Badanie algorytmów uczenia maszynowego w zastosowaniu do rozpoznawania mowy personelu medycznego

7@KSMM'2023

Opiekun Projektu:
prof. dr hab. inż. Andrzej Czyżewski
Klient projektu:
mgr inż. Szymon Zaporowski

Osoby wspomagające
mgr inż. Marta Zielonka
inż. Franciszek Górski
inż. Mateusz Żak
Zespół pracowników KSMM

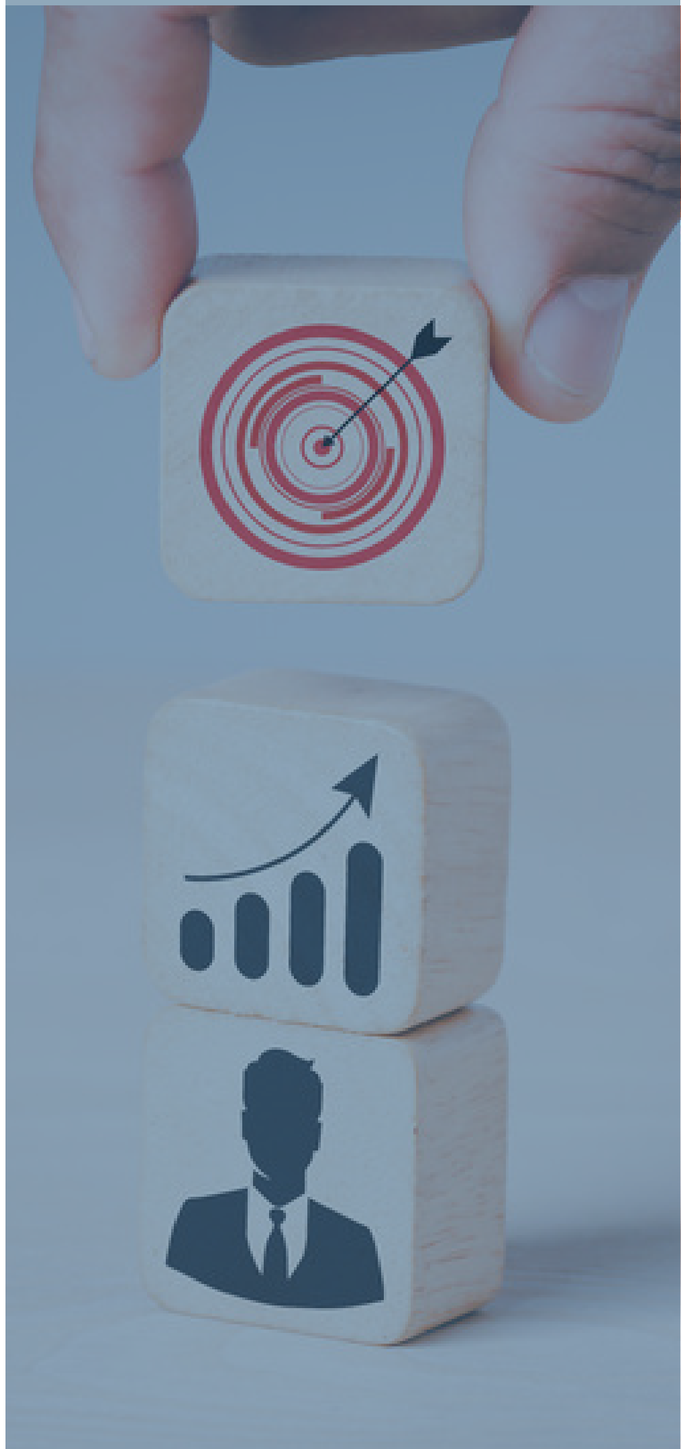
Wykonawcy
inż. Wiktor Krasiński - lider
inż. Jakub Nowak
inż. Przemysław Rośleń
inż. Jan Stopiński

Cel projektu

- Opracowanie metodyki dotrenowania i testowania dostępnych algorytmów transkrypcji mowy na tekst, np. w środowisku do uczenia głębokiego Whisper. Następnie, w oparciu o zadeklarowaną współpracę ze strony lekarzy GUMEDu należy zbudować i nagrać słownik polskich wyrażeń medycznych, które są używane przy opisywaniu chorób, kierowaniu na badania przez specjalistów, wypisaniu recept.

- Nagrywanie należy przeprowadzić w warunkach naturalnych od strony akustycznej, w typowym otoczeniu. Następnie, w ten sposób otrzymany materiał należy adnotować i użyć do dotrenowania dostępnej sieci neuronowej, która wcześniej już została wytrenowana w oparciu o duże słowniki, w tym słownik języka polskiego (np. Mozilla Polish).

