Automatyczne rozpoznawanie mowy potocznej na potrzeby wywiadów medycznych

Karolina Kuligowska[1], *Maciej Stanusch*[2], *Marek Koniew*[3]

Wprowadzenie

Rozpoznawanie mowy, znane w języku angielskim jako *speech recognition*, to zdolność programu informatycznego do identyfikowania słów w języku mówionym, a następnie przekształcania ich w czytelny dla człowieka format tekstowy. Najnowsze postępy w dziedzinie sztucznej inteligencji i uczenia maszynowego, a także rosnąca moc obliczeniowa i pojemność pamięci komputerów umożliwiły powstanie szerokiej gamy systemów do rozpoznawania ludzkiej mowy. Znane są one najczęściej pod angielskim terminem *Speech-To-Text* (STT) albo *Automatic Speech Recognition* (ASR). Przykłady takich rozwiązań stanowią systemy: Google w chmurze Google Cloud [Google ASR], Microsoft w chmurze Microsoft Azure [Microsoft ASR], Dragon w chmurze Nuance Communications [Dragon STT], Phonexia [Phonexia STT], Techmo [Techmo ASR], Rev AI [Rev AI ASR] i inne.

Systemy ASR mogą tworzyć wysokiej jakości transkrypcje, jednakże charakteryzują się różnym poziomem jakości przekładania mowy na tekst, zarówno pod względem stopnia dokładności (*accuracy*), jak i współczynnika błędów w słowie (*word error rate*, WER). Ich dokładność, wyrażająca procentowo udział poprawnie rozpoznanych słów danej wypowiedzi na tekst pisany, waha się od 70% do 95%, a ich WER waha się od 8,53% do 34,90% [Georgila i in. 2020]. Taka rozpiętość w wydajności rozpoznawania mowy ludzkiej wynika z tego, że wskaźniki błędów systemu ASR mogą potencjalnie wynikać z różnych przyczyn poza samą zrozumiałością mowy. Szybkość i dokładność automatycznego przebiegu procesu rozpoznawania mowy zależą od wielu kluczowych czynników, takich jak chociażby [Behrman 2017; Kim i in. 2019; Kuligowska, Kisielewicz, Włodarz 2018]:

[1] Uniwersytet Warszawski, kkuligowska@wne.uw.edu.pl
[2] SOVVA SA, ms@sovva.ai
[3] SOVVA SA, marek@koniew.pl

- rodzaj języka naturalnego, w którym realizowana jest wypowiedź,
- zakres tematyki wypowiedzi (np. czy dotyczy słownictwa potocznego, czy specjalistycznego),
- pochodzenie etniczne i akcent mówcy,
- potencjalne zakłócenia akustyczne w tle wypowiedzi,
- zastosowane podejścia, modele i algorytmy rozpoznawania mowy, rozmiar słownika oraz specyfikacja gramatyki.

Istniejące systemy ASR z różnym skutkiem − pod względem finalnej jakości transkrypcji − przezwyciężają problemy dotyczące rozpoznawania języka potocznego, ewentualnie gwar lub dialektów, zakresu tematyki wypowiedzi i różnorodności słownictwa [Wahyutama, Hwang 2023], a także wypowiedzi reprezentujących zróżnicowane właściwości brzmieniowe mowy, wypowiedzi niepoprawnych, zbyt długich, wreszcie wypowiedzi zawierających pauzy, powtórzenia i nagłe przerwy [Iancu 2019; Vinnarasu, Jose 2019].

