

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

21. decembar 2021.

Sadržaj

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

Uvod

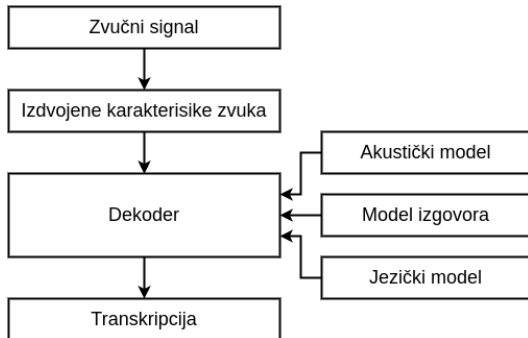
- **Automatsko prepoznavanje govora** (eng. *Automatic Speech Recognition, ASR*) je proces pretvaranja zvučnog signala govora u odgovarajući niz reči pomoću računara
- Neke od najznačajnijih primena su:
 - pametni lični asistenti (Google Assistant, Apple Siri, ...)
 - transkripcija i pretraživanje audio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

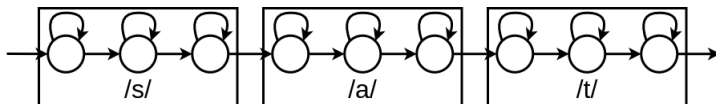
- Koriste statističke metode za određivanje najverovatnije transkripcije
- Ako je X ulazni zvuk, traži se najverovatniji niz reči
$$\hat{W} \approx \operatorname{argmax}_{W,S} P(X|S)P(S|W)P(W)$$



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

- Zasnovan na dubokim neuronskim mrežama
- Ima prostiju strukturu od statističkog modela
- On se bavi direktnim prevodenjem ulaznog signala u niz grafema, karatera ili reči
- Treniranje postaje lakše zbog jednostavnije strukture
- Sa druge strane potrebna je veća količina podataka nego ako se koristi statistički model
- U odnosu na način rešavanja, razlikujemo **CTC model** [1] i **model zasnovan na pažnji** [2]

CTC (Connectionist Temporal Classification)

- Cilj je napraviti preslikavanje h koje slika proizvoljan zvučni signal u niz labela:

$$h : (\mathbb{R}^m)^* \rightarrow L^*$$

- Ako O označava test skup, tada se funkcija greške definiše na sledeći način:

$$LER(h, O) = \frac{1}{|O|} \sum_{(\mathbf{x}, \mathbf{z}) \in O} \frac{ED(h(\mathbf{x}), \mathbf{z})}{|\mathbf{z}|}$$

- Implementiran kao rekurentna neuronska mreža koja za svaki od T trenutaka ima m ulaza, a izlaz je softmax sloj dimenzije $|L'|$, gde je L' azbuka proširena blanko karakterom

Modeli zasnovani na pažnji

- Za razliku od CTC modela, ovi modeli se oslanjaju na **mehanizam pažnje** (eng. *attention mechanism*) i odbacuju pretpostavku o nezavisnosti zvučnih segmenata
- Glavne komponente ovog modela su:
 - **Slušalac** – rekurentna neuronska mreža piramidalne strukture čiji zadatak je da napravi reprezentaciju zvuka višeg nivoa
 - **Speler** – rekurentna neuronska mreža koja proizvodi konačan niz karaktera
- Koristeći koncepte stanja i konteksta, verovatnoća odabira narednog karaktera zavisi od svih prethodnih karaktera



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.
-  J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” 2015.