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

Zastosowanie automatycznego rozpoznawania mowy w transkrypcji wywiadów medycznych – porównanie silników ASR dla języka polskiego

1. Wstęp

Technologia rozpoznawania mowy (ang. *Automatic Speech Recognition*, ASR, ang. *Speech-To-Text*, STT), posiada liczne zastosowania, od usług dyktowania i transkrypcji notatek głosowych, przez interaktywne systemy odpowiedzi głosowej, po asystentów wirtualnych. Technologia ta ewoluuje w szybkim tempie, czego efektem są coraz dokładniejsze systemy ASR z mniejszą liczbą błędów w procesie rozpoznawania. Warto jednak zauważyć, że jakość obecnych systemów rozpoznawania mowy znacznie się różni i niektóre z nich wciąż mają trudności z dokładną transkrypcją wybranych słów.

Systemy ASR obecne na rynku światowym poczyniły znaczne postępy w ostatnich latach pod kątem jakości rozpoznawania mowy. Najpopularniejsze przykłady takich rozwiązań obejmują aktualnie Google Cloud z silnikiem Google ASR [1], Microsoft Azure Cloud z silnikiem Microsoft ASR [2], Nuance Communications Cloud z silnikiem Dragon STT [3], Phonexia z silnikiem Phonexia STT [4], Techmo z silnikiem Techmo ASR [5], Rev AI z silnikiem Rev AI ASR [6]. Są one w stanie dobrze rozpoznawać codzienny język potoczny, wyrażenia idiomatyczne, kolokwializmy i regionalne akcenty, wypowiedzi bogate w słownictwo, a także przetwarzać nagrania z mową błędną lub wadliwą, o różnych właściwościach akustycznych [7, 8, 9-12]. Warto przy tym zauważyć, że na dokładność rozpoznania mowy i liczbę błędnych rozpoznań wyrazów przez systemy ASR mogą wpływać różne czynniki, poza samą klarownością mowy. Czynniki te obejmują [13, 14-16, 17]: rodzaj języka naturalnego używanego w mowie, zastosowane słownictwo (potoczne lub specjalistyczne), akcent mówcy i pochodzenie etniczne, hałas w tle i zakłócenia akustyczne, a także architekturę rozwiązań użytych w rozpoznawaniu mowy (słowniki, implementacja gramatyki itp.).

Przeprowadzona przez autorów analiza literatury wykazała brak badań dotyczących oceny jakości dostępnych na polskim rynku rozwiązań ASR pomagających w przeprowadzaniu wywiadów medycznych w potocznym języku polskim. Może to wynikać z faktu, iż rozwiązania ASR w zagranicznych podmiotach branży medycznej, które poprawiły jakość opieki nad pacjentami i zarządzanie szpitalem, są najczęściej prezentowane w źródłach komercyjnych w formie raportów firmowych, wywiadów eksperckich lub tzw. *white papers* [18]. Istnieje zatem potrzeba badań porównawczych aktualnego działania oraz potencjału funkcjonowania systemów ASR w celu

[1] kkuligowska@wne.uw.edu.pl, Wydział Nauk Ekonomicznych, Uniwersytet Warszawski, www.wne.uw.edu.pl
[2] ms@sovva.ai, SOVVA SA, http://sovva.ai
[3] marek@koniew.pl, SOVVA SA, http://sovva.ai

stymulacji pozytywnych zmian w krajowym podejściu do świadczenia opieki zdrowotnej.

Celem niniejszego badania jest zapełnienie wyżej opisanej luki badawczej przez przeprowadzenie analizy porównawczej wiodących na polskim rynku rozwiązań ASR. Kryteria oceny tych rozwiązań oparto na stopniu dokładności konwersji wypowiedzianych fraz w języku polskim na format tekstowy, przy użyciu miar dokładności dla poprawnie rozpoznanych słów (ang. *accuracy*) i wskaźnika błędnie rozpoznanych słów (ang. *Word Error Rate*, WER) oraz miar Levenstheina i Jaro-Winklera. Wyłonienie najlepszego z tych rozwiązań pozwoli w niedalekiej przyszłości na jego integrację z Platformą Omni-Chatbot (OCP) opracowaną przez SOVVA SA i udoskonalenie systemu automatycznej obsługi klienta. Docelowy system firmy SOVVA SA, przetwarzając naturalnie wypowiadany potoczny język polski, ma potrafić przeprowadzić wstępny głosowy wywiad lekarski (ang. *pre-triage*) i dzięki temu zidentyfikować przyczyny kontaktu pacjenta z placówką medyczną.

Artykuł jest zorganizowany w następujący sposób: w Rozdziale 2 przedstawiono przegląd dostępnej literatury nad badaniami w zastosowaniach systemów rozpoznawania mowy w sektorze medycznym. W Rozdziale 3 opisano metodologię przyjętą do realizacji niniejszych badań, a także ich przebieg. W Rozdziale 4 przedyskutowano wyniki badań. Podsumowanie i wnioski zawarto w Rozdziale 5.

