Marcin PACHOLCZYK
Politechnika Śląska

PRZEGLĄD I PORÓWNANIE ROZWIĄZAŃ ROZPOZNAWANIA MOWY POD KĄTEM ROZPOZNAWANIA ZBIORU KOMEND GŁOSOWYCH

Streszczenie. W artykule zaprezentowano przegląd i porównanie wybranych rozwiązań służących do rozpoznawania mowy. Przegląd miał na celu ocenę możliwości zastosowania tego typu rozwiązań w procesie szkolenia funkcjonariuszy służb specjalnych (np. Biura Ochrony Rządu).

REVIEW AND COMPARISON OF SPEECH RECOGNITION SOLUTIONS FOCUSED ON RECOGNITION OF SET OF VOICE COMMANDS

Summary. The article presents an overview and comparison of selected solutions for speech recognition. The review was focused on assessing the applicability of such solutions in the training process of special services officers (e.g. Government Protection Bureau).

1. Wprowadzenie

System przeznaczony do rozpoznawania mowy w języku polskim musi uwzględniać jej specyfikę. Powszechnie dostępne rozwiązania, rozwijane od lat 70-tych XX wieku, przeznaczone są zwykle do rozpoznawania mowy w języku angielskim. Począwszy od systemów rozpoznawania cyfr, przez wyizolowane słowa, pojedyncze zdania, aż do systemów oferujących swobodną komunikację w języku naturalnym, pomimo intensywnego rozwoju technik rozpoznawania mowy, systemy tego typu nadal nie dorównują możliwościom percepcyjnym człowieka, nawet w języku angielskim. W przypadku języka polskiego sytuacja wydaje się o wiele bardziej złożona. W języku polskim liczne są wysokoczęstotliwościowe głoski szczelinowe (frykatywne) i zwarte, a sam język jest w wysokim stopniu fleksyjny i niepozycyjny. Istnieją jednak skuteczne rozwiązania dla rozpoznawania mowy w języku polskim zarówno w dziedzinie oprogramowania przeznaczonego dla komputerów osobistych jak i sprzętowe, możliwe do wbudowania w projektowane urządzenia przenośne lub stacjonarne. W niniejszym opracowaniu skupiono się na aspekcie rozpoznawania komend głosowych czyli wyizolowanych słów, ewentualnie krótkich zdań w trybie rozkazującym. Przegląd obejmuje rozwiązania dla których możliwe było uzyskanie działającego produktu np. program SkryBot został zakupiony, ARM uzyskany od producenta w wersji testowej. Oprócz opisanych rozwiązań można wymienić program Sarmata 2.0 (Techmo Sp. z o.o.) oraz rozwiązania PrimeSpeech

(Primespeech Łukasz Brocki), jednak w przypadku pierwszego rozwiązania niemożliwe okazało się uzyskanie wersji produktu umożliwiającej przeprowadzenie opisanych testów, natomiast firma Primespeech zdecydowanie odmówiła udostępnienia wersji testowej do celów badawczych.

2. Rozwiązania w dziedzinie oprogramowania

SkryBot (Przepisywanie.pl Sp. z o.o.)
SkryBot [2] to program do zamiany mowy na tekst przeznaczony dla komputerów osobistych. Możliwości w zakresie rozpoznawania mowy to m.in. rozpoznawanie mowy z mikrofonu lub z pliku, integracja z system Windows (tzw. klawiatura głosowa), systemy dialogowe – robotyka, automatyzacja wykonywania przelewów przez telefon itp. Program może rozpoznawać w czasie rzeczywistym lub znacznie szybszym (w przypadku nagrań). SkryBot jest programem działającym w systemach z rodziny Windows Vista, 7, 8, 8.1, 10 – 32 oraz 64 bit.

System Automatycznego Rozpoznawania Mowy (Future Voice System Sp. z o.o.)
System Automatycznego Rozpoznawania Mowy (ARM) [1] to program do transkrypcji mowy dyktowanej na tekst. System według jego twóców jest szczególnie dedykowany dla służb odpowiedzialnych za bezpieczeństwo państwa. Jest szczególnie przydatny do sporządzania dokumentów formalnych – pism, umów, wniosków, raportów, uzasadnień orzeczeń sądu, jak i innych plików tekstowych, np. e-maili. ARM dysponuje szerokim i specjalistycznym słownikiem ogólnym, jak również słownictwem charakterystycznym dla środowiska prawniczego. Aktualny poziom poprawnego rozpoznawania mowy przez system ARM to ponad 95%. Future Voice System Sp. z o.o. jest dystrybutorem systemu ARM opracowanego przez Poznańskie Centrum Superkomputerowo-Sieciowe i Polską Platformę Bezpieczeństwa Wewnętrznego przy współpracy z Fundacją Uniwersytetu im. Adama Mickiewicza w Poznaniu. Funkcje ARM to m.in. rozpoznawanie i przetwarzanie mowy na żywo – rozpoznany przez system tekst jest od razu widoczny na ekranie komputera oraz funkcja rozpoznawania mowy z wcześniej przygotowanych nagrań. ARM jest programem działającym w systemach z rodziny Windows Vista, 7, 8, 8.1, 10 – 32 oraz 64 bit.

Google Cloud Speech-to-Text
Rozwiązanie firmy Google [3] jako jedyne z wymienionych wykorzystuje chmurę obliczeniową (zdalne zasoby obliczeniowe), a co za tym idzie nie wymaga od użytkownika stosowania sprzętu o określonych parametrach i może być z powodzeniem wykorzystywany z poziomu urządzeń mobilnych (tablet, smartphone działające w systemach Android lub iOS). *Google Cloud Speech-to-Text* umożliwia programistom konwertowanie dźwięku na tekst poprzez zastosowanie zaawansowanych modeli sieci neuronowych w łatwym w użyciu interfejsie API. Rozwiązanie rozpoznaje 120 języków i wariantów wymowy. Umożliwia m.in. sterowanie głosowe i sterowanie dźwiękiem w centrach telefonicznych i itp. Może przetwarzać sygnał mowy w czasie rzeczywistym lub nagrania.

