes DiaBiz ASR PolEval 22 (public domain) and SpokesBiz subsets (CC-BY-NC-ND), with monolingual data recorded using lavalier microphones in quiet settings. Notable entries include the SpokesMix series, featuring emotional and parliamentary speeches. The speech sources range from volunteers to public speakers, providing diverse acoustic environments and interaction types. The PELCRA dataset complements BIGOS by evaluating ASR performance in spontaneous conversational contexts.

RQ 7: How can the benchmark dataset be shared and available for feedback from the ASR community?

The benchmark dataset was shared and made available for feedback from the ASR community by using the Hugging Face platform, which ensured the discoverability and long-term accessibility of the BIGOS and PELCRA datasets. These datasets were shared under open licenses (CC-BY-SA for BIGOS and CC-BY-NC-ND for PELCRA), facilitating wide use and collaboration. Furthermore, the results of the content analysis of the dataset were made available on a public dashboard, providing information and helping to interpret the evaluation results. The platform also enabled tracking of downloads and provided mechanisms for users to provide direct feedback, ensuring continuous improvement based on community input.

5.3.2 Observations

Curated dataset contents

1. **Diverse licensing**: Both *BIGOS* and *PELCRA for BIGOS* datasets include a mix of open licenses (e.g., CC-BY, CC-BY-NC-ND) and proprietary licenses in the case of BIGOS. Curation of public datasets required careful verification of source licenses, and, for datasets under CC-BY-NC-ND licenses, consultation with original authors to get consent for curation and distribution. The curated datasets referenced the original licenses. Similarly to the Common Voice dataset hosted on Hugging Face Hub, the Gated datasetsmechanism was used to collect user acknowledgments of the license terms.

2. **Variety of speech types**: The datasets encompass both read and spontaneous speech, with *BIGOS* focusing on read speech and *PELCRA for BIGOS* emphasizing spontaneous speech. This ensures the datasets are comprehensive and challenging for ASR evaluation. Significant effort was invested to unify formats, such as creating consistent train/dev/test splits and assigning standard IDs to speakers and files.

3. **Wide range of domains**: Recordings from multiple domains were included in BIGOS and PELCRA datasets as presented in tables 4.14 and 4.15, respectively. This diversity offered comprehensive testing scenarios across various contexts, but also increased the curation workload to eliminate inconsistencies in transcription formats.

4. **Monolog vs. dialog interactions**: BIGOS datasets mainly include monologs, crucial for testing ASR systems in voice commands and dictation. PELCRA for

BIGOS, on the other hand, focuses on dialogs in customer service and spontaneous interactions. Manual annotation of conversational speech is labor-intensive, making the availability of such Polish datasets under permissive licenses highly valuable for the ASR community.

5. **Acoustic environments and audio devices**: Recordings range from quiet studios to public spaces and use devices from studio mics to mobile phones. BIGOS features controlled environments, while PELCRA for BIGOS includes varied ones. Extensive curation was needed, including format unification to WAV 16 bits/16 kHz, amplitude normalization, and segmenting long files, ensuring robustness in ASR performance evaluation.

6. **Extensive metadata**: Both datasets offer rich metadata (e.g., speaker gender, age, recording conditions), with PELCRA having more available metadata. The curation process involved converting text to UTF8, standardizing metadata, and generating additional metadata from audio and text transcriptions, aiding ASR performance analysis.

7. **Focus on realism**: BIGOS balances clean data with realistic noise, ensuring practical data. PELCRA for BIGOS reflects real-world complexities, requiring greater care in removal of invalid audio files and adding lexicon-based normalization .

8. **Publicly accessible analysis**: Analysis results are made available on public dashboards like Hugging Face for both datasets. This promotes transparency, community engagement, and collaborative improvements, demonstrating the significant efforts put into making these datasets a valuable resource for the ASR community.

Validation of benchmark dataset design

5.3.3 Implications

The design goals set prior to the curation of the BIGOS datasets were met. The preservation and analysis of publicly accessible datasets had several positive implications. The first is the resolution of interoperability issues. Standardizing the format and structure of the datasets makes it easier for practitioners and researchers to integrate and use these

Design Requirement	**Observation**
Task-appropriate	Varied speech types (read and spontaneous) in BIGOS and PELCRA dataset
Accessible	Requires acknowledgment of original licenses.
Discoverable	Publicly listed on popular platform. Dashboard with dataset cards.
Diverse and challenging	Wide range of domains, speech types, acoustic conditions and speakers.
Annotated	Extensive metadata, especially in PELCRA.
Optimally sized	Significant efforts in unification and curation of many formats.
Clean yet realistic	Balance of clean data and realistic noise levels.
Well-documented	Acknowledgement of original licenses and authors.Documented curation proce
Well-explained	Evaluation baselines and script examples available.

Table 5.1: Benchmark dataset design requirements validation

resources. Standardization also increases consistency for comparative analysis and benchmarking. Additionally, making curated datasets accessible on platforms like Hugging Face increases their practical utility. Users can download and use the dataset with a single line of code, streamlining access and integration with popular ML tools. This ease of access should positively impact adoption and facilitate work on automatic speech recognition (ASR) by the global ML community. Furthermore, providing a reference dataset like BIGOS sets a benchmark for future research and evaluations. It can serve as the standard for testing new models and systems, helping to track progress and ensuring advances in ASR technology are measured against a stable reference. These efforts advance Polish ASR by fostering best practices, promoting teamwork, and ensuring high-quality data availability.

It is crucial to note that the curated datasets are based on the work of original authors. Therefore, the documentation and access methods of curated datasets were designed to recognize the original creators and the licenses of the dataset. References to original documentation, articles, and authors are provided. Furthermore, the documentation for both the BIGOS and PELCRA datasets urges and supports users to cite the original work upon which the curated datasets are founded, by providing bulk citation in *BibTeX* format.

Future work includes adding noisy recordings to test ASR robustness[130] and collecting new recordings from various Polish speakers[2, 1]. Methods can also be explored to expand vocabulary coverage using curated or generated prompts and Text to Speech (TTS) engines.[123]

5.4 RO3: Survey of ASR benchmarks for Polish

5.4.1 Results overview

RQ7: How to identify and systematically categorize Polish ASR benchmarks using publicly available information?

Polish ASR benchmarks were identified and classified by reviewing public reports and publications. Forty attributes were used to categorize information, including the year, evaluated systems, best models, average WER, and conclusions. The details of the survey methodology were described in Section 3.4

RQ8: What methods, datasets, and ASR systems have been considered in the Polish ASR benchmarks so far?

The methods included only automatic evaluations and various metrics based on string matching. All benchmarks used the WER metric. Medical UW SOVVA PS 23 employed manual inspection and error classification. The datasets covered various domains such as voice control, parliamentary speech, customer support, student presentations, podcasts, and medical interviews. Evaluated ASR systems include commercial solutions such as Azure and Google, as well as academic models such as Whisper, GOLEM, ARM, and Voicelab. One benchmark incorporated a sociodemographic study focusing on the speaker's gender (Medical PG 23). The scripts and the complete dataset for evaluation were provided for one benchmark (SpokesBiz CLARIN 23). The details were described in Section 4.3.2

RQ9: What automatic speech recognition (ASR) systems supporting Polish have not yet been evaluated?

Specialized Google STT models, NVidia Nemo models, Facebook MMS models, Assembly AI systems and Whisper models other than large have not yet been comprehensively evaluated for Polish language, yet. This gap was addressed in this work.

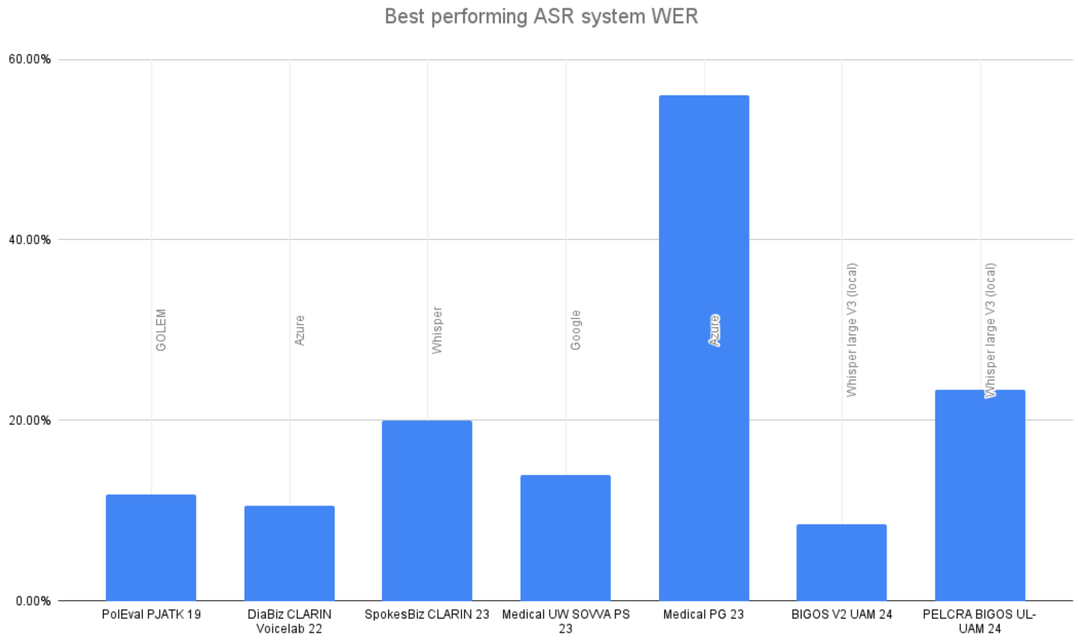


Figure 5.1: WER scores of top systems in Polish ASR benchmarks.

RQ10: Which ASR benchmarks have evaluated commercial and freely available systems?

Three benchmarks (DiaBiz, SpokesBiz, and Medical PG) have assessed both commercial and open-source systems. Two benchmarks (Medical UW SOVVA PS 23, BOR POLSL PS 18) focused solely on commercial systems. The SpokesBiz 23 benchmark was utilized to evaluate the freely available Whisper, whereas PolEval 19 was used to compare performance of community-provided systems.

RQ11: What ASR system is ranked as the best performing one?

Microsoft's Azure was identified as the top-performing ASR system in two out of the three benchmarks where it was evaluated (DiaBiz and Medical PG). In this study, the highest accuracy was achieved by the freely available Whisper large V3 model. The comparison of the best system WER reported in other benchmarks and obtained in this study is illustrated in 5.1. The benchmarks conducted in this study are named BIGOS V2 UAM 24 and PELCRA BIGOS UAM UL 24.

RQ12: How are the main conclusions derived from the ASR benchmarks so far?

The main conclusions are the following:

- Specific tasks, such as dictating medical terms, present substantial challenges for current ASR systems; however, reported WER rates may vary significantly, making it difficult to generalize findings.
- The performance of ASR systems varies across different tasks and domains.
  - Medical domain – WER of 14% (Medical UW SOVVA PS 23) and 56% (Medical PG 23)
  - Government training voice commands – mean WER of 50%
  - Customer support conversations – mean WER of 10.5%
  - Podcasts, spontaneous interview – mean WER of 20%
- Azure systems outperform other commercial systems.
- Commercial systems outperform freely available solutions.
- The impact of sociodemographic factors on ASR performance was not thoroughly investigated yet.
- PELCRA group leads in good practices; Diabiz 22 and Spokes 23 benchmarks are comprehensive in terms of vocabulary coverage, open to contributions from the community and replicable thanks to datasets availability.

RQ13: How to share the Polish ASR benchmark survey with the community?

The survey results are shared as spreadsheet[2] and HF dashboard [3] allowing the community to engage with the data, offer feedback, and collaborate on further improvements.

5.4.2 Implications

The survey on Polish ASR benchmarks provided a comprehensive overview of publicly reported ASR performance in different tasks and domains. The survey highlighted perfor-

[2] Polish ASR benchmarks catalog

[3] AMU PL ASR survey

mance variability based on recording quality, vocabulary domain, and type of speech. The development of a standardized taxonomy allowed a comparison of the benchmarks conducted so far and identification of methodological gaps. The survey results were publicly shared to promote transparency and community participation.

The methodological improvements to consider include:

1. Inclusion of larger number of ASR systems and datasets.
2. Analysis of the impact of sociodemographic factors such as age, gender, and regional accents.
3. Sharing datasets and evaluation results for manual verification and replication.

5.5 RO4: Design and implementation of system for ASR systems benchmarking

5.5.1 Results overview

RQ 15: What tools and systems exist for ASR benchmarking?

For dataset management, tools such as Pandas, Hugging Face Datasets, and SDE are widely used. Pandas is a data manipulation and analysis library, capable of handling various data formats and performing complex operations. The Hugging Face Datasets library simplifies the loading, processing, and sharing of public datasets and supports efficient handling of large datasets. SDE provides tools for exploring and analyzing speech datasets, including dataset statistics, audio data inspection, and transcription analysis.

For ASR evaluation, commonly used tools include *sclite, jiwer, asr-evaluation, fstalign, evaluate*, and *asr-evaluator*. *Sclite* calculates WER and provides speaker-level statistics and alignment. *Jiwer* offers CER, MER, WIL, and text normalization. *Asr-evaluation* by Ben Lambert calculates WER, WRR, SER, and provides confusion tables. *Fstalign* supports multiple input formats and detailed error analysis. *Evaluate* by Hugging Face offers fewer metrics but integrates well with Hugging Face Datasets and Transformers. *Asr-evaluator* from NVIDIA's NeMo toolkit includes on-the-fly data augmentation and reliability scoring.

RQ 16: What challenges arise in evaluating multiple ASR systems, and what strategies can address them?

The main challenge is standardizing datasets and managing diverse data formats. Normalizing hypotheses across systems and datasets is crucial. Specialized tools such asPandas, Hugging Face Datasets, JIWER and SDE help standardize formats and detect quality issues.

Another major challenge is ensuring reproducibility. The sharing of datasets and tools, as in this research, standardizes the evaluation procedures and references used for scoring ASR output accuracy.

RQ 17: How can the system be extended to new ASR systems, datasets, languages, metrics, and normalization methods?

