Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 585–590
DOI: 10.15439/2023F1609
ISSN 2300-5963 ACSIS, Vol. 35

BIGOS - Benchmark Intended Grouping of Open Speech Corpora for Polish Automatic Speech Recognition

Michał Junczyk
Adam Mickiewicz University
email: michal.junczyk@amu.edu.pl

***Abstract*—This paper presents a Benchmark Intended Grouping of Open Speech (BIGOS), a new corpus designed for Polish Automatic Speech Recognition (ASR) systems. This initial version of the benchmark leverages 1,900 audio recordings from 71 distinct speakers, sourced from 10 publicly available speech corpora. Three proprietary ASR systems and five open-source ASR systems were evaluated on a diverse set of recordings and the corresponding original transcriptions. Interestingly, it was found that the performance of the latest open-source models is on par with that of more established commercial services. Furthermore, a significant influence of the model size on system accuracy was observed, as well as a decrease in scenarios involving highly specialized or spontaneous speech. The challenges of using public datasets for ASR evaluation purposes and the limitations based on this inaugural benchmark are critically discussed, along with recommendations for future research. BIGOS corpus and associated tools that facilitate replication and customization of the benchmark are made publicly available.**

I. Introduction

AUTOMATIC Speech Recognition (ASR) is used in various applications and usage scenarios. Given that multiple aspects impact the difficulty of ASR tasks (vocabulary, acoustic conditions, speech type, etc.), the quality of target systems heavily depends on the effectiveness of the evaluation process. Benchmarking and evaluation ultimately aim to validate the system's ability to adapt to novel and unseen data.[1] To achieve this, multiple evaluation methods, datasets, and metrics are needed. The most commonly used metric for ASR evaluation is the Word Error Rate (WER), which quantifies word-level insertions, deletions, and substitutions between a system and reference transcriptions. WER has known limitations [2, 3]. When used on a narrow set of evaluation data, the assessment of the capabilities of the models, particularly in terms of generalization to unseen data, may be unclear.

Unlike English [4, 5, 6], German [7] and recently Hungarian [8], the Polish language lacks a common public-domain reference dataset for ASR benchmarking. Consequently, the results of Polish speech recognition studies are generally not directly comparable. Although transcribed recordings are available, it is often not practical to find or use all available public-domain datasets.

This study introduces BIGOS, a resource intended to enable systematic benchmarking and tracking of Polish ASR systems over time across a diverse range of publicly available corpora. The primary purpose of BIGOS is to alleviate the painstaking efforts required to discover and compile speech corpora from multiple sources. To ensure that original licenses are respected by BIGOS users, the corpus is distributed on the Hugging Face platform[1], which allows gated access. Alternatively, scripts for self-curation and customization of the dataset are also provided. [2] The first iteration of the benchmark presented in this work is performed using 1,900 utterances sourced from 10 corpora and 3 commercial ASR systems and 5 freely available.

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature, and Section 3 outlines the construction of the BIGOS benchmark and dataset, detailing the source speech corpora, corpus statistics, and ASR systems evaluated. Section 4 presents an exemplary application of BIGOS for the evaluation of ASR systems, Section 5 describes the limitations, and Section 6 concludes the paper by outlining the directions for future research.

II. Related work

A. ASR evaluation datasets

Prominent English-only datasets for ASR research and evaluation include the Wall Street Journal, VoxForge, Fisher, CHiME, LibriSpeech, TED-Lium, Common Voice, and Earnings. Wall Street Journal corpus covers news broadcast recordings, while SwitchBoard and Fisher include spontaneous telephone conversations. LibriSpeech [9] and MLS [10] feature narrated audiobooks, while VoxForge includes narrated Wikipedia articles. The TED-LIUM corpus [11] contains oratory educational talks, while the CHiME [12] dataset represents recordings of noisy environments in the real world. Earnings-21 and Earnings-22 contain conversational speech from earnings call recordings [4, 5]. The most voluminous dataset in terms of both the duration of speech content and language coverage is the MLS (Multilingual Librispeech), which contains 41,000 hours of material [10] for 8 languages. The Mozilla Common Voice dataset covers speech for more than 55 languages and boasts the largest number of contributing speakers, with over 10,000 as of March 2023 [13]. Both Common Voice and MLS include Polish language data. All of the aforementioned datasets offer a diverse range of speech sources, speaker demographics, and speech types,

[1] https://huggingface.co/datasets/michaljunczyk/pl-asr-bigos
[2] https://github.com/goodmike31/pl-asr-bigos-tools

providing researchers with valuable resources to investigate various aspects of ASR and to train new systems.

