ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ
Faculty of Mathematics and Computer Science

Task 3:

Polish Automatic Speech Recognition Challenge

Michał Junczyk, Iwona Christop
(Adam Mickiewicz University in Poznań)

Piotr Pęzik
(University of Łódź)

IPI PAN seminar

Warsaw, 2.12.2024

Wprowadzenie

ESB (end-to-end speech benchmark) dataset (2022.10, Hugging Face Audio Team)

Benchmark datasets

Evaluating Speech Recognition systems is a hard problem. We use the multi-dataset benchmarking strategy proposed in the ESB paper to obtain robust evaluation scores for each model.

ESB is a benchmark for evaluating the performance of a single automatic speech recognition (ASR) system across a broad set of speech datasets. It comprises eight English speech recognition datasets, capturing a broad range of domains, acoustic conditions, speaker styles, and transcription requirements. As such, it gives a better indication of how a model is likely to perform on downstream ASR compared to evaluating it on one dataset alone.

The ESB score is calculated as a macro-average of the WER scores across the ESB datasets. The models in the leaderboard are ranked based on their average WER scores, from lowest to highest.

Dataset	Domain	Speaking Style	Train (h)	Dev (h)	Test (h)	Transcriptions	License
LibriSpeech	Audiobook	Narrated	960	11	11	Normalised	CC-BY-4.0
VoxPopuli	European Parliament	Oratory	523	5	5	Punctuated	CC0
TED-LIUM	TED talks	Oratory	454	2	3	Normalised	CC-BY-NC-ND 3.0
GigaSpeech	Audiobook, podcast, YouTube	Narrated, spontaneous	2500	12	40	Punctuated	apache-2.0
SPGISpeech	Financial meetings	Oratory, spontaneous	4900	100	100	Punctuated & Cased	User Agreement
Earnings-22	Financial meetings	Oratory, spontaneous	105	5	5	Punctuated & Cased	CC-BY-SA-4.0
AMI	Meetings	Spontaneous	78	9	9	Punctuated & Cased	CC-BY-4.0

Wprowadzenie

Open ASR Leaderboard 2023 (English) (2023.10 by Hugging Face Audio team)

📐 The 🤗 Open ASR Leaderboard ranks and evaluates speech recognition models on the Hugging Face Hub.
We report the Average WER (⬇️ lower the better) and RTFx (⬆️ higher the better). Models are ranked based on their Average WER, from lowest to highest. Check the 📈 Metrics tab to understand how the models are evaluated.
If you want results for a model that is not listed here, you can submit a request for it to be included ✉️✨.
The leaderboard currently focuses on English speech recognition, and will be expanded to multilingual evaluation in later versions.

🏅 Leaderboard | 📈 Metrics | ✉️✨ Request a model here!

model	Average WER ⬆	RTFx ⬆	AMI	Earnings22	Gigaspeech	LS Clean	LS Other	SPGISpeech	Tedlium
nvidia/canary-1b	6.5	235.34	13.9	12.19	10.12	1.48	2.93	2.06	3.56
nyrahealth/CrisperWhisper	6.67	84.05	8.71	12.89	10.24	1.82	4	2.7	3.2
nvidia/parakeet-tdt-1.1b	7.01	2390.61	15.87	14.49	9.52	1.4	2.6	3.16	3.59
nvidia/parakeet-rnnt-1.1b	7.12	2053.15	17.01	13.94	9.89	1.45	2.5	2.93	3.83
nvidia/parakeet-ctc-1.1b	7.4	2728.52	15.67	13.75	10.28	1.83	3.51	4.02	3.57
openai/whisper-large-v3	7.44	145.51	15.95	11.29	10.02	2.01	3.91	2.94	3.86
nvidia/parakeet-tdt_ctc-110m	7.49	5345.14	15.89	12.37	10.52	2.4	5.22	2.54	4.07
nvidia/parakeet-rnnt-0.6b	7.5	2815.72	17.4	14.66	10.01	1.62	3.02	3.32	3.85

Wprowadzenie

PolEval 2019 Task 5 -
otwarte wyzwanie ASR dla jęz. polskiego
- 2019, Daniel Korzinek

Zbiór testowy - nagrania z sejmu RP.

Zbiory treningowe - mowa czytana,
nagrania z sejmu.

Wprowadzenie

Cel zorganizowanego wyzwania w roku 2024:

Porównanie systemów dostępnych publicznie z systemami stworzonymi przez społeczność na **wielodziedzinowym** zbiorze testowym.

Zbiory danych

- Wielodziedzinowy korpusy nagrań mowy.
  - **BIGOS V2:** Mowa czytana, monologi. link
  - **PELCRA dla BIGOS:** Mowa spontaniczna, dialogi. link
- Skompilowany z 24 otwartych zbiorów danych.
- Różnorodność dziedzin, mówców, urządzeń i warunków akustycznych.

Zbiory danych - c.d.

- Wyodrębnione podzbiory do treningu, rozwoju i testowania.
- Nagrania, transkrypcje i metadane dostępne na Hugging Face.
- Zbiór testowy - tylko nagrania, bez transkrypcji.

