

# Automatsko prepoznavanje govora

Seminarski rad u okviru kursa  
Metodologija stručnog i naučnog rada  
Matematički fakultet

Vladimir Vuksanović, Aleksa Kojadinović, Lazar Čeliković  
kontakt email prvog, drugog, trećeg autora

16. novembar 2021.

## Sažetak

U ovom tekstu je ukratko prikazana osnovna forma seminarskog rada. Obratite pažnju da je pored ove .pdf datoteke, u prilogu i odgovarajuća .tex datoteka, kao i .bib datoteka korišćena za generisanje literature. Na prvoj strani seminarskog rada su naslov, apstrakt i sadržaj, i to sve mora da stane na prvu stranu! Kako bi Vaš seminarski zadovoljio standarde i očekivanja, koristite uputstva i materijale sa predavanja na temu pisanja seminarskih radova. Ovo je samo šablon koji se odnosi na fizički izgled seminarskog rada (šablon koji *morate* da koristite!) kao i par tehničkih pomoćnih uputstava. Pročitajte tekst pažljivo jer on sadrži i važne informacije vezane za zahteve obima i karakteristika seminarskog rada.

**Ključne reči:** prepoznavanje govora

## Sadržaj

<b>1</b>	<b>Uvod</b>	<b>2</b>
<b>2</b>	<b>Izazovi</b>	<b>2</b>
<b>3</b>	<b>Statistički model</b>	<b>3</b>
3.1	Procesiranje zvucnog signala . . . . .	4
3.2	Akustički model . . . . .	4
3.3	Model izgovora . . . . .	5
3.4	Jezicki model . . . . .	5
3.5	Dekodiranje . . . . .	5
<b>4</b>	<b>End-to-end model</b>	<b>5</b>
4.1	CTC model . . . . .	5
4.2	modeli zasnovani na pažnji . . . . .	6
<b>5</b>	<b>Metrike za evaluaciju</b>	<b>6</b>
<b>6</b>	<b>Zaključak</b>	<b>6</b>
	<b>Literatura</b>	<b>6</b>
<b>A</b>	<b>Pregled skupova podataka</b>	<b>7</b>

# 1 Uvod

Govor je za ljude najintuitivniji i prirodni način komunikacije. Zbog toga je od samog nastanka kompjutera, nastala i ideja da koristimo isti način komunikacije da interagujemo sa njima. To bi znatno smanjilo potrebu predznanje za korišćenje kompjutera i učinilo ga pristupačnijim većem broju ljudi. Najveća prepreka ovim sistemima do skoro je bio kako sa velikom tačnošću prepoznati šta je korisnik rekao. Taj postupak se naziva automatsko prepoznavanje govora.

**Automatsko prepoznavanje govora** (eng. *Automatic Speech Recognition, ASR*) je proces pretvaranja zvučnog signala govora u sekvencu reči pomoću kompjutera. Neke od najznacajnijih primena ovih sistema su: pametni licni asistenti (Google Assistant<sup>1</sup>, Apple Siri<sup>2</sup>,...), transkripcija snimaka, pretraživanje audio sadržaja i pristupačnost.

Iako su istraživanja na ovu temu počela još sredinom dvadesetog veka, popularnost je počela da dobija tek u poslednjoj deceniji kada je uvođenje dubokih neuronskih mreža drastično povećalo performanse ovih sistema. Ta razlika je bila dovoljna da učini ove sisteme praktično primenljivim umesto nezgodnim za upotrebu zbog velikog broja gresaka. Jedan od najznacajnijih postignuća je ostvareno 2016. godine je kompanija Majkrosoft napravila sistem koji je ostvario iste rezultate kao ljudski eksperti na transkripciji Switchboard skupa podataka [1]. Za glavne uzroke ovog naglog poboljšanja se smatraju [2]:

1. Sakupljanje velike količine transkribovanih skupova podataka
2. Nagli porast u performansama grafičkih procesorskih jedinica (GPU)
3. Poboljšanje algoritama za učenje i arhitektura modela

U nastavku ćemo prvo navesti neke izazove koje treba da rešimo da bi smo napravili dobar sistem za prepoznavanje govora, zatim ćemo opisati način rada dva najpopularnija modela: statistički i end-to-end i na kraju ćemo predstaviti način za njihovu evaluaciju.

