Automatsko prepoznavanje govora

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

Vladimir Vuksanović, Aleksa Kojadinović, Lazar Čeliković
vladevuksan99@gmail.com, drugog, trećeg autora

20. novembar 2021.

Sažetak

Prepoznavanje govora je jedan od osnovnih zadataka iz oblasti obrade prirodnog jezika (eng. *natural laguage processing*). Iako je prirodan za ljude, ovo je uvek bio težak zadatak za mašine. Tema ovog rada je da približi čitaoca zadatku automatskog prepoznavanja govora, problemima koji se javljaju i najznačajnijim arhitekturama ovih sistema. Biće reči o klasičnim, statističkim modelima kao i modernijim modelima potpuno zasnovanim na dubokim neuronskim mrežama. Poseban deo će biti izdvojen za opisivanje načina evaluacije modela i pregled performansi za neke od najznačajnijih.

Ključne reči: prepoznavanje govora

Sadržaj

1 Uvod

Govor je za ljude najintuitivniji i najprirodniji način komunikacije. Zbog toga je u toku razvoja računarskih sistema, nastala ideja da isti vid komunikacije koristimo i za interakciju sa kompjuterima. To bi znatno smanjilo potrebno predznanje za njihovo korišćenje i učnilio ih pristupačnijim većem broju ljudi. Najveća prepreka ovoj ideji do skoro je bilo kako sa velikom tačnosti 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 odgovarajući niz reči pomoću kompjutera. Neke od najznačajnijih primena ovih sistema danas su pametni lični asistenti (Google Assistant[1], Apple Siri[2],...), transkripcija i pretraživanje audio sadržaja, automatsko titlovanje snimaka i pristupačnost.

Iako su istraživanja na ovu temu počela još sredinom dvadesetog veka, ovi sistemi su popularnost stekli tek u poslednjoj deceniji kada je uvođenje dubokih neuronskih mreža drastično povećalo njihove performanse. Ta razlika je bila dovoljna da učini ove sisteme praktično primenljivim u odnosu na prethodne koji su bili nezgodni za upotrebu zbog velikog broja grešaka koje su pravili. Prethodna decenija je videla i izjednačavanje sposobnosti računara i ljudi u prepoznavanju govora kada 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 ogromne količine tanskribovanih skupova podataka
2. Nagli porast u performansama grafičkih procesorskih jedinica
3. Poboljšanje algoritama za mašinsko učenje i arhitektura modela

U nastavku ovog rada će prvo biti navedeni neki izazovi na koje se nailazi prilikom modelovanja sistema za prepoznavanja govora, zatim će biti opisan način rada dva najpopularnija modela: statistički i end-to-end i na kraju će biti predstavljen 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 razlicitog 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 sto 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.

[1] https://assistant.google.com/
[2] https://www.apple.com/siri/

- **Stil govora** — Postoje različiti sistemi u zavisnosti od toga koji tip govora mogu da prepoznaju [6]. Tipovi govora poređani po težini propoznavanja su:
  1. Izolovane reči — reči su razdvojene dugim periodima tišine
  2. Povezane reči — reči su razdvojene kratkim pauzama
  3. Neprekidan govor — uvežbani govor, čitanje ili diktiranje
  4. Spontani govor — neuvežbani, prirodni govor

  Prve implementacije prepoznavanja govora su radile na nivou izolovanih reči i koristile su se za prepoznavanje određenih komandi ili cifara. Danas se najviše truda ulaže u poboljšanje prepoznavanja neprekidnog i spontanog govora.
- **Karakteristike govornika** — Svaki čovek ima različitu boju glasa i govori različitom brzinom. Čak i starost osobe i jačina govora bitno utiču na frekvenciju glasa. Poseban problem pravi postojanje različitih dijalekata i akcenata koji mogu da imaju potpuno različite načine za izgovaranje istih reči. 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 jedna 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. Neke vrste šumova 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 primarnig govornika, i samim tim teško ukloniti.
- **Veličina rečnika** — Povećanje broja reči koje model može da prepozna takođe povećava njegovu složenost i otežava treniranje, ali se time dobija na tačnosti. Zbog toga je potrebno naći dobar kompromis između veličine 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 opšteg govora današnji sistemi su trenirani na skupu od oko 50.000-100.000 reči.