- W toku eksperymentów należy uzyskać wyniki oceny skuteczności rozpoznawania wyrazów związanych ze słownikiem medycznym z uwzględnieniem obliczenia metryk błędów.



Scenariusze użycia

Wizyta lekarska

Usprawnienie pracy lekarza podczas wizyty

Sala operacyjna

Operacje automatycznie rejestrowane, nie wymagając zaangażowania pracowników.

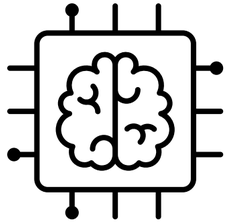
SOR

Brak potrzeby spisywania przebiegu akcji przez pracowników

Cechy charakterystyczne

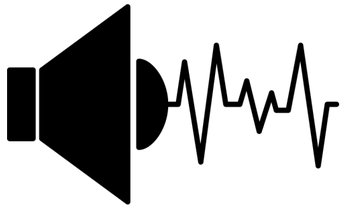
Wybraliśmy następujące narzędzia do konwersji mowy na tekst:

- **Whisper (small, medium, large)**
- **Google**
- **Microsoft Azure**



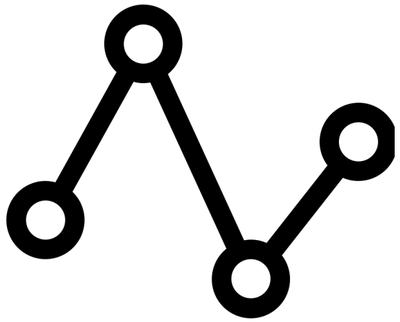
Wygenerowaliśmy 1200 próbek słownictwa medycznego (100 słów) w formacie mp3:

- **600 męskich (500 naturalnych + 100 syntezowanych)**
- **600 żeńskich (500 naturalnych + 100 syntezowanych)**



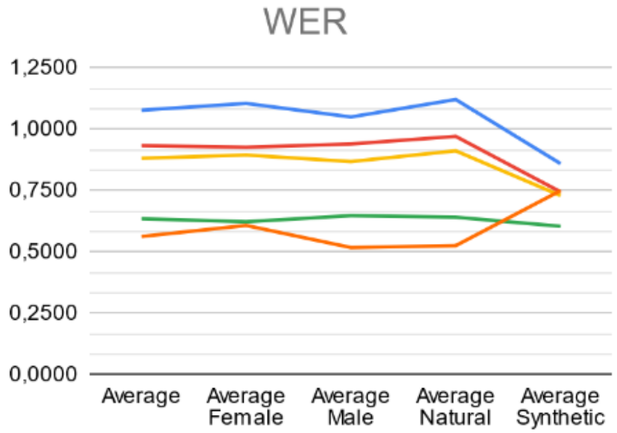
Wykorzystaliśmy następujące metryki w celu oceny rezultatów:

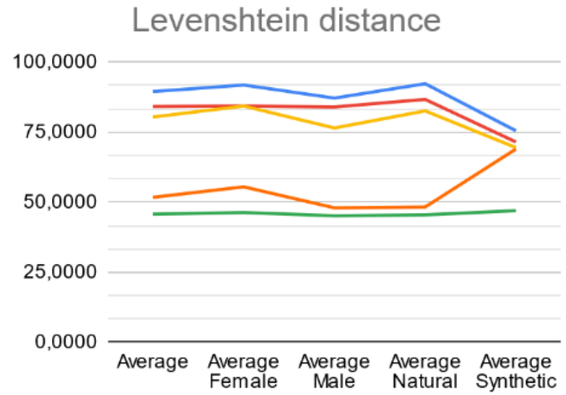
- **WER**
- **MER**
- **CER**
- **WIL**
- **Levenshtein distance**
- **Jaccard distance**

