Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2023) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

27th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2023)

Challenges of Automatic Speech Recognition for medical interviews - research for Polish language

Karolina Kuligowska[a]*, Maciej Stanusch[b], Marek Koniew[b]

[a]University of Warsaw, Długa St. 44/50, 00-241 Warsaw, Poland
[b]SOVVA SA, Karola Goduli St. 36, 41-712 Ruda Śląska, Poland

Abstract

Automatic Speech Recognition (ASR) systems are quickly becoming a crucial element in supporting healthcare providers, improving the flow of information among medical teams, and enhancing the patient's experience. However, to be fully supportive, these ASR systems must meet certain requirements dictated by market realities: high accuracy of speech recognition and low rate of errors, the possibility of additional training the model, and the possibility of on-premise system installation. Therefore, the aim of this paper is to perform a comparative analysis of leading ASR systems available on the Polish market for the needs of conducting medical interviews. We selected three systems, Google ASR, Microsoft ASR, and Techmo ASR, and we compared their performance on a prepared data set of medical-related expressions spoken in Polish. The results of our analysis indicated that there are minor discrepancies in the accuracy of speech recognition between all three evaluated ASR systems, whereas only two ASR systems met the raised requirements, in both cases partially. Still, they all exhibited specific problems in recognising word endings or word boundaries. We were able to categorise such problems into: Misrecognitions, Quality Problems, and Word Boundaries, varying in their level of influence on the further speech recognition process. Our research findings are expected to provide valuable insights to a wide range of stakeholders facilitating the development of tailored speech recognition solutions that meet the specific needs of medical sector.

Keywords: automatic speech recognition; speech-to-text; Polish speech recognition system; medical interview transcript;

* Corresponding author.
E-mail address: kkuligowska@wne.uw.edu.pl

1. Introduction

Speech recognition is the process in which a computer program identifies words in spoken language and converts them into a text format. Recent advances in artificial intelligence and machine learning, as well as the increasing computing power and memory capacity of computer technology, have enabled the development of various human speech recognition systems, also known as Speech-To-Text (STT) or Automatic Speech Recognition (ASR). Most popular examples of such solutions include [5,8,16,19,21,23,27]: Google Cloud [Google ASR], Microsoft Azure Cloud [Microsoft ASR], Nuance Communications Cloud [Dragon STT], OpenAI [Whisper ASR], Phonexia [Phonexia STT], Techmo [Techmo ASR], Rev AI [Rev AI ASR], and others.

While ASR systems can produce accurate transcriptions, their performance can vary greatly in terms of Accuracy and Word Error Rate (WER) measures. The ASR Accuracy is measured as the percentage of correctly recognised words that are transcribed in a given utterance. Typically, it falls within the range of 70% to 95%. The ASR WER is measured as the percentage of transcription errors in the recognised speech, and in a given voice sample it can range from 8.53% to as high as 34.90% [7]. It is important to note that the Accuracy and WER of ASR systems may be influenced by various factors beyond the clarity of the spoken language itself.

The processing speed and accuracy of speech recognition depend on several key factors, such as [2,11,13,14,15]:

- type of the natural language used in the speech,
- topic range of the speech (e.g. colloquial or specialised vocabulary),
- speaker's ethnic origin and accent,
- potential background noise and acoustic disturbances,
- models, algorithms and approaches used in speech recognition (e.g. the size of the vocabulary, the specification of the grammar, etc.).

ASR systems have come a long way in overcoming various challenges related to recognising everyday conversational language, colloquialisms, and regional dialects, as well as speech on diverse topics with a broad range of vocabulary. These systems have also made strides in recognising speech with varying acoustic properties, flawed speech, extended utterances, and even speech that contains pauses, repeats, and sudden interruptions [10,24,25,26]. However, the level of success in speech recognition may still vary depending on the complexity of the speech.

The literature analysis performed by the authors showed a small number of studies focusing on the comparison of the quality of various ASR systems in general medical applications, and in particular a severe lack of studies centering on the comparison of ASR systems available on the Polish market for the needs of medical interviews and taking into account interlocutors using colloquial Polish language. In the world of healthcare, there are already existing examples of cutting-edge ASR solutions that have been successfully implemented to support patients, streamline administrative processes, and improve healthcare facilities management. However, despite their potential to revolutionise the medical field, these ASR solutions are frequently presented and described by commercial sources in the form of company reports, testimonials, or white papers, yet they are rarely discussed in scientific research [20]. Therefore, there is a need for comparative studies of currently available ASR systems, with a specific emphasis on measuring their performance, scalability, and ease of access in healthcare delivery.