2. Przegląd literatury dla systemów ASR w zastosowaniach medycznych

Wykorzystanie technologii ASR w medycynie staje się w ostatnich latach coraz popularniejszym obszarem badań. Jednym z wiodących zastosowań tej technologii jest przetwarzanie elektronicznej dokumentacji klinicznej, która obejmuje notatki głosowe nagrane przez pracowników służby zdrowia oraz zapisy głosowe podsumowujące wypisy ze szpitala [19, 20]. Dokumenty te często zawierają złożoną terminologię medyczną i wymagają dużej dokładności transkrypcji, aby zapewnić pacjentom najlepszą możliwą opiekę.

Wraz ze wzrostem popularności dedykowanych systemów zapisywania notatek głosowych oraz ich dalszej transkrypcji i ostatecznie włączania do elektronicznej dokumentacji medycznej nastąpił gwałtowny wzrost badań mających na celu zrozumienie, w jaki sposób technologia ASR może pomóc świadczeniodawcom usprawnić przepływ pracy i zapewnić lepszą opiekę nad pacjentem [8, 21]. Pojawienie się rozwiązań chmurowych zainicjowało intensywne badania możliwości automatycznego przetwarzania mowy w celu jej transkrypcji w czasie rzeczywistym podczas wywiadów medycznych. Do takich badań wykorzystywane są scenariusze rozmów lekarzy z tzw. symulowanymi pacjentami (ang. *simulated patients*), czyli wyszkolonymi aktorami wcielającymi się w role pacjentów w sytuacjach medycznych [22, 23]. Oznacza to przeniesienie środka ciężkości zainteresowań naukowców, bowiem przyglądają się oni teraz poprawie dokładności i wydajności systemów ASR w wywiadach medycznych i koncentrują się na zapewnieniu asysty dla rozpoznawania w czasie rzeczywistym.

W 2020 roku grupa naukowców z University of South California przeprowadziła kompleksowe badanie porównujące dokładność różnych, publicznie dostępnych, komercyjnych i naukowych systemów ASR [24]. Badaniem objęto systemy opracowane przez Amazon, Apple, Google, IBM, Microsoft i Kaldi, które zostały

przetestowane na siedmiu zestawach angielskich danych głosowych w sześciu różnych zakresach tematycznych. Zbiory danych głosowych składały się ze spontanicznych nagrań mowy pochodzącej z nieindukowanych sztucznie rozmów, dzięki czemu badanie było wysoce reprezentatywne dla rzeczywistych scenariuszy. Naukowcy wykorzystali WER jako podstawową miarę oceny. Badanie wykazało, że wszystkie badane systemy ASR nadal popełniały znaczną liczbę błędów transkrypcyjnych, niezależnie od producenta i rodzaju wytrenowania. Co więcej, żaden z systemów rozpoznawania mowy nie dominował we wszystkich zestawach danych. Jako jeden z powodów takich wyników naukowcy wskazali rodzaj treningu, jaki przechodzą systemy ASR. Przykładowo, gorsze wyniki uzyskały te systemy, które były trenowane na próbkach głosowych pochodzących z audiobooków, które znacząco odbiegają jakością dźwięku i samej mowy od mowy spontanicznej. Równie złe wyniki wykazały systemy trenowane w kontraście do poprzednich na próbkach konwersacyjnych, ale jedynie w zakresie mowy telefonicznej o niskiej jakości, co wpłynęło negatywnie na ostateczną dokładność. Część z błędów rozpoznania można przypisać wielkości zestawu danych, jednak większość z nich jest ściśle związana z samą dziedziną, której dotyczą wypowiedzi, a więc użytego specjalistycznego słownictwa. Równie błędnie rozpoznawana była spontaniczna mowa potoczna.

Porównanie badań zespołu naukowców z University of South California z wcześniejszymi, przeprowadzonymi w 2010 r. przez X. Yao et al. [25] oraz w 2013 r. przez F. Morbini et al. [26], dokonanymi na podobnych zbiorach danych, wyraźnie wskazuje, że w ostatnich latach dokonał się znaczny postęp w dziedzinie rozpoznawania mowy. Oznacza to, że choć systemy ASR wymagają dalszych prac, szczególnie w obszarach rozpoznawania mowy specjalistycznej oraz spontanicznej mowy potocznej, to m.in. integracja technik głębokiego uczenia w technologii ASR wpłynęła na większą dokładność systemów ASR. Poprzednie edycje badań zakładały także sprawdzenie tzw. NLU (ang. *Natural Language Understanding*), co jednak jest wypadkową samej dokładności rozpoznania mowy, które prowadzi do lepszego rozumienia języka naturalnego. Kluczowa dla tej pozytywnej korelacji i zastosowania tezy na szerszą skalę jest jednak optymalizacja systemu ASR w celu uzyskania jak najdokładniejszej transkrypcji, a nie próba optymalizacji pod kątem wybranej miary dla NLU. Naukowcy podkreślają również znaczenie uwzględnienia dziedziny zastosowania i kontekstu przy wyborze najodpowiedniejszego rozwiązania do rozpoznawania mowy. Uwypukla to potrzebę opracowywania dostosowanych rozwiązań, które spełniają określone wymagania danego klienta końcowego, takie jak zakres tematyczny i kontekst rozpoznawanej mowy, w celu osiągnięcia optymalnej wydajności.

