



**AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE**

Rozpoznawanie mowy

Definicja, historia, algorytmy

Szymon Durak

**Informatyka Stosowana WIMiIP
Podstawy Sztucznej Inteligencji**

Definicja

Rozpoznawanie mowy jest technologią umożliwiającą urządzeniu elektronicznemu, np. komputerowi, interpretację ludzkiej mowy. Interpretacji może podlegać zarówno mowa jako taka (np. celem transkrypcji), jak również jej znaczenie (celem interakcji z użytkownikiem).

Klasyfikacja metod rozpoznawania mowy

- **Ze względu na segmentację przetwarzanej wypowiedzi** – od pojedynczych fonemów (najmniejsza jednostka języka różnicująca, ale sama pozbawiona znaczenia) po mowę ciągłą, a nawet spontaniczną
- **Ze względu na czas odpowiedzi systemu** – przetwarzanie w czasie rzeczywistym lub przetwarzanie zgromadzonych zasobów
- **Ze względu na stopień zależności od mówcy** – zależne od mówcy, grupy mówców lub niezależne od mówcy
- **Ze względu na rozmiar słownika** – od dziesiątek słów dla rozpoznawania cyfr po dziesiątki tysięcy i więcej dla mowy ciągłej

Historia rozpoznawania mowy

- 1952 – Fonetograf Drayfusa-Graya, służący do zapisywania fonemów, maszyna Davisa rozróżniająca 10 słów (cyfry języka angielskiego)
- 1962 – Shoebox – maszyna firmy IBM rozpoznająca 10 cyfr i 6 innych słów języka angielskiego
- 1975 – zaproponowano system „Dragon” oparty na procesach Markowa
- Lata 70-te – pierwszy użycie algorytmu BeamSearch, rozpoznawanie mowy łączonej, niekomercyjne badania nad rozpoznawaniem mowy ciągłej
- Lata 80-te – stosowanie słowników rzędu kilkunastu tysięcy wyrazów i ukrytych modeli Markowa (HMM)
- Lata 90-te – słowniki liczą dziesiątki tysięcy słów, rozpoznawanie mowy ciągłej
- Współcześnie – trenowanie modelu języka przez Google na setkach miliardów zapytań, milionowy słownik języka angielskiego

Problemy i ograniczenia

- Duża przestrzeń przeszukiwania powodowana rozmiarem danych
- Zmienność mowy ze względu na intonację, płęć, akcent, emocje, tempo, dźwięki tła itd.
- Problemy z rozumieniem mowy i reprezentacją wiedzy
- Nieuwzględnianie mowy ciała, emocji itp.
- Niska deterministyczność sygnału mowy
- Niejednoznaczności językowe, słowa wieloznaczne, słowa o jednakowym lub bardzo zbliżonym brzmieniu
- Zakłócenia przetwarzanego sygnału, utrata jakości
- Wzajemne przeszkadzanie sobie użytkowników korzystających z systemów rozpoznawania mowy
- Zmęczenie użytkownika ciągłym mówieniem
- Trudności w koncentracji na wykonywaniu zadania równocześnie z mówieniem do urządzenia
- Powolny spadek stopy błędów wraz ze wzrostem rozmiaru modelu języka potrzebne są duże pamięci i moce obliczeniowe

Dekodowanie mowy

Ze względu na małą deterministyczność sygnału mowy, do dekodowania stosuje się algorytmy z zakresu kryptoanalizy. Algorytmy przewidują prawdopodobieństwa wystąpienia słów na podstawie dużych, językowych danych statystycznych. Dane te informują o możliwych kombinacjach niewielkich kontekstów. Część danych musi zostać odrzucona przez algorytmy obcinające (pruning) ze względu na ograniczone zasoby sprzętowe i złożoność obliczeniową. Obcinanie odbywa się na poziomie drzew decyzyjnych, modelu języka, przestrzeni przeszukiwania słów. Model języka pozwala na ukierunkowanie przeszukiwania, ale sprawia, że rozwiązania mogą nie być optymalne. Dość nowe jest podejście oparte na metodzie pamięci podręcznej, w wyniku którego model stopniowo przełącza się w bardziej lokalny kontekst. Innymi rozwiązaniami są np. systemy ograniczone do pojedynczej dziedziny, lub pomocnicza obróbka danych na serwerze zewnętrznym.