The possibility for extension was a crucial design requirement as detailed in section 3.11. Detailed instructions on incorporating a new dataset, system, or metric into the evaluation process are provided in the documentation on GitHub[4].

5.6 RO5: Using a curated dataset to benchmark ASR systems for Polish

5.6.1 Results overview

RQ 18: What is the ASR accuracy for different datasets?

This section analyzes the content of the dataset and the results of the ASR evaluation. The features calculated for all splits serve as a reference for assessing linguistic diversity and recognition difficulty. Some feature values, such as vocabulary size, may be lower for the evaluation subset. However, due to the pseudorandom split method, the features of the whole dataset are accurate enough to approximate the characteristics of the test set for the purpose of this analysis. Some of the initial observations are somewhat speculative. To confirm the hypotheses of subpar performance indicated by these initial observations, it is necessary to manually inspect the data (including references, hypotheses, and audio recordings), conduct further speech corpus analysis, or carry out additional experiments

[4] AMU BIGOS Tools

with controlled variables. Section 4.4.3 presented WER scores for BIGOS and PELCRA dataset subsets.

Common Voice dataset (mozilla-common_voice_15-23)

- **Description**: Multilingual resource with global volunteers reading pre-defined sentences.
- **Evaluation and dataset statistics**:
  - **Median WER**: 10.09%
  - **Range**: 2.1% to 55.77%
  - **Unique utterances**: 36,853
  - **Unique words**: 66,815
  - **Unique characters**: 87
  - **Words per second**: 1.6
  - **Characters per second**: 9.6
  - **Avg. duration**: 5.17s
  - **Avg. length**: 8.27 words, 57.88 characters
- **Insight**: Relatively low median WER (10%) and nearly perfect accuracy for Whisper (WER 2.1%) were expected, given the wide adoption of Common Voice dataset for ASR training. Nearly perfect accuracy for Whisper may result from test data leakage. High WER variability indicates struggles of smaller models with diverse accents and pronunciations. Another factor that adversely affects accuracy is the presence of non-standard characters such as symbols, numerals, or diacritics, which are sourced from the internet and used as prompts for speakers.

Multilingual LibriSpeech (MLS) dataset (fair-mls-20)

- **Description**: Large multilingual corpus from LibriVox audiobooks.
- **Statistics**:
  - **Median WER**: 11.29%

  - **Range**: 4.42% to 52.73%
  - **Unique utterances**: 26,069
  - **Unique words**: 89,464
  - **Unique characters**: 37
  - **Words per second**: 2.28
  - **Characters per second**: 12.24
  - **Avg. duration**: 14.89s
  - **Avg. length**: 33.98 words, 216.31 characters

- **Insight**: The extensive vocabulary is counterbalanced by the limited alphabet size, which is a consequence of the test sets being cleaned and the references manually transcribed. The low WER for the best system (4.4%) could be attributed to the quiet audio environment and the precise pronunciation typical of audiobook readings. Considering that MLS is the second most widely used multilingual dataset, following Common Voice, the potential for test data leakage should also be taken into account.

Clarin Studio dataset (clarin-pjatk-studio-15)

- **Description**: Polish corpus with 13,802 short utterances recorded in a studio.
- **Statistics**:
  - **Median WER**: 10.93%
  - **Range**: 5.86% to 34.96%
  - **Unique utterances**: 13,525
  - **Unique words**: 57,853
  - **Unique characters**: 39
  - **Words per second**: 2.87
  - **Characters per second**: 8.65
  - **Avg. duration**: 14.71s
  - **Avg. length**: 42.2 words, 169.44 characters

- **Insight**: The extensive vocabulary (comparable to Common Voice) and the lengthy average utterances indicate rich linguistic content. Low median WER may result from clean speech recorded in controlled environment. Noteworthy, decent performance is achieved even for models with higher error rates (35% WER).

Clarin Mobile dataset (clarin-pjatk-mobile-15)

- **Description**: Polish dataset recorded via telephone, reflecting contemporary pronunciation.
- **Statistics**:
  - **Median WER**: 14.57%
  - **Range**: 5.57% to 52.96%
  - **Unique utterances**: 3,487
  - **Unique words**: 26,424
  - **Unique characters**: 35
  - **Words per second**: 2.03
  - **Characters per second**: 11.77
  - **Avg. duration**: 12.86s
  - **Avg. length**: 26.08 words, 177.44 characters
- **Insight**: The inherent difficulties of telephony, including background noise and inconsistent audio quality, lead to a median WER that is 5 percentage points higher than that of Clarin Studio. Despite having a smaller vocabulary and a more limited character set than Clarin Studio, the accuracy is lower.

Jerzy Sas PWR datasets (Politechnika Wrocławska)

- **Male speaker speech set (pwr-maleset-unk)**
  - **Median WER**: 9%
  - **Range**: 3.79% to 43.79%
  - **Unique utterances**: 4,006

  - **Unique words**: 12,970
  - **Unique characters**: 62
  - **Words per second**: 1.71
  - **Characters per second**: 10.05
  - **Avg. duration**: 4.85s
  - **Avg. length**: 8.3 words, 57.07 characters
  - **Insight**: The single speaker, controlled conditions and small vocabulary size result in low median WER, despite large character set,

- **Utterances containing short words (pwr-shortwords-unk)**
  - **Median WER**: 7.77%
  - **Range**: 4% to 46.06%
  - **Unique utterances**: 668
  - **Unique words**: 3,649
  - **Unique characters**: 54
  - **Words per second**: 1.76
  - **Characters per second**: 10.32
  - **Avg. duration**: 5.44s
  - **Avg. length**: 9.59 words, 65.76 characters
  - **Insight**: Lowest median WER for all subsets. Yet, short words are challenging for some ASR systems due to brief, often unclear pronunciation, which is reflected in variable WER.

- **Spoken commands (pwr-viu-unk)**
  - **Median WER**: 10.54%
  - **Range**: 0% to 260.86%
  - **Unique utterances**: 13
  - **Unique words**: 18
  - **Unique characters**: 27

  - **Words per second**: 1.27
  - **Characters per second**: 6.98
  - **Avg. duration**: 1.39s
  - **Avg. length**: 1.77 words, 11.45 characters
  - **Insight**: Although the dataset had a very small alphabet and vocabulary, it posed significant challenges for many systems. The slow pronunciation of short recordings led to a high rate of insertion errors (Hallucinations), causing considerable variability in WER. Interestingly, the Whisper Cloud model exhibited a higher tendency to hallucinate compared to locally hosted models.

M-AI Labs Speech corpus (mailabs-19)

- **Description**: Derived from LibriVox audiobooks, curated semi-automatically.
- **Statistics**:
  - **Median WER**: 10.28%
  - **Range**: 4.9% to 44.06%
  - **Unique utterances**: 14,796
  - **Unique words**: 51,144
  - **Unique characters**: 77
  - **Words per second**: 2.18
  - **Characters per second**: 12.08
  - **Avg. duration**: 7.79s
  - **Avg. length**: 16.99 words, 111.07 characters
- **Insight**: Comparable median and range of WER to MLS, even with a character set that is twice as large. The speech characteristics are alike, which is expected given the same source (public audiobooks). Nevertheless, the average audio duration and utterance lengths are roughly half of those in MLS, indicating that different segmentation techniques were used for the two datasets.

AZON Read and Spontaneous Speech datasets (pwr-azon_read-20, pwr-azon_spont-20)

- **Read Speech (pwr-azon_read-20)**
  - **Median WER**: 16.98%
  - **Range**: 5.31% to 38.47%
  - **Unique utterances**: 1,517
  - **Unique words**: 7,628
  - **Unique characters**: 32
  - **Words per second**: 1.35
  - **Characters per second**: 10.17
  - **Avg. duration**: 7.38s
  - **Avg. length**: 9.96 words, 85.06 characters
  - **Insight**: The controlled recordings lead to a moderate Word Error Rate (WER). The limited vocabulary and character set result in a narrow WER range across different systems. However, the median WER is relatively high, likely due to the presence of scientific terms.
- **Spontaneous Speech (pwr-azon_spont-20)**
  - **Median WER**: 24.03%
  - **Range**: 16.4% to 44.74%
  - **Unique utterances**: 456
  - **Unique words**: 5,004
  - **Unique characters**: 32
  - **Words per second**: 2.24
  - **Characters per second**: 12.36
  - **Avg. duration**: 16.9s
  - **Avg. length**: 37.84 words, 246.76 characters
  - **Insight**: Spontaneous and longer recordings are more challenging than read speech. Higher WER values from unpredictability and natural speech variations can be observed, despite small vocabulary size.

Google FLEURS (google-fleurs-22)

- **Description**: Parallel speech benchmark dataset in 102 languages, used for evaluating ASR and translation systems.
- **Statistics**:
  - **Median WER**: 14.52%
  - **Range**: 5.33% to 56.06%
  - **Unique utterances**: 1,919
  - **Unique words**: 13,826
  - **Unique characters**: 71
  - **Words per second**: 1.67
  - **Characters per second**: 10.06
  - **Avg. duration**: 11.04s
  - **Avg. length**: 18.45 words, 129.5 characters
- **Insight**: The moderate linguistic diversity poses a reasonable challenge. The variation in WER is comparable to other well-known multilingual datasets such as Common Voice and MLS.

PolyAI Minds14 (polyai-minds14-21)

- **Description**: Dataset for the development of spoken intent recognition systems in e-banking domain.
- **Statistics**:
  - **Median WER**: 36.61%
  - **Range**: 13.59% to 139.29%
  - **Unique utterances**: 550
  - **Unique words**: 1,636
  - **Unique characters**: 69
  - **Words per second**: 0.92

  - **Characters per second**: 4.91
  - **Avg. duration**: 19.65s
  - **Avg. length**: 18.08 words, 114.65 characters

- **Insight**: The significantly higher WER points to difficulties in accurately transcribing lengthy audio files with diverse and specific intents in e-banking speech. WER values exceeding 100% imply issues related to Hallucinations. A manual review of audio samples uncovered data quality problems such as truncated recordings and repeated utterances, leading to extremely high WER for recordings with unusually rapid speech rates.

PolEval 22 Diabiz sample (ul-diabiz_poleval-22)

- **Description**: Dialog corpus of phone-based customer-agent interactions.
- **Statistics**:
  - **Median WER**: 48.97%
  - **Range**: 30.04% to 100.27%
  - **Unique utterances**: 8,760
  - **Unique words**: 13,716
  - **Unique characters**: 72
  - **Words per second**: 2.97
  - **Characters per second**: 13.56
  - **Avg. duration**: 3.96s
  - **Avg. length**: 11.75 words, 65.42 characters
- **Insight**: The high WER and dense speech rates indicate significant challenges in recognizing rapid, complex interactions typical in customer-agent dialogues. The high variability reflects different accents, noise levels, and conversational interruptions. Considerably higher baseline WER for the best system (30%) compared to evaluation performed on official DiaBiz test set [110] points to potential issues with transcriptions or audio files requiring further examination.

CBIZ_BIO (ul-spokes_biz_bio-23)

- **Description**: Biographical interviews covering childhood, job, and family, with an informal tone.
- **Statistics**:
  - **Median WER**: 51.52%
  - **Range**: 27.89% to 114.1%
  - **Unique utterances**: 54,096
  - **Unique words**: 108,163
  - **Unique characters**: 113
  - **Words per second**: 2.57
  - **Characters per second**: 12.92
  - **Avg. duration**: 9.04s
  - **Avg. length**: 23.28 words, 140.11 characters
- **Insight**: The elevated median WER indicates the difficulty in transcribing biographical interviews. Informal, storytelling speech with an extensive alphabet and vocabulary, inconsistent pronunciation, and rapid speech rate contribute to the challenges faced by ASR systems.

CBIZ_INT (ul-spokes_biz_int-23)

- **Description**: Job interviews for potential babysitters.
- **Statistics**:
  - **Median WER**: 25.24%
  - **Range**: 12.1% to 49.4%
  - **Unique utterances**: 1,100
  - **Unique words**: 5,195
  - **Unique characters**: 68
  - **Words per second**: 2.85

  - **Characters per second**: 14.62
  - **Avg. duration**: 7.31s
  - **Avg. length**: 20.85 words, 127.72 characters

- **Insight**: The limited vocabulary and short recordings durations results in relatively low median WER, despite high speech rates typical to conversational speech.

CBIZ_LUZ (ul-spokes_biz_luz-23)

- **Description**: Unrestricted conversations among friends and families, characterized by their free and natural flow.
- **Statistics**:
  - **Median WER**: 38.15%
  - **Range**: 19.67% to 61.29%
  - **Unique utterances**: 41,600
  - **Unique words**: 87,990
  - **Unique characters**: 105
  - **Words per second**: 2.94
  - **Characters per second**: 13.85
  - **Avg. duration**: 6.37s
  - **Avg. length**: 18.74 words, 107.01 characters
- **Insight**: Informal, spontaneous conversations exhibit higher WER due to diverse vocabulary and natural speech patterns. The large size of alphabet and rapid speech rate add to ASR difficulties.

CBIZ_POD (ul-spokes_biz_pod-23)

- **Description**: Internet podcasts focusing on board games, nature photography, society, traveling, and international affairs.
- **Statistics**:

  - **Median WER**: 37.22%
  - **Range**: 20.25% to 59.47%
  - **Unique utterances**: 22,753
  - **Unique words**: 69,735
  - **Unique characters**: 101
  - **Words per second**: 3.06
  - **Characters per second**: 15.38
  - **Avg. duration**: 8.68s
  - **Avg. length**: 26.56 words, 160.07 characters

- **Insight**: The varied topics and informal tone in podcasts result in high WER, reflecting challenges in handling diverse and complex content at fast speech rates.

CBIZ_PRES (ul-spokes_biz_pres-23)

- **Description**: Student presentations on topics including culture, literature, parenting, and gender roles.
- **Statistics**:
  - **Median WER**: 27.7%
  - **Range**: 10.54% to 53.22%
  - **Unique utterances**: 17,155
  - **Unique words**: 47,352
  - **Unique characters**: 100
  - **Words per second**: 2.17
  - **Characters per second**: 11.98
  - **Avg. duration**: 6.76s
  - **Avg. length**: 14.66 words, 95.66 characters
- **Insight**: The structured nature of presentations and moderate speech rates contributes to lower WER, though the variety in topics and presentation styles presents challenges.