3. Mobilne stanowisko do pobierania próbek głosowych

W celu pozyskiwania próbek głosowych wysokiej jakości utworzone zostało mobilne stanowisko złożone z:

- urządzenia do rejestracji dźwięku Olympus Linear PCM Recorder LS-P1 (rys. 2) umożliwiający nagrywanie dźwięku w postaci nieskompresowanej (próbkowanie PCM 96 kHz/24 bit);
- zestawu mikrofonowego Kenwood KHS-8BL (rys. 1).

Urządzenie LS-P1 umożliwia pozyskiwanie najwyższej jakości stereofonicznych nagrań dźwiękowych o bardzo małym poziomie szumów. Pozwala również na inteligentny automatyczny dobór poziomu nagrania bez ryzyka przesterowania. Dobór zestawu mikrofonowego wynikał z sugestii funkcjonariuszy BOR. Zestaw ten jest typowym wyposażeniem funkcjonariusza pełniącego czynną służbę i współpracuje z wykorzystywanym sprzętem do komunikacji radiowej.

Rys. 1. Zestaw mikrofonowy Kenwood KHS-8BL (źródło: materiały producenta)

4. Bazy próbek głosowych typowych komend używanych przez funkcjonariuszy Biura Ochrony Rządu

Bazę próbek głosowych utworzono dla 13 mówców (pracowników Instytutu Automatyki, Politechniki Śląskiej) i następujących typowych komend używanych przez funkcjonariuszy BOR w czasie pełnienia obowiązków służbowych:

- „BOR! Stój, rzuć broń!” (K1)
- „Stój, bo strzelam!” (K2)
- „Biuro Ochrony Rządu! Proszę się zatrzymać!” (K3)
- „Biuro Ochrony Rządu! Proszę pokazać bagaż!” (K4)
- „BOR! Proszę się zachować zgodnie z prawem!” (K5)

Próbki głosowe zostały pobrane w typowych warunkach panujących w zamkniętych pomieszczeniach. Do pobrania próbek wykorzystano stanowisko mobilne opisane w sekcji 3.

Rys. 2. Rejestrator dźwięku Olympus LS-P1 (źródło: materiały producenta)

W bazie próbek znalazły się również próbki przygotowane przez funkcjonariuszy BOR i zarejestrowane dla pięciu mówców w trzech różnych sytuacjach (warunkach środowiskowych). Próbki nagrane przez funkcjonariuszy nie zostały opisane. Z odsłuchu nagrań można jedynie wnioskować, że zostały zarejestrowane w pomieszczeniu zamkniętym oraz na zewnątrz (np. słyszalne podmuchy wiatru). Próbki zostały zarejestrowane przy użyciu tego samego stanowiska mobilnego, które zostało wypożyczone funkcjonariuszom.

5. Procedura testowa dla algorytmów rozpoznawania mowy

W celu oceny programów służących do rozpoznawania mowy stworzono następujące kryteria:

- Skuteczność rozpoznawania kompletnej wypowiedzi głosowej (wszystkie słowa).
- Skuteczność rozpoznawania poszczególnych słów w kompletnej wypowiedzi głosowej.
- Skuteczność rozpoznawania poszczególnych słów wypowiadanych w sposób wyizolowany (testy przeprowadzono dla jednego mówcy).

Testy dla wielu mówców przeprowadzono dla podstawowej wersji modelu rozpoznawania (niepoprawionej). Programy SkryBot i ARM oferują funkcję poprawiania modelu rozpoznawania mowy, testy dla pojedynczego mówcy przeprowadzono również dla modelu poprawionego (przystosowanego do mówcy).

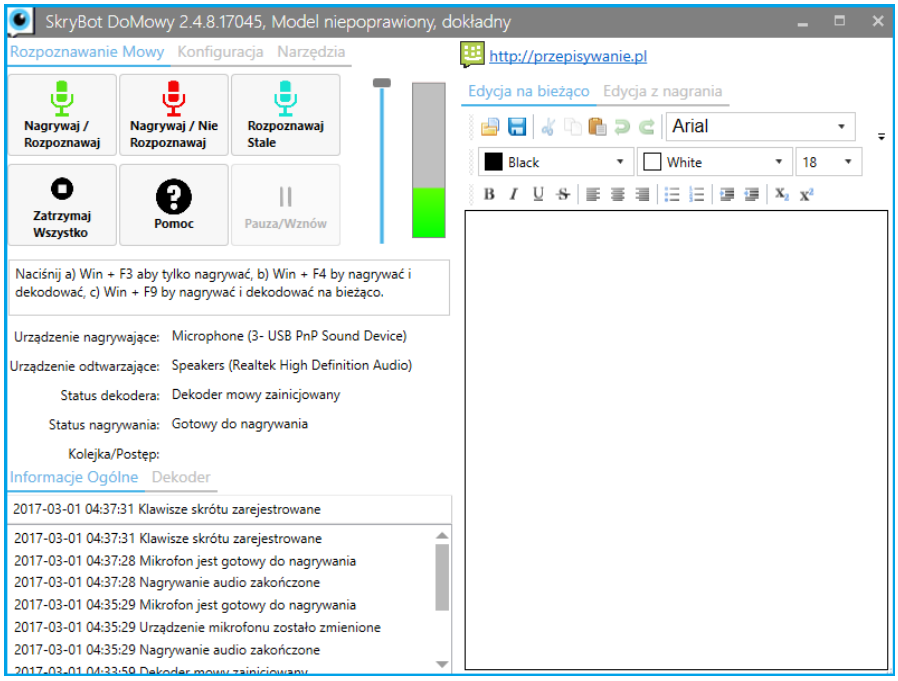
6. Testy funkcjonalne programu SkryBot DoMowy