Napędzany inteligentnymi i autonomicznymi systemami postęp technologiczny w Przemyśle 4.0, który rewolucjonizuje tradycyjne praktyki produkcyjne i przemysłowe, wymusza przeplatanie się różnych systemów przetwarzania języka naturalnego oraz mowy potocznej i zaczyna czerpać z ekspansji najnowszych rozwiązań w środowiskach przemysłowych [Mah, Skalna, Muzam 2022]. Zastosowanie ASR w inżynierii produkcji obarczone jest kolejnymi wyzwaniami choćby dlatego, że hale produkcyjne są głośne, więc wdrożona w nich technologia musi być zdolna do odpornego na hałas i cechującego się niskimi opóźnieniami rozpoznawania mowy [Lugosch i in. 2019]. Powinna być również intuicyjna w obsłudze i skalowalna, aby można ją było łatwo wdrożyć w fabrykach na innych rynkach i w innych regionach świata, dla rozpoznania różnych języków i akcentów pracowników fabryki. Takiego zastosowania systemu ASR w inżynierii produkcji dokonała firma BSH, europejski producent sprzętu AGD, która wprowadziła do jednej ze swoich niemieckich fabryk rozwiązanie do sterowania głosowego w trybie offline [Tomar 2021]. Rozwiązanie to wykorzystuje opatentowaną przez kanadyjską firmę Fluent.ai technologię rozpoznawania mowy w hałaśliwym otoczeniu [Tomar, Desruisseaux, Seetzen 2020]. Przytoczony przykład ukazuje tylko jedną z wielu możliwości zastosowań systemów ASR w inżynierii produkcji i obsługi maszyn fabrycznych. Wydawanie głosowych komend przez pracowników linii produkcyjnych zwiększa możliwości wspierania wydajności obsługi linii montażowej i ergonomii hali produkcyjnej oraz eliminuje potrzebę fizycznego manipulowania urządzeniami sterującymi. Zaprezentowane w artykule badanie może wspierać decydentów zarządzania produkcją w wyborze właściwego rozwiązania ASR w zależności od zidentyfikowanych potrzeb danego sektora funkcjonowania, w tym wypadku − pod kątem rozpoznawania polskiej mowy potocznej w sektorze medycznym.

Wykonana przez autorów analiza literaturowa wykazała nieliczny zasób opracowań dotyczących porównania jakości różnych rozwiązań ASR w ogólnych zastosowaniach medycznych, w szczególności zaś dotkliwy brak opracowań dotyczących

porównania rozwiązań ASR dostępnych na polskim rynku z perspektywy wywiadów medycznych i z uwzględnieniem rozmówców posługujących się polskim językiem potocznym. Choć można przytoczyć przykłady już zaimplementowanych rozwiązań ASR w zagranicznych systemach opieki zdrowotnej, zarówno z sukcesem wspierających samych pacjentów, jak i procesy administracyjne czy zarządzanie wewnątrzszpitalne, to najczęściej są one opisywane przez komercyjne źródła w postaci np. raportów firm lub tzw. *white papers* aniżeli analizowane w badaniach naukowych [Repka 2022].

Dlatego też cel niniejszego badania stanowi przeprowadzenie analizy porównawczej trzech wiodących na polskim rynku rozwiązań ASR (Google, Microsoft, Techmo), a następnie wybór spośród nich takiego rozwiązania, które jest optymalne pod kątem rozpoznawania polskiej mowy potocznej w tematyce wywiadów medycznych.

W artykule przedstawiono przegląd dostępnej literatury w kontekście systemów rozpoznawania mowy z perspektywy ogólnych zastosowań medycznych, opisano metodologię przyjętą na potrzeby realizacji niniejszego badania, a także przebieg badań. Następnie zaprezentowano wyniki badań, podsumowanie i wnioski.

Systemy rozpoznawania mowy w zastosowaniach medycznych — przegląd literatury

Niezmiennie intensywny rozwój technologii rozpoznawania mowy sprawia, że systemy ASR są coraz bardziej dokładne i popełniają coraz mniej błędów w trakcie działania. Równocześnie jednak istniejące systemy rozpoznawania głosu nie są sobie równe, a wiele z nich ma trudności z wierną transkrypcją niektórych słów.

Naukowcy z University of South California przeprowadzili w 2020 r. badanie porównujące stopień dokładności kilku publicznie dostępnych, komercyjnych i naukowych, systemów ASR dla sześciu różnych tematycznych dziedzin dialogowych na siedmiu różnych zestawach danych głosowych w języku angielskim [Georgila i in. 2020]. Porównali systemy ASR wykonane przez Amazon, Apple, Google, IBM, Microsoft i Kaldi, przy czym wszystkie zestawy danych głosowych składały się z nagrań mowy spontanicznej rzeczywistych rozmów. Głównym miernikiem oceny był współczynnik błędów w słowie (WER). Otrzymane przez kalifornijskich naukowców wyniki wskazują, że obecnie nawet najnowocześniejsze systemy do rozpoznawania mowy ujawniają duży odsetek błędów transkrypcji w dziedzinach wymagających rozpoznawania dedykowanego słownictwa lub spontanicznej mowy potocznej.