B. ASR benchmarks

The idea of using available speech datasets to benchmark the quality of ASR systems was first implemented nearly a decade ago. Gaida et al. [14] were the first to conduct a comprehensive evaluation of several open-source speech recognition tools. Dernoncourt developed a framework to evaluate seven ASR systems in two different collections and provided scripts to format Common Voice and LibriSpeech.[3] Moore et al. [15] introduced a meta-dataset containing reference text, hypotheses from two separate ASR systems, the Word Error Rate (WER), and annotations about speech intelligibility. Ulasik created a multilingual CEASR dataset for English and German[7], based on reference transcriptions from popular public-domain datasets and transcripts from four undisclosed ASR systems. Siegert et al. [16] performed a longitudinal study and found no significant changes in WER for 4 commercial systems over 8 months. Aksenova et al. [1] conducted a comprehensive survey on existing ASR benchmarking methodologies and proposed a systematic benchmarking framework for the most common use cases. Xu et al.[17] compared 4 commercial ASR services with respect to robustness to acoustic background noise. Varod et al. highlighted that ASR performance is language and system specific and that low-resource languages such as Hebrew can have a performance comparable to high-resource languages such as German.[18] The ASR4REAL benchmark [19] revealed significant accuracy variations depending on the accent of the speaker and socioeconomic status. Papadopoulou evaluated four commercial ASR systems in the context of translation post-editing effort [20]. The challenges associated with the recognition of spontaneous and accented speech were further analyzed in the benchmarks organized by the Rev and Google companies. [4, 5, 21]. Pasandi et al. highlighted that conversational speech is the most challenging and environmentally relevant type of data for speech recognition. Pires et al. constructed the Portuguese Evaluation Benchmark[22] using the Mozilla Common Voice and Voxforge datasets and five commercial ASR engines. Mihajlik et al. conducted an evaluation of open-source Hungarian ASR systems using a comprehensive linguistic dataset [8]. Extending the studies by Ulasik et al. for English and German, Wirth et al. [3] questioned the prevailing statistical ASR evaluation paradigm by performing a manual recognition error assessment. Of paramount importance, the study identified that 18% of the ASR errors originated from flawed ground-truth transcriptions and another 18% from flawed or ambiguous audio within publicly accessible datasets.

C. Polish ASR benchmarks

The first evaluation of commercial ASR systems for the Polish language was carried out in 2018 [23]. The first open benchmark for ASR systems was organized by Korzinek [24]. In 2019, Unai et al. [25] evaluated a self-developed Polish ASR system using 223 hours of speech collected from six datasets, including the Clarin-PL Studio Corpus (EMU)[26], the PELCRA family of corpora [27, 28], the Polish Senate recordings corpus [29], the Simple4All Tundra Corpus, and the test results for the PolEval 2019 competition [24]. The most extensive benchmark to date is Diabiz *Diabiz* performed using a set of 400 dialogs in eight domains and three commercial ASR systems. [30, 31].

III. BIGOS CORPUS DESIGN AND CURATION

As indicated by the Polish ASR Speech Data catalog [4] as of March 2023, approximately 5300 hours of speech in 51 datasets are available for Polish ASR development. Roughly 1000 hours of transcribed speech spread across 13 datasets is freely accessible under permissive licenses, facilitating the curation of a new evaluation dataset detailed in the following section.

A. BIGOS corpus overview

Table III-A summarizes the properties of the BIGOS dataset.

Table I
BIGOS DATASET PROPERTIES

Attribute	Value
Datasets sourced	10
Speech material (hours)	4.5
Test cases total	1900
Speakers	71

B. Sourcing and pre-analysis

Polish ASR Speech Data Catalog was used to identify suitable datasets to be included in the benchmark. The following mandatory criteria were considered:

- Dataset must be downloadable.
- The license must allow for free, noncommercial use.
- Transcriptions must be available and align with the recordings.
- The sampling rate of audio recordings must be at least 8 kHz.
- Audio encoding using a minimum of 16 bits per sample.