Hence, the aim of this study is to fill the research gap described above, as well as significantly improve the system for the automatic customer service functioning within the Omni-Chatbot Platform (OCP) developed by SOVVA SA, by conducting a comparative analysis of Google ASR, Microsoft ASR, and Techmo ASR - systems selected among the leading ones on the Polish market. The criteria for selecting these three ASR systems were based on their potential for wide application in the healthcare industry, specifically, the recognition of medical-related utterances while assuming that Polish colloquial language would be used by the interlocutors. The outcomes of this comparative analysis will facilitate the selection of the most optimal ASR solution, not only considering the performance for Polish expressions' recognition, scalability, and accessibility, but also addressing the need of further integration with the above-mentioned OCP system. The target OCP system aims to conduct a preliminary medical interview (pre-triage) and to identify the patient's intention of contact by analyzing their natural, colloquial language.

This paper is organised as follows. Section 2 presents an overview of the available literature in the context of speech recognition systems from the perspective of general medical applications. Section 3 describes the methodology adopted for the implementation of this study, as well as the course of the research. Section 4 reveals the results of the research. The summary and conclusions are included in Section 5.

2. ASR systems in medical applications – related work

The field of speech recognition technology is evolving at a rapid pace, resulting in increasingly accurate and reliable ASR systems. As these systems continue to undergo intensive development, the rate of errors during recognition process is decreasing. However, it is important to note that the quality of existing voice recognition systems varies widely, with some of them struggling to accurately transcribe certain words.

In 2020, a group of researchers at the University of South California conducted a comprehensive study comparing the accuracy of various, publicly available, commercial and scientific ASR systems [7]. These included systems developed by Amazon, Apple, Google, IBM, Microsoft, and Kaldi, which were tested on seven sets of English voice data across six different dialogue domains. The voice data sets consisted of spontaneous speech recordings of real conversations, making the study highly representative of real-life scenarios. The researchers used WER as the primary assessment measure. The study revealed that even the most sophisticated ASR systems still make a significant number of transcription errors, especially in areas that require recognition of specialised vocabulary or spontaneous colloquial speech.

A comparison of these recent studies with earlier research [17,28] on similar data sets clearly indicates that over the last few years significant advancements have been made, particularly with the integration of deep learning techniques in the ASR technology. However, researchers emphasise the importance of considering the field of application and context when selecting the most optimal speech recognition solution. This highlights the need for developing tailored systems that cater to specific requirements, such as the application domain and context of using speech recognition, in order to select the most efficient solution, which consequently enhances service quality.

For the researchers exploring innovative ways to streamline medical documentation and ultimately improve the quality of patient care, the ASR systems have been the subject of research with a particular focus on the processing of digital clinical records. This includes voice notes recorded by doctors and nurses, as well as voice records summarising hospital discharges [1,9]. With the rise in adoption of voice recorders supported by transcription systems for digital clinical documentation, there has been a surge of interest in the impact of ASR technology on medical practice and healthcare professionals. This has led to extensive research aimed at understanding how ASR technology can help healthcare providers streamline their workflows and deliver better patient care [22]. With the advent of cloud solutions, the possibilities of automatic speech processing for real-time transcription during medical interviews are intensely being explored by researchers. The hypothetical scenarios of conversations between doctors and so-called simulated patients, i.e. trained actors playing the role of patients in medical situations, are being used for such research [4,6]. This development marks a shift in focus, with researchers now looking at improving the accuracy and efficiency of ASR systems in processing real-time medical conversations.

A recent study conducted by a team of researchers led by Kim, Liu, and Calvo [12] examined the performance of five ASR systems for transcription quality in the medical field. The study evaluated the online platforms - Google Cloud, IBM Watson, Microsoft Azure, Trint, and YouTube – in the scope of their embedded ASR systems. The voice data sets were derived from 12 medicine students' interactions with two simulated patients, resulting in a total of 24 teleconsultations. Since the beginning the researchers assumed that manual transcriptions would outperform automatic transcriptions in terms of accuracy, and the study's findings confirmed this hypothesis. Still, automatic processing, despite its error-proneness, offers advantages in terms of efficiency and cost-effectiveness. Interestingly, among evaluated ASR systems, YouTube's ASR system surpassed other cloud-based solutions in terms of word recognition accuracy, highlighting the platform's superior performance in the field of medical-related transcription.

