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 – pierwsz 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, płeć, 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 przewidywanie kolejnego elementu sekwencji, stosowane głównie do słów. Sprowadza się do zliczania wystąpienia sekwencji o długości N celem predykcji kolejnego elementu na podstawie N dotychczasowych. Siłą N-gramów jest prostota i skalowalność, wadą – ogromne dane językowe potrzebne do uzyskania dobrych N-gramów.
- **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

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 przewidywanie 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 wydawancych 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