3 Statistički model

Dugo vremena statistički pristup je bio dominantan za sisteme za prepoznavanje govora. Iako je u skorije vreme pao u senku modela zasnovanih na dubokim neuronskim mrežama opisanim u poglavlju 4 ovaj model je jos uvek u širokoj upotrebi i veoma vredan izučavanja.

Cilj ovih sistema je da pronađu najverovatniju transkripciju za zadati ulaz. Formalno, neka je $\hat{W}$ optimalan niz reči za transkripciju nekog zvučnog signala X. Cilj je optimizovati formulu [8]:

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

primenom Bajesove formule to se može zapisati kao:

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

a kako je $P(X)$ konstantno za konkretan ulaz, može da se eliminiše:

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(X|W)P(W) \tag{1}$$

Ideja je umesto modeliranja $P(W|X)$ što je teško, odvojeno modelirati verovatnoće iz prethodne formule jer za to postoje bolje tehnike.

Najprirodniji način za računanje $P(X|W)$ bi bio da se W podeli na reči i za svaku od njih računa verovatnoća da je izgovorena. U nekim slučajevima, kao na primer kada treba prepoznati samo neki mali skup komandi, tada se i koristi ovaj pristup. Problem nastaje u sistemima sa velikim vokabularom zato što postoji veliki broj varijacija u izgovoru reči, a trening skup sadrži možda par primera za svaku od njih što nije dovoljno da se dobro nauči njeno prepoznavanje. Dakle, umesto na reči potrebna je finija podela i za tu potrebu se koriste foneme. **Foneme** su najmanje jezičke jedinice na osnovu kojih mogu da se razlikuju značenja većih jedinica. One postoje samo kao apstraktna ideja, a njihova fizicka realizacija se zove glas.

Ako je S niz fonema, $P(X|W)$ iz formule 1 se razlaže i dobija se:

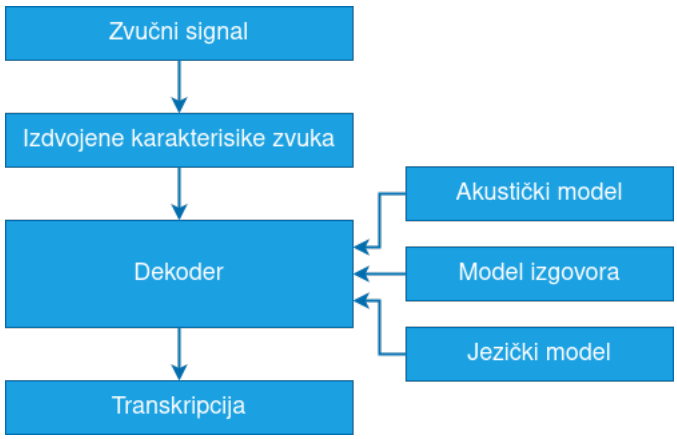
$$\hat{W} = \underset{W}{\operatorname{argmax}} \sum_{S} P(X, S|W)P(W)$$

što se može aproksimirati kao:

$$\hat{W} \approx \underset{W,S}{\operatorname{argmax}} P(X|S)P(S|W)P(W) \tag{2}$$

Veličine iz prethodne formule imaju svoja imena na osnovu komponente koja ih računa: $P(X|S)$ se naziva **akusticki model** (eng. *acoustic model*), $P(S|W)$ je **model izgovora** (eng. *pronunciation model*), a $P(W)$ se zove **jezički model** (eng. *language model*)[3].

Na slici 1 je prikazana cela struktura statističkog modela od zvučnog signala do transkripcije:



Slika 1: Statisticki model

U nastavku će biti opisana svaka od prikazanih komponenti, njena uloga i način rada.

[3]Neka literatura $P(X|W)$ naziva akustičkim modelom, a model izgovora tretira kao njegov deo. U ovom radu će biti podrazumevano da su oni odvojeni modeli.

3.1 Obrada zvučnog signala

Sirovi zvučni signal je veoma nepogodan za korišćenje zato što sadrži veliku količinu nebitnih informacija i šuma. Zbog toga se, pre prosleđivanja akustičkom modelu, signal prvo obrađuje tako da ostanu samo ključne karakteristike i smanji šum i veličina reprezentacije.