3. Proposed method

The main research objectives of this paper, aiming at the comparative analysis of three ASR systems selected among the leading ones on the Polish market, are:

- verification of the quality of speech recognition measured on a prepared data set of medical-related expressions recorded in Polish colloquial language,
- verification of the possibility of additional training the speech recognition model with the specificity of Polish everyday language in medical-related applications,
- verification of the possibility of installing the speech recognition model as on-premise software.

For evaluating the quality of speech recognition we applied the following metrics: Accuracy, WER, Levenshtein editing distance algorithm and Jaro-Winkler similarity. The Accuracy measures the percentage of correctly recognized words in a given utterance. The WER indicates the percentage of incorrect words in the transcription of a given utterance. Both the Levenshtein and Jaro-Winkler methods measure the degree of similarity of the two strings, however they have different underlying algorithms and approaches. The Levenshtein algorithm calculates the minimum number of changes required to transform two comparable character strings into identical ones by inserting, replacing, or deleting characters. The lower the editing distance, the higher the degree of similarity of the two strings [18]. The Jaro–Winkler similarity is the weighted sum of percentage of matched characters from each string and is not based on the editing distance like Levenshtein measure [3]. The higher the Jaro-Winkler similarity, the higher the similarity of the two strings.

We took into consideration several factors while selecting three ASR systems among the leading systems on the Polish market. These factors included the system's popularity in research practice, its proven performance, scalability potential, and ease of access. Based on these considerations, and after analyzing the publicly available technical data and documentation, we have selected for comparative analysis:

- Google ASR system (manufacturer of Google Inc.),
- Microsoft ASR system (manufacturer Microsoft Inc.),
- Techmo ASR system (manufacturer Techmo sp. z o.o.).

In order to accomplish the research objectives outlined earlier, we initially implemented the Application Programming Interface (API) to facilitate the testing of each ASR system. Following this, a data set consisting of medical-related expressions was prepared for a comprehensive comparative analysis. The data set was created on the basis of a text collection of 19,000 categorised medical-related expressions provided and owned by SOVVA, the company leading the research project. Examples of such expressions include: "odczuwam różne dolegliwości jelitowe", "boli mnie opuchlizna po ukąszeniu", "w ranie jest jakaś sącząca i krwista wydzielina", "czuję taki dziwny ucisk w klatce", "mam czerwoną krostkę na wewnętrznej powierzchni policzka", "codziennie rano mam okropny napad duszności", "występują u mnie wahania ciśnienia krwi", "widzę takie tęczowe koła i mroczki przed oczami", "czy można ten lek dostać w syropie", "proszę o zapisanie leków na ból głowy" etc.

From this collection, medical experts (participating in the research project) selected the 1,000 expressions most commonly raised by patients during visits to doctors' offices. They encompassed both colloquial terms and those conforming to medical-related nomenclature, indicating various medical conditions such as pain, injury, rash, shiver, swelling, tingling, burning etc. Examples of such expressions include: "narastający ból", "opuchlizna", "rany na skórze", "krosty z ropą na skórze", "duszności", "spadek ciśnienia", "szumy w uszach", "mroczki przed oczami", "drżenie rąk", "częste oddawanie moczu", "uczucie mrowienia na przedramieniu", "pieczenie w przełyku", "problem z krzepnięciem krwi", "problemy ze snem" etc. These selected expressions were then recorded into 1,000 audio files and uploaded to an FTP server for testing. Recordings were prepared according to a set of following rules:

- Each file contained a recording of a single phrase (expression or word), recorded by three individuals of different genders, representing distinct regions (north and south of Poland), and varying in age (between 30 and 50 years),
- The files were saved in the WAV format with the following parameters, commonly used for speech recognition tasks: channels – 1 (mono), sample format – 16-bit PCM encoding, sampling frequency – 16,000 Hz. The audio sampling rate for utterances was set to 16 kHz to facilitate precise laboratory analysis. While the standard rate of 8 kHz is sufficient for basic voice recognition over the phone, the higher rate of 16 kHz provides more accurate representation of the audio signal [11],
- For each recording, a TXT file was created with the mapping: file_name – recorded_phrase, e.g.: 001.wav – osteoporosis or 002.wav – erosion.