Split	No. samples from BIGOS	No. samples from PELCRA	Total
train	82 025	229 150	311 175
dev-0	14 254	28 532	42 786
test-A	1 002	1 167	2 169
test-B	991	1 178	2 169
Total	98 272	260 027	358 299

Przykłady - mowa czytana

- Fragmenty Wikipedii oraz artykułów prasowych
  - *policja twierdzi że kierowca pojazdu który potrącił fotografa raczej nie usłyszy zarzutów*
- Literatura klasyczna np. “*Lalka*” B. Prusa
  - *nie słucha zgraja ten już wóz wyprzęga zabiera konie a drugi pieniędzy krzyczy i buławą sięga ów z mieczem wpada na sługi*
- Specjalistyczna terminologia
  - *solwatacja dotyczy elektrolitów oba jony mogą być solwatowane preferencyjnie przez ten sam składnik*

Przykłady - mowa spontaniczna

- Wypowiedzi z sejmu:
  - *w związku z tym po uzyskaniu jednolitej opinii Konwentu Seniorów podjąłem decyzję o uzupełnieniu porządku dziennego*
- Podcasty specjalistyczne:
  - *Dzisiaj porozmawiamy o **Pleiads Conflict.** I o tej grze będą mówili wam...*
- Wywiady:
  - ***yy** dzisiaj jest całkiem nieźle zważywszy na to że spędziłem **yy** dwie i pół godziny u dentysty*
- Nagrania obsługi klienta*:*
  - *Wie pan co ja nie jestem zbyt biegły komputerowo. **Czy czy** mogę prosić pana o pomoc i przesłanie mi tych dokumentów na mój adres mailowy*

Szczegóły wyzwania

- Uczestnicy mieli dostarczyć automatycznie wygenerowane transkrypcje dla nagrań ze zbiorów testowych A oraz B.
- Uczestnicy mogli opracować swoje lub dostroić istniejące systemy ASR.
- Metryki:
  - Word Error Rate (WER)
  - Character Error Rate (CER)
- Ograniczenia:
  - Bez użycia zewnętrznych danych.
  - Bez ręcznej transkrypcji przykładów testowych.

Metryki do oceny jakości

- **Słowna** stopa błędów *(ang. **Word** Error Rate - WER)*

$$WER = \frac{S + D + I}{N}$$

. S - liczba podstawionych **słów** (substitutions),
. D - liczba usuniętych **słów** (deletions),
. I - liczba wstawionych **słów** (insertions),
. N - całkowita liczba **słów** w prawidłowej transkrypcji (referencji)

Referencja (transkrypcja)*: **Ala ma kota i psa.***
Hipoteza systemu ASR*: **Sala ma kota psota i.***

Metryki do oceny jakości

- **Znakowa** stopa błędów *(ang. Character Error Rate - CER)*

$$\mathrm{CER} = \frac{\text{number of errors}}{\text{reference text length in characters}}$$

- Number of errors (liczba błędów) = S + D + I
- S - liczba podstawionych **znaków** (substitutions),
- D - liczba usuniętych **znaków** (deletions),
- I - liczba wstawionych **znaków** (insertions),
- N - całkowita liczba **znaków** w prawidłowej transkrypcji (referencji)

Referencja (transkrypcja)*: **Ala ma kota i psa.***

Hipoteza systemu ASR*: **Sala ma kota i bz.***

Normalizacja tekstu

- Normalizacja referencji oraz hipotez systemów ASR ze zgłoszeń:
  - Ujednolicenie wielkości liter (ang. *case folding)*
  - Usunięcie znaków interpunkcyjnych

Wyniki bazowe

9 systemów, 25 wariantów modeli.

System komercyjne:

- Whisper cloud
- Google ASR V1
- Google ASR V2
- Microsoft Azure
- Assembly AI

System darmowo dostępne:

- NVidia Nemo
- Whisper local
- Multilingual Speech (MMS)
- Wav2Vec2

Wyniki wyzwania

- Najlepszy wynik: LIT-MR
  - Test A/B - **WER 11.07%** i **CER 6.85%**.

- Baza odniesienia:
  - Test A - Whisper-large-v3: **WER 14.51 %, CER 8.41%**
  - Test B - Whisper-large-v3: **WER 14.02 %, CER 7.98%**

Wyniki zespołu LIT-MR

- Dostrojenie modeli z wykorzystaniem danych dostarczonych w ramach wyzwania.

- Redukcja stopy błędów w porównaniu do modeli bazowych:
  - Test A - WER 3.44 pp, CER 1.56 pp
  - Test B - WER 2.95 pp, CER 1.13 pp

Wyniki bazowe (top12) vs LIT-MR - test set A

System	WER [%]	CER [%]
poleval_best_lit_mr	11.07	6.85
whisper_large_v3	14.51	8.41
whisper_cloud	15.28	8.72
assembly_best	16.47	9.84
whisper_medium	17.61	9.48
google_v2_long	19.54	12.26
google_long	20.27	13.49
google_short	20.69	13.86
nemo_multilang	22.59	12.02
whisper_small	23.39	11.74
mms_all	25.14	10.48
azure_latest	25.85	19.26
w2v-1b-pl	26.62	9.22

Polish ASR leaderboard

Wyniki bazowe (top12) vs LIT-MR - test set B

System	WER [%]	CER [%]
poleval_best_lit_mr	11.07	6.85
whisper_large_v3	14.02	7.98
whisper_cloud	14.57	8.19
assembly_best	15.97	9.35
whisper_medium	17.21	9.50
google_short	20.17	13.35
google_v2_long	21.12	14.19
google_long	21.59	15.24
nemo_multilang	22.05	11.72
whisper_small	24.14	12.64
mms_all	24.66	10.11
azure_latest	25.59	19.08
w2v-1b-pl	26.21	9.02

Polish ASR leaderboard

Dziękujemy za uwagę.

michal.junczyk@amu.edu.pl
iwona.christop@amu.edu.pl
piotr.pezik@uni.lodz.pl