6.1. Opis funkcjonalny programu SkryBot DoMowy

W opisie uwzględniono w szczególności te funkcje programu SkryBot Domowy, które zdaniem autora opracowania mają znaczenie w przypadku zastosowania programu podczas szkolenia funkcjonariuszy Biura Ochrony Rządu. Podstawową funkcją programu SkryBot DoMowy jest zamiana mowy na tekst. Należy zaznaczyć, że każdy program tego typu, dotyczy to również programu SkryBot DoMowy, posiada pewien ograniczony słownik, który może być przystosowany do konkretnych zastosowań. Producent programu SkryBot, oprócz wersji ogólnej oferuje również

wersje wyposażone w słowniki zawierające wyrażenia prawnicze (SkryBot Prawo), urzędowe (SkryBot Administracja) i medyczne (SkryBot Medycyna Rodzinna). Według deklaracji producenta, może on przygotować słownik przeznaczony do innego ściśle określonego celu. Brak określonego słowa w słowniku skutkuje przypisaniem innego o podobnym brzmieniu. Program SkryBot DoMowy umożliwia pracę w następujących trybach:

- Rozpoznawanie stałe.
- Rozpoznawanie przerywane (najpierw nagrywanie, potem rozpoznawanie - naprzemiennie).
- Rozpoznawanie z plików nagranych za pomocą SkryBota (tryb dyktafonu).
- Rozpoznawanie z plików nagranych poza SkryBotem.



Rys. 3. Główne okno programu SkryBot DoMowy (źródło: materiały producenta)

6.2. Skuteczność zamiany mowy na tekst w programie SkryBot DoMowy dla wielu mówców

Testy dla wielu mówców dla zbioru zawierającego wypowiedzi 13 mówców, przygotowanego na Politechnice Śląskiej (zbiór POLSL) przeprowadzono w trybie rozpoznawania z plików dźwiękowych nagranych poza SkryBotem. Pliki przed rozpoznawaniem poddano konwersji oferowanej przez program SkryBot. Program wymaga, aby przed uruchomieniem procesu zamiany mowy na tekst pliki dźwiękowe skonwertować do formatu WAV 16 kHz / 16 bit, mono. W Tabela 1 zestawiono skuteczność rozpoznawania kompletnych wypowiedzi (wszystkich słów w wypowiedzi) – komend (dla zbioru POLSL).

Tabela 1

Skuteczność rozpoznawania kompletnych wypowiedzi dla wielu mówców w programie SkryBot (zbiór POLSL)

Komenda	Liczba rozpoznanych pełnych wypowiedzi	%
K1	0/13	0
K2	7/13	54
K3	3/13	23
K4	6/13	46
K5	4/13	30

Tabela 2

Skuteczność rozpoznawania poszczególnych słów w wypowiedzi w programie SkryBot (zbiór POLSL)

Komenda		%
K1		
BOR	0/13	0
stój	0/13	0
Rzuć	1/13	7
broń	1/13	7
K2		
Stój	8/13	61
bo	9/13	69
strzelam	7/13	54
K3		
Biuro	12/13	92
Ochrony	13/13	100
Rządu	13/13	100
Proszę	5/13	38
Się	7/13	54
Zatrzymać	10/13	77
K4		
Biuro	12/13	92
Ochrony	12/13	92
Rządu	12/13	92
Proszę	11/13	85
Pokazać	10/13	77
Bagaż	8/13	61
K5		
BOR	0/13	0
Proszę	4/13	31
Zachować	8/13	61
Się	8/13	61
Zgodnie	13/13	100
Z	13/13	100
Prawem	13/13	100

W tabeli 2 zestawiono skuteczność rozpoznawania poszczególnych słów wchodzących w skład wypowiedzi – komendy (dla zbioru POLSL).

Osobnej serii testów poddano materiały (próbki głosowe) przygotowane przez funkcjonariuszy Biura Ochrony Rządu. Komunikaty przygotowane przez BOR zostały nagrane dla pięciu mówców w trzech różnych sytuacjach (warunkach środowiskowych). W Tabela 3 zestawiono skuteczność rozpoznawania kompletnych wypowiedzi (wszystkich słów w wypowiedzi) – komend (dla zbioru BOR).

Tabela 3

Skuteczność rozpoznawania kompletnych wypowiedzi dla wielu mówców w programie SkryBot (zbiór BOR)

Komenda	Sytuacja (warunki)	Liczba rozpoznanych pełnych wypowiedzi	%
K1	S1	0	0%
	S2	0	0%
	S3	0	0%
K2	S1	1/5	20%
	S2	0/5	0%
	S3	0/5	0%
K3	S1	0/5	0%
	S2	0/5	0%
	S3	0/5	0%
K4	S1	1/5	20%
	S2	1/5	20%
	S3	0/5	0%
K5	S1	0/5	0%
	S2	0/5	0%
	S3	0/5	0%

W Tabela 4 zestawiono skuteczność rozpoznawania poszczególnych słów wchodzących w skład wypowiedzi – komendy (dla zbioru BOR).