The following is an overview of 10 datasets that meet the criteria and were chosen as sources for the BIGOS dataset.

- The Common Voice dataset *(mozilla-common-voice-19)*, developed by Mozilla, is an open source multilingual resource [13]. This project aims to democratize voice technology by providing a wide-ranging, freely available dataset that covers many 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

[3]https://github.com/Franck-Dernoncourt

[4]https://github.com/goodmike31/pl-asr-speech-data-survey

three months under a permissive Creative Commons 0 license.

- The Multilingual LibriSpeech (MLS) dataset *(fair-mls-20)* is a large, multilingual corpus created for speech research by Facebook AI Research (FAIR)[10]. This dataset is derived from audiobooks from LibriVox and covers eight languages, including about 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. Humans have evaluated the transcriptions in the test sets.
- The Clarin Studio Corpus *(clarin-pjatk-studio-15)* is provided by CLARIN-PL, a subsection of CLARIN 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 Corpus *(clarin-pjatk-mobile-15)* is a Polish speech corpus of read speech recorded over 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.
- The Jerzy Sas PWR datasets (Politechnika Wrocławska) *(pwr-viu-unk, pwr-shortwords-unk, pwr-maleset-unk)*. According to the documentation available online[5] speech was collected using a variety of microphones and in relatively noise-free acoustic conditions. Three datasets are available: short words, very important utterance (VIU), and male AM set.
- The M-AI Labs Speech corpus *(mailabs-19)*, similar 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 mission of *"enabling (European) companies to take advantage of AI & ML without having to give up control or know-how."* [6] 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. The training data consists of nearly a thousand hours of audio for all languages, including 53.5 hours for Polish.
- The AZON Read and Spontaneous Speech Corpora [7] *(pwr-azon-spont-20, pwr-azon-read-20)* 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 5 and 2 hours of transcribed speech, respectively. The AZON database is available under a CC-BY-SA license.

Two additional corpora, the Spelling and Numbers Voice database (SNUV) from the University of Łódź's PELCRA group and the CLARIN Cyfry corpus, initially met the necessary requirements for this study. However, their unique transcription conventions led to high error rates during initial tests. For example, the word *"pstrąg"* in *SNUV* corpus is transcribed as *"py sy ty ry ą gy"*. The conventional normalization employed by most ASR systems is *"p s t r ą g"*. In the case of *Cyfry* corpus, only numeric expressions are transcribed, hence high error rates are produced for correctly recognized nonnumeric expressions. As such, these corpora will be included in the next iteration of the benchmark, following a thorough manual retranscription process to mitigate these issues.

C. Curation and selection

Necessary preprocessing parameters were consolidated into specific configuration files for each dataset, including download links, metadata fields to be extracted, etc. Subsequently, the text data and audio were extracted and encoded in a unified format. Dataset-specific transcription norms are preserved, including punctuation and casing. To strike a balance in the evaluation dataset and to facilitate the comparison of Word Error Rate (WER) scores across multiple datasets, 200 samples are randomly selected from each corpus. The only exception is 'pwr-azon-spont-20', which contains significantly longer recordings and utterances, therefore only 100 samples are selected. Finally, the first version of the BIGOS corpus contains 1900 recordings of the 115,915 available in the 10 datasets (1. 64% of the total available transcribed speech). The table II provides detailed information on the composition of the BIGOS 1.0 corpus.

Table II
NUMBER OF RECORDINGS AND AVERAGE DURATIONS

Dataset	Size[h]	Files	Average length[s]
fair-mls-20	0.81	200	14.52 ±2.82
clarin-pjatk-mobile-15	0.72	200	13.05 ±3.51
pwr-azon-spont-20	0.72	100	25.75 ±7.12
clarin-pjatk-studio-15	0.56	200	10.10 ±4.32
pwr-azon-read-20	0.48	200	8.72 ±1.95
mailabs-19	0.42	200	7.60 ±3.10
mozilla-common-voice-19	0.27	200	4.89 ±1.50
pwr-shortwords-unk	0.24	200	4.41 ±1.42
pwr-maleset-unk	0.19	200	3.44 ±0.44
pwr-viu-unk	0.08	200	1.49 ±0.22
Total	4.49	1900	-
Average	-	-	9.39 ±2.64