Badanie z 2022 r. przeprowadzone przez zespół naukowców pod kierownictwem Kima, Liu i Calvo [27] dotyczyło oceny wydajności pięciu systemów ASR pod kątem jakości transkrypcji mowy o kontekście medycznym. W badaniu oceniono platformy internetowe – Google Cloud, IBM Watson, Microsoft Azure, Trint i YouTube – w zakresie ich wbudowanych systemów ASR. Zbiory danych głosowych pochodziły z 24 przeprowadzonych telekonsultacji medycznych (interakcje dwunastu studentów medycyny, każdy z dwoma symulowanymi pacjentami). Naukowcy zakładali wstępnie, że ręczne transkrypcje przewyższają te automatyczne. Wyniki ich badań

potwierdziły tę hipotezę, bowiem ręczne transkrypcje okazały się dokładniejsze. Co ciekawe, spośród ocenianych systemów, system ASR YouTube przewyższył inne rozwiązania oparte na chmurze pod względem dokładności rozpoznawania słów, wykazując zaskakującą przydatność platformy w dziedzinie transkrypcji medycznej.

Pomimo korzyści z zastosowań technologii ASR w dziedzinie medycyny, wciąż istnieje szereg wyzwań, którym należy sprostać. Obejmują one potrzebę większej standaryzacji w rejestrowaniu i transkrypcji dokumentacji medycznej, a także opracowania skalowalnych i bardziej dokładnych systemów ASR, które będą w stanie poradzić sobie z szeroką gamą dialektów i akcentów używanych w placówkach opieki zdrowotnej. Ponadto istnieje potrzeba opracowań w zakresie porównania działania systemów ASR, które pomogą zidentyfikować najskuteczniejsze rozwiązania dla określonych zastosowań medycznych.

3. Metodologia

Głównym celem badawczym niniejszego artykułu jest wykonanie analizy porównawczej poziomu dokładności rozpoznawania mowy wiodących na polskim rynku systemów ASR na specjalnie do tego celu wyselekcjonowanej próbce danych. Dokonując selekcji wiodących na polskim rynku systemów ASR do dalszego wykonania analizy porównawczej, wzięto pod uwagę kilka czynników. Były nimi mianowicie: dostępność systemu ASR i popularność tego narzędzia w praktyce badawczej, skalowalność, dostępność danych technicznych oraz danych wejściowych/wyjściowych realizowanych projektów. Na podstawie oceny tych cech zostały wyłonione do weryfikacji trzy systemy ASR dostępne na polskim rynku, a mianowicie:

- system Google ASR (producent Google Inc.);
- system Microsoft ASR (producent Microsoft Inc.);
- system Techmo ASR (producent Techmo sp. z oo).

Przygotowana na cele badania próbka danych zawiera 1000 nagrań krótkich zdań i fraz wypowiedzianych w języku polskim, które zawierają wyrażenia medyczne - zarówno określenia potoczne, jak i te zgodne z nomenklaturą medyczną. Wypowiedzi nagrane zostały przez osoby różnej płci (kobiety i mężczyzn), zróżnicowane regionalnie pod względem pochodzenia (północ i południe Polski) oraz w różnym wieku (przedział 30-50 lat).

Dla szczegółowego porównania oceny jakości rozwiązań ASR zastosowano ocenę dokładności dla poprawnie rozpoznanych słów, wskaźnik błędnie rozpoznanych słów WER, a także miarę odległości edycyjnej Levenshteina i miarę podobieństwa Jaro-Winklera. Dokładność pokazuje procent poprawnie rozpoznanych słów w danej wypowiedzi, a wskaźnik WER pokazuje procent błędnych słów w transkrypcji danej wypowiedzi. Miara odległości edycyjnej Levenshteina jest obliczana na poziomie pojedynczych znaków w słowie i wskazuje minimalną liczbę zmian wymaganych do przekształcenia dwóch łańcuchów znaków w identyczne przez wstawianie, zastępowanie lub usuwanie znaków w jednym z nich. Im niższa odległość Levenshteina, tym wyższe podobieństwo dwóch łańcuchów znaków [28]. Kolejna miara odległości między łańcuchami znakowymi, znana szerzej jako miara podobieństwa Jaro-Winklera, to z kolei metoda oparta na odległości Jaro [29], ale udoskonalona przez preferowanie identycznych znaków na początku każdego łańcucha

i niebazująca na odległości związanej z edycją [30]. Im wyższe podobieństwo Jaro-Winklera, tym wyższe podobieństwo dwóch łańcuchów znaków.