CBIZ_VC & CBIZ_VC2 (ul-spokes_biz_vc-23 & ul-spokes_biz_vc2-23)

- **Description**: Thematic discussions on society and lifestyle.
- **Statistics**:
  - **Median WER**: 27.39% (VC) and 36.5% (VC2)
  - **Range**: 11.75% to 58.35% (VC) and 22.46% to 62.59% (VC2)
  - **Unique utterances**: 44,647 (VC) and 25,567 (VC2)
  - **Unique words**: 63,913 (VC) and 79,725 (VC2)
  - **Unique characters**: 96 (VC) and 114 (VC2)
  - **Words per second**: 3.03 (VC) and 2.59 (VC2)
  - **Characters per second**: 14.83 (VC) and 12.93 (VC2)
  - **Avg. duration**: 4.14s (VC) and 11.31s (VC2)
  - **Avg. length**: 12.56 words, 73.97 characters (VC) and 29.3 words, 175.44 characters (VC2)
- **Insight**: Thematic conversations exhibit moderate to high WER because of the diverse and intricate vocabulary and fast speech rates.

CBIZ_WYW (ul-spokes_biz_wyw-23)

- **Description**: Interviews with a fixed set of questions on personal preferences and experiences.
- **Statistics**:
  - **Median WER**: 39.78%
  - **Range**: 17.14% to 68.61%
  - **Unique utterances**: 11,192
  - **Unique words**: 39,147
  - **Unique characters**: 94
  - **Words per second**: 2.56
  - **Characters per second**: 12.74

  - **Avg. duration**: 8.94s
  - **Avg. length**: 22.85 words, 136.74 characters
- **Insight**: Structured interviews occurred to be challenging for many of ASR systems, resulting in high median WER of almost 40%. The varied vocabulary and moderate speech rate impact ASR performance.

PELCRA_EMO (ul-spokes_mix_emo-18)

- **Description**: Focused interviews of people reflecting on their emotions.
- **Statistics**:
  - **Median WER**: 13.92%
  - **Range**: 5.27% to 62.46%
  - **Unique utterances**: 20,798
  - **Unique words**: 15,485
  - **Unique characters**: 67
  - **Words per second**: 2.74
  - **Characters per second**: 12.23
  - **Avg. duration**: 3.79s
  - **Avg. length**: 10.37 words, 56.71 characters
- **Insight**: Emotional speech shows relatively lower WER, possibly due to smaller vocabulary size, speech rate and distinct prosody aiding recognition.

PELCRA_LUZ (ul-spokes_mix_luz-18)

- **Description**: Open interviews representing conversational speech.
- **Statistics**:
  - **Median WER**: 43.33%
  - **Range**: 24.15% to 76.4%
  - **Unique utterances**: 19,526

  - **Unique words**: 20,101
  - **Unique characters**: 83
  - **Words per second**: 3.03
  - **Characters per second**: 13.75
  - **Avg. duration**: 3.22s
  - **Avg. length**: 9.78 words, 54.13 characters

- **Insight**: Spontaneous conversational speech shows a high word error rate because of its inherent variability and fast speaking pace. The extensive vocabulary further complicates automatic speech recognition performance.

PELCRA_PARL (ul-spokes_mix_parl-18)

- **Description**: Spoken parliamentary content.
- **Statistics**:
  - **Median WER**: 23.39%
  - **Range**: 13.33% to 72.62%
  - **Unique utterances**: 8,502
  - **Unique words**: 15,338
  - **Unique characters**: 78
  - **Words per second**: 2.29
  - **Characters per second**: 12.86
  - **Avg. duration**: 5.1s
  - **Avg. length**: 11.67 words, 77.31 characters
- **Insight**: Parliamentary speech, while formal and structured, involves complex and potentially jargon-filled language, contributing to moderate difficulty for ASR systems. The moderate speech rate and rich vocabulary reflect the nature of formal oratory.

RQ 19: What is the accuracy gap between commercial and free systems?

Conversational speech (PELCRA) has higher error rates due to its spontaneous nature, with more variability in style, speed, and pauses. Read speech (BIGOS) is more structured and consistent, resulting in lower WERs. The more detailed analysis can be found in section 4.3.2

RQ 20: Does ASR accuracy vary with speech features?

Both `whisper_large_v3` and `whisper_cloud` perform similarly across speech rates. For rates between 1.5 and 5, most WERs are below 30%. Severe errors occur at lower rates, while higher rates increase WER, indicating limited robustness for faster speech. Outliers suggest challenging scenarios or truncated audio files or transcriptions. More details can be found in section 4.4.3

RQ 21: Is there an accuracy difference by age or sex?

Evaluation on both datasets revealed gender bias in evaluated ASR systems. Male speakers generally experience lower WERs, but the degree of bias varies among systems. Given small size further research is required to validate the findings. The summary of reqults is presented below and details in sections 4.4.3 and 4.4.3

- **BIGOS Dataset:**
  - Evaluation dataset size: 1200 samples (442 female, 758 male).
  - Significant gender bias observed in WER, with male speakers typically achieving lower WERs.
  - Largest disparity: "azure_latest" system (-15.19 p.p. lower WER for males than females).
  - Minimal bias in systems like "whisper_large_v3" and "assembly_best" (-0.82% and -0.8%).
  - Median difference for all systems: -0.92 p.p.
  - Average difference for all systems: 0.82 p.p.
- **PELCRA Dataset:**

- Evaluation dataset size: 1597 samples (908 female, 689 male).
- Gender bias also present, with male speakers generally having lower WERs.
- Largest disparity: "assembly_nano" system (-26.23 p.p. lower WER for males than females).
- Some systems showed a positive difference favoring females, such as "google_v2_short" and "azure_latest".
- Median difference for all systems: -4.7 p.p.
- Average difference for all systems: -4.59 p.p.

5.6.2 Observations

Impact of normalization

Normalization techniques resulted in significant reductions in error rates for all types of metrics (SER, WER, MER, CER). Applying all methods reduced WER by 16.07 p.p. for the PELCRA dataset and 15.52 p.p. for the BIGOS dataset, highlighting the sensitivity of lexical metrics to spelling and formatting variations.

Impact of model size on accuracy

- `whisper_large v2`, `whisper_large`, and `whisper_large v3` show the best performance with the lowest WERs and the largest model sizes.
- `whisper_tiny` is the second smallest model and has the highest WER among all evaluated.
- `nemo_pl_quartznet` and `nemo_pl_multilang` are relatively small models with reasonably low WERs, indicating that they are efficient given their size.

Detailed results are presented in Section 4.4.3

5.6.3 Implications

The evaluation of Polish ASR systems is the largest to date in terms of systems and datasets. Its benefits include:

Benchmark	Year	Datasets	Models	Dataset-model comb.
BOR POLSL PS 18	2018	2	3	6
PolEval PJATK 19	2019	1	6	6
DiaBiz CLARIN Voicelab 22	2022	7	3	21
Medical PG 23	2023	1	3	3
Medical UW SOVVA PS 23	2023	1	3	3
SpokesBiz CLARIN 23	2023	8	1	8
BIGOS V2 UAM 24	2024	12	25	300
PELCRA BIGOS UL-UAM 24	2024	12	25	300

Table 5.2: Evaluated models, datasets, and their combinations.

- Informing the public about the strengths and weaknesses of ASR systems supporting Polish.

- Quantifying the impact of normalization and the limitations of string-distance metrics.

- Highlighting the superior performance of Whisper models and new Assembly AI services.

- Showcasing the value of speaker, recording, and utterance metadata availability for evaluation purposes

- Encouraging researchers and companies to showcase superior performance on a public benchmark.

5.6.4 Methodological gaps in ASR benchmarking addressed in this study

The benchmarks executed using BIGOS and PELCRA BIGOS datasets expanded the scope of Polish language model evaluations. Comparison to previous benchmarks is presented in table 5.2 and the respective figures below.

Number of test datasets :

- Previous: Minimum 1, Median 2.5, Maximum 8

- BIGOS and PELCRA: 12 datasets each

- Figure 5.2

Number of evaluated models:

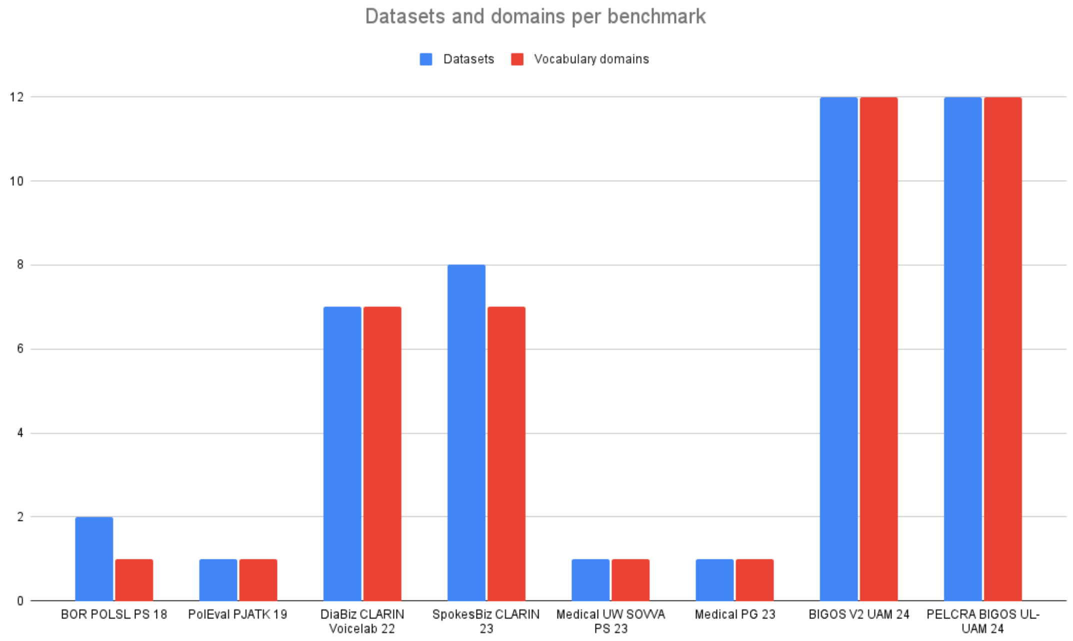


Figure 5.2: Number of datasets and vocabulary domains in Polish ASR benchmarks.

- Previous: Minimum 1, Median 3, Maximum 6
- New: 25 models each
- Figure: 5.3

Number of dataset-model combinations:

- Previous: Minimum 3, Median 6, Maximum 21
- New: 300 combinations each
- Figure: 5.4

This research broadened datasets, models, and their combinations, providing a more detailed evaluation of ASR systems for Polish compared to previous studies.

5.6.5 Limitations and future work

Utilizing more diverse and representative datasets

To better represent the Polish language, more dialects, sociolects, accents, and domains (e.g., formal speech, technical discussions, medical dictation) should be included. Future

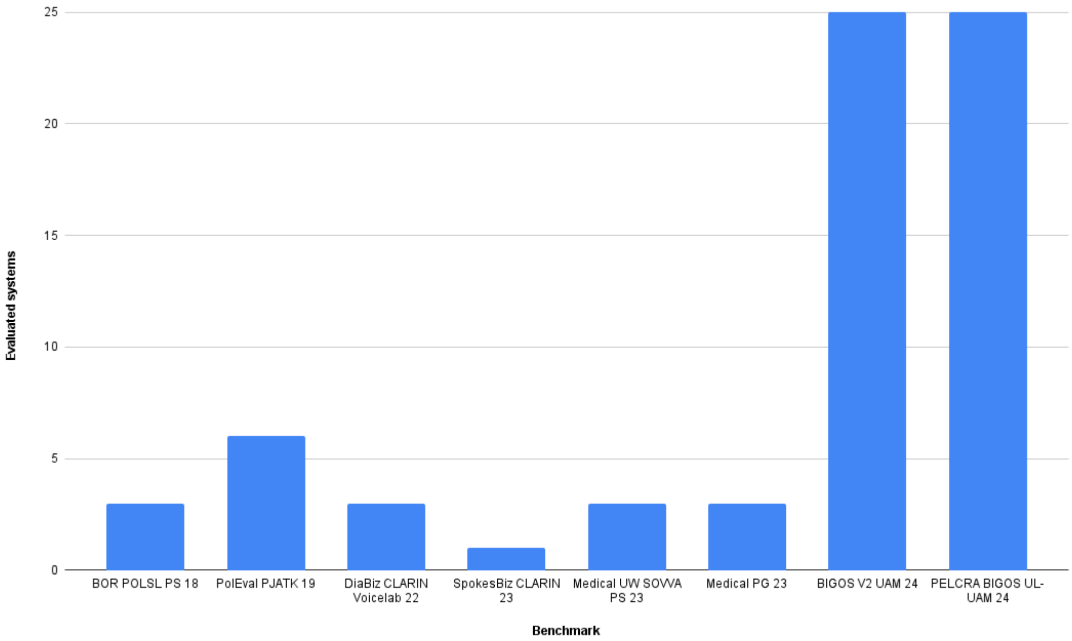


Figure 5.3: Number of evaluated models in Polish ASR benchmarks.