Tabela 4

Skuteczność rozpoznawania poszczególnych słów w wypowiedzi w programie SkryBot (zbiór BOR)

Komenda / Warunki	S1	%	S2	%	S3	%
K1						
BOR	0	0%	0	0%	0	0%
stój	0	0%	0	0%	0	0%
Rzuć	0	0%	0	0%	0	0%
broń	0	0%	0	0%	0	0%
K2						
Stój	1	20%	0	0%	0	0%
bo	1	20%	0	0%	0	0%
strzelam	1	20%	0	0%	0	0%
K3						
Biuro	1	20%	1	20%	1	20%

Ochrony	1	20%	1	20%	1	20%
Rządu	2	40%	3	60%	3	60%
Proszę	0	0%	0	0%	0	0%
Się	2	40%	0	0%	0	0%
Zatrzymać	3	60%	1	20%	2	40%
K4						
Biuro	1	20%	1	20%	1	20%
Ochrony	1	20%	1	20%	0	0%
Rządu	2	40%	3	60%	1	20%
Proszę	1	20%	1	20%	0	0%
Pokazać	2	40%	1	20%	0	0%
Bagaż	3	60%	3	60%	1	20%
K5						
BOR	0	0%	0	0%	0	0%
Proszę	0	0%	3	60%	0	0%
Zachować	2	40%	2	40%	1	20%
Się	1	20%	1	20%	0	0%
Zgodnie	3	60%	2	40%	1	20%
Z	4	80%	2	40%	2	40%
Prawem	4	80%	2	40%	2	40%

7. Testy funkcjonalne programu ARM

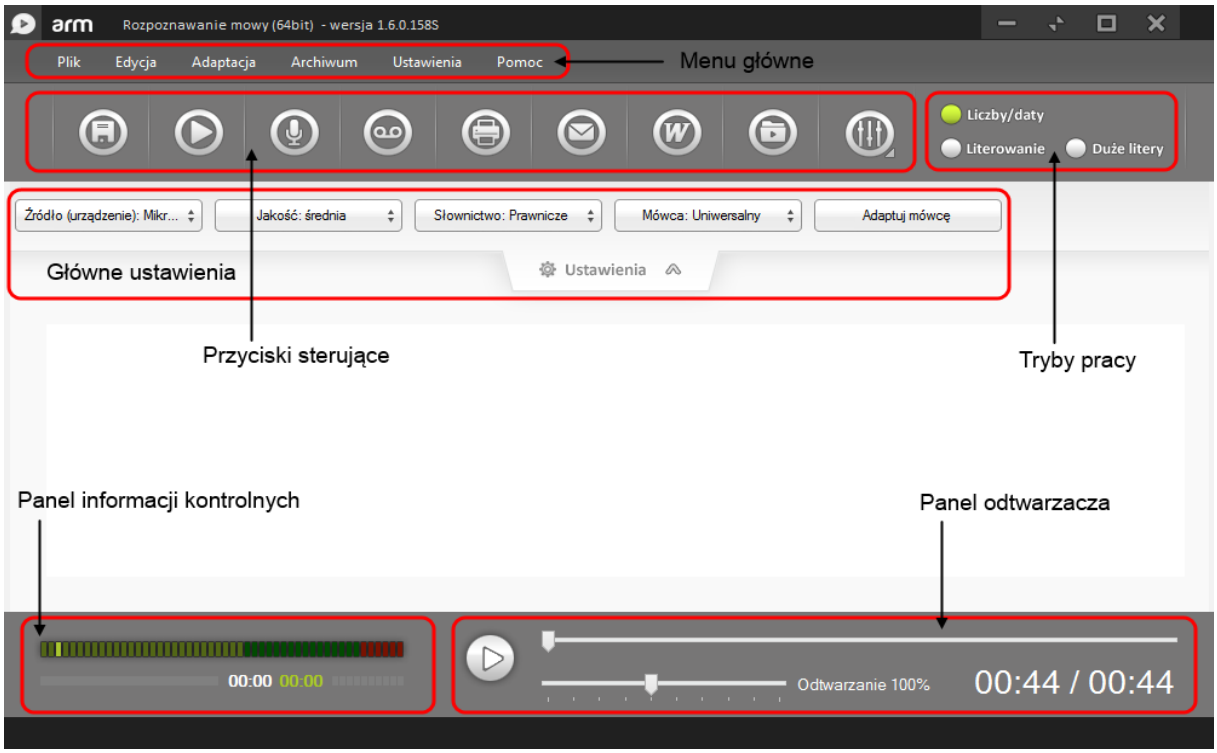
7.1. Opis funkcjonalny programu ARM

W opisie uwzględniono w szczególności te funkcje programu *ARM*, które zdaniem autora opracowania mają znaczenie w przypadku zastosowania programu podczas szkolenia funkcjonariuszy Biura Ochrony Rządu (zadania rozpoznawania krótkich komend głosowych, wygłaszanych w sposób nacechowany emocjonalnie i w niekorzystnych warunkach środowiskowych). Podstawową funkcją programu *ARM* jest zamiana mowy na tekst. Należy zaznaczyć, że każdy program tego typu, dotyczy to również programu *ARM*, posiada pewien ograniczony słownik, który może być przystosowany do konkretnych zastosowań. Brak określonego słowa w słowniku skutkuje przypisaniem innego o podobnym brzmieniu. Wersja próbna programu udostępniona przez producenta do testów, zawiera słownik wyrażeń prawniczych. Program *ARM* umożliwia dodawanie słów do słownika, wraz z określeniem jak często one występują.

Program *ARM* umożliwia pracę w następujących trybach:

- Rozpoznawanie z mikrofonu.
- Rozpoznawanie z plików nagranych za pomocą *ARM* (tryb dyktafonu).
- Rozpoznawanie z plików nagranych poza programem *ARM*.

Program *ARM* oferuje siedem ustawień jakości rozpoznawania (najniższa, niższa, niska, średnia, wyższa, wysoka, najwyższa), które różnią się szybkością procesu rozpoznawania. Wszelkie testy programu prowadzono dla ustawień domyślnych (wartość „średnia”).