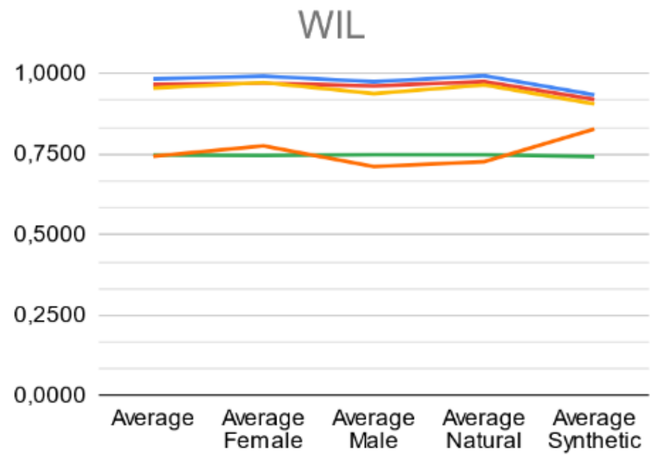


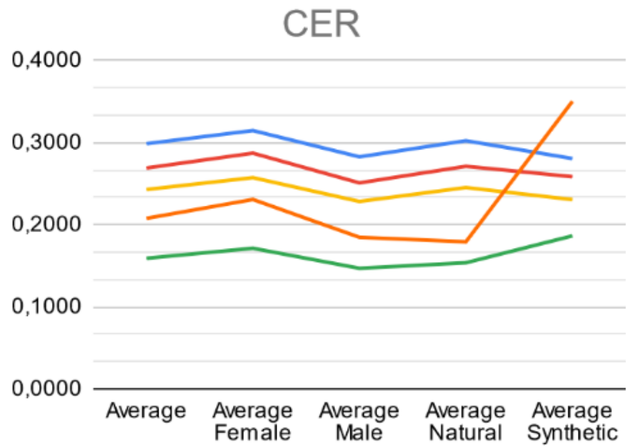
Name	If recommended	Scope	Ease of use (1-5; 1-difficult, 5-easy)	Price	How annotations are stored?	Limitations/disadvantages	API (please double-check)
Prodi.gy	yes	Span Categorization, Dependencies & Relations, Audio & Video, Zero-Shot Prompts, Named Entity Recognition, Text Classification, Computer Vision, A/B Evaluation	5	$390 (Personal) $490 (Company; price per seat; available in packs of 5 seats)	standoff format	-	-
LLama Alpaka	no	-	-	-	-	-	no
TagEditor	no	text	4	Full version minimum donation of 20$. Demo version allows for 30 sentences.	-	requires Windows 10, 64-bit	no
Inception	yes	text	3	Open Source	standoff format	requiers Java 11 or higher	yes
Brat	yes	text	4	Open Source	standoff format	preferable OS: Linux, but will work on Windows in WSL, Python 2.5+ required (No Python 3.x!), Fully supported on Google Chrome and Safari.	no
Doccano	yes	-	5	Open Source	standoff format	Python 3.8+ Offers only simple annotation capabilities	yes BETA
LightTag	yes	-	4	Free for individual use and teams*. *Free for teams up to 10k annotations/month, then $0.02/additional annotations	standoff format	takes quite time to start annotating and to learn a tool, number of functionalities may be overwhelming, tags for annotating not visible all the time (only when hovering over), interface may be unintuitive for beginners	yes
Audino	yes	Voice Activity Detection (VAD), Diarization, Speaker Identification, Automated Speech Recognition, Emotion Recognition tasks	5	Open Source	standoff format and database	Simple and nice to cooperate, Need to host a web application	only for upload data
Label Studio	yes	Computer Vision, Natural Language Processing, Audio/Speech Processing, Conversational AI, Ranking & Scoring, Structured Data Parsing, Time Series Analysis, Videos	5	The community edition is free, for the enterprise edition you need to contact the authors.	-	all functionalities available only when using enterprise	yes

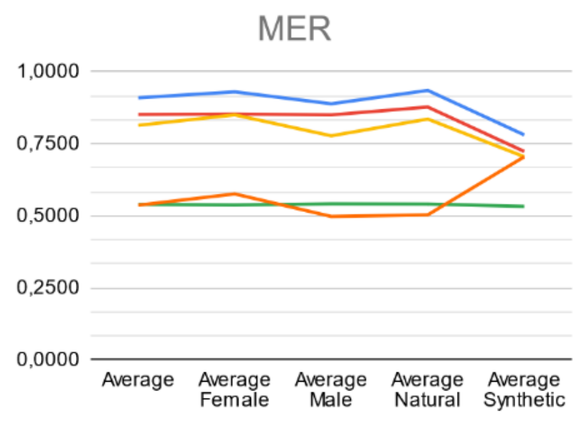
Wyniki cz.1

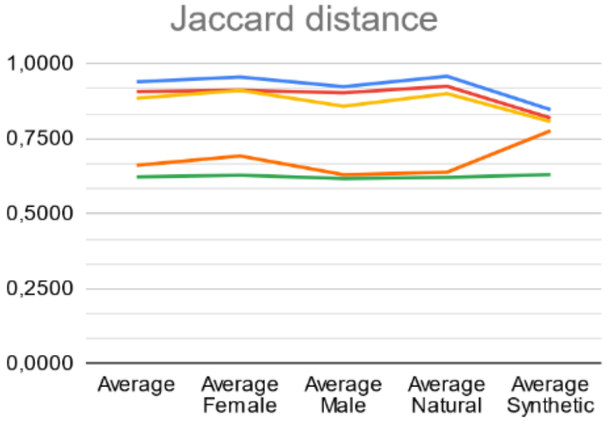












- Whisper (small)
- Whisper (medium)
- Whisper (large)
- Microsoft Azure speech-to-text
- Google speech-to-text