Subsequently, the audio files created according to the aforementioned rules were processed by each ASR system using the pre-implemented API. The resulting text transcripts were then compared and analysed using proposed measures (Accuracy, WER, Levenshtein editing distance algorithm and Jaro-Winkler similarity) to evaluate the overall quality of speech recognition of each ASR system.

4. Research results

Firstly, we present the combined results of the technical capabilities of the ASR systems, that aim at accessibility for the benefit of the end customer. The following Table 1 summarises the research findings, including: 1) verification of the possibility of additional training the speech recognition model, and 2) the verification of the possibility of on-premise installation of the ASR systems. These findings are based on the analysis of technical specification and publicly accessible documentation provided for each ASR system, as well as experimental attempts to additionally train these systems and install them in the infrastructure of the end customer.

Table 1. Technical capabilities available for the end customer by individual ASR system.

	Google ASR	Microsoft ASR	Techmo ASR
Possibility of additional training	No	Yes	Yes/No*
Possibility of on-premise installation	No	No	Yes

Source: own elaboration.

* In the case of the Techmo ASR, it is possible to additionally train the speech recognition model only as a paid service provided by the Techmo team. The user of the system has no possibility of fine-tuning one's own model.

Table 1 shows that only the Techmo ASR system offers a possibility of additional training of the model (available as a paid service provided by Techmo) and an on-premise installation. Although Microsoft ASR allows additional training of the model, it cannot be installed on-premise. Google ASR, on the other hand, does not fulfill either of the two technical aspects. Based on additional experiments with ASR systems, it was also found that the detection of specific expressions in the Google ASR system can be improved by controlling the built-in pre-defined dictionaries of spoken expressions. Microsoft ASR does not support such functionality at all, while Techmo ASR does not provide any information on the possibilities in this regard.

Subsequently, we present the findings from the examined ASR systems in terms of the overall quality using proposed measures. The results obtained from a data set of 1,000 recordings of medical-related expressions indicate a speech recognition Accuracy of over 86% by all the tested ASR systems, with a discrepancy of only 1.7% between the best and worst result. However, it is important to note that these tests were conducted under controlled laboratory conditions, without taking into consideration external factors such as noise or interferences that could occur during a real-life conversation in a room or over the phone. The following Table 2 summarises the Accuracy levels exhibited by ASR systems.

Table 2. The Accuracy of speech recognition by individual ASR system.

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

Source: own elaboration.

The Accuracy measure, as well as WER that represents the percentage of incorrectly recognized words, concerns speech recognition at the whole word level. In order to measure the recognition at the single character level, the Levenshtein and Jaro-Winkler methods were further applied. In principle, Levenshtein's measure of editing distance is a value expressed in the number of changes needed to unify two strings of characters, while the Jaro-Winkler measure is the degree of similarity of two strings expressed in the range from 0 to 1. As values of all applied measures are variously expressed, in order to be able to compare them using a single indicator, the results were standardized to ultimately show the percentage similarity between original recorded utterance and recognized transcription. According to the result of the Kolmogorov-Smirnov test, there is no significant difference between the distributions of recognition similarity for the three ASR systems in terms of the Levenstein, Jaro-Winkler and WER measures, and all p-values are less than 0.05. In the analysis of recognized expressions, the WER and Jaro-Winkler measures gave identical results and were each time higher than the Levenshtein-based similarity percentage. The best, average and worst result remained the same, subsequently: Google ASR, Techmo ASR and Microsoft ASR, regardless of the measure used, as shown in the Table 3.

Table 3. Percentage similarity of recognised expressions by individual ASR system.

	Levenshtein	Jaro-Winkler	WER
Google ASR	83.35%	90.83%	90.83%
Microsoft ASR	81.75%	88.37%	88.37%
Techmo ASR	82.86%	89.40%	89.40%

Source: own elaboration.