Porównanie tych wyników z badaniami lat poprzednich [Morbini i in. 2013; Yao i in. 2010] na podobnych zestawach danych pokazuje, że w ciągu ostatnich kilku lat nastąpił duży postęp w technologii ASR, zwłaszcza w zakresie wykorzystania technik głębokiego uczenia. W swoich wynikach badań naukowcy podkreślają równocześnie, jak ważne jest rozważenie dziedziny zastosowania i kontekstu wykorzy-

stania rozpoznawania mowy w celu wybrania najodpowiedniejszego rozwiązania.

W dotychczasowych badaniach systemów ASR dla zastosowań medycznych literatura koncentrowała się głównie na przetwarzaniu elektronicznej dokumentacji klinicznej: notatek głosowych lekarzy i pielęgniarek oraz zapisów głosowych streszczających wypisy ze szpitali [Altar, Sison 2019; Joseph i in. 2020]. Ponieważ medyczna dokumentacja głosowa wspomagana systemami transkrypcji jest coraz bardziej powszechna, nastąpił jednocześnie wzrost badań nad wpływem technologii ASR na pracę klinicystów i praktykę medyczną [Saxena i in. 2018]. Jednak wraz z rozwojem rozwiązań chmurowych naukowcy zaczęli także badać możliwości automatycznego przetwarzania mowy na potrzeby transkrypcji konwersacji w trakcie wywiadów medycznych w czasie rzeczywistym. Na potrzeby takich badań tworzono hipotetyczne scenariusze rozmów lekarzy z tzw. symulowanymi pacjentami (*simulated patients*), czyli aktorami przeszkolonymi do odgrywania roli pacjentów w sytuacji medycznej [Blackley, Huynh, Wang 2019; Fareez, Parikh, Wavell 2022].

W najnowszym badaniu systemów ASR w dziedzinie medycyny zespół naukowców pod przewodnictwem Kim, Liu i Calvo [2022] przetestował pięć internetowych platform pod kątem jakości transkrypcji wbudowanych w nie systemów ASR. Porównali oni chmurowe systemy ASR: Google Cloud, IBM Watson, Microsoft Azure, Trint oraz YouTube, natomiast zestawy danych głosowych składały się z interakcji 12 studentów medycyny z 2 symulowanymi pacjentami (łącznie 12 diadycznych telekonsultacji medycznych). Zgodnie ze wstępnymi hipotezami badaczy transkrypcje ręczne okazały się znacznie dokładniejsze niż transkrypcje automatyczne, a wśród systemów ASR automatyczne transkrypcje napisów YouTube'a znacznie przewyższały pozostałe chmurowe rozwiązania pod kątem dokładności rozpoznania wyrazów.

Metodologia badawcza

Niniejszy artykuł prezentuje cząstkowe wyniki prac badawczych realizowanych przez spółkę SOVVA SA w ramach projektu *Opracowanie systemu wykorzystującego uczenie maszynowe do automatycznego przeprowadzania wstępnej diagnostyki i wywiadów medycznych w jednostkach służby zdrowia* współfinansowanego ze środków Narodowego Centrum Badań i Rozwoju, którego celem jest istotne ulepszenie posiadanego przez beneficjenta systemu (platformy OCP) umożliwiającego automatyczną obsługę klientów (bez wsparcia człowieka lub z jego minimalnym udziałem) poprzez stworzenie funkcjonalności istotnych dla potrzeb podmiotów działających na rynku służby zdrowia w warunkach pandemicznych/epidemicznych i/lub przy ograniczonych zasobach organizacyjnych. Docelowy system będzie mógł przeprowadzić wstępny wywiad medyczny (*pre-triage*) oraz zidentyfikować intencję kontaktu pacjenta, posługującego się językiem naturalnym (potocznym), przyspieszając proces rejestracji pacjenta oraz ułatwiając zbieranie danych niezbędnych do świadczenia usługi medycznej.

W ramach przedmiotowego projektu jednym z zadań zespołu badawczego był „Wybór optymalnego rozwiązania przetwarzającego mowę na tekst" spośród dostępnych na polskim rynku rozwiązań z perspektywy przyszłego zastosowania, jakim jest zamiana mowy na tekst zagadnień związanych z medycyną przy założeniu częstego posługiwania się językiem potocznym przez rozmówców.

Na podstawie przeglądu literatury można stwierdzić, że jakość systemów rozpoznawania mowy (STT, ASR) jest różnorodna w zależności od rodzaju mowy, którą należy zamienić na postać tekstową, oraz że nie istnieją aktualne opracowania naukowe dotyczące jakości ww. rozwiązań w przedmiocie niniejszego projektu (wyrażenia medyczne wypowiadane w języku potocznym).