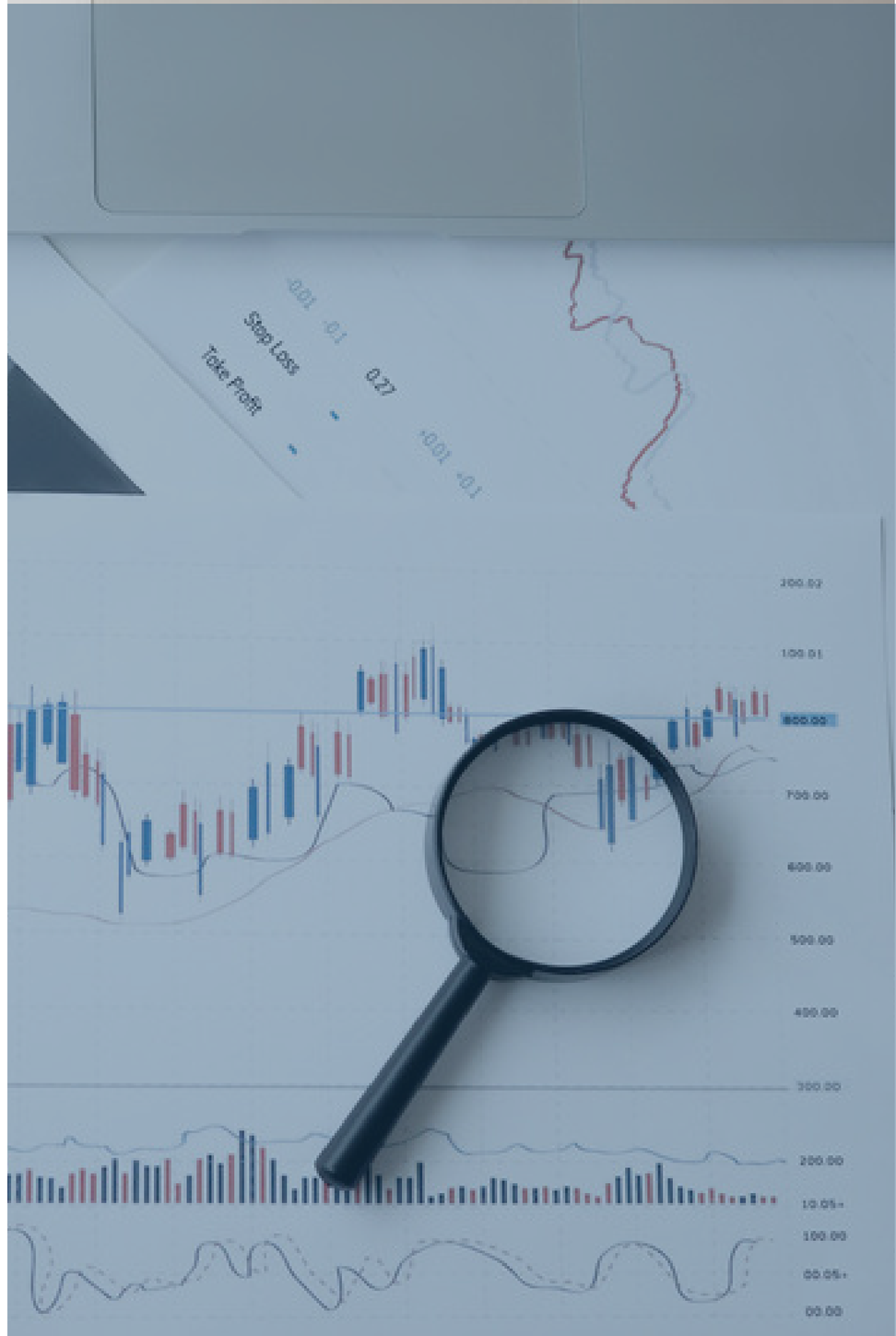
Wyniki cz.2

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

Podsumowanie

- Wykonaliśmy systematyczny przegląd literatury.
- Zbadaliśmy narzędzia do przekształcania mowy na tekst (ang. speech-to-text).
- Przejrzeliśmy dostępnych narzędzi do adnotacji.
- Wygenerowaliśmy próbki słownictwa medycznego.
- Przetestowaliśmy dostępne narzędzia speech-to-text na nagranych próbkach.
- Zapoznaliśmy się z metrykami oceny jakości transkrypcji mowy na tekst.
- Przeanalizowaliśmy wyniki transkrypcji medycznej przy użyciu wybranych narzędzi.
- Przygotowaliśmy i zgłosiliśmy artykuł na konferencję.