The obtained speech recognition results also contained incorrect or incomplete words as output from the ASR systems. This created the opportunity to study various sorts of errors that appeared during the transcription of the speech to text. These ASR problems were categorised into three following groups: Misrecognitions (complex and difficult to handle in further processing), Quality Problems (complex but possible to handle in further processing), and Word Boundaries (relatively easy to handle in further processing).

Misrecognitions refer to cases where the ASR system failed to recognise the recorded expression, either entirely or with a different meaning. Quality Problems refer to cases where the ASR system recognised the recorded expression with minor discrepancies that did not affect the meaning of the whole phrase. Finally, Word Boundaries refer to cases where the ASR system recognised the recorded expression properly, but some words (precisely beginnings or endings of words) were not correctly identified. Most representative examples of each category of ASR problems are listed below.

Misrecognitions:

- not recognised at all: "omdlenia", "ścieńczenia", "pęcherzyca"
- recognised with different meaning: "brak łaknienia" recognised as: "brakło tchnienia"
- recognised with different meaning: "problem z ukrwieniem oczu" recognised as: "problem z ukrwienie moczu"
- recognised with different meaning: "wściekłość" recognised as: "zaległość"

Quality Problems:

- original: "zubożona mowa" recognised as: "zburzona mowa"
- original: "żwir w woreczku żółciowym" recognised as: "żwir po woreczku żółciowym"
- original: "plamy rumieniowe" recognised as: "plamy rumień owe"
- original: "złogi w pęcherzyku żółciowym" recognised as: "złogi pęcherzyku żółciowym"

Word Boundaries:

- original: "stres pourazowy" recognised as: "stres po urazowy"
- original: "problemy z chodzeniem" recognised as: "problemy schodzeniem"
- original: "śpię w dzień" recognised as: "śpiew dzień"
- original: "niedosłyszenie" recognised as: "nie do słyszenia"

The following Table 4 summarises the composition of all these ASR problems.

Table 4. Percentage of specific recognition problem by individual ASR system.

	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%

Source: own elaboration.

According to the presented data, Techmo ASR had the lowest percentage of Misrecognitions as compared to Microsoft ASR and Google ASR. Specifically, Techmo ASR presented 61,7% of Misrecognitions towards 71,2% for Microsoft ASR and 73,2% for Google ASR. At the same time Microsoft ASR had the lowest percentage of Quality Problems as compared to Google ASR and Techmo ASR. Specifically, Microsoft ASR presented 14,8% of Quality Problems towards 23,6% for Google ASR and 26,6% for Techmo ASR. On the other hand, Google ASR had the lowest percentage of Word Boundaries as compared to Microsoft ASR and Techmo ASR. Specifically, Google ASR presented 3,2% of Word Boundaries towards 14,0% for Microsoft ASR and 11,7% for Techmo ASR.

Overall, Techmo ASR performed better in terms of the relation of Misrecognitions to the rest of recognition problems that do not affect further processing, i.e. Quality Problems and Word Boundaries. However it is worth mentioning that the data set was relatively small, comprising 1,000 expressions, hence, we were unable to discern more detailed patterns in the types of problems that each of the three ASR systems had.

It is also beneficial to focus on the expressions that posed difficulties for all systems. Our research identified a set of words that were confused or falsely recognised by all three ASR systems simultaneously, i.e.: "świąd", "ból", "krosty", "krwiaki", "sinienie", "płaczliwość", "leki", "zauszne", "białka", "płonica", "omamy", and "urojenia". Another set of words were confused or falsely recognised by two ASR systems at the same time (in various pairs): "poronienie", "problem z", "plamy", "słyszę", "uraz", "kłykciny", "utrata", "przerosła", "poparzeniowa", "rozpadowe", "rozpadająca", "poty", "wściekłość", "punkcja", "łuski", "duże", "przymglone", "zmniejszceni", "rogowiec", "kolka", "przeczos", "kolki", "płacz", "krosta", "strup", "kiła", "ostre", "wady", "polipy", "sine", "ptsd", "usnąć", "dreszcze", "dokrwiona", "trudność", "przestawić", "oziębienie", "stres", and "schizofrenia". Furthermore, we identified errors that were mostly repeated by all three ASR systems, i.e.: "świąd", "krosty", "ból", "poronienie", "problem z", "białka", "kłykciny", "urojenia", "omamy", and "płonica".

