

# Automatsko prepoznavanje govora

Seminarski rad u okviru kursa  
Metodologija stručnog i naučnog rada

Vladimir Vuksanović, vladevuksan99@gmail.com  
Aleksa Kojadinović, kojadinovic.aleksa98@gmail.com  
Lazar Čeliković, celikoviclazar@hotmail.com

Matematički fakultet

18. decembar 2021.

# Sadržaj

- 1 Uvod
- 2 Izazovi
- 3 Statistički model
- 4 End-to-end model
- 5 Metrike za evaluaciju
- 6 Literatura

# Uvod

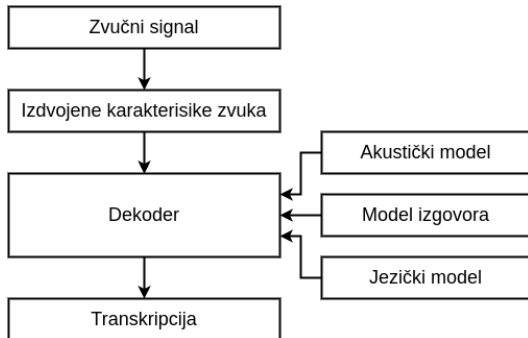
- **Automatsko prepoznavanje govora** (eng. *Automatic Speech Recognition, ASR*) je proces pretvaranja zvučnog signala govora u odgovarajući niz reči pomoću računara
- Neke od najznačajnijih primena su:
  - pametni lični asistenti (Google Assistant, Apple Siri, ...)
  - transkripcija i pretraživanje audio sadržaja
  - automatsko titlovanje snimaka
  - pristupačnost (eng. *accessibility*)

# Izazovi

- ① Mala količina podataka za trening
- ② Stil govora
  - Izolovane reči
  - Povezane reči
  - Neprekidan govor
  - Spontani govor
- ③ Karakteristike govornika (pol, starost, brzina govora...)
- ④ Okruženje govornika (pozadinska buka, oprema za snimanje)
- ⑤ Veličina rečnika

# Statistički model

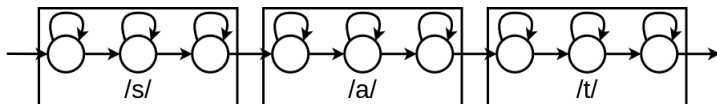
- Koriste statističke metode za određivanje najverovatnije transkripcije
- Ako je  $X$  ulazni zvuk, traži se najverovatniji niz reči  
$$\hat{W} \approx \operatorname{argmax}_{W,S} P(X|S)P(S|W)P(W)$$



Slika: Struktura statističkog modela

# Akustički model i model izgovora

- Akustički model
  - Predviđa verovatnoće koliko ulazni zvuk odgovara nizu fonema
  - Foneme su najmanje jezičke jedinice na osnovu kojih mogu da se razlikuju značenja većih jedinica
  - Implementiran skrivenim Markovljevim modelom
- Model izgovora
  - Mapira reči u njihov način izgovora (fonetski zapis)
  - Definisan od strane eksperta za jezik
  - Određuje način povezivanja modela fonema u model reči



Slika: Primer skrivenog Markovljenog modela za reč "sat"

# Jezički model

- Određuje verovatnoću predviđanja rečenice na osnovu:
  - relativne učestalosti reči
  - redosleda reči
  - sintaksne ispravnosti
  - semantičke ispravnosti
- Implementiran pomoću n-grama
- Dužina n-grama je obično 3 i smanjuje se dok se ne pronade prvo pojavljivanje u trening skupu

# End-to-end model

[1] [2]



# CTC (Connectionist Temporal Classification)

# Modeli zasnovani na pažnji



# Word Error Rate (WER)

$$WER = \frac{I + D + S}{N}$$

gde je:

- $I$  broj umetnutih reči
- $D$  broj obrisanih reči
- $S$  broj zamenjenih reči
- $N$  ukupan broj reči u referenci

# Literatura

-  A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, (New York, NY, USA), pp. 369–376, Association for Computing Machinery, 2006.
-  W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, “Listen, attend and spell,” 2015.