Signal se deli na kratke segmente koji se zovu **okviri** (eng. *frame*). Svaki od njih je fiksne dužine (obično 10-30 milisekundi) sa kratkim preklapanjem sa susednim okvirima radi smanjenja naglih promena prilikom prelaska iz jednog u drugi. Pretpostavka je da je u svakom 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 najčešće zasnovana na Furijeovoj transformaciji kojom se izdvajaju samo njegove najbitnije karakteristike. Tačna reprezentacija koja se koristi varira u zavisnosti od modela, ali jedna od najpopularnijih je **MFCC** (Mel-Frequency Cepstral Coefficients) [9] zasnovana na Mel skali koja oponaša ljudski slušni sistem. Ovako obrađen signal se prosleđuje akustičkom modelu.

3.2 Akustički model

Akustički model je zadužen da pretvori obrađeni zvučni signal u niz fonema. Ovaj zadatak predstavlja idealan slučaj za primenu skrivenih Markovljevih modela [10].

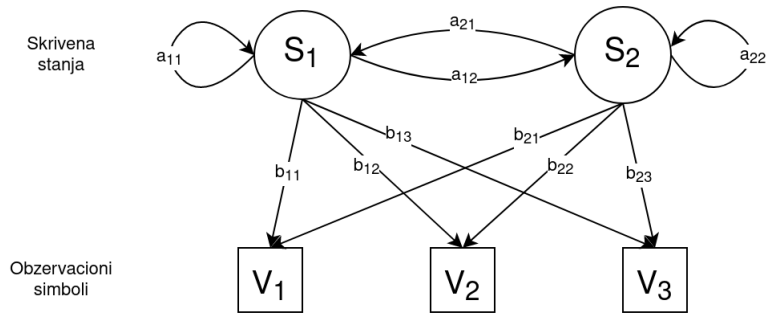
Skriveni Markovljev model (eng. *hidden Markov model*) je dinamički sistem kojeg karakteriše sledeće:

1. N skrivenih stanja $S = \{S_1, S_2, \ldots, S_N\}$ pri cemu q_t označava stanje u trenutku t
2. M obzervacionih simbola $V = \{V_1, V_2, \ldots, V_M\}$ pri cemu o_t označava obzervaciju u trenutku t
3. Raspodela verovatnoća promene stanja predstavljena matricom $A = \{a_{ij}\}$ dimenzije $N \times N$ gde važi:
$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \quad 1 \leq i, j \leq N$$
4. Raspodela verovatnoća obzervacionih simbola iz stanja j predstavljena matricom $B = \{b_j(k)\}$ dimenzije $N \times M$ gde važi:
$$b_j(k) = P(o_t = V_k | q_t = S_j) \quad 1 \leq j \leq N, 1 \leq k \leq M$$
5. Raspodela inicijalnog stanja $\pi = \{\pi_i\}$ gde važi:
$$\pi_i = P(q_1 = S_i) \quad 1 \leq i \leq N$$

Primer jednog modela je prikazan na slici 2.

Konkretno za prepoznavanje govora, skriveni Markovljevi modeli se koriste za opisivanje svake foneme. Svaki od njih se sastoji od nekog zadatog broja stanja (obično 3 ili 5), a obzervacije su okviri zvučnog signala čije su verovatnoće predstavljene kao mešavina Gausovih raspodela. Raspodele verovatnoća su dobijene treniranjem modela na skupu podataka. Za svaku rečenicu iz trening skupa se konstruiše model spajanjem modela fonema u reči, pa reči u rečenicu. Na njega se onda primeni Baum-Welch algoritam koji popravlja verovatnoće da više odgovaraju datim podacima.

Istrenirani model posle može da predvidi najverovatniji niz fonema (stanja) za datu obzervaciju koristeći Vetrebi algoritam za pretragu.