Rys. 4. Widok okna głównego programu ARM (źródło: materiały producenta)

7.2. Skuteczność zamiany mowy na tekst w programie ARM dla wielu mówców

W

Tabela 5 zestawiono skuteczność rozpoznawania kompletnych wypowiedzi (wszystkich słów w wypowiedzi) w programie ARM – komend (dla zbioru POLSL).

Tabela 5

Skuteczność rozpoznawania kompletnych wypowiedzi dla wielu mówców w programie ARM (zbiór POLSL)

Komenda	Liczba rozpoznanych pełnych wypowiedzi	%
K1	0/13	0
K2	7/13	54
K3	6/13	46
K4	9/13	69
K5	2/13	15

W Tabela 6 zestawiono skuteczność rozpoznawania poszczególnych słów wchodzących w skład wypowiedzi – komendy w programie ARM (dla zbioru POLSL).

Tabela 6

Skuteczność rozpoznawania poszczególnych słów w wypowiedzi w programie ARM (zbiór POLSL)

Komenda		%
K1		
BOR	2/13	15
stój	2/13	15

Rzuć	4/13	30
broń	7/13	53
K2		
Stój	8/13	61
bo	7/13	54
strzelam	7/13	54
K3		
Biuro	11/13	85
Ochrony	11/13	85
Rządu	11/13	85
Proszę	6/13	46
Się	5/13	38
Zatrzymać	11/13	85
K4		
Biuro	10/13	77
Ochrony	10/13	77
Rządu	11/13	85
Proszę	10/13	77
Pokazać	10/13	77
Bagaż	11/13	85
K5		
BOR	3/13	23
Proszę	3/13	23
Zachować	9/13	69
Się	3/13	23
Zgodnie	12/13	92
Z	12/13	92
Prawem	12/13	92

Osobnej serii testów poddano materiały (próbki głosowe) przygotowane przez funkcjonariuszy Biura Ochrony Rządu (zadania rozpoznawania krótkich komend głosowych, wygłaszanych w sposób nacechowany emocjonalnie i w niekorzystnych warunkach środowiskowych). Komunikaty przygotowane przez BOR zostały nagrane dla czterech mówców w trzech różnych sytuacjach (warunkach środowiskowych).
W Tabela 7 zestawiono skuteczność rozpoznawania kompletnych wypowiedzi (wszystkich słów w wypowiedzi) – komend w programie ARM (dla zbioru BOR).

Tabela 7

Skuteczność rozpoznawania kompletnych wypowiedzi dla wielu mówców w programie ARM (zbiór BOR)

Komenda	Sytuacja (warunki)	Liczba rozpoznanych pełnych wypowiedzi	%
K1	S1	0/5	0
	S2	0/5	0
	S3	0/5	0
K2	S1	1/5	20

	S2	0/5	0
	S3	0/5	0
K3	S1	0/5	0
	S2	0/5	0
	S3	0/5	0
K4	S1	1/5	20
	S2	1/5	20
	S3	0/5	0
K5	S1	0/5	0
	S2	0/5	0
	S3	0/5	0

W Tabela 8 zestawiono skuteczność rozpoznawania poszczególnych słów wchodzących w skład wypowiedzi – komendy w programie ARM (dla zbioru BOR).

Tabela 8

Skuteczność rozpoznawania poszczególnych słów w wypowiedzi w programie ARM (zbiór BOR)

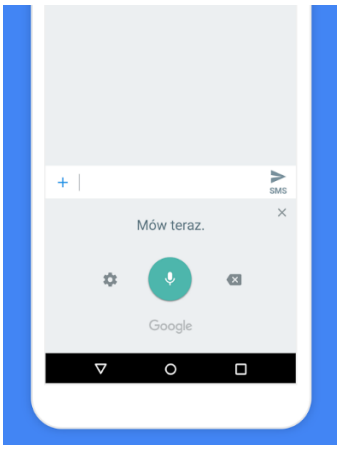
Komenda / Warunki	S1	%	S2	%	S3	%
K1						
BOR	0/5	0	0/5	0	0/5	0
stój	0/5	0	0/5	0	0/5	0
Rzuć	0/5	0	0/5	0	0/5	0
broń	0/5	0	0/5	0	0/5	0
K2						
Stój	1/5	20	0/5	0	0/5	0
bo	1/5	40	0/5	0	0/5	0
strzelam	2 /5	40	0/5	0	0/5	0
K3						
Biuro	1/5	20	1/5	20	1/5	20
Ochrony	1/5	20	2/5	40	1/5	20
Rządu	2/5	40	2/5	40	1/5	20
Proszę	0/5	0	1/5	20	2/5	40
Się	0/5	0	0/5	0	2/5	40
Zatrzymać	4/5	80	0/5	0	2/5	40
K4						
Biuro	1/5	20	1/5	20	1/5	20
Ochrony	1/5	20	1/5	20	1/5	20
Rządu	1/5	20	4/5	80	1/5	20
Proszę	2/5	40	1/5	20	1/5	20
Pokazać	2/5	40	1/5	20	1/5	20
Bagaż	5/5	100	2/5	40	1/5	20
K5						
BOR	1/5	20	0/5	0	2/5	40
Proszę	1/5	20	0/5	0	1/5	20
Zachować	1/5	20	1/5	20	1/5	20

Się	1/5	20	0/5	0	1/5	20
Zgodnie	3/5	60	4/5	80	4/5	80
Z	3/5	60	5/5	100	4/5	80
Prawem	3/5	60	5/5	100	4/5	80

8. Testy funkcjonalne Google Cloud Speech-to-Text

8.1. Opis funkcjonalny Google Speech API

W opisie uwzględniono w szczególności te funkcje *Google Cloud Speech-to-Text*, które zdaniem autora mają znaczenie w przypadku zastosowania programu podczas szkolenia funkcjonariuszy Biura Ochrony Rządu . Podstawową funkcją Google Speech API jest zamiana mowy na tekst z wykorzystaniem sieci neuronowych. Rozpoznawanie mowy realizowane jest w 120 językach. *Google Cloud Speech-to-Text* zrealizowane jest w architekturze klient - serwer. Próbka przesyłana jest przez Internet od serwera, który odsyła rozpoznany tekst. Testy przeprowadzono wykorzystując mechanizm klawiatury głosowej w urządzeniu z systemem Android (smartfon Samsung Galaxy S5 – rys.5).