Głównymi celami badań przedmiotowego projektu, a jednocześnie głównymi celami niniejszego artykułu są:

- weryfikacja stopnia dokładności rozpoznawania mowy wiodących na polskim rynku rozwiązań ASR na potrzeby przedmiotowego projektu poprzez wykonanie analizy komparatystycznej na określonej próbce danych,
- weryfikacja możliwości dotrenowania dostępnych rozwiązań ASR,
- weryfikacja możliwości instalacji dostępnych rozwiązań ASR w infrastrukturze klienta końcowego (instalacja typu on-premise).

Analizę komparatystyczną wykonano każdorazowo na tym samym zbiorze wypowiedzi testowych, co pozwoliło wyłonić najlepsze rozwiązanie dla celów realizacji projektu. Ocenę jakości badanych rozwiązań wykonano, stosując algorytm odległości edycyjnej Levenshteina, wyliczający najmniejszą możliwą liczbę zmian, które należy wykonać na dwóch porównywalnych ciągach znaków, by osiągnąć ich identyczność poprzez wstawienie znaku, zamianę znaku na inny, usunięcie znaku. Im niższa jest odległość edycyjna między dwoma ciągami, tym bardziej podobne są one do siebie [Niewiarowski, Stanuszek 2013].

Biorąc pod uwagę wiele czynników, takich jak popularność narzędzi w praktyce badawczej, ich skalowalność, a także jakość weryfikowaną poprzez analizę upublicznionych danych technicznych oraz danych wejściowych i wyjściowych realizowanych projektów, jak również dostępność samych systemów ASR, autorzy podjęli decyzję o weryfikacji trzech dostępnych na polskim rynku rozwiązań zamieniających mowę na tekst (ASR), a mianowicie:

- systemu ASR firmy Google (producent Google Inc.),
- systemu ASR firmy Microsoft (producent Microsoft Inc.),
- systemu ASR firmy Techmo (producent Techmo sp. z o.o.).

W celu realizacji przedstawionych celów badawczych zostały wykonane następujące prace poprzedzające właściwą analizę komparatystyczną wyżej wymienionych systemów ASR:

- Wykonanie interfejsu programistycznego aplikacji (*Application Programming Interface*, API) do wyżej wymienionych systemów ASR w celu umożliwienia przeprowadzenia testów poszczególnych rozwiązań.

- Weryfikacja możliwości trenowania trzech wyżej wymienionych systemów ASR pod kątem uzupełnienia modelu rozpoznawania mowy o specyfikę języka potocznego w zastosowaniach medycznych.
- Weryfikacja możliwości instalacji trzech wyżej wymienionych systemów ASR w infrastrukturze nabywcy (jako oprogramowanie lokalne, *on-premise*).
- Stworzenie pliku testowego (datasetu) 1000 zagadnień medycznych, sformułowanych w języku potocznym (w formie tekstowej), w celu przeprowadzenia właściwej analizy komparatystycznej trzech wyżej wymienionych systemów ASR, a więc pod kątem zdolności do rozpoznawania pytań w języku potocznym.

Plik testowy, niezbędny do przeprowadzenia właściwej analizy komparatystycznej, został przygotowany na podstawie posiadanego przez SOVVA (spółkę realizującą projekt badawczy) zestawu 19 tysięcy skategoryzowanych wyrażeń medycznych. Spośród nich zatrudnieni w rolach ekspertów lekarze wyłonili 1000 najpopularniejszych wg nich zagadnień używanych przez pacjentów w trakcie wizyt w gabinetach lekarskich.

Powyższy zestaw wyrażeń medycznych został nagrany w postaci 1000 plików dźwiękowych i umieszczony na przygotowanym do dalszych testów serwerze FTP. Pliki z nagraniami zostały stworzone w oparciu o następujące reguły:

- Każdy plik zawierał nagranie pojedynczej frazy (wyrażenia/słowa),
- Pliki zapisane zostały w formacie WAV o następujących parametrach: częstotliwość próbkowania: 16 000 Hz, kanały: 1 (mono), format próbki: PCM, 16-bitowy,
- Do każdego zestawu nagrań stworzono plik TXT z mapowaniem: nazwa_pliku -> nagrana_fraza, np. 001.wav -> osteoporoza lub 002.wav -> nadżerka.