Slika 2: Primer skrivenog Markovljevog modela sa 2 skrivena stanja (krugovi) i 3 obzervaciona simbola (kvadrati)

3.3 Model izgovora

U prethodnoj sekciji je rečeno da se prilikom treniranja modeli fonema spajaju u modele reči, ali nije objašnjeno tačno kako se to radi. To je baš zadatak modela izgovora. On je u suštini veliki rečnik koji za svaku reč čuva niz fonema kako se ona izgovara. Ako postoji više varijacija izgovora, one se smatraju kao ražlicite stavke u rečniku. Sada konstrukcija modela reči postaje trivijalna, samo se pronađe njen izgovor u rečniku i nadovežu odgovarajuci modeli fonema. Ukoliko se reč ne nalazi u rečniku, sistem za prepoznavanje govora neće biti u stanju da je prepozna.

Prirodno sledeće pitanje je kako se određuju ova preslikavanja. Ovo je zapravo jedan od najtežih zadataka za modeliranje zato što se on ne uči na skupu podataka nego ga konstruišu eksperti iz tog domena. Za svaku reč, neko je morao da zapiše na koji način se izgovara uzimajući u obzir da potencijalno postoji više izgovora. Sa obzirom da današnji sistemi razlikuju oko 100.000 reči, ovo nimalo nije lak posao.

3.4 Jezički model

Povratkom na formulu 2, jezički model dodeljuje verovatnoću pojavljivanja $P(W)$ svakoj mogućoj sekvenci reči W. Ovde se uzima u obzir relativna učestalost reči, verovatnoća da se reči nađu jedna za drugom, i mogu da se vrše dodatne sintaksne i semantičke provere. Postoje rečenice koje zvuče slično ali nemaju sve semantičko znacenje. Tada se od njih bira ona koja ima najviše smisla.

Kako $P(W)$ ne zavisi od zvučnog signala, može se odvojeno trenirati na samo tekstualnom skupu podataka kojih postoji dosta više i imaju veći broj primera od skupova sa transkribovanim snimcima. U nekim slučajevima trenirana raspodela se može promeniti u zavisnosti od korisnika (npr. pametni lični asistenti prepoznaju kontakte na korisnikovom telefonu).

Najčešći vid implementacije ovog modela je pomoću n-grama. Neka se transkripcija W razdvaja na reči $W = \{w_1, w_2, \ldots, w_m\}$ i neka je n dužina n-grama. To znači da pri računanju verovatnoće pojavljivanja neke reči u obzir uzimamo samo $n-1$ njenih prethodnika. Tada se $P(W)$ može

predstaviti kao:

$$P(W) = \prod_{i=1}^{m} P(w_i|w_1, \ldots, w_{i-1}) = \prod_{i=1}^{m} P(w_i|w_{i-n+1}, \ldots, w_{i-1})$$

Ako $C(x)$ označava broj pojavljivanja sekvence x u trening skupu, verovatnoća da je w_i sledeća reč se može proceniti kao:

$$P(w_i|w_{i-n+1}, \ldots, w_{i-1}) = \frac{C(w_{i-n+1}, \ldots, w_i)}{C(w_{i-n+1}, \ldots, w_{i-1})}$$

Radi jednostavnije implementacije na početak W se dodaje $n-1$ "prazna"reč.

Veliki problem sa ovim modelom je što onemogućava predviđanje n-grama koji nisu prethodno viđeni u trening skupu jer bi im bila dodeljena verovatnoća 0. Čak i za male vrednosti n ovo nije nerealna situacija. Kao kompenzacija za to moguće je umesto fiksnog, koristiti promenljivo n koje bi krenulo od neke zadate vrednosti i smanjivalo se sve dok ne pronađe bar jedna instanca tog n-grama. Onda je moguće direktno korisiti tu vrednost ili na neki način uračunati koliko se parametar smanjio pre nego što je n-gram pronađen.