Rys. 5. Klawiatura głosowa w systemie Android

8.2. Skuteczność zamiany mowy na tekst w Google Cloud Speech-to-Text dla wielu mówców

W Tabela 9 zestawiono skuteczność rozpoznawania kompletnych wypowiedzi (wszystkich słów w wypowiedzi) w *Google Cloud Speech-to-Text* – komend (dla zbioru POLSL)

Tabela 9

Skuteczność rozpoznawania kompletnych wypowiedzi dla wielu mówców w *Google Cloud Speech-to-Text* (zbiór POLSL)

Komenda	Liczba rozpoznanych pełnych wypowiedzi	%
K1	2/13	15
K2	12/13	92
K3	10/13	77
K4	12/13	92
K5	9/13	69

W Tabela 10 zestawiono skuteczność rozpoznawania poszczególnych słów wchodzących w skład wypowiedzi – komendy w *Google Cloud Speech-to-Text* (dla zbioru POLSL).

Tabela 10

Skuteczność rozpoznawania poszczególnych słów w wypowiedzi w *Google Cloud Speech-to-Text* (zbiór POLSL)

Komenda		%
K1		
BOR	0/13	0
stój	6/13	46
Rzuć	5/13	38
broń	7/13	54
K2		
Stój	12/13	92
bo	12/13	92
strzelam	12/13	92
K3		
Biuro	13/13	100
Ochrony	13/13	100
Rządu	13/13	100
Proszę	12/13	92
Się	9/13	69
Zatrzymać	13/13	100
K4		
Biuro	12/13	92
Ochrony	13/13	100
Rządu	13/13	100
Proszę	12/13	92
Pokazać	13/13	100
Bagaż	13/13	100
K5		
BOR	0/13	0
Proszę	10/13	77
Zachować	11/13	85
Się	11/13	85
Zgodnie	12/13	92
Z	12/13	92
Prawem	12/13	92

Osobnej serii testów poddano materiały (próbki głosowe) przygotowane przez funkcjonariuszy Biura Ochrony Rządu. Komunikaty przygotowane przez BOR zostały nagrane dla czterech mówców w trzech różnych sytuacjach (warunkach środowiskowych).

W tabeli 11 zestawiono skuteczność rozpoznawania kompletnych wypowiedzi (wszystkich słów w wypowiedzi) – komend w programie ARM (dla zbioru BOR).

W Tabela 12 zestawiono skuteczność rozpoznawania poszczególnych słów wchodzących w skład wypowiedzi – komendy w *Google Cloud Speech-to-Text* (dla zbioru BOR).

Tabela 11

Skuteczność rozpoznawania kompletnych wypowiedzi dla wielu mówców w *Google Cloud Speech-to-Text* (zbiór BOR)

Komenda	Sytuacja (warunki)	Liczba rozpoznanych pełnych wypowiedzi	%
K1	S1	0/5	0
	S2	1/5	20
	S3	0/5	0
K2	S1	2/5	40
	S2	2/5	40
	S3	3/5	60
K3	S1	1/5	20
	S2	3/5	60
	S3	0/5	0
K4	S1	1/5	20
	S2	2/5	40
	S3	2/5	40
K5	S1	1/5	20
	S2	2/5	40
	S3	0/5	0

Tabela 12

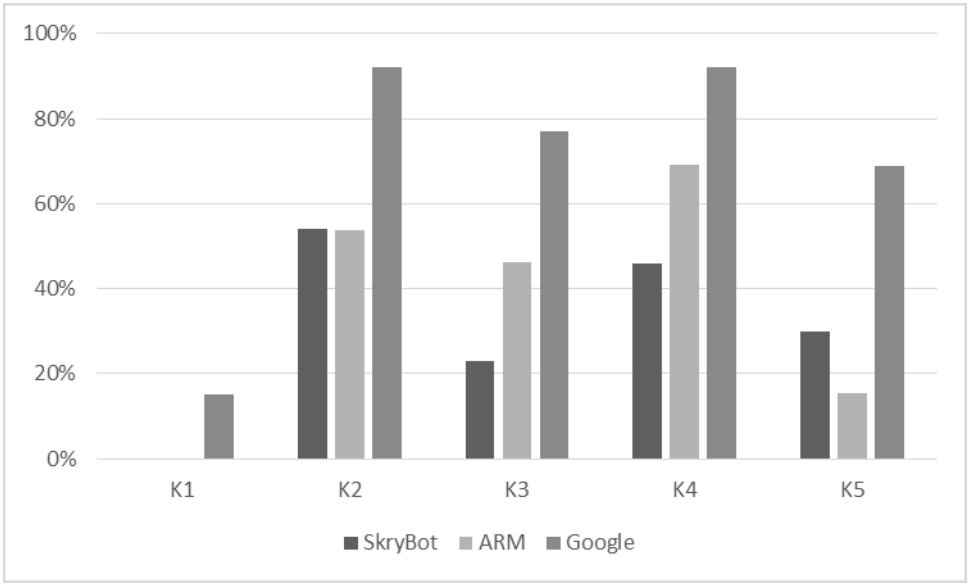
Skuteczność rozpoznawania poszczególnych słów w wypowiedzi w *Google Cloud Speech-to-Text* (zbiór BOR)