Aby osiągnąć zaprezentowany cel badawczy, przed właściwą analizą porównawczą próbki nagrań wypowiedzianych języku polskim, wykonano wymienione poniżej prace.

1. Zaimplementowanie API (ang. *Application Programming Interface*) w celu umożliwienia testowania każdego systemu ASR.
2. Zebranie inicjalnego zestawu danych na bazie 19000 skategoryzowanych tekstowych wyrażeń medycznych, których właścicielem jest SOVVA SA (firma realizująca niniejszy projekt badawczy współfinansowany przez NCBiR). Z tego zestawu eksperci medyczni wybrali próbkę 1000 zagadnień najczęściej poruszanych przez pacjentów podczas wizyt w gabinetach lekarskich.
3. Wybraną próbkę 1000 tekstowych zagadnień najczęściej poruszanych przez pacjentów następnie poddano nagraniom, a każdy pojedynczy plik dźwiękowy zawierał nagranie pojedynczej frazy (wyrażenia lub słowa). Pliki zostały zapisane w formacie WAV o następujących parametrach: częstotliwość próbkowania: 16000 Hz, kanały: 1 (mono), format próbki: PCM, 16-bitowy. Nagrania zostały przeprowadzone w warunkach laboratoryjnych, a warunki zewnętrzne, takie jak hałas w tle lub zakłócenia, nie zostały uwzględnione.
4. Otrzymane 1000 plików audio przesłano na serwer FTP. Następnie zostały one przetworzone przez każdy system ASR za pomocą wcześniej zaimplementowanego API.
5. Uzyskane transkrypcje poddano następnie badaniom w celu finalnego wykonania analizy porównawczej poziomu dokładności rozpoznawania mowy badanych systemów ASR.

Jako dodatkowe badanie, dotyczące ulepszenia każdego z badanych systemów ASR, zweryfikowano możliwość dostosowania ich do indywidualnych potrzeb różnych podmiotów medycznych. Chodzi tutaj o potencjał tkwiący w dodatkowym trenowaniu swojego zbioru danych oraz w swobodnej instalacji danego rozwiązania ASR w infrastrukturze klienta końcowego (tzw. instalacja *on-premise*).

4. Rezultaty badań

Wyniki uzyskane z badania na próbce 1000 głosowych wyrażeń i fraz najczęściej wypowiadanych przez pacjentów pokazują wysoką dokładność (powyżej 86%) rozpoznawania mowy przez badane systemy ASR, przy czym rozbieżność pomiędzy najlepszym i najgorszym wynikiem wyniosła zaledwie 1,7%. Wykonana analiza porównawcza systemów ASR została podsumowana w zamieszczonej dalej Tabeli 1.

Tabela 1. Dokładność rozpoznawania mowy przez poszczególne systemy ASR

	Google ASR	Microsoft ASR	Techmo ASR
Dokładność (*accuracy*)	88,1%	86,4%	87,5%

Źródło: opracowanie własne

Miara odległości edycyjnej Levenshteina jest to wartość wyrażona w liczbie zmian potrzebnych do ujednolicenia dwóch łańcuchów znaków, z kolei miara Jaro-Winklera

stanowi stopień podobieństwa dwóch łańcuchów znaków wyrażony w przedziale od 0 do 1, wreszcie wskaźnik WER przedstawia się jako procent błędnie rozpoznanych słów. Dla ujednolicenia otrzymanych wyników i możliwości porównania silników ASR z zastosowaniem różnych miar, wszystkie trzy miary zostały na końcu przekształcone w procentowo wyrażone podobieństwo między dwoma łańcuchami znaków (wypowiedź oryginalna i tekst transkrypcji). Zgodnie z wynikiem testu Kołmogorowa-Smirnowa nie ma istotnej różnicy między rozkładami podobieństwa rozpoznania dla trzech silników w zakresie miary Levensteina, Jaro-Winklera i WER, a wszystkie wartości p-value są mniejsze niż 0,05. Co więcej, kwantyle rozkładów podobieństw są w większości identyczne dla wszystkich trzech silników, co sugeruje, że wykazują podobne rezultaty w zakresie rozpoznawania terminologii medycznej. W analizie rozpoznanych fraz, miary WER i Jaro-Winklera dały identyczne wyniki i były każdorazowo wyższe od procentowo wyrażonego podobieństwa na bazie Levenshteina, ale najlepszy, średni i najgorszy wynik dla danego silnika ASR pozostał taki sam bez względu na zastosowaną miarę, co pokazano w zamieszczonej dalej Tabeli 2.