## 2 Izazovi

Prepoznavanje govora je veoma težak zadatak zato što je potrebno da radi podjednako dobro u veoma različitim uslovima. Neki od najvećih izazova su:

- **Mala količina podataka za trening** — Za ostvarivanje dobrih rezultata potrebno je sakupiti više stotina ili čak hiljada sati labeliranih zvučnih snimaka koji treba da sadrže više govornika različitog pola i starosti, koji govore različitim akcentima. Dok u skorije vreme jeste nastao porast u količini dostupnih podataka, veliki problem još uvek predstavlja reprezentativnost različitih varijacija u govoru i nedostatak podataka za jezike sa manjim brojem govornika. Zbog toga se istražuju alternativni načini za treniranje kao što su samotreniranje (eng. *self-training*) [3], iterativno treniranje [4] ili treniranje koristeći kompjuterski generisan glas [5]. U dodatku A se može naći tabela sa pregledom nekih od najpopularnijih trening skupova na engleskom jeziku.

---

<sup>1</sup><https://assistant.google.com/>

<sup>2</sup><https://www.apple.com/siri/>

- **Stil govora** — Postoje različiti sistemi u zavisnosti od toga koji tip govora mogu da prepoznaju [6]. Tipovi govora poredjani po težini prepoznavanja su:

1. Izolovane reci — reci su razdvojene dugim periodima tisine
2. Povezane reci — reci su razdvojene kratkim pauzama
3. Nепrekidan govor — uvezbani govor, citanje ili diktiranje
4. Spontan govor — neuvezbani, prirodni govor

Prve implementacije prepoznavanja govora su radile na nivou izolovanih reci i koristile su se za prepoznavanje određenih komandi ili cifara. Danas se najviše truda ulaze u poboljšanje neprekidnog i spontanog govora.

- **Karakteristike govornika** — Svaki čovek ima različitu boju glasa i govori različitom brzinom. Čak i starost osobe i jčina govora bitno uticu na frekvenciju glasa. Poseban problem pravi postojanje različitih dijalekata i akcenata koji mogu da imaju potpuno različite načine za izgovaranje istih reci. Jedan način za rešavanje ovog problema je treniranje sistema na glasu govornika koji će ga koristiti. To su sistemi zavisni od korisnika (eng. *speaker dependent*) i koriste se u slučajevima da samo jedno osoba treba da ih koristi. Sa druge strane postoje sistemi nezavisni od korisnika (eng. *speaker independent*) koji treba da rade podjednako dobro za sve govornike.
- **Okruženje govornika** — Ovi sistemi će retko biti korišćeni u potpuno tihim prostorijama sa profesionalnom opremom za snimanje, zbog toga treba da budu tolerantni na različite vrste pozadinske buke ili kvaliteta mikrofona (koliko je to moguće). Neke vrste sumova je moguće otkloniti analizom zvuka ili naprednijim metodama [7], ali jedan od najvećih problema predstavlja postojanje drugih govornika u okolini. Te signale je često teško razlikovati od glasa primarnog govornika, i samim tim teško ukloniti.
- **Velicina rečnika** — Povećanje broja reci koje model može da prepozna takodje povećava njegovu složenost i otežava treniranje, ali se time dobija na tačnosti. Zbog toga je potrebno naći dobar kompromis između velicine rečnika i složenosti modela. U slučajevima kada je potrebno pouzdano prepoznati samo neki skup komandi koriste se mali rečnici i oni su često veoma pouzdani, ali za prepoznavanje opsteg govora današnji sistemi su trenirani na skupu od oko 50.000-100.000 reci.

### 3 Statisticki model

Dugo vremena statisticki pristup je bio dominantan za sisteme za prepoznavanje govora. Cilj ovih sistema je da pronadju najverovatniju transkripciju za zadati signal. Formalno, neka je  $\hat{W}$  optimalan niz reci za transkripciju nekog zvučnog signala  $X$ . Cilj je optimizovati formulu [8]:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|X)$$

primenom Bajesove formule to možemo da zapisemo kao:

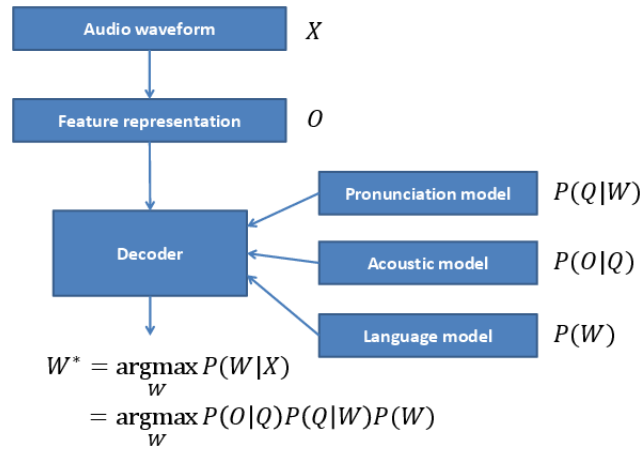
$$\hat{W} = \underset{W}{\operatorname{argmax}} \frac{P(X|W)P(W)}{P(X)}$$

a kako je  $P(X)$  konstantno za konkretan ulaz, mozemo da ga eliminisemo:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(X|W)P(W)$$