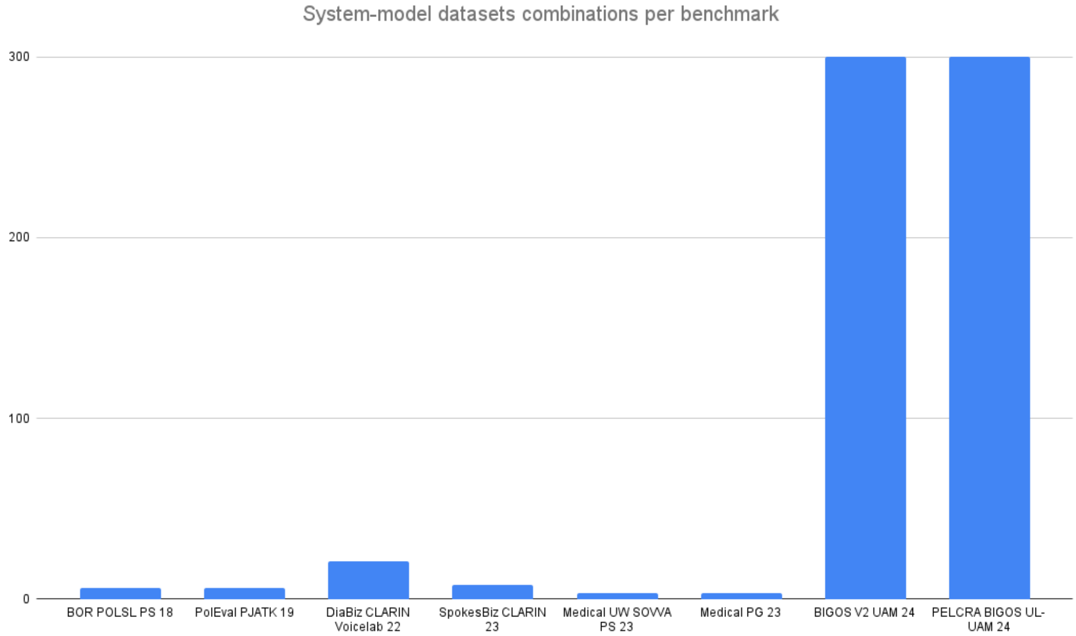


Figure 5.4: Number of dataset-system-model combinations in Polish ASR benchmarks

research should focus on gathering additional datasets from diverse regions, age groups, professions, and contexts. Evaluations can then use this metadata for analysis of ASR performance in various sociodemographic groups, as suggested by Aksenova et.al.[2]

Incorporating additional systems

The independent public leaderboard provides motivation for commercial ASR service providers to demonstrate the excellence of their ASR systems in practical applications. In addition to expanding the coverage of commercial systems, new freely accessible models should also be included.

Automatic quality assessment with language models

Considering the swift advancements in LLMs (Large Language Models)[5] and the effectiveness of embedding-based metrics to evaluate machine translation quality (COMET)[120], it should be valuable to investigate the incorporation of semantic metrics [125] or the automatic identification of invalid test samples. [148].

5.7 RO6: Organization of competition for the ASR community

At the time of writing this thesis, the competition has just started, and no results have been submitted yet. The competition is scheduled to end in September 2024, with the results to be discussed in October 2024. [6] Subsequently, the community-submitted results will be added to the AMU ASR Leaderboard[7]. As a result, participants and the public will be able to compare these with the updated results of commercial models and newly released open models.

[5] Polish LLM Leaderboard

[6] PolEval 2024 dates

[7] AMU ASR Leaderboard

Chapter 6

Conclusion

This study improved the assessment of Polish ASR systems by addressing data management gaps. The goal of improving the utility of public speech datasets for Polish ASR evaluation was achieved by curating a comprehensive survey, catalog, and benchmark dataset. By developing a benchmarking framework and using the curated dataset, this research has also provided the most comprehensive comparison of Polish ASR systems to date. The framework enables systematic evaluations and can be expanded to include new datasets, systems, scenarios, and analysis dimensions, effectively advancing Polish ASR evaluation towards state-of-the-art practices. Standardized data formats and processes enable the use of developed data management and evaluation frameworks in other languages.

6.1 Main Research Questions and Answers

- **RQ1**: How to systematically categorize Polish ASR speech datasets using public information? **Answer**: The datasets were systematically categorized using a taxonomy of 65 attributes, extracted from the original documentation, and manually annotated.

- **RQ2**: What is the current state of Polish ASR speech datasets? **Answer**: A comprehensive survey identified 53 distinct datasets, with 83% available through public domain resources, providing over 27,000 hours of speech data, including 6,000 hours of transcribed speech.

- **RQ3**: How can the survey findings be shared for community feedback? **Answer**: The findings were shared through publicly accessible digital repositories and platforms like GitHub and Hugging Face, allowing for direct user feedback.

- **RQ4**: What factors are crucial in designing and curating an ASR benchmark dataset? **Answer**: Crucial factors include task appropriateness, accessibility, discoverability, diversity, annotation, optimal size, cleanliness, and proper documentation.

- **RQ5**: What are the data curation steps required to create a benchmark dataset from publicly available speech datasets? **Answer**: Steps include selecting datasets, cleaning, normalizing, formatting, and organizing data into standardized formats with accessible public documentation and proper licensing.

- **RQ6**: Which public Polish speech datasets can be used as benchmarks? **Answer**: The datasets cataloged in the survey, comprising those with sufficient licensing. quality, and diversity, were selected for benchmarking purposes.

- **RQ7**: How can the curated dataset be shared with the community? **Answer**: The curated dataset was shared via platforms like Hugging Face, ensuring discoverability and accessibility under open licenses.

- **RQ8**: How to categorize Polish ASR benchmarks using public information? **Answer**: Polish ASR benchmarks were categorized using a literature-based taxonomy with 40 attributes on datasets, systems, tasks, and evaluation metrics.

- **RQ9**: What methods, datasets, and ASR systems have been used in Polish ASR benchmarks? **Answer**: Various methods, datasets, and systems are cataloged in the survey results. 2 out 6 benchmarks are reproducible.

- **RQ10**: Which Polish ASR systems have not been evaluated? **Answer**: The survey identified 20 free and open models not previously evaluated.

- **RQ11**: Which benchmarks evaluated commercial and free systems? **Answer**: Several benchmarks evaluated both commercial and free systems, comparing their performance.

- **RQ12**: Which ASR system performs best? **Answer**: Microsoft and Google led in older benchmarks, while newer systems like Whisper and Assembly AI now dominate.

- **RQ13**: What are the main conclusions from the ASR benchmarks? **Answer**: Performance varied significantly across systems, datasets, and speaker demographics, challenging assumptions of system superiority. Differences in similar benchmarks suggest a need for more consistent evaluation procedures.

- **RQ14**: How to share the survey results with the community? **Answer**: Through public repositories and dashboards on popular platforms, allowing community access and feedback.

- **RQ15**: What tools and systems exist for ASR benchmarking? **Answer**: Various tools for ASR benchmarking were identified, including those from Hugging Face and NVIDIA's NeMo toolkit, supporting multiple evaluation metrics and formats.

- **RQ16**: What challenges arise in evaluating multiple ASR systems, and what strategies can address them? **Answer**: Challenges include standardizing protocols and managing diverse datasets, addressed by a comprehensive benchmarking framework.

- **RQ17**: How can the system be extended to new ASR systems, datasets, languages, metrics, and normalization methods? **Answer**: Modular design and established tools enable extending benchmarks to new systems, datasets, languages, and metrics.

- **RQ18**: What is the ASR accuracy for different datasets? **Answer**: ASR accuracy varied significantly across different datasets, with performance assessed in various practical scenarios and datasets.

- **RQ19**: What is the accuracy gap between commercial and free systems? **Answer**: Initial assessments showed a gap, but newer free systems demonstrated competitive performance against commercial offerings. High efficiency of free NVidia Nemo models was discovered.

- **RQ20**: Does ASR accuracy vary with speech features? **Answer**: Yes, accuracy varied with features such as audio duration, speaking rate, and spontaneity of speech.

- **RQ21**: Is there an accuracy difference by age or gender? **Answer**: Yes, ASR systems showed accuracy differences by age and gender, with some systems favoring male speakers.

- **RQ22**: How to share evaluation results with the community? **Answer**: Via public leaderboards and popular platforms such as GitHub and Hugging Face Hub.

- **RQ23**: How to compare community solutions with state-of-the-art ASR systems? **Answer**: Organize community competitions and integrate results with public leaderboards for open and commercial systems.

6.2 Contributions and achievements

1. **Creation of speech data survey and catalog**: A catalog of 53 datasets was created. Datasets were categorized by 65 attributes extracted from original documentation and manually annotated. This survey and catalog assessed Polish ASR speech data and enabled selection of datasets for benchmarking.

2. **Curation of benchmark dataset:** A benchmark dataset was created from 24 openly available datasets. It includes audio samples from various sources of read and spontaneous speech. Analysis and integration of data sources were performed to ensure consistency and reliability. This involved automatic standardization of format and contents. The dataset is openly available and actively maintained. All curated subsets and splits comprises nearly 400,000 recordings and over 800 hours of transcribed speech.

3. **Survey of Polish ASR benchmarks:** A thorough survey of existing Polish ASR benchmarks was conducted, cataloging methods, datasets, and systems used in previous evaluations. This helped identify key gaps and informed the development of a more comprehensive benchmarking framework.

4. **Development of a benchmark framework:** The framework supports various datasets, systems, and metrics, ensuring consistent ASR evaluation with standardized protocols. It can be reused to replicate findings from this study or perform new benchmarks for other datasets or languages.

5. **Evaluation of ASR systems:** Using a curated dataset, 7 ASR systems and 25 models, both commercial and open-source, were compared. Significant variations across different systems, datasets, and speaker demographics were discovered. The superior

performance of Azure and Google systems reported previously was challenged by the improved results of newer systems like Whisper and Assembly AI.

6. **Open sharing of resources:** All datasets, tools, and evaluation results have been made openly available to the research community. This promotes transparency, reproducibility, and collaboration, enabling other researchers to build upon the work, either by developing ASR systems for Polish based on evaluation results or applying the framework to other languages.

7. **Organization of open challenge:** An open challenge was organized to engage the ASR community, encouraging the adoption of the curated benchmark dataset and facilitating a comparative evaluation of state-of-the-art ASR systems with community-developed solutions.

6.3 Future Directions

Future studies should focus on:

- **Existing dataset limitations:** Improving dataset quality, consistency, and coverage of curated datasets through new techniques such as automatic annotation and tools for streamlining dataset curation.

- **New data collection methods:** Applying innovative methods to create more diverse speech datasets.

- **Metrics expansion:** Utilizing accuracy metrics based on language models, rich-annotated datasets for advanced weighting, latency measurements and efficiency-related metrics.

- **Expanding community engagement:** Engaging through platforms like Hugging Face and open competitions for iterative improvements and feedback.

6.4 Research Impact

This research addressed several shortcomings in data management and ASR benchmarking methods.

Data utilization

- **Discoverability and accessibility**: A comprehensive speech data survey and catalog made Polish ASR speech datasets more discoverable and accessible (Research Objective 1).

- **Dataset utility**: Curated datasets BIGOS and PELCRA provided convenient access to 800 hours of speech and nearly 400,000 recordings from 5000 speakers. (Research Objective 2)

- **Dataset utilization**: The number of Polish ASR speech datasets utilized for ASR benchmarking purposes was increased 3 times compared to previous studies. (Research Objectives 2 and 5).

Data quality

- **Understanding of test data**: Organizing available documentation and analyzing the contents of curated datasets improved understanding of the test data used for evaluation (Research Objectives 1 and 2).

Evaluation reproducibility

- **Replication feasibility**: The sharing of evaluation datasets and tools improved the feasibility of replicating benchmarking results (Research Objectives 2 and 4).

- **Cross-study validation**: Publicly accessible benchmark datasets improved the feasibility of validating ASR research results between studies (Research Objective 2).

Evaluation scope

- **Ecological validity**: The creation of a data management framework positively impacted the feasibility of performing ecologically valid ASR evaluations (Research Objectives 2 and 4).

- **Performance understanding**: Conducting the largest evaluation of available models and commercial ASR systems improved understanding of ASR for Polish (Research Objective 5).

- **Benchmarking feasibility**: Developing a system for evaluations and competition improved the feasibility of comparing new systems, whether public, commercial, or community-contributed (Research Objectives 3, 4 and 6).

The author hopes that the developed research artifacts will benefit both academia and industry by advancing data curation and ASR benchmarking research for the Polish language and other languages.

Chapter 7

Appendix

7.1 ASR speech datasets survey

7.1.1 Attributes of speech datasets catalog

1. **Dataset name:** Full name of a speech dataset consisting of alphanumeric characters underscores and hyphens.

2. **Dataset ID:** Dataset's unique identifier for reporting composed of lowercase letters and hyphens.

3. **Access type:** Dataset access type from the cost perspective with possible values including free paid and no-info.

4. **Access link:** Web reference for accessing or purchasing a dataset provided in URL format.

5. **Available online:** Validated access status as of March 2023 with possible values of yes and no.

6. **License:** Dataset license type which can be Apache CC-0 CC-BY CC-BY-SA ELRA HZSK-PUB LDC or Proprietary.

7. **Publisher:** Creator or publisher of a dataset composed of alphabetical characters and hyphens.

8. **Repository:** Main repository hosting a dataset consisting of alphabetical characters and hyphens.

9. **Languages:** Language and country code of the recorded speakers represented as language code (ISO-639-1) and country code (ISO-3166-2) possibly including multiple languages.

10. **Creation year:** Year a dataset was created or published represented as a four-digit number.

11. **ISLRN:** International Standard Language Resource Number provided as ISRLN.

12. **ISBN:** International Standard Book Number provided as ISBN.

13. **LR catalog ID:** Language data repository ID represented as a combination of a URL or a string containing alphanumeric characters hyphens and underscores.

14. **Reference publication:** Link to a relevant publication describing a dataset provided in URL format.

15. **Contact point:** Contact point referenced in the documentation composed of alphanumeric characters hyphens underscores and the '@' symbol.

16. **Latest version:** The latest version of the dataset released expressed as a decimal number.

17. **Last update year:** Date (year) of the last update represented as a four-digit number.

18. **Sponsor:** Institution that funded the creation of the dataset consisting of alphanumeric characters hyphens and underscores.

19. **Price — non-commercial usage:** Price for noncommercial usage with possible values including free or a numerical value.

20. **Price — commercial usage:** Price for commercial usage with possible values including free or a numerical value.

21. **Purpose and split:** Target usage and available data splits with possible values being *train valid test* or none.

22. **Size audio total [hours]:** Total amount of audio data in hours represented as a decimal number.

23. **Size of audio transcribed [hours]:** Total amount of speech data transcribed expressed as a decimal number.

24. **Size [GB]:** Size of a dataset in gigabytes represented as a decimal number.

25. **Speakers:** Number of unique speakers who contribute speech recordings expressed as an integer.

26. **Audio recordings:** Number of voice recordings in the corpus represented as an integer.

27. **Audio segmentation:** Indicates whether audio recordings are segmented with possible values of yes or no.

28. **Tokens:** Number of tokens in the corpus represented as an integer.