Algorytmy i modele przydatne w rozpoznawaniu mowy ludzkiej

- **Ukryte modele Markowa** (Hidden Markov Models, HMM) – model uznajemy za łańcuch Markowa z ukrytymi stanami. Można je przedstawić jako najprostszą dynamiczną sieć Bayesa.
- **N-gramy** – statystyczne modele językowe służące przewidywaniu kolejnego elementu sekwencji, stosowane głównie do słów. Sprowadza się do zliczania wystąpienia sekwencji o długości N celem predykcji kolejnego elementu na podstawie N dotychczasowych.
- **Sieci neuronowe**, często łączone z modelami Markowa
- **Dynamiczne sieci Bayesa** – przedstawiające zależności bazując na rachunku prawdopodobieństwa i modelowane za pomocą skierowanych, acyklicznych grafów
- **Analiza cepstralna** – cepstrum możemy rozumieć jako informację o prędkości zmian w poszczególnych pasmach widma dźwiękowego, co przydatne jest w przetwarzaniu mowy
- **Transformacja Fouriera** (algorytm FFT)
- **Nieliniowa transformacja czasowa** (DTW)

Możliwe zastosowania rozpoznawania mowy

- **Sterowanie głosem komputerem osobistym** przez osobę niepełnosprawną, która nie może korzystać z innych interfejsów
- **Sterowanie urządzeniami** o znacznej miniaturyzacji, a zatem ubogich lub w ogóle pozbawionych klasycznych interfejsów interakcji, przy pomocy prostych poleceń głosowych
- **Transkrypcja mowy ciągłej** do tekstu, zarówno przez algorytmy dziedzinowe o ograniczonym słowniku, jak i całościowe, bardziej zaawansowane rozwiązania
- **Systemy dialogowe** – działające w sposób naturalny interfejsy głosowe, nie podlegające ograniczeniu do ustalonych możliwych opcji
- **Translacja speech-to-speech**, czyli pomiędzy dwoma lub więcej językami naturalnymi, może służyć np. do automatycznej translacji komunikatów głosowych wydawanych podróżnym
- **IVR** - Systemy sterowania głosem w biurach obsługi klienta, zastępujące listy dostępnych opcji i konieczność wyboru najbardziej pasującej do problemu

Źródła

- https://pl.wikipedia.org/wiki/Rozpoznawanie_mowy
- <https://pl.wikipedia.org/wiki/Cepstrum>
- <https://pl.wikipedia.org/wiki/N-gram>
- https://en.wikipedia.org/wiki/Hidden_Markov_model
- https://pl.wikipedia.org/wiki/Sie%C4%87_bayesowska
- <https://pl.wikipedia.org/wiki/Fonem>



**AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE**

Rozpoznawanie mowy

**Zastosowanie, aplikacje, rozwiązania
komercyjne**

Karolina Czechowicz

**Informatyka Stosowana WIMiIP
Podstawy Sztucznej Inteligencji**

Rozpoznawanie mowy polskiej

- **VoiceNote II** – umożliwia rozpoznawanie różnych języków, w tym języka polskiego. Współpracuje z Dropboxem oraz Google Drive. Prawidłowe rozpoznanie mowy uzależnione jest od powolnego, wyraźnego czytania tekstu.
- **SkryBot** – dedykowany dla języka polskiego. Współpracuje z wieloma programami zarówno Microsoftu, Google'a, jak i Skypem, Gadu-Gadu czy Open Office. Personalizacja głosu użytkownika. Istnieją wersje dedykowane dla konkretnych grup zawodowych.
- **MagicScribe** – pozwala na minimalizację czasu poświęconego na tworzenie dokumentacji m.in. poprzez eliminację konieczności manualnego posługiwania się komputerem. Implementacja baz słowników branżowych, możliwość przeprowadzenia treningu głosowego.