Prvi deo,  $P(X|W)$ , racuna akusticki model, a drugi,  $P(W)$ , racuna jezicki model. Ideja je da odvojeno optimizujemo obe prethodne velicine i ocekujemo da ce se tako maksimizirati ukupna verovatnoca. U praksi,  $P(X|W)$  je tesko precizno odrediti na nivou reci zbog velikih varijacija u izgovoru, pa se cesce odredjuje na nivou glasova ili fonema. Zbog toga se dodatno uvodi i model izgovora (eng.) koji kombinuje te manje jedinice u reci.

Na slici 1 je prikazana cela struktura statistickog modela od zvucnog signala do transkripcije: U nastavku ce biti opisana svaka od pomenutih



Slika 1: Statisticki model

komponenti, njena uloga i nacin rada.

### 3.1 Procesiranje zvucnog signala

Sirovi zvucni signal je veoma nepogodan za koriscenje zato sto sadrzi veliku kolicinu nepotrebnih informacija i šuma. Zbog toga se pre prosledjivanja akustickom modelu, signal prvo obradjuje tako da ostanu samo kljucne karakteristike i smanji sum i velicina reprezentacije.

Signal delimo na kratke segmente koje zovemo okviri (eng. frame). Svaki od njih je fiksne duzine (obicno 10-30 milisekundi) sa kratkim preklapanjem susednim okvirima radi smanjenja naglih promena prilikom prelaska iz jednog u drugi. Pretpostavka je da je u svakom od okviru glas konstantan, to jest da se glasovi mogu menjati samo prelaskom iz jednog okvira u drugi. Na svaki od tih novodobijenih delova se zatim primenjuje neka vrsta spektralne analize kojom se izdvajaju samo njegove najbitnije karakteristike. Tacna reprezentacija koja se koristi varira u zavisnosti od modela, ali jedna od najpopularnijih je MFCC [?].

### 3.2 Akusticki model

Akusticki model je zaduzen da pretvori sekvencu okvira u sekvencu fonema. Za modelovanje svake foneme se koriste skriveni Markovljevi modeli (eng. *Hidden Markov Models*) [9]. Ti modeli se zatim nadovezuju da grade slova, reci...

### 3.3 Model izgovora

U slučaju da želimo da prepoznamo samo malu količinu reci, prethodni model bi bio sasvim dovoljan pod uslovom da imamo dobar trening skup za te reci. Problem nastaje ako želimo da prepoznamo veću količinu reci, tada ćemo za neku rec imati možda samo par primera u celom skupu. To nije dovoljno da se istrenira model.

$$P(X|W) = P(X|Q) \sum_Q P(Q|W)$$

$P(Q|W)$  zovemo model izgovora (eng. *Pronunciation Model*). On mapira niz fonema u odgovarajuću rec ili niz slova koja se tako izgovara.

### 3.4 Jezicki model

Odredjuje najverovatniju kombinaciju reci

### 3.5 Dekodiranje

U praksi je rečnik prilično veliki pa nije praktično računati verovatnoću za svaku kombinaciju reci u rečenici. Umesto toga se koristi beam search algoritam.

## 4 End-to-end model

### 4.1 CTC model

Neka je ulazni zvuk uzorkovan u proizvoljnom broju jednakih vremenskih intervala, gde je svaki od njih predstavljen vektorom realnih brojeva dužine  $m$ . Neka je  $L$  konačna azbuka labela (oznaka). Cilj je napraviti preslikavanje  $h$  koje slika proizvoljan zvučni signal u niz labela:

$$h : (\mathbb{R}^m)^* \rightarrow L^* \quad (1)$$

U praksi fiksiramo broj vremenskih trenutaka na neku vrednost  $T$ . Iako je broj trenutaka fiksiran za zvučne signale, nizovi labela ne moraju biti iste dužine za svaku ulaznu instancu, stoga jednu trening instancu predstavlja par  $(\mathbf{x}, \mathbf{z})$  gde je  $\mathbf{z}$  vektor labela dužine najviše  $T$ . Ako bismo test skup označili sa  $O$ , tada funkciju greske možemo definisati na sledeći način:

$$LER(h, O) = \frac{1}{|O|} \sum_{(\mathbf{x}, \mathbf{z}) \in O} \frac{ED(h(\mathbf{x}), \mathbf{z})}{|\mathbf{z}|} \quad (2)$$

$ED$  predstavlja edit distancu (minimalni broj izmena koji dovodi jedan niz karaktera do drugog, pri čemu dozvoljene izmene podrazumevaju brisanje, supstituciju i umetanje karaktera). Prethodna mera naziva se stopa greske labela (eng. *label error rate* -  $LER$ ).