Tabela 2. Podobieństwo rozpoznanych fraz przez poszczególne systemy ASR

	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%

Źródło: Opracowanie własne

Uzyskane wyniki rozpoznawania mowy z badanych silników ASR często zawierały błędnie zapisane lub niekompletne słowa, co daje możliwość zbadania różnorodności błędów, które wystąpiły podczas transkrypcji mowy na tekst. Można je mianowicie podzielić na trzy kategorie: a) błędne rozpoznania, b) problemy z jakością, c) granice słów.

Błędne rozpoznania pojawiają się, gdy silnik ASR nie rozpoznaje wypowiedzi w całości lub z nadanym zupełnie innym znaczeniem. Przykłady wykrytych błędnych rozpoznań, w których silnik ASR nie rozpoznał poprawnie frazy, co uniemożliwia dalszą optymalizację przetwarzania:

• w ogóle nierozpoznane: „omdlenia", „ścieńczenia", „pęcherzyca";

• rozpoznane w innym znaczeniu:

• oryginał „leukoplakia" rozpoznany jako „lenko platja";

• oryginał „problem z ukrwieniem oczu" rozpoznany jako „problem z ukrwienie moczu";

• oryginał „wściekłość" rozpoznany jako „zaległość".

Kolejna kategoria to problemy z jakością rozpoznania, które pojawiają się, gdy silnik ASR rozpoznaje frazę z niewielkimi rozbieżnościami, co jednakże nie zmienia znaczenia rozpoznanej frazy. Przykłady wykrytych problemów z jakością:

• oryginał „zubożona mowa" rozpoznany jako „zburzona mowa";

• oryginał „żwir w woreczku żółciowym" rozpoznany jako „żwir po woreczku żółciowym";

• oryginał „wrażenie że nie słyszę” rozpoznany jako „teraz wrażenie że nie słyszę”;
• oryginał „złogi w pęcherzyku żółciowym” rozpoznany jako „złogi pęcherzyku żółciowym”.

Wreszcie ostatni problem stanowią granice słów, w przypadku których silniki ASR poprawnie rozpoznały frazę, ale niektóre słowa nie mają poprawnie rozpoznanych granic, co można stosunkowo łatwo zoptymalizować w dalszym przetwarzaniu. Przykłady wykrytych problemów z granicami słów:

• oryginał „stres pourazowy” rozpoznany jako „stres po urazowy”;
• oryginał „problemy z chodzeniem” rozpoznany jako „problemy schodzeniem”;
• oryginał „śpię w dzień” rozpoznany jako „śpiew dzień”.

W zamieszczonej dalej Tabeli 3 przedstawiono zestawienie procentowe wymienionych problemów w rozpoznawaniu mowy napotkanych dla każdego silnika ASR.

Tabela 3. Procent problemów ASR dla każdego silnika

	błędne rozpoznania	problemy z jakością	granice słów
Google ASR	73,2%	23,6%	3,2%
Microsoft ASR	71,2%	14,8%	14,0%
Techmo ASR	61,7%	26,6%	11,7%

Źródło: Opracowanie własne

Według przedstawionych danych, Techmo ASR miało najniższy odsetek błędnych rozpoznań w porównaniu z Microsoft ASR i Google ASR: Techmo ASR wykazywał 61,7% błędnych rozpoznań wobec 71,2% dla Microsoft ASR i 73,2% dla Google ASR. Jednocześnie Microsoft ASR miał najniższy odsetek problemów z jakością w porównaniu z Google ASR i Techmo ASR: Microsoft ASR wykazywał 14,8% problemów z jakością wobec 23,6% dla Google ASR i 26,6% dla Techmo ASR. Z kolei Google ASR miał najniższy odsetek błędnie rozpoznanych granic słów w porównaniu z Microsoft ASR i Techmo ASR: Google ASR wykazywał 3,2% granic słów wobec 14,0% dla Microsoft ASR i 11,7% dla Techmo ASR. Można zauważyć, że Techmo ASR wypadło lepiej pod względem stosunku błędnych rozpoznań do pozostałych problemów ASR (problemów z jakością i granic słów), które można zoptymalizować w dalszym przetwarzaniu.

Dodatkowo odkryto, że część wykrytych problemów i nieścisłości była spowodowana wielkością wejściowego zestawu danych. Aby rozwiązać ten problem, wykorzystano redukcję do rdzenia oraz bazę danych dotyczących synonimów. Pozwoliło to zmniejszyć pewne rozbieżności gramatyczne w danych wejściowych, które miały wpływ na uzyskane wyniki. Ponadto wejściowy zbiór danych był niewielkich rozmiarów, gdyż obejmował jedynie 1000 fraz. W związku z tym nie można było wykryć innych problemów, które mogą być potencjalnie wykazywane przez każdy z trzech badanych silników automatycznego rozpoznawania mowy.