29. **Unique tokens:** Number of unique tokens expressed as an integer.

30. **Automatic QA:** Type of automatic quality assurance process applied with possible values of yes or no.

31. **Manual QA:** Type of manual quality assurance process applied with possible values of yes or no.

32. **Manual QA scope:** Application of manual QA consisting of alphanumeric characters and spaces.

33. **Transcription coverage:** Proportion of transcribed recordings expressed as a percentage.

34. **Transcription protocol:** Specifies whether a transcription protocol is described with possible values being yes no or its description.

35. **Denormalized transcriptions:** Indicates whether available transcriptions are without abbreviations numerals punctuation etc. with possible values of yes or no.

36. **Transcription and annotation format:** Format of transcription files consisting of alphanumeric characters and periods.

37. **Domain:** Domain of utterances which can include academic lecture books broadcast conversations customer service digits general interview multi-domain news numbers parliament speech or public transport.

38. **Speech type:** Type of speech with possible values including conversational read public speech or various.

39. **Audio collection process:** Audio collection process with potential values controlled corpus or various.

40. **Speech recordings source:** Speech recordings source which can include volunteers university employees crowd public speakers or paid contributors.

41. **Acoustic environment:** Acoustic conditions under which audio was collected with possible values of broadcast car home mixed quiet space office public space or various.

42. **Audio device:** Audio devices used for speech collection such as a condenser mic headset mobile phone landline phone or various.

43. **Device model:** Recording device(s) and model(s) represented by a combination of alphanumeric characters and hyphens.

44. **Audio format:** Audio storage format with potential values including flac mp3 raw riff or wav.

45. **Audio codec:** Audio encoding format with possible values being mp3 ogg opus or vorbis.

46. **Audio channels:** Number of audio recording channels represented as an integer ranging from 1 to 16.

47. **Sampling rate [Hz]:** Sampling rate of recorded audio expressed as a four- or five-digit number.

48. **Bits per sample:** Number of bits used to encode each audio sample with possible values 8 16 24 or 32.

49. **Speaker info:** Anonymous information that recording originates from specific speaker.

50. **Age info:** Annotation of the age of the speakers with potential values of yes or no.

51. **Age balance:** Indicates whether the age distribution of the speakers is balanced between demographic groups with potential values of yes or no.

52. **Age distribution notes:** Information about the characteristics of age distribution represented as free text.

53. **Gender info:** Annotation of the gender of the speakers with potential values of yes or no.

54. **Gender balance:** Indicates whether the gender distribution of the speakers is balanced between demographic groups with potential values of yes or no.

55. **Gender distribution notes:** Information about the characteristics of gender distribution represented as free text.

56. **Nativity info:** Annotation of the nativity of the speakers with potential values of yes or no.

57. **Accent info:** Annotation of the accent of the speakers with potential values of yes or no.

58. **Accent representative:** Indicates whether the accent distribution of the speakers is representative of the target population groups with potential values of yes or no.

59. **Accent distribution notes**: Information about the characteristics of accent distribution represented as free text.

60. **Education info:** Annotation of the education of the speakers with potential values of yes or no.

61. **Occupation info:** Annotation of the occupation of the speakers with potential values of yes or no.

62. **Health info:** Annotation of the health condition of the speakers with potential values of yes or no.

63. **Speech signal time-alignment annotation:** Annotation of the duration of speech segments with potential values of yes or no.

64. **Speaker diarization annotation:** In the case of speech recordings containing speech of multiple speakers annotation of the duration of specific speaker speech segments with potential values of yes or no.

65. **Named entities annotation:** Annotation of named entities in utterances with potential values of yes or no.

66. **Part of speech annotation:** Annotation of part of speech information in utterances with potential values of yes or no.

7.1.2 Attributes of ASR benchmarks survey

1. **Benchmark:** Codename of the benchmark. Allowed values: Text.

2. **Date of the last update of the catalog entry:** When was the catalog entry updated? Allowed values: Date.

3. **Relevant publication:** Link to the publication with benchmark description. Allowed values: URL.

4. **Year:** What year the benchmark took place? Allowed values: Numeric.

5. **Systems-models evaluated:** What systems and models were evaluated? Allowed values: Text.

6. **Best system-model:** What was the best performing system-model variant? Allowed values: Categorical.

7. **Best system-model average WER:** What was the average WER for the best performing system-model variant? Allowed values: Percent.

8. **Major conclusion:** What are the major observations derived from the benchmark? Allowed values: Text.

9. **Benchmark limitations:** What are the benchmark methodological limitations? Allowed values: Text.

10. **Commercial systems:** Were commercial systems evaluated? Allowed values: yes, no.

11. **Freely available systems:** Were freely available systems evaluated? Allowed values: yes, no.
12. **Community provided:** Were community provided systems evaluated? Allowed values: yes, no.
13. **Replication recipe and resources:** Were the tools/scripts/data enabling benchmark replication made available to the public? Allowed values: yes, no.
14. **Evaluation dataset availability:** Was the evaluation dataset made available to the public? Allowed values: yes, no, partial(audio), partial(sample).
15. **Frequency:** Was benchmark performed systematically or only once? Allowed values: one-time, systematic.
16. **Automatic evaluation:** Was automatic evaluation used? Allowed values: yes, no.
17. **Human evaluation:** Was human evaluation used? Allowed values: yes, no.
18. **Lexical metrics:** What lexical metrics were used? Allowed values: Text.
19. **Language model based metrics:** What text embedding based metrics were used? Allowed values: Text.
20. **Annotation based metrics:** What annotation derived metrics were used? Allowed values: Text.
21. **Use cases:** What ASR use-cases are covered in the benchmark scope? Allowed values: Text.
22. **Socio-demographic analyses:** What analyses are performed in the socio-demographic dimensions? Allowed values: Text.
23. **Vocabulary domain:** What domains are represented in the evaluation data? Allowed values: Text.
24. **Speech types:** What types of speech are represented in the evaluation data? Allowed values: read, spontaneous.
25. **Audio sources:** What is the source of audio recordings? Allowed values: Text.

26. **Recording devices:** What recording devices were used to collect speech data? Allowed values: Text.

27. **Acoustic conditions:** What types of acoustic conditions are represented in the evaluation data? Allowed values: clean, noisy, mixed.

28. **Recordings annotations:** What recording-level annotations are available? Allowed values: Text.

29. **Speaker meta-data:** What speaker-level meta data is available? Allowed values: Text.

30. **Datasets:** How many datasets were used for evaluation? Allowed values: integer.

31. **Vocabulary domains:** How many vocabulary domains are represented in the evaluation dataset? Allowed values: integer.

32. **Recordings:** How many unique recordings are present in the dataset? Allowed values: integer.

33. **Speakers:** How many unique speakers are represented in the dataset? Allowed values: integer.

34. **Dataset size [hours]:** What is the total size of evaluation data? Allowed values: float.

35. **Systems-models evaluated:** How many system-model variants were evaluated? Allowed values: integer.

36. **Automatic metrics:** How many automatically calculated metrics were used? Allowed values: integer.

37. **Annotation metrics:** How many human annotation derived metrics were used? Allowed values: Numeric.

38. **Acoustic scenarios:** How many acoustic scenarios were considered? Allowed values: Numeric.

39. **Socio-demographic scenarios:** How many socio-demographic analysis variants were considered? Allowed values: Numeric.

40. **System-model dataset pairs:** How many system model dataset combinations were evaluated? Allowed values: Numeric.

7.1.3 Freely available speech datasets for Polish ASR

License	Dataset ID	Audio [hours]	Recordings	Speakers
CC-0	fair-voxpopuli-pl-21	111		282
CC-0	mozilla-comm-voice-20	148		3062
CC-BY	pjatk-clarin_mobile-15	13		
CC-BY	ul-pelcra_emo-18	26	40	80
CC-BY	clarin-radio-21		192	200
CC-BY	fair-mls-20	137		16
CC-BY	clarin-sejm_senat-18	97		516
CC-BY	pjatk-clarin_studio-15	56		
CC-BY	polyai-minds14-21	1	578	
CC-BY	pjatk-clarin_pinc-21	32		
CC-BY	google-fleurs-22	12.1	3937	
CC-BY	ul-pelcra_mmk-18	2	4	11
CC-BY	ul-pelcra_yt2-20	5	23	45
CC-BY	ul-pelcra_yt1-20	5	25	106
CC-BY	ul-pelcra_luz-18	20	21	42
CC-BY	ul-pelcra_mmw-18	7	14	65
CC-BY	ul-pelcra_snuv-12	220		210
CC-BY	ul-pelcra_emi-18	9	22	44
CC-BY	ul-pelcra_parl-15	12	48	251
CC-BY	ul-pelcra_mmw2-18	7	14	38
CC-BY-NC-ND	ul-pelcra_spokesbiz-23	650	925	590
CC-BY-SA	clarin-pjatk-cyfry-16	1	488	25
CC-BY-SA	pwr-azon-spontaneous-20	2	456	27
CC-BY-SA	pwr-azon_read-20	5		29
HZSK-PUB	hzsk-hamcopolig-11			
no info	ul-pelcra_plec-11			
no info	pwr-maleset-unk	6		
no info	pwr-viu-unk	1		
no info	pwr-shortwords-unk	1	939	
no info	pjatk-poleval-19	1	29	
Proprietary	mailabs-19	54		

Table 7.1: Publicly and freely available speech datasets for Polish.

7.1.4 Commercially available speech datasets for Polish ASR

License	Dataset ID	Audio [hours]	Recordings	Speakers
Proprietary	shaip-mobile-speech-21	1482		2049
Proprietary	clarin-diabiz-22	410		196
Proprietary	appen-mobile-unk	293		353
Proprietary	shaip-media-corpus-21	269		533
Proprietary	appen-speechdat-10	78		1000
Proprietary	clarin-diabiz-eval-22	41		146
Proprietary	appen-gphone-02	25		99
Proprietary	clarin-diabiz-sample-22	1	18	
LDC	ldc-polish-speech-db-19	280		200
LDC	ldc-clsu-pl-05	4		
ELRA	elra-speecon-pl-05	248		600
ELRA	elra-gphone-elra-02	25		99
ELRA	elra-babel-98	16		60

Table 7.2: Commercially available speech datasets for Polish.

7.1.5 Dataset subsets cards

pjatk-clarin_mobile-15

Metric	Train	Validation	Test	Total
Samples	2861	242	392	3495
Audio [hours]	10.3	0.83	1.35	12.48
Speakers	96	8	13	117
Words	74634	6286	10222	91142
Chars	507238	43079	69841	620158

Table 7.3: Dataset size per split — Clarin Mobile.

Metric	Train	Validation	Test
Unique utterances	2857	242	391
Unique words	23166	3465	5071
Unique chars	36	34	34
Words per second	2.01	2.09	2.1
Characters per second	11.66	12.26	12.27
Average audio duration [seconds]	12.97	12.4	12.4
Average utterance length [words]	26.09	25.98	26.08
Average utterance length [chars]	177.29	178.01	178.17
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.4: Dataset features per split — Clarin Mobile.

pjatk-clarin_studio-15

Metric	**Train**	**Validation**	**Test**	**Total**
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.5: Dataset size per split — Clarin Studio.

Metric	**Train**	**Validation**	**Test**
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.6: Dataset features per split — Clarin Studio.

fair-mls-20

Metric	**Train**	**Validation**	**Test**	**Total**
Samples	25042	511	519	26072
Audio [hours]	103.65	2.07	2.14	107.86
Speakers	16	4	4	24
Words	852851	16199	16996	886046
Chars	5425676	102012	111981	5639669

Table 7.7: Dataset size per split — MLS.

Metric	**Train**	**Validation**	**Test**
Unique utterances	25041	511	519
Unique words	86582	6727	7360
Unique chars	38	36	35
Words per second	2.29	2.17	2.21
Characters per second	12.26	11.5	12.33
Average audio duration [seconds]	14.9	14.61	14.84
Average utterance length [words]	34.06	31.7	32.75
Average utterance length [chars]	216.66	199.63	215.76
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.8: Dataset features per split — MLS.

mailabs-corpus_librivox-19

Metric	**Train**	**Validation**	**Test**	**Total**
Samples	11834	1527	1501	14862
Audio [hours]	25.61	3.34	3.19	32.14
Speakers	87	87	86	260
Words	201233	26210	25036	252479
Chars	1315543	171714	163415	1650672

Table 7.9: Dataset size per split — Munich AI Labs Librivox.

Metric	**Train**	**Validation**	**Test**
Unique utterances	11783	1526	1500
Unique words	44265	10606	10158
Unique chars	78	74	76
Words per second	2.18	2.18	2.18
Characters per second	12.09	12.08	12.06
Average audio duration [seconds]	7.79	7.89	7.65
Average utterance length [words]	17.0	17.16	16.68
Average utterance length [chars]	111.17	112.45	108.87
Meta coverage sex	100.0	100.0	100.0
Meta coverage age	N/A	N/A	N/A

Table 7.10: Dataset features per split — Munich AI Labs Librivox.

mozilla-common_voice_15-23

Metric	Train	Validation	Test	Total
Samples	19119	8895	8896	36910
Audio [hours]	27.95	12.55	12.5	53
Speakers	76	544	2300	2920
Words	166153	72502	66678	305333
Chars	1195204	502667	438631	2136502

Table 7.11: Dataset size per split — Common Voice.

Metric	Train	Validation	Test
Unique utterances	19119	8892	8886
Unique words	40615	25409	24823
Unique chars	83	81	83
Words per second	1.65	1.61	1.48
Characters per second	10.23	9.52	8.27
Average audio duration [seconds]	5.26	5.08	5.06
Average utterance length [words]	8.69	8.15	7.5
Average utterance length [chars]	62.51	56.51	49.31
Meta coverage sex	88.43	58.54	16.64
Meta coverage age	87.95	59.65	16.66

Table 7.12: Dataset features per split — Common Voice.

pwr-azon_read-20

Metric	Train	Validation	Test	Total
Samples	1820	382	586	2788
Audio [hours]	3.78	0.68	1.26	5.72
Speakers	19	4	6	29
Words	18131	3523	6113	27767
Chars	154286	30653	52222	237161

Table 7.13: Dataset size per split — AZON read.