Opracowane w powyższy sposób pliki zostały następnie przeprocesowane przez poszczególne systemy ASR za pomocą wcześniej przygotowanego interfejsu API. Następnie dokonano analizy jakości otrzymanych transkrypcji. Do analizy tej wykorzystano algorytm obliczania odległości Levenshteina.

Wyniki badań

Przedstawione wyniki stanowią podsumowanie zrealizowanych celów badawczych wg założeń projektu, a także niniejszego artykułu. Wyniki w obszarze możliwości dotrenowania poszczególnych systemów ASR oraz instalacji w infrastrukturze klienta (*on-premise*) stanowią zestawienie informacji pozyskanych na podstawie analizy dokumentacji danego rozwiązania ASR, jak i weryfikacji w postaci prób dotrenowania systemów i ich instalacji *on-premise*. Natomiast wyniki analizy komparatystycznej są sumą zrealizowanych prac wymienionych wcześniej. W tablicy 1 zaprezentowano wyniki badań w obszarze możliwości dotrenowania poszczególnych systemów ASR oraz instalacji w infrastrukturze klienta (*on-premise*).

TABLICA 1. Wyniki badań w obszarze możliwości dotrenowania poszczególnych systemów ASR oraz instalacji w infrastrukturze klienta (*on-premise*)

Rozwiązanie	Google ASR	Microsoft ASR	Techmo ASR
Możliwość dotrenowania	Nie	Tak	Tak/nie*
Możliwość instalacji *on-premise*	Nie	Nie	Tak

* W p rzypadku rozwiązania Techmo ASR istnieje możliwość dotrenowania modeli wyłącznie jako usługa płatna realizowana przez zespół Techmo. Użytkownik produktu nie ma możliwości samodzielnego dotrenowania modelu.

Ź r ó d ł o: Opracowanie własne.

Jak widać w zestawieniu w tablicy 1, tylko rozwiązanie Techmo ASR spełnia oba założenia jednocześnie, a mianowicie możliwość dotrenowania systemu (choć jedynie w wersji płatnej, jako usługa realizowana przez zespół Techmo) oraz instalacji w infrastrukturze klienta (*on-premise*). Rozwiązanie Microsoft ASR co prawda oferuje możliwość dotrenowania systemu, ale nie można go zainstalować lokalnie. Natomiast system Google ASR nie spełnia żadnego z powyższych założeń. Dodatkowo sprawdzono, że w przypadku systemu Google ASR możliwe jest polepszenie wykrywania określonej frazy lub fraz poprzez sterowanie słownikami. System Google ASR udostępnia predefiniowane słowniki, które pozwalają na lepsze rozumienie fraz i otrzymywanie tekstu w konkretnym formacie. System Microsoft ASR nie daje żadnej z powyższych możliwości. Odnotowano brak danych dotyczących możliwości systemu Techmo ASR w tym zakresie.

W tablicy 2 podsumowano analizę komparatystyczną jakości rozpoznawania mowy przez poszczególne systemy ASR w postaci procentowej.

TABLICA 2. Podsumowanie analizy komparatystycznej jakości rozpoznawania mowy przez poszczególne systemy ASR

Rozwiązanie	Google ASR	Microsoft ASR	Techmo ASR
Stopień dokładności rozpoznania mowy	88,1%	86,4%	87,5%

Ź r ó d ł o: Opracowanie własne.

Otrzymane rezultaty wskazują na relatywnie niewielkie różnice pomiędzy poszczególnymi rozwiązaniami: przy próbie n = 1000 pomiędzy najgorszym a najlepszym rozwiązaniem różnica w stopniu dokładności rozpoznania mowy wynosi zaledwie 1,7%. Należy jednak zwrócić uwagę, iż powyższe rezultaty badań zostały otrzymane w warunkach laboratoryjnych, bez uwzględnienia szumów zewnętrznych oraz zakłóceń, które mogą powstać w przypadku prowadzenia rozmowy w niewygłuszonym pomieszczeniu lub przez telefon.