Komenda / Warunki	S1	%	S2	%	S3	%
K1						
BOR	0/5	0	0/5	0	0/5	0
stój	0/5	0	0/5	0	0/5	0
Rzuć	1/5	20	1/5	20	0/5	0
broń	2/5	40	1/5	20	0/5	0
K2						
Stój	2/5	40	2/5	40	2/5	40
bo	2/5	40	2/5	40	2/5	40
strzelam	2/5	40	2/5	40	2/5	40
K3						
Biuro	2/5	40	3/5	60	3/5	60
Ochrony	2/5	40	3/5	60	3/5	60
Rządu	3/5	60	3/5	60	3/5	60
Proszę	3/5	60	3/5	60	1/5	20
Się	1/5	20	2/5	20	0/5	0
Zatrzymać	4/5	80	3/5	60	2/5	40

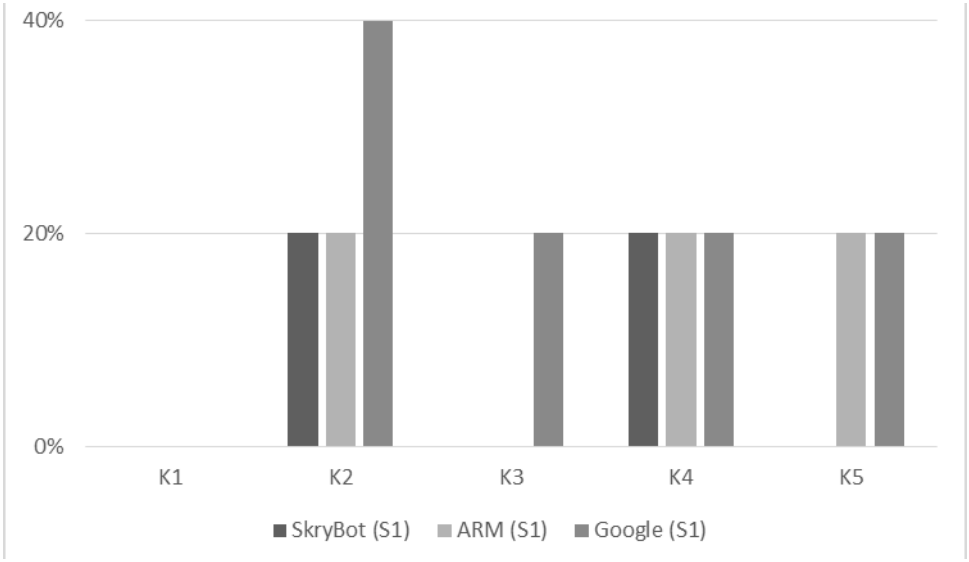
K4						
Biuro	1/5	20	2/5	40	5/5	100
Ochrony	1/5	20	2/5	40	4/5	80
Rządu	2/5	40	2/5	40	4/5	80
Proszę	4/5	80	4/5	80	4/5	80
Pokazać	4/5	80	4/5	80	4/5	80
Bagaż	2/5	40	4/5	80	3/5	60
K5						
BOR	0/5	0	1/5	20	0/5	0
Proszę	3/5	60	3/5	60	0/5	0
Zachować	4/5	80	3/5	60	1/5	20
Się	1/5	20	3/5	60	1/5	20
Zgodnie	3/5	60	4/5	80	3/5	60
Z	3/5	60	4/5	80	3/5	60
Prawem	3/5	60	4/5	80	3/5	60

9. Porównanie skuteczności rozpoznawania mowy dla różnych technologii i wielu mówców

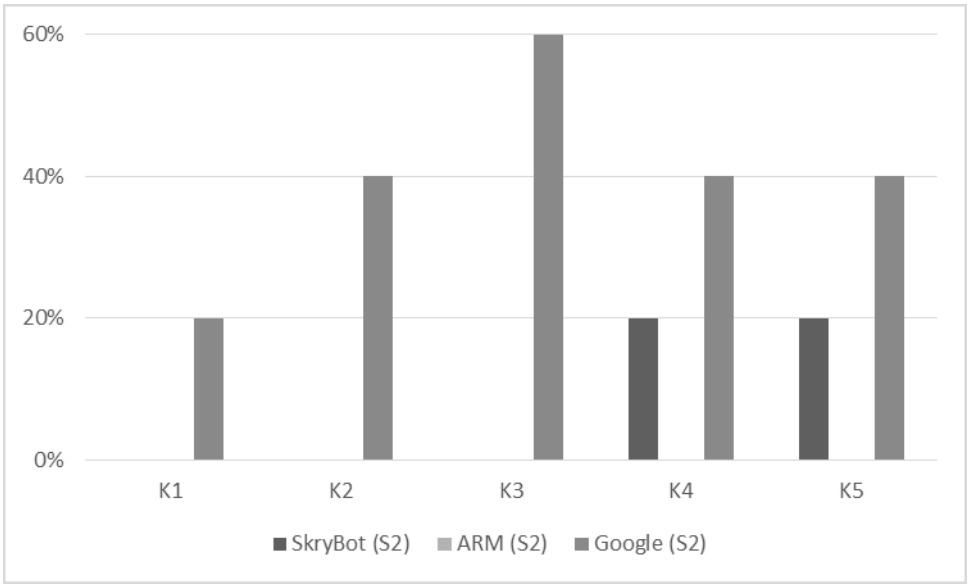
Poniższe wykresy przedstawiają porównanie skuteczności rozpoznawania kompletnych komend (K1-K5) dla zbioru POLSL (rys. 6) oraz dla zbioru BOR z podziałem na różne warunki panujące podczas rejestracji nagrań (sytuacje). Rysunki 7 ÷ 9 odnoszą się do sytuacji S1 do S3.