Warto również zwrócić uwagę na usterki rozpoznawania mowy na poziomie poszczególnych słów. Wykonane badania pozwoliły zidentyfikować zestaw słów mylonych przez wszystkie trzy badane silniki ASR. Były to mianowicie: „świąd”, „ból”, „krosty”, „krwiaki”, „sinienie”, „płaczliwość”, „leki”, „zauszne”, „białka”,

„płonica", „omamy" i „urojenia". Ponadto większość słów z powyższego zestawu zidentyfikowano jako błędy najczęściej występujące w całym zbiorze. Dziesięć najczęściej mylonych słów to kolejno: „świąd", „krosty", „ból", „poronienie", „problem z", „białka", „kłykciny", „urojenia", „omamy" oraz „płonica".

Z kolei inny zestaw słów został błędnie rozpoznany przez dwa silniki ASR jednocześnie (w różnych parach). Chodzi tutaj o następujące słowa: „poronienie", „problem z", „plamy", „słyszę", „uraz", „kłykciny", „utrata", „przerosła", „poparzeniowa", „rozpadowe", „rozpadająca", „poty", „wściekłość", „punkcja", „łuski", „duże", „przymglone", „rogowiec", „kolka", „przeczos", „kolki", „płacz", „krosta", „strup", „kiła", „ostre", „wady", „polipy", „sine", „ptsd", „usnąć", „dreszcze", „dokrwiona", „trudność", „przestaw", „oziębienie", „stres", „schizofrenia".

Jako dodatkowy aspekt analizy porównawczej, zbadano też potencjał ulepszenia każdego z badanych systemów ASR poprzez weryfikację możliwości dodatkowego trenowania własnego zbioru danych i swobodnej instalacji *on-premise* danego rozwiązania ASR u klienta końcowego. Zbadano te możliwości w oparciu na kompleksowej analizie dokumentacji i próbach eksperymentalnych. Wyniki, podsumowane w Tabeli 4, dostarczają cennych informacji na temat technicznych możliwości i ograniczeń tych rozwiązań i mogą przyczynić się do rozwoju bardziej wydajnych technologii ASR dla różnych zastosowań.

Tabela 4. Dodatkowe aspekty technicznych możliwości systemu ASR u klienta końcowego

	Google ASR	Microsoft ASR	Techmo ASR
Możliwość dodatkowego trenowania własnego zbioru danych	Nie	Tak	Tak (dodatkowo płatna usługa Techmo)
Możliwość swobodnej instalacji *on-premise* u klienta końcowego	Nie	Nie	Tak

Źródło: opracowanie własne

Z Tabeli 4 wynika, że tylko rozwiązanie Techmo ASR oferuje możliwość dodatkowego trenowania własnego zbioru danych oraz swobodnej instalacji *on-premise*. Chociaż Microsoft ASR umożliwia dodatkowe trenowanie własnego zbioru danych, nie można go jednak zainstalować lokalnie. Z drugiej strony Google ASR nie spełnia żadnego z obu aspektów technicznych. Na bazie dodatkowych eksperymentów przeprowadzonych z oprogramowaniem silników ASR zauważono również, że wykrywanie określonych fraz w systemie Google ASR można usprawnić przez kontrolę wbudowanych predefiniowanych słowników, które zapewniają lepsze rozpoznanie wypowiadanych fraz. Natomiast Microsoft ASR nie obsługuje wcale takiej funkcjonalności, z kolei Techmo ASR nie udostępnia informacji na temat możliwości w tym zakresie.

5. Podsumowanie

W artykule przedstawiono analizę porównawczą dostępnych na polskim rynku silników Google ASR, Microsoft ASR i Techmo ASR, pod kątem oceny poziomu ich dokładności w rozpoznawaniu potocznej mowy polskiej związanej z tematyką

medyczną. Na podstawie wyników badania stwierdzono, że poziomy dokładności systemów były bardzo podobne i przekraczały 86% dokładności rozpoznania dla wszystkich trzech silników ASR. Natomiast zbadana możliwość ulepszenia danego systemu ASR przez dodatkowe trenowanie własnego zbioru danych oraz instalację *on-premise* okazały się czynnikami do dodatkowego rozważenia przez klienta końcowego jako ważne z perspektywy posiadania kontroli w zastosowaniu technologii w podmiotach medycznych. Otrzymane wyniki wypełniają lukę badawczą w zakresie analiz porównawczych silników ASR oraz podkreślają znaczenie starannego rozważenia zarówno czynników technicznych, jak i praktycznych przy wyborze rozwiązania ASR do konkretnego zastosowania, wskazując zarówno Microsoft ASR, jak i Techmo ASR jako optymalne rozwiązanie do transkrypcji wywiadów medycznych prowadzonych w języku polskim.

