nedeljainformatike v8.0

Može li AI da prepozna emocije? Razumevanje govora

Kristina Nikolić

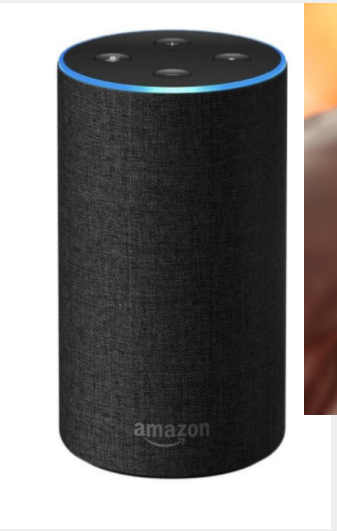
Matematička gimnazija

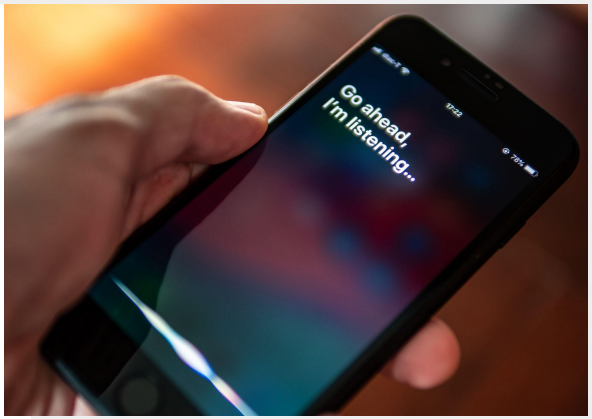
16. 05. 2023.