3.5 Dekodiranje

Konstrukcijom svih prethodno opisanih komponenti i njihovim treniranjem, model je gotov i spreman za upotrebu. Jedino što ostaje je pretražiti prostor dopustivih rečenica da se pronađe ona koja najviše odgovara glasovnom signalu. U praksi, taj prostor je veoma veliki i njegova celokupna pretraga je eksponencijalne složenosti, stoga nije izvodljivo tražiti egzaktno rešenje. Umesto toga se koristi heuristički, **beam search algoritam**. Ideja je da se rešenje gradi iterativno i u svakom trenutku umesto testiranja svih mogućih puteva gramzivo se bira samo b najverovatnijih. Parametar b se određuje tako da balansira veličinu prostora za pretraživanje i vreme potrebno za njegov obilazak. Dakle, algoritam počinje od niza sa jednom praznom rečenicom. Jedan korak se sastoji od produživanja svakog elementa iz niza svim mogućim rečima i brisanja svih osim b najverovatnijih. Ta verovatnoća se jednostavno računa množenjem verovatnoće nove reči sa verovatnoćom rečenice koja je dopunjena. Ovaj postupak se ponavlja sve do kraja svih rečenica iz niza kada se najverovatnija od njih proglašava za konačnu transkripciju govora.

4 End-to-end model

End-to-end model je ceo zasnovan na dubokim neuronskim mrežama. Za razliku od statističkog modela koji proces prepoznavanja govora deli u više složenijih celina, end-to-end model je po strukturi prostiji. Umesto razdvajanja formule 1 pomoću Bajesovog pravila, on direktno računa verovatnoću $P(W|X)$, to jest bave se direktnim prevođenjem ulaznog zvučnog signala u niz grafema, karatera ili reči. To donosi velike prednosti, naime: više nije potreban ekspert za jezik nego se sve uči iz podataka, treniranje postaje lakše zbog jednostavnije strukture i performanse postaju bolje zato što se optimizuje samo jedna vrednost. Sa druge strane, pošto model treba sam sve da nauči, potrebna je veća količina podataka nego ako se koristi statistički model.

Glavni problem ovim modelima je kako odrediti kom delu zvuka odgovara deo transkripcije. To je posao koji su u statističkom pristupu radili skriveni Markovljevi modeli. U odnosu na način rešavanja ovog problema razlikuju se **CTC** [11] i **modeli zasnovani na pažnji** (eng. *attention based models*) [12].

4.1 CTC model

CTC (Connectionist Temporal Classification) rešava problem poravnanja zvuka i teksta tako što tretira izlaz iz mreže kao raspodelu verovatnoća za svaku labelu ili blanko karakter. Ideja iza uvođenja ovog karaktera je da se pomoću njega vrši poravnanje sa zvukom, ali da se on ignoriše prilikom predviđanja labela.

Neka je ulazni zvuk podeljen na segmente jednakih dužina, gde je svaki od njih obrađen u vektor realnih brojeva dužine m[4] i neka je L konačna azbuka labela (oznaka). Cilj je napraviti preslikavanje h koje slika proizvoljan zvučni signal u niz labela:

$$h : (\mathbb{R}^m)^* \to L^* \tag{3}$$

U praksi je broj vremenskih trenutaka fiksiran 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 najvise T.

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

$$LER(h, O) = \frac{1}{|O|} \sum_{(\mathbf{x},\mathbf{z}) \in O} \frac{ED(h(\mathbf{x}), \mathbf{z})}{|\mathbf{z}|} \tag{4}$$

gde ED predstavlja edit rastojanje[5]. Prethodna mera naziva se stopa greske labela (eng. *label error rate - LER*). Formula 4 je prirodna ocena greške za probleme koji za cilj imaju minimizaciju greške prevođenja.

Model se implementira 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, $L' = L \cup \{\epsilon\}$. Svaki izlaz se interpretira kao raspodela verovatnoća za pojavljivanje odgovarajuće labele iz L' u tom vremenskom trenutku. Neformalno rečeno, prolaskom kroz izlaze u svakom od T trenutaka dobija se putanja $\pi \in (L')^T$ koja predstavlja jedan mogući odabir labela. Ako $y^t_{\pi_t}$ predstavlja softmax vrednost u trenutku t oznake π_t, tada je verovatnoća odabira kompletne putanje:

$$P(\pi|\mathbf{x}) = \prod_{t=1}^{T} y^t_{\pi_t}$$

U praksi, zvuk je segmentovan na veoma sitne vremenske intervale (oko 10ms), stoga je pojava blanko ili dupliciranih oznaka veoma česta. Iz tog razloga se uvodi preslikavanje $\mathcal{B}$ čija je uloga prečišćavanje nizova labela uklanjanjem blanko oznaka i susednih duplikata.