Przedstawione badania mogą pomóc ukierunkować projektowanie systemów konwersacyjnych wyspecjalizowanych dziedzinowo i interfejsów głosowych zdolnych do rozpoznawania potocznej mowy polskiej w celu przeprowadzania wstępnej diagnostyki i wywiadu lekarskiego przy minimalnej ingerencji człowieka, co jest istotną zaletą w czasach epidemii i/lub przy ograniczonych zasobach organizacyjnych.

6. Finansowanie badania

Praca współfinansowana przez Narodowe Centrum Badań i Rozwoju (NCBiR) przedstawia cząstkowe wyniki projektu badawczego pt. „Opracowanie systemu wykorzystującego uczenie maszynowe do automatycznego przeprowadzania wstępnej diagnostyki i wywiadów lekarskich w jednostkach służby zdrowia" realizowanego przez SOVVA SA.

Literatura

1. Google ASR, https://cloud.google.com/speech-to-text [dostęp 04.2023].
2. Microsoft ASR, https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text [dostęp 04.2023].
3. Dragon STT, https://www.nuance.com/en-gb/dragon.html [dostęp 04.2023].
4. Phonexia STT, https://www.phonexia.com/product/speech-to-text/ [dostęp 04.2023].
5. Techmo ASR, https://techmo.pl/technologia/rozpoznawanie-mowy [dostęp 04.2023].
6. Rev AI ASR, https://www.rev.ai/ [dostęp 04.2023].
7. Iancu B., *Evaluating Google Speech-to-Text API's Performance for Romanian e-Learning Resources,* Informatica Economica, vol. 23, no. 1, 2019, s. 17-25.
8. Scholz M.L., Collatz-Christensen H., Blomberg S.N.F., Boebel S., Verhoeven J., Kraft T., *Artifcial intelligence in Emergency Medical Services dispatching: assessing the potential impact of an automatic speech recognition software on stroke detection taking the Capital Region of Denmark as case in point,* Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine vol. 30, no. 36, 2022, s. 1-17.
9. Tomar V., Desruisseaux M., Seetzen H., *System and method for implementing a vocal user interface by combining a speech to text system and a speech to intent system*, 2020, https://patentimages.storage.googleapis.com/1d/05/1d/014c820a9a7b7b/US10878807.pdf [dostęp 02.2023].
10. Tomar V., *How Speech Technology Is Optimizing Factory Lines*, 2021, https://industrytoday.com/how-speech-technology-is-optimizing-factory-lines/ [dostęp 02.2023].

11. Vinnarasu A., Jose D.V., *Speech to text conversion and summarization for effective understanding and documentation*, International Journal of Electrical and Computer Engineering (IJECE), vol. 9, issue 5, 2019, s. 3642-3648.
12. Wahyutama A.B., Hwang M., *Auto-Scoring Feature Based on Sentence Transformer Similarity Check with Korean Sentences Spoken by Foreigners*, Applied Sciences 13, vol. 373, no. 1, MDPI 2023, s. 1-16.
13. Kim J.Y., Liu C., Calvo R.A., McCabe K., Taylor S.C.R., Schuller B.W., Wu K., *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, 2019, s. 1-13.
14. Kuligowska K., Kisielewicz P., Włodarz A., *Wady i ograniczenia systemów rozpoznawania mowy*, Roczniki Kolegium Analiz Ekonomicznych, nr 49/2018, Szkoła Główna Handlowa, Warszawa 2018, s. 307-317.
15. Lugosch L., Ravanelli M., Ignoto P., Tomar V., Bengio Y., *Speech Model Pre-training for End-to-End Spoken Language Understanding*, Preprint arXiv, Audio and Speech Processing (eess.AS), 2019, doi.org/10.48550/arXiv.1904.03670, s. 1-5.
16. Mah P.M., Skalna I., Muzam J., *Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0,* Applied Sciences, vol. 12, no. 18: 9207, 2020, s. 1-26.
17. Zue V.W., *The use of speech knowledge in automatic speech recognition*, Proceedings of the IEEE, 1985, vol. 73, no. 11, s. 1602-1615.
18. Repka A., *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 2022, s. 104.
19. Altar H.S., Sison R.C., *Medical Transcriptionist's Experience with Speech Recognition Technology*, Proceedings of Australasian Conference on Information Systems (ACIS 2019), Perth, Western Australia 2019, s. 915- 924.
20. Joseph J., Moore Z.E.H., Patton D., O'Connor T., Nugent L.E., *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, 2020, s. 2125-2137.
21. Saxena K., Diamond R., Conant R.F., Mitchell T.H., Gallopyn G., Yakimow K.E., *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. 2017 AMIA Joint Summits on Translational Science, 2018, s. 186-195.
22. 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., *A dataset of simulated patient-physician medical interviews with a focus on respiratory cases*, Scientific Data, vol. 9, no. 313, 2022, s. 1-7.
23. Kim D., Oh J., Heeju Im H., Yoon M., Park J., Lee J., *Automatic Classification of the Korean Triage Acuity Scale in Simulated Emergency Rooms Using Speech Recognition and Natural Language Processing: a Proof of Concept Study*, Journal of Korean Medical Science vol. 36, no. 27, 2021, s. 1-13.
24. Georgila K., Leuski A., Yanov V., Traum D., *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 2020, s. 6469-6476.