5. Summary and conclusions

In this paper, after conducting a literature review and selecting three ASR systems among the leading systems on the Polish market, we compared the performance of Google ASR, Microsoft ASR, and Techmo ASR systems in recognising colloquial Polish speech related to medical topics. The results reveal a speech recognition Accuracy of over 86% exhibited by all three ASR systems, with a discrepancy of only 1.7% between the best and worst result. In the comparative analysis using not only Accuracy measure, but also WER, Levenshtein editing distance algorithm and Jaro-Winkler similarity of recognized expressions, the best, average and worst result remained the same, subsequently: Google ASR, Techmo ASR and Microsoft ASR, regardless of the measure used.

As there were only minor differences in overall speech recognition performance among all ASR systems, thus the possibility of additional training the speech recognition model of the individual ASR system and the possibility of its on-premise installation became critical considerations. We found that only two ASR systems met the raised requirements, however in both cases partially: Microsoft ASR and Techmo ASR. The first has no possibility of on-premise installation, and the latter offers a paid service provided by the Techmo team for an additional training. We hereby concluded, that it is beneficial to use both Microsoft ASR and Techmo ASR systems for medical interviews conducted in Polish language, and the final choice of the system depends on the very specific needs of the end customer, namely the preference for on-premise installation or the need to additionally train the speech recognition model.

The presented study may efficiently support the selection of the appropriate ASR system in the medical sector. It indicates further directions to automatically conduct initial diagnostics and medical interviews in healthcare units with minimal human support, which can be highly beneficial especially during epidemics and/or with limited organizational resources. We hope that our findings will prove valuable to consumers, beneficiaries, and designers of dedicated ASR solutions, especially in terms of recognising Polish colloquial speech in the conversational systems and voice interfaces used depending on the identified needs of a given sector of operation.

6. Acknowledgements

This work was co-financed by the Polish National Center for Research and Development (NCBiR), and presents partial results of a research project entitled “Development of a system using machine learning to automatically conduct initial diagnostics and medical interviews in healthcare units” carried out by SOVVA SA.

References

[1] Altar H. S., Sison R.C. (2019). Medical Transcriptionist’s Experience with Speech Recognition Technology, Proceedings of Australasian Conference on Information Systems (ACIS 2019), Perth, Western Australia, p. 915 - 924.

[2] Behrman A. (2017). A Clear Speech Approach to Accent Management. American journal of speech-language pathology, vol. 26, no. 4, p. 1178 - 1192.

[3] Black P. E. (2022). Jaro-Winkler, [in:] Dictionary of Algorithms and Data Structures [online], Black P. E. (ed.), 2022, https://www.nist.gov/dads/HTML/jaroWinkler.html [accessed 05.2023]

[4] Blackley S. V., Huynh J., Wang L., Korach Z., Zhou L. (2019). Speech recognition for clinical documentation from 1990 to 2018: a systematic review, Journal of the American Medical Informatics Association (JAMIA), vol. 26, no. 4, p. 324–338.

[5] Dragon STT, https://www.nuance.com/en-gb/dragon.html [accessed 05.2023]

[6] Fareez F., Parikh T., Wavell C., Shahab S., Chevalier M., Good S., De Blasi I., Rhouma R., McMahon C., Lam J. P., Lo T., Smith C. W. (2022). A dataset of simulated patient-physician medical interviews with a focus on respiratory cases, Scientific Data, vol. 9, no. 313, p. 1 - 7.

[7] Georgila K., Leuski A., Yanov V., Traum D. (2020). Evaluation of Off-the-shelf Speech Recognizers Across Diverse Dialogue Domains. Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020), Marseille, France, p. 6469 - 6476.

[8] Google ASR, https://cloud.google.com/speech-to-text [accessed 05.2023]

[9] Joseph J., Moore Z. E. H., Patton D., O'Connor T., Nugent L.E. (2020). The impact of implementing speech recognition technology on the accuracy and efficiency (time to complete) clinical documentation by nurses: A systematic review, Journal of Clinical Nursing, vol. 29, issue 13-14, p. 2125 - 2137.

[10] Iancu B. (2019). Evaluating Google Speech-to-Text API's Performance for Romanian e-Learning Resources. Informatica Economica, vol. 23, no. 1/2019, p. 17 - 25.