Rozpoznawanie mowy polskiej

- **Dictation Pro** – jest szybki. Zawiera specjalne polecenia do formatowania, usuwania i kopiowania fragmentów tekstu
- **Google Web Speech API** – w języku polskim w fazie beta; i **Google Voice Search** dla użytkowników mobilnych
- **Sarmata** – stworzona przez Zespół Przetwarzania Sygnałów AGH. Obsługa ograniczonej ilości komend, interfejs głosowy wybranych programów lub urządzeń (ilość obsługiwanych komend zależna jest od warunków rejestracji nagrań), wyszukiwarka akustyczna.

Rozpoznawanie mowy angielskiej

- **Dragon Naturally Speaking** – narzędzie do codziennej pracy z komputerem i dokumentacją, dodatkowe wyspecjalizowane narzędzia dla prawa i służb mundurowych
- **IBM ViaVoice** – wymaga treningu, polegającego na czytaniu określonych słów i fraz, co umożliwia adaptację programu do użytkownika, obsługuje różne języki: angielski, chiński, niemiecki, japoński
- **Call Steering** – oprogramowanie IVR, które wyklucza żmudne nawigowanie klienta po skomplikowanym menu i pozwala mu na opisanie problemu własnymi słowami i przekierowuje go do odpowiedniego działu, usprawnia działanie call center

Inteligentny osobisty asystent

- W smartfonach marki Samsung – **S Voice**
- **Cortana** Microsoftu – obsługiwana także przez Androida i iOS
- Na systemie Apple iOS – **Siri**
- **Google Now** – dostępny w systemie Android oraz jako aplikacja na iOS, Windows i macOS

Jak to działa?

1. Rozpoznanie mowy. Porównanie ze wzorcem, nauka nowych wzorców
2. Interpretacja komunikatów. System posiada słownik, na podstawie którego przyporządkowuje słowa do danego znaczenia lub konkretnego serwisu.
3. Komunikat zwrotny. Wyświetlenie informacji z innych serwisów lub przetworzenie i przeczytanie odpowiedniego tekstu.

Źródła

- https://pl.wikipedia.org/wiki/Rozpoznawanie_mowy
- <http://dsp.agh.edu.pl/sarmata/>
- https://en.wikipedia.org/wiki/IBM_ViaVoice
- <http://www.sestek.com/call-steering/>



**AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE**

Rozpoznawanie mowy

Najpopularniejsze API

Grzegorz Chromik

**Informatyka Stosowana WIMiIP
Podstawy Sztucznej Inteligencji**

CLOUD SPEECH API (OD GOOGLE)

- Umożliwia konwersję audio do tekstu dla ponad 80 języków i ich wariantów dzięki wykorzystaniu sieci neuronowych.
- Potrafi zapisywać mowę rejestrowaną przez mikrofon jaki i z podesłanych plików
- możliwe do wykorzystania na każdym urządzeniu potrafiącym wysłać REST lub gRPC request.
- Cena- do godziny miesięcznie darmowa, powyżej \$0.006 za 15 sekund.

Bing Speech API

- Obsługuje tylko 7 języków mowa -> tekst
- 17 języków i ich odmian tekst -> mowa
- Cennik :

Za darmo	5K rozmów/miesiąc	
Tekst -> mowa	1000 znaków/rozmowa	\$4/1000 rozmów
Krótką mowa ->tekst	15s/rozmowa	\$4/1000 rozmów
Długa mowa ->tekst	2 min/rozmowa	0-10h : \$9/h, 10-100h : \$7.50/h, pow.100h : \$5.50/h

- Nastawione na odczytywanie poleceń wydawanych głosowo
- Przykład dla inteligentnego domu:
„Set thermostat to 72 degrees.”
{
 "action": "thermostat.setTemperature",
 "degrees": "72",
 "units": "F",
}
- Cennik: wersja płatna „PREFERRED Let's Talk”

Speechmatics

- 18 języków, deklarują, że ich AI potrafi nauczyć się nowego w czasie kilku tygodni
- Przetwarzanie mowy na tekst(dla dużych plików)
- Ceny:

Kwota	Minuty	Cena za minutę
£10.00	166	6p
£100.00	1,666	6p
£1,000.00	20,000	5p
£10,000.00	250,000	4p

Źródła

- <https://cloud.google.com/speech/>
- <https://www.microsoft.com/cognitive-services/en-us/speech-api>
- <https://api.ai/>
- <https://speechmatics.com/>