Metric	Train	Validation	Test
Unique utterances	1237	353	479
Unique words	6465	2123	3019
Unique chars	33	33	33
Words per second	1.33	1.44	1.34
Characters per second	10.02	11.11	10.13
Average audio duration [seconds]	7.47	6.39	7.77
Average utterance length [words]	9.96	9.22	10.43
Average utterance length [chars]	84.77	80.24	89.12
Meta coverage sex	100.0	100.0	100.0
Meta coverage age	N/A	N/A	N/A

Table 7.14: Dataset features per split — AZON read.

pwr-azon_spont-20

Metric	Train	Validation	Test	Total
Samples	357	51	48	456
Audio [hours]	1.67	0.26	0.21	2.14
Speakers	23	2	2	27
Words	12984	2672	1598	17254
Chars	85731	16297	10493	112521

Table 7.15: Dataset size per split — AZON spontaneous.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.16: Dataset features per split — AZON spontaneous.

pwr-maleset-unk

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.17: Dataset size per split — PWR Maleset.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.18: Dataset features per split — PWR Maleset.

pwr-shortwords-unk

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.19: Dataset size per split — PWR Shortwords.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.20: Dataset features per split — PWR Shortwords

pwr-viu-unk

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.21: Dataset size per split — PWR Very Important Utterances.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.22: Dataset features per split — PWR Very Important Utterances.

google-fleurs-22

Metric	**Train**	**Validation**	**Test**	**Total**
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.23: Dataset size per split — Google FLEURS.

Metric	**Train**	**Validation**	**Test**
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.24: Dataset features per split — Google FLEURS.

polyai-minds14-21

Metric	**Train**	**Validation**	**Test**	**Total**
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.25: Dataset size per split — Minds-14.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.26: Dataset features per split — Minds-14.

ul-diabiz_poleval-22

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.27: Dataset size per split — PolEval 22.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.28: Dataset features per split — PolEval 22.

ul-spokes_mix_emo-18

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.29: Dataset size per split — Spokes Mix Emo.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.30: Dataset features per split — Spokes Mix Emo.

ul-spokes_mix_luz-18

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.31: Dataset size per split — Spokes Mix Luz.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.32: Dataset features per split — Spokes Mix Luz.

ul-spokes_mix_parl-18

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.33: Dataset size per split — Spokes Mix Parl.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.34: Dataset features per split — Spokes Mix Parl.

ul-spokes_biz_bio-23

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.35: Dataset size per split — Spokes Biz Bio.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.36: Dataset features per split — Spokes Biz Bio.

ul-spokes_biz_int-23

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.37: Dataset size per split — Spokes Biz Interviews.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.38: Dataset features per split — Spokes Biz Interviews.

ul-spokes_biz_luz-23

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.39: Dataset size per split — Spokes Biz Luz.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.40: Dataset features per split — Spokes Biz Luz.

ul-spokes_biz_pod-23

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.41: Dataset size per split — Spokes Biz Podcasts.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.42: Dataset features per split — Spokes Biz Podcasts.

ul-spokes_biz_pres-23

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.43: Dataset size per split — Spokes Biz Presentations

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.44: Dataset features per split — Spokes Biz Presentations

ul-spokes_biz_vc-23

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.45: Dataset size per split — Spokes Biz Various 1.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.46: Dataset features per split — Spokes Biz Various 1.

ul-spokes_biz_vc2-23

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.47: Dataset size per split — Spokes Biz Various 2.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.48: Dataset features per split — Spokes Biz Various 2.

ul-spokes_biz_wyw-23

Metric	Train	Validation	Test	Total
Samples	10999	1407	1404	13810
Audio [hours]	44.98	5.81	5.64	56.43
Speakers	440	57	56	553
Words	464421	58587	59832	582840
Chars	1864416	235157	240398	2339971

Table 7.49: Dataset size per split — Spokes Biz Interviews.

Metric	Train	Validation	Test
Unique utterances	10815	1402	1399
Unique words	50454	12343	12578
Unique chars	40	36	36
Words per second	2.87	2.8	2.95
Characters per second	8.65	8.44	8.9
Average audio duration [seconds]	14.72	14.88	14.45
Average utterance length [words]	42.22	41.64	42.62
Average utterance length [chars]	169.51	167.13	171.22
Meta coverage sex	N/A	N/A	N/A
Meta coverage age	N/A	N/A	N/A

Table 7.50: Dataset features per split — Spokes Biz Interviews.

7.1.6 Commercial ASR systems pricing

The cost as of May 15th, 2024 are provided in the table 7.51 The cost is provided for 1 hour of processed audio material. Pricing was retrieved from publicly available sources on 16 May 2024.

- Azure Speech Services Pricing
- Google Cloud Speech-to-Text Pricing
- OpenAI API Pricing
- AssemblyAI Pricing

System	Audio per month	Speed SLA	Audio log.	Unit cost
azure_latest	<5 hours	Yes	Yes	free
	>5 hours	Yes	Yes	$0.18
google_v2	<8333 hours	No	Yes	$0.18
	>8333 hours	Yes	Yes	$0.96
google_v1	<1 hours	Yes	N/A	free
	>1 hours	Yes	Yes	$0.96
	>1 hours	Yes	No	$1.44
whisper_cloud	N/A	Yes	Yes	$0.36
assembly_ai_default	<100 hours	Yes	Yes	Free
	>100 hours	Yes	Yes	$0.12
assembly_ai_best	>100 hours	Yes	Yes	$0.37

Table 7.51: Commercial ASR services pricing

7.1.7 Freely available ASR models sizes

Size of freely available models is presented in table 7.52

System-model variant	Parameters [million]
mms_1b-all	1000
mms_1b-fl102	1000
mms_1b-l1107	1000
nemo_fastconformer	118
nemo_multilang_fastconformer	114
nemo_quartznet	19
whisper_local_tiny	39
whisper_local_base	74
whisper_local_small	244
whisper_local_medium	769
whisper_local_large-v1	1550
whisper_local_large-v2	1550
whisper_local_large-v3	1550
wav2vec_xls-r-1b-polish	1000
wav2vec_large_xlsr	300

Table 7.52: Size of freely available ASR models

7.1.8 Call for participation in 2024 Polish ASR challenge

Polish Automatic Speech Recognition Challenge

Introduction

Automatic Speech Recognition (ASR) has made significant progress over the last decade. Improvements in deep learning and increased data availability have resulted in levels of accuracy for artificial speech transcription that are on par with human transcription, at least in specific domains, tasks, and speech characteristics. ASR technology has expanded to cover many new languages, use cases, user demographics, and devices. However, achieving robust speech recognition remains a challenge for many low-resource languages, specific speaker groups, application domains, and acoustic conditions.

To gauge the technological advances in Polish ASR technology, we are introducing the Open Challenge for Polish ASR. This initiative draws inspiration from the Multi-Domain End-to-End Speech Recognition Benchmark for the English language [1].

In order to promote multi-domain evaluation across a wide array of speech datasets, a new test dataset named BIGOS was introduced [2]. It comprises recordings from 12 open datasets and has been manually curated to ensure reliable evaluation results.

PELCRA benchmark dataset contains selected corpora from PELCRA repository [3]

(SpokesMix, SpokesBiz and Diabiz sample) in the BIGOS format. The author of curated PELCRA corpora hopes that standardized formatting and distribution via Hugging Face platform will simplify access and use of publicly available ASR speech datasets for Polish. PELCRA corpora significant contributions are spontaneous and conversational speech. Combined with BIGOS corpora, it enables the most comprehensive publicly available evaluation of Polish ASR systems in terms of number of speakers, devices, and acoustic conditions.

Task Definition

The goal of this challenge is to benchmark open Polish ASR systems against commercial services on a wide range of datasets.

Participants are provided with training, development, and test sets, from BIGOS and PELCRA corpora. Both datasets are available on Hugging Face [4, 5]. While scores for `test-A` will be visible from the beginning, final ranking will be based on systems' performance on `test-B` set and provided after the submissions are closed.

The participants are allowed to both create their own system, and fine-tune an existing solution. However, they are obligated to provide a relevant description for the submission.

For each audio recording, the system is supposed to generate a transcription of the utterance. Participants are forbidden to use data outside the provided training and validation sets to develop their systems. It is also prohibited to manually transcribe the test examples.

Dataset

The dataset is divided into four splits, and each one of them is stored in the corresponding directory. The files for each split have the same structure.

The `in.tsv` file is a tab-separated file with four columns:

1. `dataset` - name corresponding to the dataset available on Hugging Face,
   i.e. `amu-cai/pl-asr-bigos-v2` or `pelcra/pl-asr-pelcra-for-bigos`,

2. `subset` - name corresponding to the subset of the dataset, as on Hugging Face,

3. `split` - name corresponding to the split of the subset, as on Hugging Face,

4. `audioname` - name corresponding to the file id, as on Hugging Face.

Example of `in.tsv` file:

	BIGOS	PELCRA	Total
No. samples	82 025	229 150	311 175

	BIGOS	PELCRA	Total
No. samples	14 254	28 532	42 786

```
amu-cai/pl-asr-bigos-v2 fair-mls-20 train fair-mls-20-train-0022-00001

amu-cai/pl-asr-bigos-v2 fair-mls-20 train fair-mls-20-train-0022-00002
```

Although the text data are provided in TSV format, the audio files are available on Hugging Face platform [4, 5]. For `train` and `dev-0` splits, also `expected.tsv` files are provided. It is a tab-separated file with one column, where each row is a transcription for the matching audio recording from the `in.tsv` file.

Example of `expected.tsv` file: *szum mnoży w skałach okolicznych staje się rzeką a w gwałtownym pędzie pieni się huczy i zżyma w bałwany tym sroższy w biegu im dłużej wstrzymany lecą sandały i trepki i pasy wrzawa powszechna przeraża i głuszy zdrętwiał hyacynt na takie hałasy chciałby uniknąć bitwy z całej duszy a przeklinając nieszczęśliwe czasy resztę kaptura nasadził na uszy*

Training Data

The `train` set consists of 311 175 samples.

Development Data

The `dev-0` set consists of 42 786 samples.

Participants are allowed to use `dev-0` set to develop their systems.

Test Data

The `test-A` set consists of 20 284 samples and `test-B` - 20 285 samples.

Participants are forbidden to manually transcribe the test examples.

Downloading Datasets

The text data are provided in the TSV format and the audio files are available on Hugging Face [4, 5].

Evaluation

Submission Format

	BIGOS	PELCRA	Total
`test-A`	7 386	12 898	20 284
`test-B`	7 607	12 678	20 285
Total	14 993	25 576	40 569

The goal of the task is to generate an accurate transcription for each utterance. The submission should consist of a single tab-separated file with one column. Each line in the `out.tsv` file should contain hypothesis for the matching audio recording from the `in.tsv` file.

Example of `out.tsv` file:

szum mnoży w skałach okolicznych staje się rzeką a w gwałtownym pędzie pieni się huczy i zżyma w bałwany tym sroższy w biegu im dłużej wstrzymany lecą sandały i trepki i pasy wrzawa powszechna przeraża i głuszy zdrętwiał hyacynt na takie hałasy chciałby uniknąć bitwy z całej duszy a przeklinając nieszczęśliwe czasy resztę kaptura nasadził na uszy

Metrics

For each provided submission two measures of accuracy will be calculated:

Word Error Rate (WER) - number of incorrectly transcribed words divided by the total number of tokens in the reference sentences.

Character Error Rate (CER) - number of inccorectly transcribed characters divided by the total number of characters in the reference sentences.

Both metrics range from 0 to 1, where 0 is the best score.

Text Normalization

As some references do not contain punctuation or capitalization, evaluation is performed on normalized text to minimize the probability of false errors. All punctuation marks are removed and the case folding is applied.

Since normalization is performed during evaluation, there is no need for post-processing on the participant's side.

Baseline

The scores achieved by the baseline systems are available on the leaderboard. The scripts used to generate the results are also provided.

References

1. Sanchit Gandhi, Patrick von Platen, Alexander M. Rush. 2022. *ESB: A Benchmark For Multi-Domain End-to-End Speech Recognition.* [Paper] [Hugging Face]

2. Michał Junczyk. *BIGOS - Benchmark Intended Grouping of Open Speech Corpora for Polish Automatic Speech Recognition.* [Paper] [Hugging Face]

3. PELCRA Tools [Documentation]

4. amu-cai/pl-asr-bigos-v2 · Datasets at Hugging Face [Hugging Face]

5. pelcra/pl-asr-pelcra-for-bigos · Datasets at Hugging Face [Hugging Face]

Bibliography

[1] Alëna Aksënova, Zhehuai Chen, Chung-Cheng Chiu, Daan van Esch, Pavel Golik, Wei Han, Levi King, Bhuvana Ramabhadran, Andrew Rosenberg, Suzan Schwartz, and Gary Wang. Accented Speech Recognition: Benchmarking, Pre-training, and Diverse Data. *arXiv preprint*, 5 2022.

[2] Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. How Might We Create Better Benchmarks for Speech Recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. *arXiv preprint*, 12 2019.

[4] Łukasz Augustyniak, Kamil Tagowski, Albert Sawczyn, Denis Janiak, Roman Bartusiak, Adrian Szymczak, Marcin Wątroba, Arkadiusz Janz, Piotr Szymański, Mikołaj Morzy, Tomasz Kajdanowicz, and Maciej Piasecki. This is the way: designing and compiling LEPISZCZE, a comprehensive NLP benchmark for Polish. *arXiv preprint*, 11 2022.

[5] M. Bacchiani. Automatic transcription of voicemail at AT&amp;T. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 1, pages 25–28. IEEE, 2001.

[6] Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. A Toolbox for Construction and Analysis of Speech Datasets. *arXiv preprint*, 4 2021.

[7] Catherine Barrett, Patricia McCabe, Sarah Masso, and Jonathan Preston. Pro-

tocol for the Connected Speech Transcription of Children with Speech Disorders: An Example from Childhood Apraxia of Speech. *Folia Phoniatrica et Logopaedica*, 72(2):152–166, 2020.

[8] Emily M. Bender and Batya Friedman. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 12 2018.

[9] Kathrin Blagec, Jakob Kraiger, Wolfgang Frühwirt, and Matthias Samwald. Benchmark datasets driving artificial intelligence development fail to capture the needs of medical professionals. *Journal of Biomedical Informatics*, 137:104274, 1 2023.