Formula 2 jeste prirodna ocena greske za probleme koji za cilj imaju minimizaciju greske prevodjenja.

Koristeći sva prethodna razmatranja konstruisemo rekurentnu neuron-sku mrežu koja na ulazu ima  $mT$  ulaza, dok se na izlazu dobija  $T$  vektora dimenzija  $L' = L \cup \{\epsilon\}$  pri čemu svaki predstavlja raspodelu verovatnoća oznaka za svaki trenutak proširujući azbuku blanko labelom  $\epsilon$ .

Neformalno receno, prolaskom kroz izlazne softmax nizove dobijamo putanju  $\pi \in (L')^T$  koja predstavlja jedan moguci odabir labela. Ako  $y_{\pi_t}^t$  predstavlja softmax ???? vrednost  $t$ -tog trenutka oznake  $\pi_t$  tada verovatnocu odabira kompletne putanje dobijamo kao proizvod:

$$p(\pi|x) = \prod_{t=1}^T y_{\pi_t}^t$$

U praksi uzorkovanje zvuka se vrši u veoma sitnim vremenskim intervalima (oko 10ms). Stoga je pojava blanko ili dupliciranih oznaka veoma česta. Iz tog razloga uvodimo preslikavanje  $\beta$  čija je uloga preciscavanje nizova labela uklanjanjem blanko oznaka i susednih duplikata.

$$\beta : (L')^T \rightarrow L^U, U \leq T$$

Prisetimo da za jednu preciscenu putanju  $l$  može postojati više mogućih izvornih putanja, pa je verovatnoca njenog odabira jednaka sumi po svim izvornim putanjama.

$$p(l|x) = \sum_{\pi \in \beta^{-1}(l)} p(\pi|x) \quad (3)$$

Imajuci u vidu sve prethodno navedeno, zadatak preslikavanja  $h$  je odabir najverovatnije preciscene putanje za dati ulaz.

$$h(x) = \operatorname{argmax}_{l \in L^U} p(l|x) \quad (4)$$

Najjednostavniji algoritam jeste pohlepni odabir najbolje oznake za svaki vremenski trenutak ponaosob. Medjutim, ovakav pristup ne garantuje optimalnost. Naravno, postoje bolji algoritmi za resavanje datog problema. Isti neće biti obradjivani u ovom radu, ali se mogu naci u [].

## 4.2 modeli zasnovani na paznji

## 5 Metrike za evaluaciju

Standardna mera za procenu kvaliteta sistema za prepoznavanje govora je stopa pogresnih reci (eng. *Word Error Rate (WER)*).

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

gde je  $S$  broj zamenjenih reci,  $D$  broj obrisanih reci,  $I$  broj umetnutih reci, i  $N$  ukupan broj reci u referentnoj recenici. Minimalna vrednost koju može da dobije je 0, dok maksimalna vrednost može da bude preko 1 (npr. veliki broj umetnutih reci). Stopa pogresnih reci se efikasno racuna dinamickim programiranjem, pomocu Vagner-Fiserovog algoritma (eng.)

Cilj sistema za prepoznavanje govora je da minimizuje ovu vrednost.

## 6 Zaključak

## Literatura

- [1] Microsoft, "Historic achievement: Microsoft researchers reach human parity in conversational speech recognition," 2016.

- [2] A. Hannun, “The history of speech recognition to the year 2030,” 2021.
- [3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” 2020.
- [4] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, “Improved noisy student training for automatic speech recognition,” *Interspeech 2020*, Oct 2020.
- [5] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, “Deep speech: Scaling up end-to-end speech recognition,” 2014.
- [6] M. A. Anusuya and S. K. Katti, “Speech recognition by machine, a review,” 2010.
- [7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 23, p. 7–19, jan 2015.
- [8] U. Kamath, J. Liu, and J. Whitaker, *Deep Learning for NLP and Speech Recognition*. Springer Publishing Company, Incorporated, 1st ed., 2019.
- [9] L. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

## A Pregled skupova podataka

TIMIT	cell2	cell3
SwitchBoard	cell5	cell6
LibriSpeech	cell8	cell9
CommonVoice	a	a