$$\mathcal{B} : (L')^T \to L^U, U \leq T$$

[4] Ovo može da bude MFCC kao u poglavlju 3.1, ali je češće nešto jednostavnije

[5] minimalni broj izmena koji dovodi jedan niz karaktera do drugog, pri čemu dozvoljene izmene podrazumevaju brisanje, zamenu i umetanje karaktera

Kako za jednu prečišćenu putanju l može postojati više mogućih izvornih putanja, verovatnoća njenog odabira jednaka je sumi po svim izvornim putanjama.

$$P(l|x) = \sum_{\pi \in \mathcal{B}^{-1}(l)} p(\pi|\mathbf{x}) \tag{5}$$

Imajući u vidu sve prethodno navedeno, zadatak preslikavanja h je odabir najverovatnije prečišćene putanje za dati ulaz.

$$h(x) = \underset{l \in L^U}{\operatorname{argmax}} P(l|\mathbf{x}) \tag{6}$$

Ovo je moguće na više načina. Najjednostavniji, ali neoptimalan, jeste pohlepni odabir najbolje oznake za svaki vremenski trenutak ponaosob, ali postoje i drugi algoritmi koji daju bolje rezultate.

4.2 Modeli zasnovani na paznji

Za razliku od CTC modela, ovi modeli se oslanjaju na **mehanizmam pažnje** (eng. *attention mechanism*) za poravnjanje teksta sa zvukom i odbacuju problematičnu pretpostavku o nezavisnosti zvučnih segmenata na kojoj se CTC zasniva. Ova grupu modela će biti predstavljena na primeru LAS (eng. *Listen, Attend and Spell*) modela [13].

Glavne komponente ovog modela su **slušalac** (eng. *listener*) i **speler** (eng. *speller*). Okvirno gledano, uloga slušaoca jeste da transformiše ulazni signal u karakteristike višeg nivoa, koje se prosleđuju speler komponenti da proizvede konačan niz karaktera.

Slično kao u poglavlju 4.1, neka je $\mathbf{x} = (x_1, x_2, ..., x_T)$ ulazni signal, a $\mathbf{y}$ niz mogućih izlaznih karaktera (uključujući razmak, tačku, zarez, apostrof) i specijalnih karaktera (početak i kraj rečenice i nepoznat simbol). Cilj je modelovati $\mathbf{y}$ kao uslovnu raspodelu u zavisnosti od ulaznog signala $\mathbf{x}$ i prethodnih izlaza $y_j, j \in (1, i-1)$

$$P(y|x) = \prod_i P(y_i|x, y_j, j < i) \tag{7}$$

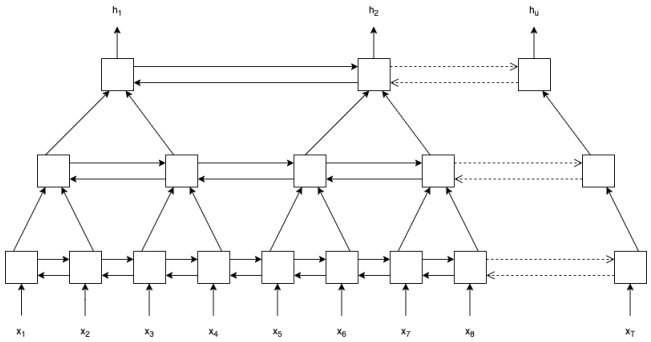
Skušalac je rekurentna neuronska mreža piramidalne strukture čiji zadatak je da napravi reprezentaciju zvuka višeg nivoa. Kao što se može primetiti na slici 3, broj neurona se polovi u svakom sloju do izlaznog. Ova osobina omogućava smanjenje složenosti izračunavanja, što je veoma poželjno u slučaju kada je dužina zvučnog segmenta mala.