[11] Kim J. Y., Liu C., Calvo R. A., McCabe K., Taylor S. C. R., Schuller B. W., Wu K. (2019). A Comparison of Online Automatic Speech Recognition Systems and the Nonverbal Responses to Unintelligible Speech, Preprint arXiv, Computing Research Repository (CoRR), doi.org/10.48550/arXiv.1904.12403, p. 1 - 13.

[12] Kim J. Y., Liu C., Calvo R. A., McCabe K., Taylor S. C. R., Schuller B. W., Wu K. (2022). Comparison of Automatic Speech Recognition Systems, [in:] Stoyanchev S., Ultes S., Li H. (eds.), Conversational AI for Natural Human-Centric Interaction, Lecture Notes in Electrical Engineering, vol. 943, Springer, p. 123 - 131.

[13] Kuligowska K., Kisielewicz P., Włodarz A. (2018). Wady i ograniczenia systemów rozpoznawania mowy, Roczniki Kolegium Analiz Ekonomicznych, nr 49/2018, Szkoła Główna Handlowa, Warszawa, p. 307 - 317.

[14] Lugosch L., Ravanelli M., Ignoto P., Tomar V., Bengio Y. (2019). Speech Model Pre-training for End-to-End Spoken Language Understanding, Preprint arXiv, Audio and Speech Processing (eess.AS), doi.org/10.48550/arXiv.1904.03670, p. 1 - 5.

[15] Mah P.M., Skalna I., Muzam J. (2022). Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0, Applied Sciences, vol. 12, no. 18: 9207, p. 1 - 26.

[16] Microsoft ASR, https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text [accessed 05.2023]

[17] Morbini F., Audhkhasi K., Sagae K., Artstein R., Can D., Georgiou P., Narayanan S., Leuski A., Traum D. (2013). Which ASR should I choose for my dialogue system? Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2013), Metz, France, p. 394 - 403.

[18] Niewiarowski A., Stanuszek M. (2013). Mechanizm analizy podobieństwa krótkich fragmentów tekstów, na bazie odległości Levenshteina, Studia Informatica, Vol. 34, No. 1 (110), Politechnika Krakowska, Instytut Informatyki, p. 108.

[19] Phonexia STT, https://www.phonexia.com/product/speech-to-text/ [accessed 05.2023]

[20] Repka A. (2022). Chatboty w służbie e-zdrowia – ewolucja telemedycyny w stronę konwersacyjnej sztucznej inteligencji, [w:] Karolina Kuligowska (red.), Chatboty w informatyce ekonomicznej: implementacja, miary, zastosowania, Laboratorium Wiedzy Artur Borcuch, Kielce, p. 104.

[21] Rev AI ASR, https://www.rev.ai/ [accessed 05.2023]

[22] Saxena K., Diamond R., Conant R. F., Mitchell T. H., Gallopyn G., Yakimow K. E. (2018). Provider Adoption of Speech Recognition and its Impact on Satisfaction, Documentation Quality, Efficiency, and Cost in an Inpatient EHR, AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science, 2017, p. 186 - 195.

[23] Techmo ASR, https://techmo.pl/technologia/rozpoznawanie-mowy [accessed 05.2023]

[24] Tomar V., Desruisseaux M., Seetzen H. (2020). System and method for implementing a vocal user interface by combining a speech to text system and a speech to intent system, https://patentimages.storage.googleapis.com/1d/05/1d/014c820a9a7b7b/US10878807.pdf [accessed 05.2023]

[25] Tomar V. (2021). How Speech Technology Is Optimizing Factory Lines, https://industrytoday.com/how-speech-technology-is-optimizing-factory-lines/ [accessed 05.2023]

[26] Vinnarasu A., Jose D. V. (2019). Speech to text conversion and summarization for effective understanding and documentation. International Journal of Electrical and Computer Engineering (IJECE), vol. 9, issue 5, p. 3642 - 3648.

[27] Whisper ASR, https://openai.com/research/whisper [accessed 05.2023]

[28] Yao X., Bhutada P., Georgila K., Sagae K., Artstein R., Traum D. (2010). Practical evaluation of speech recognizers for virtual human dialogue systems. Proceedings of the 7th International Conference on Language Resources and Evaluation , Valletta, Malta, p. 1597 - 1602.