

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

15. 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

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

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 potrebno 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 najznamenitijih primena ovih sistema su: pametni licni asistenti (Google Assistant¹, Apple Siri²,...), 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 najznamenitijih postignuća je ostvareno 2016. godine je kompanija Majkrosoft napravila sistem koji je ostvario iste rezultate kao ljudski eksperti na Switchboard skupu 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 resimo 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 različitim uslovima. Ideal kojem se teži kod ovih sistema je nezavisnost od govornika (speaker independence).

Neki od najvećih izazova su:

- Mala količina podataka za trening - Za ostvarivanje dobrih rezultata, potrebno je sakupiti više stotina sati labeliranih zvucnih snimaka koji treba da sadrže više govornika. Ovako velike skupove je teško naći, i još teže verifikovati. Zbog toga se istražuju alternativni načini za treniranje kao što su samo-treniranje [3], iterativno treniranje [4] ili treniranje koristeći kompjuterski generisan glas [google?].
- Stil govora
- Varijacije u brzini govora, jačini i frekvenciji glasa, starosti
- Varijacije u dijalektu i akcentima
- Pozadinska buka
- Velicina rečnika

¹<https://assistant.google.com/>

²<https://www.apple.com/siri/>

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 zvucnog signala X . Cilj je optimizovati formulu [5]:

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

primenom Bajesove formule to mozemo da zapisemo kao:

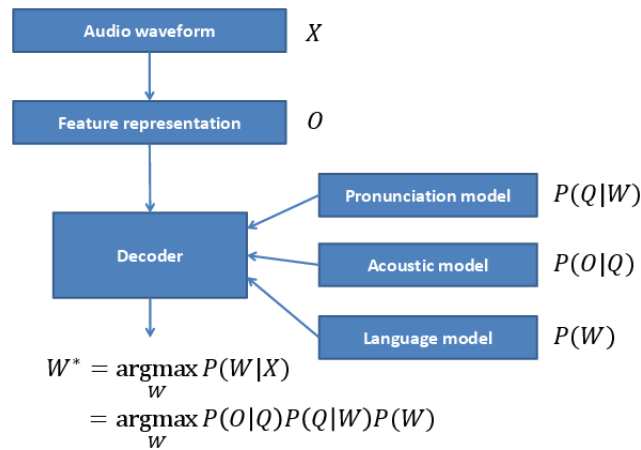
$$\hat{W} = \operatorname{argmax}_W \frac{P(X|W)P(W)}{P(X)}$$

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

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

Ideja je da odvojeno optimizujemo obe prethodne velicine. $P(X|W)$ se zove akusticki model, a $P(W)$ se zove jezikki model.

Na slici 1 je prikazana cela struktura statistickog modela od zvucnog signala do transkripcije:



Slika 1: Statisticki model

U nastavku ce biti opisana svaka komponenta statistickog modela, njihova uloga i nacin rada.

3.1 Procesiranje zvucnog signala

Sirovi zvucni signal je veoma nepogodan za koriscenje zato sto sadrzi veliku kolicinu informacije koje nam nisu potrebne. Umesto direktnog koriscenja, iz signala se prvo izvlace kljucne karakteristike. Signal delimo na kratke segmente koje zovemo prozori (eng. window). Svaki prozor je duzine od nekoliko milisekundi (obicno 10-30) pri cemu se susedni prozori delom preklapaju.

Zatim na svaki prozor primenimo neku vrstu 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 zadužen da pretvori sekvencu prozora u sekvencu fonema. Foneme. Za modelovanje svake foneme se koriste skriveni Markovljevi modeli (eng. *Hidden Markov Models*) [6]. 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

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 rečenici. 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 računa dinamičkim programiranjem, pomoću 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] U. Kamath, J. Liu, and J. Whitaker, *Deep Learning for NLP and Speech Recognition*. Springer Publishing Company, Incorporated, 1st ed., 2019.
- [6] 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