D. Preprocessing and format standardization

The following curation methods were applied to the baseline version of the BIGOS dataset:

- validation of audio file availability and validity,
- unification of audio format to WAV 16 bits/16 kHz,
- normalization of audio amplitude to -3 dBFS,
- unification of text encoding to UTF8,

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

[6]https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/

[7]https://zasobynauki.pl/zasoby/korpus-nagran-probek-mowy-do-celow-budowy-modeli-akustycznych-dla-automatycznego-rozpoznawania-mowy,53293/

Table III
ATTRIBUTES IN THE BIGOS UTTERANCE DATA OBJECT

Attribute	Description
id_file_pproc	Standardized file identifier
id_file_source	Original file identifier
id_dataset_source	Source dataset identifier
subset_source	Subset in source dataset (train, test, valid)
path_audio_source	Path to original audio file
path_trans_source	Path to original transcription file
path_audio_pproc	Path to audio file after standardization
meta_spkid_source	Original speaker identifier
meta_spkid_pproc	Standardized speaker identifier
meta_spk_age_source	Speaker age info from source
ref_original	Original transcription (reference)
hyp_whisper_cloud	Hypothesis of Whisper cloud service
hyp_google_default	Hypothesis of Google cloud service default model
hyp_azure_default	Hypothesis of Azure cloud service default model
hyp_whisper_tiny	Hypothesis of Whisper local tiny model
hyp_whisper_base	Hypothesis of Whisper local base model
hyp_whisper_small	Hypothesis of Whisper local small model
hyp_whisper_medium	Hypothesis of Whisper local medium model
hyp_whisper_large	Hypothesis of Whisper local large model

- extraction of original transcription,
- removal of redundant characters
- extraction and unification of metadata.

E. Validation and ASR transcripts generation

Upon completing the preprocessing of the entire dataset, the number of obtained recordings, transcriptions, and metadata records in the compiled dataset were checked for consistency. If the validation was successful, the ASR hypotheses for the locally hosted Whisper models were generated. ASR transcriptions for cloud services like Google, Azure, and Whisper were obtained via respective APIs. Table III presents the object of the resulting BIGOS utterance data.

IV. ASR SYSTEMS EVALUATION

A. Evaluated ASR systems

Below is an overview of the ASR systems evaluated in the first iteration of the BIGOS benchmark.

- Google Cloud Speech-to-Text [8] supports more than 125 languages and variants. The "default" model from May 2023 was used for this benchmark.
- Microsoft's Azure Speech Service [9] as of May 2023 supports more than 100 languages and variants. The "default" model from May 2023 was used for this benchmark.
- Whisper 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 [32]. The web-hosted model available via API and the locally hosted models from May 2023 were used for this benchmark.[10]

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

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

[10]https://github.com/openai/whisper/blob/main/model-card.md

B. Metrics

ASR systems predictions were evaluated against the target transcriptions using 3 industry-standard metrics:

- Sentence Error Rate (SER) calculates the proportion of sentences that are not perfectly recognized, i.e., sentences that contain at least one error.
- Word Error Rate (WER) 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.
- Character Error Rate (CER) metric calculates the minimum number of character-level operations (substitutions, insertions, and deletions) needed to change the system's output into the reference transcript, divided by the total number of characters in the reference.

V. BENCHMARK RESULTS

This section provides an overview and analysis of the results obtained.

A. Quality per system and model type

The performance of various systems was evaluated using average SER, WER, and CER values obtained from ten test datasets available in BIGOS. The "large" model of the Whisper system achieved the highest accuracy, outperforming all other systems in every metric. The "medium" model of the Whisper system came second, and the "cloud" model of the same system came third. Google and Azure's services followed these, with the remaining Whisper models trailing behind.

Interestingly, the two most accurate systems are both freely available. Despite using the same "large-v2" model, the cloud-based variant was outperformed by the locally hosted "large" variant and, even more surprisingly, by the "medium" variant, which theoretically should be less advanced. On average, free systems outperformed well-established paid services.

To understand why this is the case, a more detailed and manual examination of the evaluation results is required. However, it is crucial to note that lower scores in this evaluation do not necessarily indicate inferior performance in real-world scenarios.