[10] Łukasz Borchmann, Michal Pietruszka, Tomasz Stanislawek, Dawid Jurkiewicz, Michał Turski, Karolina Szyndler, and Filip Gralinski. DUE: End-to-End Document Understanding Benchmark. In *NeurIPS Datasets and Benchmarks*, 2021.

[11] Guillermo Cámbara, Alex Peiró-Lilja, Mireia Farrús, and Jordi Luque. English Accent Accuracy Analysis in a State-of-the-Art Automatic Speech Recognition System. *arXiv preprint*, 5 2021.

[12] William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network. 4 2021.

[13] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Yujun Wang, Zhao You, and Zhiyong Yan. GigaSpeech: An Evolving, Multi-domain ASR Corpus with 10,000 Hours of Transcribed Audio. *arXiv preprint*, 6 2021.

[14] Agnieszka Chmiel, Przemysław Janikowski, Marta Kajzer-Wietrzny, Danijel Koržinek, and Dariusz Jakubowski. EU Parliament Speech corpus, 2021.

[15] Christopher Cieri. Linguistic Resources, Development, and Evaluation of Text and Speech Systems. In *Evaluation of Text and Speech Systems*, pages 221–261. Springer Netherlands, Dordrecht, 2007.

[16] Christopher Cieri and Mark Liberman. More Data and Tools for More Languages and Research Areas: A Progress Report on LDC Activities. 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, 5 2006. European Language Resources Association (ELRA).

[17] Ann Clifton, Aasish Pappu, Sravana Reddy, Yongze Yu, Jussi Karlgren, Ben Carterette, and Rosie Jones. The Spotify Podcast Dataset. *arXiv preprint*, 4 2020.

[18] Michelle Cohn, Melina Sarian, Kristin Predeck, and Georgia Zellou. Individual Variation in Language Attitudes Toward Voice-AI: The Role of Listeners' Autistic-Like Traits. In *Interspeech 2020*, pages 1813–1817, ISCA, 10 2020. ISCA.

[19] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. *arXiv preprint*, 5 2022.

[20] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. Real Time Speech Enhancement in the Waveform Domain. *arXiv preprint*, 6 2020.

[21] Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Zelasko, and Miguel Jette. Earnings-21: A Practical Benchmark for ASR in the Wild. *arXiv preprint*, 4 2021.

[22] Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. Earnings-22: A Practical Benchmark for Accents in the Wild. *arXiv preprint*, 3 2022.

[23] Graŝyna Demenko, Stefan Grocholewski, Katarzyna Klessa, Jerzy Ogórkiewicz, Agnieszka Wagner, Marek Lange, Daniel Śledziński, and Natalia Cylwik. JURISDIC-Polish Speech Database for taking dictation of legal texts. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Ed. European Language Resources Association (ELRA)*, pages 1280–1287, 2008.

[24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet:

A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 6 2009.

[25] Lukas Drude, Jahn Heymann, Andreas Schwarz, and Jean-Marc Valin. Multi-channel Opus compression for far-field automatic speech recognition with a fixed bitrate budget. *arXiv preprint*, 6 2021.

[26] Jiayu Du, Jinpeng Li, Guoguo Chen, and Wei-Qiang Zhang. SpeechColab Leaderboard: An Open-Source Platform for Automatic Speech Recognition Evaluation. *arXiv preprint*, 3 2024.

[27] Peter D. Dueben, Martin G. Schultz, Matthew Chantry, David John Gagne, David Matthew Hall, and Amy McGovern. Challenges and Benchmark Datasets for Machine Learning in the Atmospheric Sciences: Definition, Status, and Outlook. *Artificial Intelligence for the Earth Systems*, 1(3), 7 2022.

[28] Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. ASR Errors Detection and Correction: A Review. *Procedia Computer Science*, 128:0, 2018.

[29] Lingyun Feng, Jianwei Yu, Deng Cai, Songxiang Liu, Haitao Zheng, and Yan Wang. ASR-GLUE: A New Multi-task Benchmark for ASR-Robust Natural Language Understanding. *arXiv preprint*, 8 2021.

[30] Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying Bias in Automatic Speech Recognition. *arXiv preprint*, 3 2021.

[31] Rita Frieske and Bertram E. Shi. Hallucinations in Neural Automatic Speech Recognition: Identifying Errors and Hallucinatory Models. *arXiv preprint*, 1 2024.

[32] Dennis Fucci, Marco Gaido, Matteo Negri, Mauro Cettolo, and Luisa Bentivogli. No Pitch Left Behind: Addressing Gender Unbalance in Automatic Speech Recognition through Pitch Manipulation. *arXiv preprint*, 10 2023.

[33] Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage. *arXiv preprint*, 11 2021.

[34] Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. ESB: A Benchmark For Multi-Domain End-to-End Speech Recognition. *arXiv preprint*, 10 2022.

[35] Gonçal V. Garcés Díaz-Munío, Joan-Albert Silvestre-Cerdà, Javier Jorge, Adrià Giménez Pastor, Javier Iranzo-Sánchez, Pau Baquero-Arnal, Nahuel Roselló, Alejandro Pérez-González-de Martos, Jorge Civera, Albert Sanchis, and Alfons Juan. Europarl-ASR: A Large Corpus of Parliamentary Debates for Streaming ASR Benchmarking and Speech Data Filtering/Verbatimization. In *Interspeech 2021*, pages 3695–3699, ISCA, 8 2021. ISCA.

[36] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. Datasheets for Datasets. *arXiv preprint*, 3 2018.

[37] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 3 2017.

[38] Daniela Gerz, Pei-Hao Su, Razvan Kusztos, Avishek Mondal, Michał Lis, Eshan Singhal, Nikola Mrkšić, Tsung-Hsien Wen, and Ivan Vulić. Multilingual and Cross-Lingual Intent Detection from Spoken Data. *arXiv preprint*, 4 2021.

[39] Calbert Graham and Nathan Roll. Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2), 2 2024.

[40] Stefan Grocholewski. Corpora - Speech Database for Polish Diphones. *EUROSPEECH '97 5th European Conference on Speech Communication and Technology*, 1997.

[41] Lauren Harrington. Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. *Frontiers in Communication*, 8, 7 2023.

[42] Staffan Hedström, David Erik Mollberg, Ragnheiur órhallsdóttir, and Jón Gunason. Samrómur: Crowd-sourcing large amounts of data. In Nicoletta Calzolari, Frédéric

Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2311–2316, Marseille, France, 6 2022. European Language Resources Association.

[43] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11096 LNAI(May):198–208, 5 2018.

[44] Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, Nishchal Bhandari, and Miguel Jette. Accented Speech Recognition: A Survey. *arXiv preprint*, 4 2021.

[45] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv preprint*, 5 2018.

[46] Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training. *arXiv preprint*, 4 2021.

[47] Shengshan Hu, Xingcan Shang, Zhan Qin, Minghui Li, Qian Wang, and Cong Wang. Adversarial Examples for Automatic Speech Recognition: Attacks and Countermeasures. *IEEE Communications Magazine*, 57(10):120–126, 10 2019.

[48] Jing Huang, B. Kingsbury, L. Mangu, Mukund Padmanabhan, George Saon, and Geoffrey Zweig. Recent improvements in speech recognition performance on large vocabulary conversational speech (voicemail and switchboard). In *6th International Conference on Spoken Language Processing (ICSLP 2000)*, pages 338–341, ISCA, 10 2000. ISCA.

[49] Xuedong Huang. Microsoft researchers achieve new conversational speech recognition milestone, 2017.

[50] Magdalena Igras-Cybulska, Bartosz Ziółko, Piotr Żelasko, and Marcin Witkowski. Structure of pauses in speech in the context of speaker verification and classification of speech type. *EURASIP Journal on Audio, Speech, and Music Processing*, 2016(1):18, 12 2016.

[51] Javier Iranzo-Sanchez, Joan Albert Silvestre-Cerda, Javier Jorge, Nahuel Rosello, Adria Gimenez, Albert Sanchis, Jorge Civera, and Alfons Juan. Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233. IEEE, 5 2020.

[52] Dorota Iskra, Beate Grosskopf, Krzysztof Marasek, Henk Van Den Heuvel, Frank Diehl, and Andreas Kiessling. SPEECON-Speech Databases for Consumer Devices: Database Specification and Validation. *LREC Proceedings*, 3rd LREC 2002, 2002.

[53] Michał Junczyk. Polish ASR Speech Datasets Catalog. https://github.com/goodmike31/pl-asr-speech-data-survey, 2023.

[54] Naoyuki Kanda, Guoli Ye, Yu Wu, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. Large-Scale Pre-Training of End-to-End Multi-Talker ASR for Meeting Transcription with Single Distant Microphone. *arXiv preprint*, 3 2021.

[55] Phillip Keung, Wei Niu, Yichao Lu, Julian Salazar, and Vikas Bhardwaj. Attentional Speech Recognition Models Misbehave on Out-of-domain Utterances. *arXiv preprint*, 2 2020.

[56] Suyoun Kim, Duc Le, Weiyi Zheng, Tarun Singh, Abhinav Arora, Xiaoyu Zhai, Christian Fuegen, Ozlem Kalinli, and Michael Seltzer. Evaluating User Perception of Speech Recognition System Quality with Semantic Distance Metric. In *Interspeech 2022*, pages 3978–3982, ISCA, 9 2022. ISCA.

[57] Andreas Kirkedal, Marija Stepanović, and Barbara Plank. FT Speech: Danish Parliament Speech Corpus. *Interspeech 2020*, pages 442–446, 5 2020.

[58] Seonmin Koo, Chanjun Park, Jinsung Kim, Jaehyung Seo, Sugyeong Eo, Hyeonseok Moon, and Heuiseok Lim. KEBAP: Korean Error Explainable Benchmark Dataset for ASR and Post-processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4798–4815, Stroudsburg, PA, USA, 2023. Association for Computational Linguistics.

[59] Danijel Koržinek. Task 5: Automatic speech recognition PolEval 2019 competition. Technical report, Polish-Japanese Academy of Information Technology, 2019.

[60] Piotr KOZIERSKI. Kaldi Toolkit in Polish Whispery Speech Recognition. *PRZEGLĄD ELEKTROTECHNICZNY*, 1(11):303–306, 11 2016.

[61] Piotr Kozierski, Talar Sadalla, Szymon Drgas, Adam Dąbrowski, Joanna Ziętkiewicz, and Wojciech Giernacki. Acoustic Model Training, using Kaldi, for Automatic Whispery Speech Recognition. In *Position Papers of the 2018 Federated Conference on Computer Science and Information Systems*, Annals of Computer Science and Information Systems, pages 109–114. Institute of Electrical and Electronics Engineers (IEEE), 2018.

[62] Jonáš Kratochvil, Peter Polák, and Ondřej Bojar. Large Corpus of Czech Parliament Plenary Hearings. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6363–6367, Marseille, France, 5 2020. European Language Resources Association.

[63] Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat,

Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 1 2022.

[64] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. QuartzNet: Deep Automatic Speech Recognition with 1D Time-Channel Separable Convolutions. *arXiv preprint*, 10 2019.

[65] Krzysztof Marasek, Danijel Korzinek, and Łukasz Brocki ˇ. System for Automatic Transcription of Sessions of the Polish Senate. *Archives of Acoustics*, 39(4), 2014.

[66] Jan Oldřich Krůza. Czech parliament meeting recordings as ASR training data. pages 185–188, 9 2020.

[67] Baybars Kulebi, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. ParlamentParla: A Speech Corpus of Catalan Parliamentary Sessions. In Darja Fišer, Maria Eskevich, Jakob Lenardič, and Franciska de Jong, editors, *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 125–130, Marseille, France, 6 2022. European Language Resources Association.

[68] Karolina Kuligowska, Maciej Stanusch, and Marek Koniew. Challenges of Automatic Speech Recognition for medical interviews - research for Polish language. *Procedia Computer Science*, 225:1134–1141, 2023.

[69] Ajinkya Kulkarni, Anna Tokareva, Rameez Qureshi, and Miguel Couceiro. The Balancing Act: Unmasking and Alleviating ASR Biases in Portuguese. *arXiv preprint*, 2 2024.

[70] Quentin Lhoest, Albert del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis,

Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, 11 2021. Association for Computational Linguistics.

[71] Hank Liao, Golan Pundak, Olivier Siohan, Melissa K Carroll, Noah Coccaro, Qi-Ming Jiang, Tara N Sainath, Andrew W Senior, Françoise Beaufays, and Michiel Bacchiani. Large vocabulary automatic speech recognition for children. In *Interspeech*, 2015.

[72] Thomas Liao, Rohan Taori, Deborah Raji, and Ludwig Schmidt. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1*, 2021.

[73] Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. Rethinking Evaluation in ASR: Are Our Models Robust Enough? In *Interspeech 2021*, pages 311–315, ISCA, 8 2021. ISCA.

[74] Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. Towards Measuring Fairness in Speech Recognition: Casual Conversations Dataset Transcriptions. *arXiv preprint*, 11 2021.

[75] Mingkuan Liu, Chi Zhang, Hua Xing, Chao Feng, Monchu Chen, Judith Bishop, and Grace Ngapo. Scalable Data Annotation Pipeline for High-Quality Large Speech Datasets Development. *arXiv preprint*, 9 2021.

[76] Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. ParlaSpeech-HR - a Freely Available ASR Dataset for Croatian Bootstrapped from the ParlaMint Corpus. In Darja Fišer, Maria Eskevich, Jakob Lenardič, and Fran-

ciska de Jong, editors, *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 111–116, Marseille, France, 6 2022. European Language Resources Association.

[77] Duncan Macho, Laurent Mauuary, Bernhard Noé, Yan Ming Cheng, Doug Ealey, Denis Jouvet, Holly Kelleher, David Pearce, and Fabien Saadoun. Evaluation of a noise-robust DSR front-end on Aurora databases. In *7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 17–20, ISCA, 9 2002. ISCA.

[78] Anirudh Mani, Shruti Palaskar, and Sandeep Konam. Towards Understanding ASR Error Correction for Medical Conversations. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 7–11, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

[79] Kate Margetson, Sharynne McLeod, Sarah Verdon, and Van H. Tran. Transcribing multilingual children's and adults' speech. *Clinical Linguistics & Phonetics*, 37(4-6):415–435, 6 2023.