25. Yao X., Bhutada P., Georgila K., Sagae K., Artstein R., Traum D., *Practical evaluation of speech recognizers for virtual human dialogue systems*. Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta 2010, s. 1597-1602.
26. Morbini F., Audhkhasi K., Sagae K., Artstein R., Can D., Georgiou P., Narayanan S., Leuski A., Traum D., *Which ASR should I choose for my dialogue system?*, Proceedings of the 14th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2013), Metz, France 2013, s. 394-403.
27. Kim J.Y., Liu C., Calvo R.A., McCabe K., Taylor S.C.R., Schuller B.W., Wu K., *Comparison of Automatic Speech Recognition Systems*, [w:] Stoyanchev S., Ultes S., Li H. (red.), *Conversational AI for Natural Human-Centric Interaction, Lecture Notes in Electrical Engineering*, vol. 943, Springer 2022, s. 123-131.
28. Niewiarowski A., Stanuszek M., *Mechanizm analizy podobieństwa krótkich fragmentów tekstów, na bazie odległości Levenshteina*, Studia Informatica, Vol. 34, No. 1 (110), Politechnika Krakowska, Instytut Informatyki 2013, s. 108.
29. Cohen W.W., Ravikumar P., Fienberg S.E., *A Comparison of String Metrics for Matching Names and Records*, KDD Workshop on Data Cleaning and Object Consolidation, Carnegie Mellon University, 2003, s. 1-6.
30. Kamińska A.M., *Miary podobieństw łańcuchów znakowych a deduplikacja rekordów w bibliograficznych bazach danych*, Przegląd Biblioteczny, z. 4, 2017, s. 477-495.

Zastosowanie automatycznego rozpoznawania mowy w transkrypcji wywiadów medycznych – porównanie silników ASR dla języka polskiego

Streszczenie
W dążeniu do bardziej wydajnej i skoncentrowanej na pacjencie opieki zdrowotnej kluczowym elementem wspierającym podmioty lecznicze stają się automatyczne systemy rozpoznawania mowy. By jednak uznać je za użyteczne, systemy te muszą spełniać określone wymagania podyktowane realiami rynkowymi. Dlatego też celem niniejszego artykułu była analiza porównawcza wiodących na polskim rynku silników ASR, wykonana na zbiorze nagranych w języku polskim fraz najczęściej wypowiadanych przez pacjentów podczas wizyt w gabinetach lekarskich. Wyniki naszej analizy wykazały, że między badanymi silnikami ASR istnieją niewielkie różnice w dokładności rozpoznawania mowy. Mimo to wszystkie prezentowały specyficzne problemy, które zostały podzielone na trzy grupy: błędne rozpoznania, problemy z jakością i granice słów. Wyniki badań dostarczają cennych informacji szerokiemu gronu interesariuszy, ułatwiając rozwój rozwiązań do rozpoznawania mowy polskiej dla specyficznych potrzeb sektora medycznego.
Słowa kluczowe: automatyczne rozpoznawanie mowy, speech-to-text, silniki ASR dla języka polskiego, transkrypcja wywiadów medycznych

Application of automatic speech recognition in the transcription of medical interviews - comparison of ASR engines for the Polish language

Abstract
Automatic speech recognition systems are becoming a key element supporting healthcare entities in pursuing a more efficient and patient-centred healthcare system,. However, these systems must meet certain requirements dictated by market reality to be considered beneficial. Therefore, the purpose of this paper was a comparative analysis of the leading ASR engines on the Polish market, performed on a set of phrases recorded in Polish covering the most frequently spoken utterances by patients during visits in doctors' offices. The results of our analysis showed that there are slight differences in speech recognition accuracy among the tested ASR engines. Despite this, they all presented specific problems, divided into three groups: misidentifications, quality problems and word boundaries. The research results provide valuable information to a wide range of stakeholders, facilitating the development of Polish speech recognition solutions for the specific needs of the medical sector.
Keywords: automatic speech recognition, speech-to-text, ASR engines for the Polish language, medical interview transcript