One hypothesis is that commercial systems, despite their ability to handle advanced normalization conventions, might actually perform worse when evaluated on publicly available datasets that use written forms of numerals (e.g., "one", "six o'clock") instead of numeric forms (e.g., "1", "6:00"). This paradox suggests that the use of automated evaluation metrics and publicly available datasets used "as-is" (without transcription unification) may not fully represent real-world performance and capabilities.

Tables IV present the average SER, WER, and CER scores for the Azure, Google, and Whisper systems.

B. Quality per dataset

The best overall performance was observed for the PWR corpora, which contain recordings from a single speaker in

Table IV
AVERAGE SER, WER AND CER OF EVALUATED POLISH ASR SYSTEMS

Service	System	Variant	SER	WER	CER
paid	Azure	default	64.3 ±39.2	18.3 ±12.6	10.2 ±9.3
paid	Google	default	59.9 ±38.5	16.1 ±10.0	7.4 ±5.7
paid	Whisper	large	58.7 ±34.6	11.0 ±8.0	4.1 ±3.9
free	Whisper	tiny	90.3 ±14.6	46.8 ±9.4	14.7 ±6.1
free	Whisper	base	83.7 ±19.9	32.9 ±9.9	10.1 ±4.7
free	Whisper	small	67.6 ±28.2	16.5 ±6.5	5.5 ±3.0
free	Whisper	medium	55.4 ±34.1	9.3 ±4.6	3.7 ±2.6
free	**Whisper**	**large**	**50.1 ±34.4**	**7.6 ±3.8**	**3.1 ±2.4**

Table V
AVERAGE WER PER DATASET FOR SELECTED SYSTEMS

	Paid			Free	
Dataset	**A**	**G**	**W**	**W**	**WER avg**
pwr-maleset	6.6	**3.2**	6.1	**3.2**	4.8±1.8
pwr-shortwords	7.1	**4.4**	7.8	4.8	6.0±1.7
pwr-viu	0.3	24.4	**0.0**	7.9	8.1±11.4
common-voice-19	**10.2**	19.9	11.2	10.3	12.9±4.7
mailabs-19	19.5	19.6	**8.4**	8.5	14.0±6.4
mls-20	30.0	22.9	5.9	**4.6**	15.8±12.6
azon-read-20	35.9	4.1	23.4	**3.8**	16.8±15.7
pjatk-mobile-15	26.9	30.7	11.8	**10.7**	20.0±10.3
azon-spont-20	28.2	15.7	24.2	**14.3**	20.6±6.7
WER average	18.3±12.6	16.1±10.0	11.0±8.0	**7.6±3.8**	13.2 ±4.9

a quiet acoustic environment. This limited variability led to perfect performance for the Whisper Cloud and Azure systems in *PWR VIUa* and the best average WER for the *Male* set. Interestingly, for single-word utterances, the limited context led the Google and Whisper local systems to recognize foreign language words instead of Polish words. For example, the word 'zapisz' was recognized as a Russian word 'Запись', the word 'zakończ' as the English word 'the coins', and words 'małe litery' and 'duże litery' as the Italian words 'ma vedi tere' and 'due lettere', respectively. The PWR male set dataset had the second-best performance. A median WER of 6% suggests that modern Polish ASR systems handle short utterances from contemporary literature quite effectively.

Slightly worse performance (average WER over 10% for all systems) was observed for the *MLS, M-AI Labs*, and *Common Voice* datasets. Given the widespread use and accessibility of the *MLS* and *Common Voice* datasets within the global ASR community, it is likely that these datasets were used during training, allowing all systems to efficiently handle in-domain recordings and transcriptions. This hypothesis is supported by the performance of Whisper systems family on the *MLS* corpus; however, Google's performance on the *Common Voice* dataset was nearly twice as bad as other systems. Given that Whisper is trained mostly on publicly available data, while commercial systems leverage proprietary datasets, the impact of training and evaluation data leakage is more significant in the case of Whisper.

Performance for the CLARIN mobile dataset was slightly inferior, possibly due to longer utterances and the use of commercial *default* models, which are not optimized to handle speech recorded with an 8 kHz sampling frequency.