[80] Nina Markl and Stephen Joseph McNulty. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. *arXiv preprint*, 2 2022.

[81] David Martínez, Eduardo Lleida, Phil Green, Heidi Christensen, Alfonso Ortega, and Antonio Miguel. Intelligibility Assessment and Speech Recognizer Word Accuracy Rate Prediction for Dysarthric Speakers in a Factor Analysis Subspace. *ACM Transactions on Accessible Computing*, 6(3):1–21, 6 2015.

[82] Shannon M. McCrocklin. Pronunciation learner autonomy: The potential of Automatic Speech Recognition. *System*, 57:25–42, 4 2016.

[83] Wes McKinney and others. pandas: a foundational Python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9, 2011.

[84] Valentin Mendelev, Tina Raissi, Guglielmo Camporese, and Manuel Giollo. Improved Robustness to Disfluencies in RNN-Transducer Based Speech Recognition. *arXiv preprint*, 12 2020.

[85] Péter Mihajlik, András Balog, Tekla Etelka Gráczi, Anna Kohári, Balázs Tarján, and Katalin Mády. BEA-Base: A Benchmark for ASR of Spontaneous Hungarian. *2022 Language Resources and Evaluation Conference, LREC 2022*, pages 1970–1977, 2 2022.

[86] Madison Milne-Ives, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando, and Edward Meinert. The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review. *Journal of Medical Internet Research*, 22(10):e20346, 10 2020.

[87] Adam S. Miner, Albert Haque, Jason A. Fries, Scott L. Fleming, Denise E. Wilfley, G. Terence Wilson, Arnold Milstein, Dan Jurafsky, Bruce A. Arnow, W. Stewart Agras, Li Fei-Fei, and Nigam H. Shah. Assessing the accuracy of automatic speech recognition for psychotherapy. *npj Digital Medicine*, 3(1):82, 6 2020.

[88] Ashish Mittal, Rudra Murthy, Vishwajeet Kumar, and Riyaz Bhat. Towards understanding and mitigating the hallucinations in NLP and Speech. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, pages 489–492, New York, NY, USA, 1 2024. ACM.

[89] Mehryar Mohri, Fernando Pereira, and Michael Riley. Speech Recognition with Weighted Finite-State Transducers. In *Springer Handbook of Speech Processing*, pages 559–584. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[90] Andrew Cameron Morris, Viktoria Maier, and Phil Green. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pages 2765–2768, ISCA, 10 2004. ISCA.

[91] Nathan Lambert. In defense of the open LLM leaderboard, 9 2023.

[92] Mikel K. Ngueajio and Gloria Washington. Hey ASR System! Why Aren't You More Inclusive? Automatic Speech Recognition Systems' Bias and Proposed Bias Mitigation Techniques. A Literature Review. pages 421–440, 11 2022.

[93] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv preprint*, 3 2021.

[94] Adam Nowakowski and Włodzimierz Kasprzak. Automatic speaker's age classification in the Common Voice database. pages 1087–1091, 9 2023.

[95] Maciej Ogrodniczuk, Piotr Pęzik, Marek Łaziński, and Marcin Miłkowski. Language Report Polish. In Rehm Georg and Way Andy, editors, *European Language Equality*, pages 191–194. Springer, 2023.

[96] Randal S. Olson, William La Cava, Patryk Orzechowski, Ryan J. Urbanowicz, and Jason H. Moore. PMLB: a large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, 10(1):36, 12 2017.

[97] Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. *arXiv preprint*, 4 2021.

[98] Douglas O'Shaughnessy. Trends and developments in automatic speech recognition research. *Computer Speech & Language*, 83:101538, 1 2024.

[99] Marcin Pacholczyk. Przegląd I porównanie rozwiązań rozpoznawania mowy pod kątem rozpoznawania zbioru komend głosowych. In Jolanta Krystek and Świerniak Andrzej, editors, *Automatyzacja procesów dyskretnych: Teoria i zastosowania*, pages 147–164. Politechnika Śląska, 2018.

[100] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 4 2015.

[101] Martha Maria Papadopoulou, Anna Zaretskaya, and Ruslan Mitkov. Benchmarking ASR Systems Based on Post-Editing Effort and Error Analysis. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 199–207, Held Online, 7 2021. INCOMA Ltd.

[102] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. Augmented

Datasheets for Speech Datasets and Ethical Decision-Making. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 881–904, New York, NY, USA, 6 2023. ACM.

[103] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pages 2613–2617, ISCA, 9 2019. ISCA.

[104] Hannaneh B. Pasandi and Haniyeh B. Pasandi. Evaluation of Automated Speech Recognition Systems for Conversational Speech: A Linguistic Perspective. *arXiv preprint*, 11 2022.

[105] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 11 2021.

[106] David Pearce and Hans-Günter Hirsch. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *6th International Conference on Spoken Language Processing (ICSLP 2000)*, pages 29–32, ISCA, 10 2000. ISCA.

[107] Mikel Penagarikano, Amparo Varona, Germán Bordel, and Luis Javier Rodriguez-Fuentes. Semisupervised Speech Data Extraction from Basque Parliament Sessions and Validation on Fully Bilingual Basque–Spanish ASR. *Applied Sciences*, 13(14):8492, 7 2023.

[108] Valeriia Perepelytsia and Volker Dellwo. Acoustic compression in Zoom audio does not compromise voice recognition performance. *Scientific Reports*, 13(1):18742, 10 2023.

[109] Piotr Pezik. Increasing the Accessibility of Time-Aligned Speech Corpora with Spokes Mix. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, 5 2018. European Language Resources Association (ELRA).

[110] Piotr Pęzik and Michał Adamczyk. Automatic Speech Recognition for Polish in 2022. Technical report, University of Łódź, Łódź, 4 2022.

[111] Piotr Pęzik, Sylwia Karasińska, Anna Cichosz, Łukasz Jałowiecki, Konrad Kaczyński, Małgorzata Krawentek, Karolina Walkusz, Paweł Wilk, Mariusz Kleć, Krzysztof Szklanny, and Szymon Marszałkowski. SpokesBiz – an Open Corpus of Conversational Polish. *arXiv preprint*, 12 2023.

[112] Piotr Pezik, Gosia Krawentek, Sylwia Karasińska, Paweł Wilk, Paulina Rybińska, Anna Cichosz, Angelika Peljak-Łapińska, Mikołaj Deckert, and Michał Adamczyk. DiaBiz – an Annotated Corpus of Polish Call Center Dialogs. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 723–726, Marseille, France, 6 2022. European Language Resources Association.

[113] Miroslaw Plaza, Lukasz Pawlik, and Stanislaw Deniziak. Call Transcription Methodology for Contact Center Systems. *IEEE Access*, 9:110975–110988, 2021.

[114] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling Speech Technology to 1,000+ Languages. *arXiv preprint*, 5 2023.

[115] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757–2761, ISCA, 10 2020. ISCA.

[116] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, New York, NY, USA, 6 2022. ACM.

[117] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and

Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint*, 12 2022.

[118] Francis McCann Ramirez, Luka Chkhetiani, Andrew Ehrenberg, Robert McHardy, Rami Botros, Yash Khare, Andrea Vanzo, Taufiquzzaman Peyash, Gabriel Oexle, Michael Liang, Ilya Sklyar, Enver Fakhan, Ahmed Etefy, Daniel McCrystal, Sam Flamini, Domenic Donato, and Takuya Yoshioka. Anatomy of Industrial Scale Multilingual ASR. *arXiv preprint*, 4 2024.

[119] Reece Randall, Yeonjung Hong, and Hosung Nam. The Effect of Real-time Score Feedback on L2 English Learners' Pronunciation and Motivation in an ASR-based CAPT System. *Korean Journal of Applied Linguistics*, 37(4):7–50, 12 2021.

[120] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. 9 2020.

[121] Joseph D Romano, Trang T Le, William La Cava, John T Gregg, Daniel J Goldberg, Praneel Chakraborty, Natasha L Ray, Daniel Himmelstein, Weixuan Fu, and Jason H Moore. PMLB v1.0: an open-source dataset collection for benchmarking machine learning methods. *Bioinformatics*, 38(3):878–880, 1 2022.

[122] Andrew Rosenberg. Rethinking the corpus: moving towards dynamic linguistic resources. In *Interspeech 2012*, pages 1392–1395, ISCA, 9 2012. ISCA.

[123] Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. Speech Recognition with Augmented Synthesized Speech. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 996–1002. IEEE, 12 2019.

[124] Jennifer Rowley and Frances Slack. Conducting a literature review. *Management Research News*, 27(6):31–39, 6 2004.

[125] Somnath Roy. Semantic-WER: A Unified Metric for the Evaluation of ASR Transcript for End Usability. *arXiv preprint*, 6 2021.

[126] Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. KLEJ: Comprehensive Benchmark for Polish Language Understanding. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 1191–1201, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

[127] Karpagavalli S. and Chandra E. A Review on Automatic Speech Recognition Architecture and Approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9:393–404, 2016.

[128] Thomas Schmidt. A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, (Issue 1), 6 2011.

[129] Sailik Sengupta, Jason Krone, and Saab Mansour. On the Robustness of Intent Classification and Slot Labeling in Goal-oriented Dialog Systems to Real-world Noise. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 68–79, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics.

[130] Muhammad A. Shah, David Solans Noguero, Mikko A. Heikkila, and Nicolas Kourtellis. Speech Robust Bench: A Robustness Benchmark For Speech Recognition. *arXiv preprint*, 3 2024.

[131] K. Shikano. Improvement of word recognition results by trigram model. In *ICASSP '87. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 12, pages 1261–1264. Institute of Electrical and Electronics Engineers, 1987.

[132] M.A. Siegler and R.M. Stern. On the effects of speech rate in large vocabulary speech recognition systems. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, pages 612–615. IEEE, 1995.

[133] Per Erik Solberg, Pierre Beauguitte, Per Egil Kummervold, and Freddy Wetjen. A Large Norwegian Dataset for Weak Supervision ASR. In Nikolai Ilinykh, Felix Morger, Dana Dannélls, Simon Dobnik, Beáta Megyesi, and Joakim Nivre, editors, *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 48–52, Tórshavn, the Faroe Islands, 5 2023. Association for Computational Linguistics.

[134] Matthias Sperber and Matthias Paulik. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are. *arXiv preprint*, 4 2020.

[135] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramı́rez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martı́nez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy

Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint*, 6 2022.

[136] Espen James Stokke. *Semantic Word Error Rate: A Metric Based on Semantic*

Distance. PhD thesis, The University of Bergen, 2023.

[137] Piotr Szymański, Piotr Żelasko, Mikolaj Morzy, Adrian Szymczak, Marzena Żyła-Hoppe, Joanna Banaszczak, Lukasz Augustyniak, Jan Mizgajski, and Yishay Carmiel. WER we are and WER we think we are. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3290–3295, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

[138] Edmondo Trentin and Marco Gori. A survey of hybrid ANN/HMM models for automatic speech recognition. *Neurocomputing*, 37(1-4):91–126, 4 2001.

[139] Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy T. Liu, Cheng-I Jeff Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB-SG: Enhanced Speech processing Universal PERformance Benchmark for Semantic and Generative Capabilities. *arXiv preprint*, 3 2022.

[140] Malgorzata Anna Ulasik, Manuela Hürlimann, Fabian Germann, Esin Gedik, Fernando Benites, and Mark Cieliebak. CEASR: A corpus for evaluating automatic speech recognition. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pages 6477–6485, 2020.

[141] S. Uma Maheswari, A. Shahina, and A. Nayeemulla Khan. Understanding Lombard speech: a review of compensation techniques towards improving speech based recognition systems. *Artificial Intelligence Review*, 54(4):2495–2523, 4 2021.

[142] Nahuel Unai, Roselló Beneitez, Albert Sanchis, Navarro Cotutor, and Jorge Civera Saiz. *Development and evaluation of a Polish Automatic Speech Recognition system using the TLK toolkit*. PhD thesis, Universitat Politècnica de València, Valencia, 2019.

[143] Anja Virkkunen, Aku Rouhe, Nhan Phan, and Mikko Kurimo. Finnish parliament ASR corpus. *Language Resources and Evaluation*, 57(4):1645–1670, 12 2023.

[144] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A stickier benchmark

for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[145] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[146] Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. *arXiv preprint*, 1 2021.

[147] Pete Warden. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. *arXiv preprint*, 4 2018.

[148] Johannes Wirth and Rene Peinl. Automatic Speech Recognition in German: A Detailed Error Analysis. In *2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS)*, pages 1–8. IEEE, 8 2022.

[149] Xinyu Yang, Weixin Liang, and James Zou. Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on Hugging Face. *arXiv preprint*, 1 2024.

[150] Fan Yu, Shiliang Zhang, Pengcheng Guo, Yihui Fu, Zhihao Du, Siqi Zheng, Weilong Huang, Lei Xie, Zheng-Hua Tan, DeLiang Wang, Yanmin Qian, Kong Aik Lee, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. Summary on the ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Grand Challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9156–9160, 2022.

[151] Piotr Żelasko, Sonal Joshi, Yiwen Shao, Jesus Villalba, Jan Trmal, Najim Dehak, and Sanjeev Khudanpur. Adversarial Attacks and Defenses for Speech Recognition Systems. *arXiv preprint*, 3 2021.

[152] Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang,

Quoc V. Le, and Yonghui Wu. Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition. *arXiv preprint*, 10 2020.

[153] Marta Zielonka, Wiktor Krasiński, Jakub Nowak, Przemysław Rośleń, Jan Stopiński, Mateusz Żak, Franciszek Górski, and Andrzej Czyżewski. A survey of automatic speech recognition deep models performance for Polish medical terms. In *2023 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 19–24. IEEE, 9 2023.

[154] Bartosz Ziółko, Tomasz Jadczyk, Dawid Skurzok, Piotr Żelasko, Jakub Gałka, Tomasz Pedzimaż, Ireneusz Gawlik, and Szymon Pałka. SARMATA 2.0 automatic Polish language speech recognition system. In *Proc. Interspeech 2015*, pages 1062–1063, 2015.