Speler je takođe rekurentna neuronska mreža. Kao sto je već pomenuto, verovatnoća odabira svakog karaktera zavisi od odabira svih prethodnih. Na izlaz y_i utiče stanje s_i i kontekst c_i. Stanje se računa rekurentno na osnovu prethodnog stanja, izlaza i konteksta, a kontekst se izračunava korišćenjem standardnog mehanizma pažnje. Intuitivno, kontekst prikuplja relevantno znanje o okolnim zvučnim segmentima potrebnim za predikciju narednog karaktera, na taj način određujući koliko znanje o okolini utiče na znanje o trenutnom karakteru. Struktura ove mreže je prikazana na slici 4.

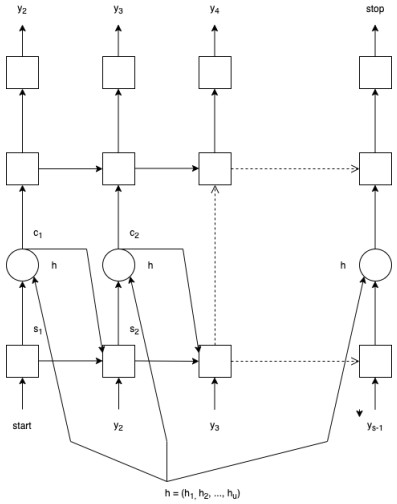
Treniranje se sprovodi određivanjem parametara koji maksimizuju sumu logaritama uslovnih verovatnoća za svaki karakter.

$$\max_\theta \sum_i \log P(y_i|\mathbf{x}, \bar{y}_j, j < i; \theta) \tag{8}$$

gde je $\bar{y_j}$ objektivna vrednost karaktera j.



Slika 3: Listener neuronska mreža



Slika 4: Speller neuronska mreža

Konačno dekodiranje se vrši heuristički, kao u poglavlju 3.5 po prostoru pretrage određenom raspodelama verovatnoća iz svakog koraka speler rekurentne mreže.

5 Metrike za evaluaciju

Standardna mera za procenu kvaliteta sistema za prepoznavanje govora je **stopa pogrešnih reci** (eng. *word error rate, WER*). Za njeno izračunavanje potrebno je da imamo generisan tekst koji evaluiramo i tačnu transkripciju kao referencu. Formula je izvedena normalizacijom Levenštajnovog rastojanja i izgelda ovako:

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

gde je:

- I broj umetnutih reči

- D broj obrisanih reči
- S broj zamenjenih reči
- N ukupan broj reči u referenci

Minimalna vrednost koju može da ima je 0, dok maksimalna vrednost može da bude preko 1 (npr. izgovorena je jedna reč a prepoznate dve). Stopa pogrešnih reči se efikasno računa dinamičkim programiranjem, pomoću Vagner-Fišerovog algoritma (eng. *Wagner-Fischer algorithm*).

Cilj sistema za prepoznavanje govora je da minimizuje ovu vrednost.

6 Zaključak

Kroz ovaj rad predstavljeni su osnovni pojmovi i najznačajniji modeli za automatsko prepoznavanje govora. Uprkos tome što su end-to-end modeli dosta korišćeniji u praksi zbog generalno boljih performansi, statistički modeli su i dalje široko rasprostranjeni i još uvek su predmet brojnih istraživanja.

Ovaj rad je bio samo uvod u automatsko prepoznavanje govora i čitalac se poziva da samostalno nastavi istraživanje na ovu temu. Radovi iz literature sadrže više informacija o konceptima i implementaciji pomenutih modela koje su bile izvan opsega ovog rada i predstavljaju dobar početak za dalje istraživanje.

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] N. Dave, "Feature extraction methods lpc, plp and mfcc in speech recognition," *International journal for advance research in engineering and technology*, vol. 1, no. 6, pp. 1–4, 2013.

[10] 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.

[11] 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), p. 369–376, Association for Computing Machinery, 2006.

[12] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," 2015.

[13] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals, "Listen, attend and spell," 2015.

A Dodatak: Pregled skupova podataka

Tabela 1: Pregled nekih od najznačajnijih skupova podataka transkribovanog govora.

Naziv skupa	Dužina (sati)	Broj govornika	Nivo transkripcije
TIMIT	5,4	630	foneme, reči
Switchboard-1	260	543	reči
LibriSpeech	982,1	1166	reči
CommonVoice	2015	75.879	reči
GigaSpeech	10.000	nepoznato	reči
VoxPopuli	543	413.581	reči