As expected, performance declined for the AZON read and spontaneous corpora, which contain scientific vocabulary from the chemistry field. However, the Google and Whisper local systems handled both types of AZON corpora proficiently, despite containing fillers and hesitations.

Table V-B presents the median WER for specific datasets sourced in BIGOS for Azure, Google, Whisper Cloud and Large systems.

VI. LIMITATIONS

The initial version of the benchmark comes with several limitations. First, the quantity and specificity of the datasets, along with the metadata about speakers and acoustic conditions, are limited. To examine ASR performance for particular sociodemographic groups, such as non-native Polish speakers or specific types of speech, such as whispery speech, dedicated datasets[33] should be used. Second, the unification of normalization relies solely on automatic methods and does not involve manual re-transcription. Lastly, the initial evaluation uses a limited number of test recordings, systems, and models, which constrains the precision and breadth of the benchmark.

VII. CONCLUSION AND FUTURE WORK

This work addresses the lack of a publicly available ASR evaluation suite for Polish by providing BIGOS, Benchmark Intended Grouping of Open Speech corpora. BIGOS, as its name suggests, was compiled from 10 existing publicly accessible Polish speech corpora. A test sample comprising 1900 recordings from 71 distinct speakers was used to gauge the performance of 3 commercial ASR systems against 5 freely available ones. Through automatic evaluation metrics, it was discovered that Whisper Cloud consistently outperforms more established services from Google and Azure on the test set representing publicly available speech datasets for Polish. Interestingly, the largest and second largest of the Whisper models exhibit superior performance compared to its paid version. The BIGOS corpus[11] and tools[12] for corpus curation and evaluation of ASR systems are available to the community, allowing reproduction and extension of this benchmark.

As indicated in the Limitations and Related Work sections, there are many interesting research directions to explore. The primary objective of the next BIGOS iteration is to include a subset of manually verified reference transcriptions. Comparison of error rates, calculated using original and manually verified transcriptions, will reveal the evaluation bias resulting from differences in normalization standards in various public-domain corpora. Furthermore, the reliability and informativeness of the evaluation could be significantly improved if the evaluation results were manually annotated, similar to the German study [3], which revealed that the evaluation errors may be caused by the poor quality of the evaluation data and that not all errors are of equal importance. Lastly, it will be interesting to measure the robustness of the systems using larger samples, new data sources, and automatically perturbed recordings.

[11] https://huggingface.co/datasets/michaljunczyk/pl-asr-bigos
[12] https://github.com/goodmike31/pl-asr-bigos-tools

REFERENCES

[1] Alëna Aksënova et al. "How Might We Create Better Benchmarks for Speech Recognition?" In: Association for Computational Linguistics, 2021, pp. 22–34. DOI: 10.18653/v1/2021.bppf-1.4.

[2] Piotr Szymański et al. "WER we are and WER we think we are". In: Association for Computational Linguistics, 2020, pp. 3290–3295. DOI: 10.18653/v1/2020.findings-emnlp.295.

[3] Johannes Wirth and Rene Peinl. "ASR in German: A Detailed Error Analysis". In: (2022). DOI: 10.48550/arXiv.2204.05617.

[4] Miguel Del Rio et al. "Earnings-21: A Practical Benchmark for ASR in the Wild". In: (2021).

[5] Miguel Del Rio et al. "Earnings-22: A Practical Benchmark for Accents in the Wild". In: (Mar. 2022). DOI: 10.48550/arXiv.2203.15591.

[6] Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. "ESC: A Benchmark For Multi-Domain End-to-End Speech Recognition". In: (Oct. 2022). DOI: 10.48550/arXiv.2210.13352.

[7] Malgorzata Anna Ulasik et al. "CEASR: A corpus for evaluating automatic speech recognition". In: 2020, pp. 6477–6485.

[8] Péter Mihajlik et al. "BEA-Base: A Benchmark for ASR of Spontaneous Hungarian". In: *2022 Language Resources and Evaluation Conference, LREC 2022* (Feb. 2022), pp. 1970–1977. DOI: 10.48550/arXiv.2202.00601.

[9] Vassil Panayotov et al. *LIBRISPEECH: AN ASR CORPUS BASED ON PUBLIC DOMAIN AUDIO BOOKS*.