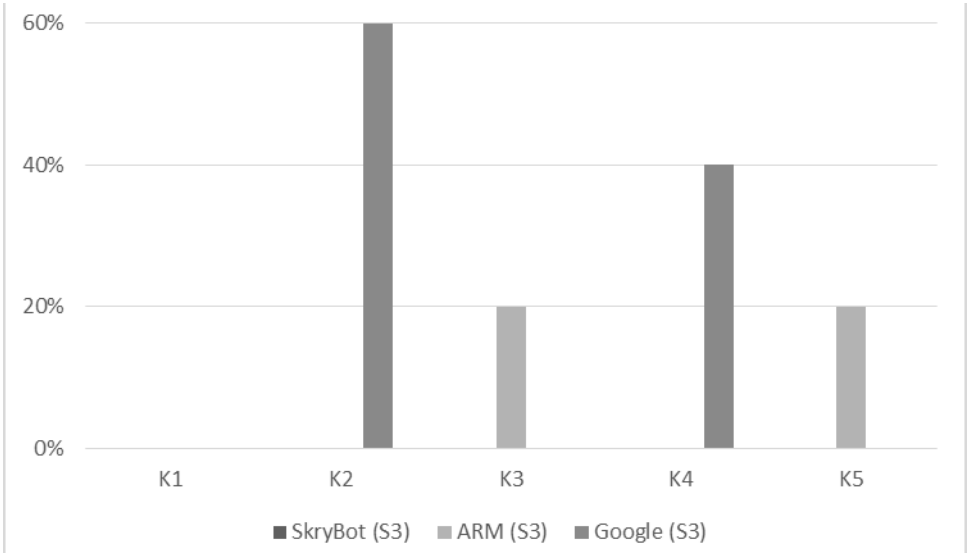
Rys. 6. Porównanie skuteczności rozpoznawania mowy dla zbioru POLSL



Rys. 7. Porównanie skuteczności rozpoznawania mowy dla zbioru BOR – sytuacja 1



Rys. 8. Porównanie skuteczności rozpoznawania mowy dla zbioru BOR – sytuacja 2



Rys. 9. Porównanie skuteczności rozpoznawania mowy dla zbioru BOR – sytuacja 3

10. Podsumowanie i wnioski

Wypowiedzi słowne w zbiorze POLSL to próbki zebrane wśród pracowników Politechniki Śląskiej. Wypowiedzi zwykle nie są nacechowane emocjonalnie, a słowa wypowiadane wyraźnie. Nagrań dokonano w typowych warunkach panujących w pomieszczeniach biurowych.

Wypowiedzi w zbiorze BOR są silnie nacechowane emocjonalnie, słowa wypowiadane są w sposób szybki i zdecydowany, tak jakby mówca chciał wypowiedzieć jak najwięcej słów w jak najkrótszym czasie. Nagrań dokonano w trzech różniących się sytuacjach (warunkach środowiskowych) o nieznanym charakterze.

Przeprowadzona analiza pokazuje, że stan zaawansowania algorytmów i technologii rozpoznawania mowy polskiej wydaje się niedostateczny dla planowanych zastosowań w procesie szkolenia funkcjonariuszy BOR. Pomijając przypadki w których programy rozpoznały poprawnie całą wypowiedź, często zdarzały się rozpoznania całkowicie błędne lub zawierające słowa o podobnym brzmieniu. Testowane algorytmy są przystosowane do rozpoznawania mowy dyktowanej przez pojedynczego mówcę. W takiej sytuacji mówca intuicyjnie stara się wypowiadać słowa wyraźnie, wolno i w sposób wyizolowany. W przypadku programów *SkryBot* i *ARM* producenci specjalnie podkreślają, że należy zachować jak najbardziej idealne warunki środowiskowe (brak szumów i innych dźwięków mogących zniekształcić wypowiadane słowa), oferują również procedury przystosowania modelu do konkretnego mówcy. Jedynie producenci *Google Cloud Speech-to-Text* deklarują, że ich produkt jest odporny na typowe warunki środowiskowe (takie jak podmuchy wiatru, szum występujący w poruszającym się pojeździe itp.). Trudno sobie wyobrazić zachowanie kompletnej ciszy w trakcie typowego szkolenia funkcjonariuszy BOR, czy dostosowanie modelu do pojedynczego mówcy. Należy podkreślić, że w niesprzyjających warunkach środowiskowych również człowiek miałby trudności w rozpoznawaniu wypowiadanych komunikatów, szczególnie pozbawionych jakiegokolwiek szerszego kontekstu.

Najlepszy w zestawieniu okazał się system *Google Cloud Speech-to-Text*, jednak jego wykorzystanie może okazać się niemożliwe ze względu na poufność treści przekazywanych za pomocą komunikatów głosowych.

Praca finansowana ze środków Narodowego Centrum Badań i Rozwoju: DOB-BIO6/11/90/2014, Wirtualny symulator działań ochronnych Biura Ochrony Rządu.

LITERATURA

1. Demenko G., Cecko R., Szymański M., Owsianny M., Francuzik P., Lange M.: Polish Speech Dictation System as an Application of Voice Interfaces. Multimedia Communications, Services and Security, 2012, p. 68-76.
2. Pawlaczyk L., Bosky P.: Skrybot – A System for Automatic Speech Recognition of Polish Language. Man-Machine Interactions, 2009, p. 381-387.

3. Schalkwyk J. et al.: Your Word is my Command: Google Search by Voice: A Case Study. Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, A. Neustein, Ed. Boston, MA: Springer US, 2010, p. 61-90.