Podsumowanie

Na polskim rynku jak dotąd nie ma rozwiązania spełniającego jednocześnie wszystkie trzy wymagania, które zostały określone w projekcie i na podstawie których powstał niniejszy artykuł. Były nimi: najwyższy możliwy stopień dokładności rozpoznawania fraz, możliwość samodzielnego dotrenowania modeli oraz możliwość instalacji *on-premise*. Na podstawie dokonanego przeglądu literatury i wyboru wiodących na polskim rynku rozwiązań autorzy porównali w trzech wyżej wymienionych wymaganiach silniki Google ASR, Microsoft ASR i Techmo ASR w rozpoznawaniu polskiej mowy potocznej w tematyce medycznej. Ze względu na niewielkie różnice w stopniu dokładności rozpoznawania mowy przez poszczególne systemy większe znaczenie zyskują pozostałe wymagania: możliwość samodzielnego dotrenowania systemu ASR oraz instalacji *on-premise*.

Trzy wymagania określone w projekcie częściowo spełniają systemy: Microsoft ASR (poza możliwością instalacji systemu *on-premise*) oraz Techmo ASR (choć możliwość dotrenowania systemu wiąże się z opłatą oraz może zostać zrealizowana jedynie poprzez zespół Techmo). Można zatem wysunąć wniosek, że korzystne jest użycie dwóch różnych podsystemów ASR, a ostateczny wybór rozwiązania jest każdorazowo uzależniony od wymagań klienta końcowego (np. konieczność instalacji *on-premise* lub chęć korzystania z usługi chmurowej typu SaaS, a także potrzeba opcji dotrenowania systemu).

Potrzeba automatycznych metod zbierania i dokumentowania danych, w tym również głosowych, które wspierają procesy poznawcze klinicystów oraz usprawniają przepływ informacji pośród zespołów medycznych, a także poprawiają komfort pacjenta podczas udzielanych świadczeń, staje się centralnym elementem rozwoju nowoczesnego systemu opieki zdrowotnej. Mamy zatem nadzieję, że przeprowadzona analiza komparatystyczna okaże się przydatna dla konsumentów, beneficjentów oraz projektantów dedykowanych systemów ASR w tematyce wywiadów medycznych, w szczególności systemów konwersacyjnych oraz interfejsów głosowych zastosowanych w branży medycznej.

Bibliografia

Altar, H.S., Sison, R.C. [2019]. *Medical Transcriptionist's Experience with Speech Recognition Technology*. Proceedings of Australasian Conference on Information Systems (ACIS 2019), 915−924.

Behrman, A. [2017]. A Clear Speech Approach to Accent Management. *American Journal of Speech-Language Pathology*, 26(4), 1178−1192.

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*(4), 324–338.

Dragon STT [2022]. https://www.nuance.com/en-gb/dragon.html

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, 9*(313), 1−7.

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), 6469−6476.

Google ASR [2022]. https://cloud.google.com/speech-to-text

Iancu, B. [2019]. Evaluating Google Speech-to-Text API's Performance for Romanian e-Learning Resources. *Informatica Economica, 23*(1), 17−25.

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, 29*(13−14), 2125−2137.

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), 1−13. doi.org/10.48550/arXiv.1904.12403

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 S. Stoyanchev, S. Ultes, H. Li (eds.), *Conversational AI for Natural Human-Centric Interaction, Lecture Notes in Electrical Engineering, 943* (pp. 123−131), Springer.

Kuligowska, K., Kisielewicz, P., Włodarz, A. [2018]. Wady i ograniczenia systemów rozpoznawania mowy. *Roczniki Kolegium Analiz Ekonomicznych, 49*, 307−317, Szkoła Główna Handlowa.

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), 1−5. doi.org/10.48550/arXiv.1904.03670

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, 12*(18), 1−26.

Microsoft ASR [2022]. https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text

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), 394−403.

Niewiarowski, A., Stanuszek, M. [2013]. Mechanizm analizy podobieństwa krótkich fragmentów tekstów, na bazie odległości Levenshteina. *Studia Informatica, 34*(1/110), 108, Politechnika Krakowska.

Phonexia STT [2022]. https://www.phonexia.com/product/speech-to-text/

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* (s. 104). Laboratorium Wiedzy Artur Borcuch.

Rev AI ASR [2022]. https://www.rev.ai/

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, 186−195.

Techmo ASR [2022]. https://techmo.pl/technologia/rozpoznawanie-mowy

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

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

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), 9*(5), 3642−3648.

Wahyutama, A.B., Hwang, M. [2023]. Auto-Scoring Feature Based on Sentence Transformer Similarity Check with Korean Sentences Spoken by Foreigners. *Applied Sciences 13, 373*(1), 1−16.

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 (LREC 2010), 1597–1602.