[10] Vineel Pratap et al. "MLS: A Large-Scale Multilingual Dataset for Speech Research". In: *Proc. Interspeech 2020*. 2020, pp. 2757–2761. DOI: 10.21437/Interspeech.2020-2826.

[11] François Hernandez et al. "TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation". In: (2018). DOI: 10.1007/978-3-319-99579-3_21.

[12] Heidi Christensen et al. "The CHiME corpus: a resource and a challenge for computational hearing in multi-source environments". In: ISCA, 2010, pp. 1918–1921. DOI: 10.21437/Interspeech.2010-552.

[13] Rosana Ardila et al. "Common Voice: A Massively-Multilingual Speech Corpus". In: (2020). DOI: 10.48550/arXiv.1912.06670.

[14] Christian Gaida et al. "Comparing Open-Source Speech Recognition Toolkits". In: 2014.

[15] Meredith Moore et al. "Say What? A Dataset for Exploring the Error Patterns That Two ASR Engines Make". In: 2019, pp. 2528–2532. DOI: 10.21437/Interspeech.2019-3096.

[16] Ingo Siegert et al. *Recognition Performance of Selected Speech Recognition APIs – A Longitudinal Study*. 2020. DOI: 10.1007/978-3-030-60276-5_50.

[17] Binbin Xu et al. "A Benchmarking on Cloud based Speech-To-Text Services for French Speech and Background Noise Effect". In: (2021).

[18] Vered Silber Varod et al. "A cross-language study of speech recognition systems for English, German, and Hebrew". In: *Online Journal of Applied Knowledge Management* (2021), pp. 1–15. DOI: 10.36965/OJAKM.2021.9(1)1-15.

[19] Morgane Riviere, Jade Copet, and Gabriel Synnaeve. "ASR4REAL: An extended benchmark for speech models". In: (2021).

[20] Martha Maria Papadopoulou, Anna Zaretskaya, and Ruslan Mitkov. "Benchmarking ASR Systems Based on Post-Editing Effort and Error Analysis". In: INCOMA Ltd., 2021, pp. 199–207.

[21] Alëna Aksënova et al. "Accented Speech Recognition: Benchmarking, Pre-training, and Diverse Data". In: (2022). DOI: 10.48550/arXiv.2205.08014.

[22] Regis Pires Magalhães et al. "Evaluation of Automatic Speech Recognition Approaches". In: *Journal of Information and Data Management* 13 (3 Sept. 2022). DOI: 10.5753/jidm.2022.2514.

[23] Marcin Pacholczyk. *Przegląd I porównanie rozwiązań rozpoznawania mowy pod kątem rozpoznawania zbioru komend głosowych*. 2018.

[24] Danijel Koržinek. "Task 5: Automatic speech recognition PolEval 2019 competition". In: (2019). URL: http://2019.poleval.pl/files/2019/11.pdf.

[25] Nahuel Unai et al. "Development and evaluation of a Polish ASR system using the TLK toolkit". 2019.

[26] Danijel Koržinek, Krzysztof Marasek, and Łukasz Brocki. *Polish Read Speech Corpus for Speech Tools and Services*. 2016.

[27] Piotr Pęzik. "Spokes – a search and exploration service for conversational corpus data". In: 2015.

[28] Piotr Pęzik. "Increasing the Accessibility of Time-Aligned Speech Corpora with Spokes Mix". In: European Language Resources Association (ELRA), 2018.

[29] Krzysztof Marasek, Danijel Korzinek, and Łukasz Brocki ˇ. "System for Automatic Transcription of Sessions of the Polish Senate". In: (2014).

[30] Piotr Pęzik et al. *DiaBiz - an Annotated Corpus of Polish Call Center Dialogs*, pp. 20–25.

[31] Piotr Pęzik and Michał Adamczyk. *Automatic Speech Recognition for Polish in 2022*. University of Łódź, 2022. URL: https://clarin-pl.eu/dspace/bitstream/handle/11321/894/ASR_PL_report_2022.pdf.

[32] Alec Radford et al. "Robust Speech Recognition via Large-Scale Weak Supervision". In: (2022). DOI: 10.48550/arXiv.2212.04356.

[33] Piotr Kozierski et al. "Acoustic Model Training, using Kaldi, for Automatic Whispery Speech Recognition". In: 2018. DOI: 10.15439/2018F255.