Referat zgłoszony na konferencję ECAI'2023

A survey of automatic speech recognition deep models performance for Polish medical terms

Marta Zielonka[a,*], Wiktor Krasiński[a], Jakub Nowak[a], Przemysław Rośleń[a], Jan Stopiński[a], Mateusz Żak[a], Franciszek Górski[a] and Andrzej Czyżewski[a]

[a]Multimedia Systems Department, Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology
ORCiD ID: Marta Zielonka https://orcid.org/0000-0003-1407-6770,
Franciszek Górski https://orcid.org/0000-0001-7537-0039,
Andrzej Czyżewski https://orcid.org/0000-0001-9159-8658

Abstract. Among the numerous applications of speech-to-text technology is the support of documentation created by medical personnel. There are many available speech recognition systems for doctors. Their effectiveness in languages such as Polish should be verified. In connection with our project in this field, we decided to check how well the popular speech recognition systems work, employing models trained for the general Polish language. For this purpose, we selected 100 words from the International Classification of Diseases dictionary, the Polish-language version of the International Statistical Classification of Diseases and Health Problems. The words were read into a microphone by five women and five men and also generated with a speech synthesizer using a male and a female voice. This resulted in 1,200 recordings tested with the following systems: Whisper, Google speech-to-text, and Microsoft Azure speech-to-text. The achieved word recognition performance is reflected by the calculated metrics: WER, WIL, Levenshtein distance, Jaccard distance, MER, and CER. Results 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, because the number of errors made when converting speech to text is too high.

1 Introduction

Over the past decade, speech-to-text (STT) systems, also called automatic speech recognition (ASR) systems, have rapidly developed. The development was made possible by advances in deep learning theory and the growing demand for speech transcription systems or intelligent voice assistants. It included growth in the accuracy of these systems and an application for more languages than just English. Nowadays, the dominant approach to developing speech-to-text systems is based on neural networks, which achieve outstanding results in the transcription of recorded text. Systems such as DeepSpeech [11] or Recurrent Neural Network Transducer [5] can be given as examples. However, solutions using the Transformer-type architecture introduced in the [18] have received the most attention in recent years. Based on this idea, solutions such as OpenAI Whisper [15] and Conformer [10] have been developed that have achieved a Word Error Rate (WER) of less than 5% on different English datasets. Moreover, the authors in [15] show that their model can work for languages other than just English. As a result, high accuracy has been achieved in transcribing everyday speech, but there are still solutions tailored to some domains of applications that are insufficient. To address this issue in our work, we tested 3 systems: OpenAI Whisper, Google Speech-To-Text [3], and Azure Speech-To-Text [4] on 100 words from the International Classification of Diseases (ICD) dictionary for the Polish language. A list of medical terms derived from the dictionary is included for interested readers as supplementary material. Our team made recordings of these 100 words for 5 male voices, 5 female voices, 1 synthetic male, and 1 synthetic female voice, giving us 1200 recordings. We evaluated each STT system on this dataset and, for each of them, calculated a set of 6 different metrics designed for automatic speech recognition tasks. These metrics are WER, WIL, Levenshtein distance, Jaccard distance, MER, and CER. These metrics reflect how tested models are ready for voice-filling medical records, describing cases, or prescribing treatment, since the number of errors made when converting speech to text is crucial in this application domain.

* Corresponding Author. Email: martaz@multimed.org

2 Methods

This section describes methods; it includes a description of off-the-shelf, most common speech-to-text (STT) engines followed by metrics used in the experiments. The first subsection is dedicated to STT tools description, in which Whisper, Google STT, and Azure STT are included. The next subsection presents the metrics used, a description, and corresponding mathematical formulas.

2.1 Speech-to-text tools

There are various ready-to-use speech-to-text engines. Many of them support multiple languages. It is easy to notice that they work well for English on data that does not contain specialized domain terms. This study tests three of the most well-known tools with speech excerpts pronounced in Polish using medical terminology only. All these tools were used in a configuration that supports the Polish language; in other words, their authors trained them on datasets also in this language. These engines are Whisper (small, medium, large versions),

Dziękujemy za uwagę