Agenda

1. Uvod u obradu govora
2. Metode dubokog učenja
3. Prepoznavanje emocija

Primene razumevanja govornog signala

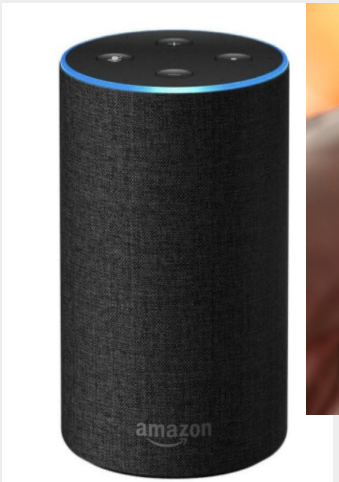


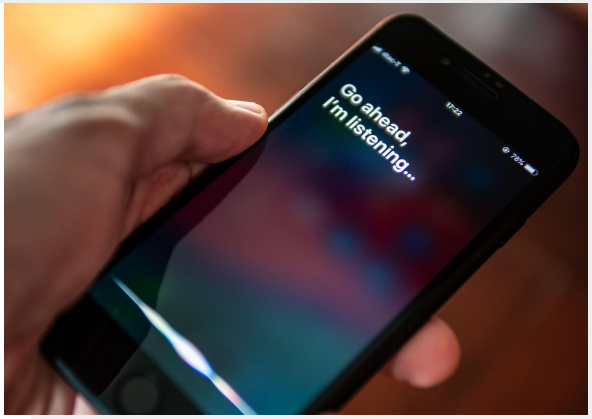


Govorni asistenti

Primene razumevanja govornog signala

Hands free sistem u automobilima



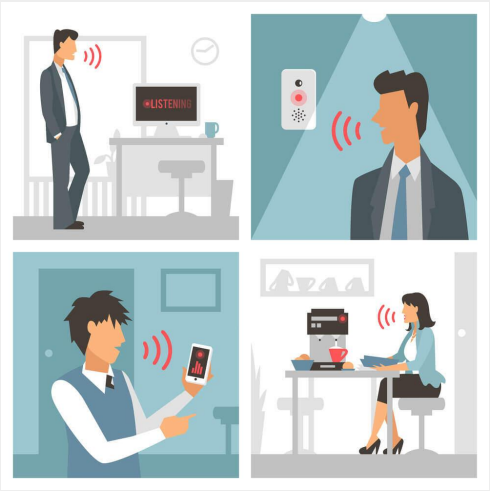


Govorni asistenti

Primene razumevanja govornog signala

Hands free sistem u automobilima

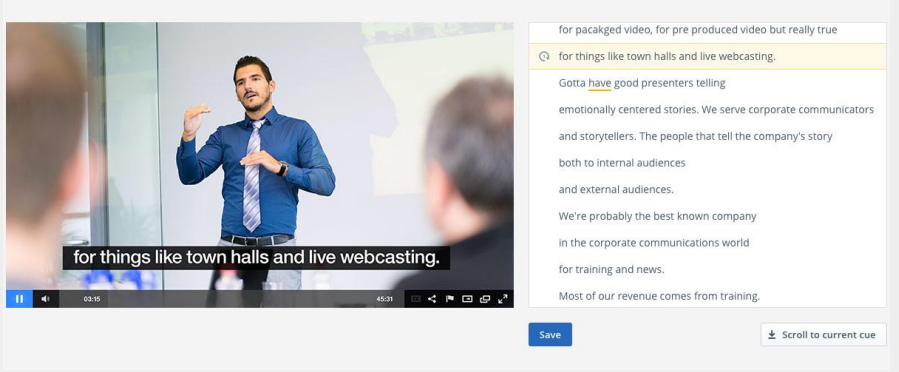
Govorni asistenti



Smart home

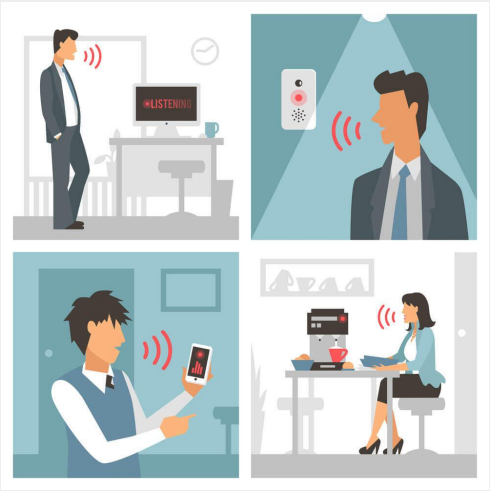
Primene razumevanja govornog signala

Hands free sistem u automobilima



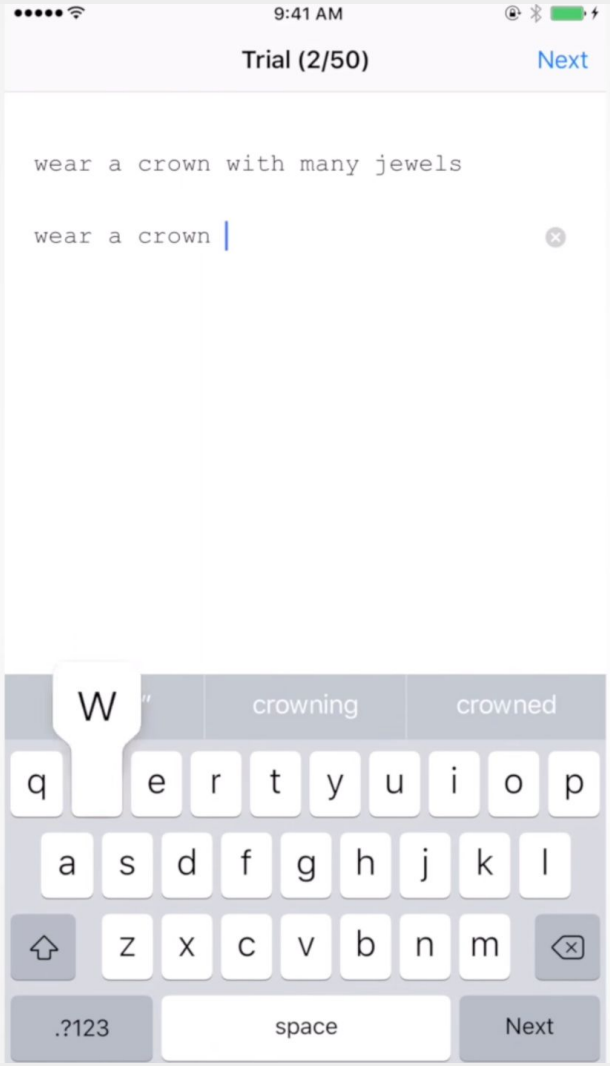
Automatski transkript

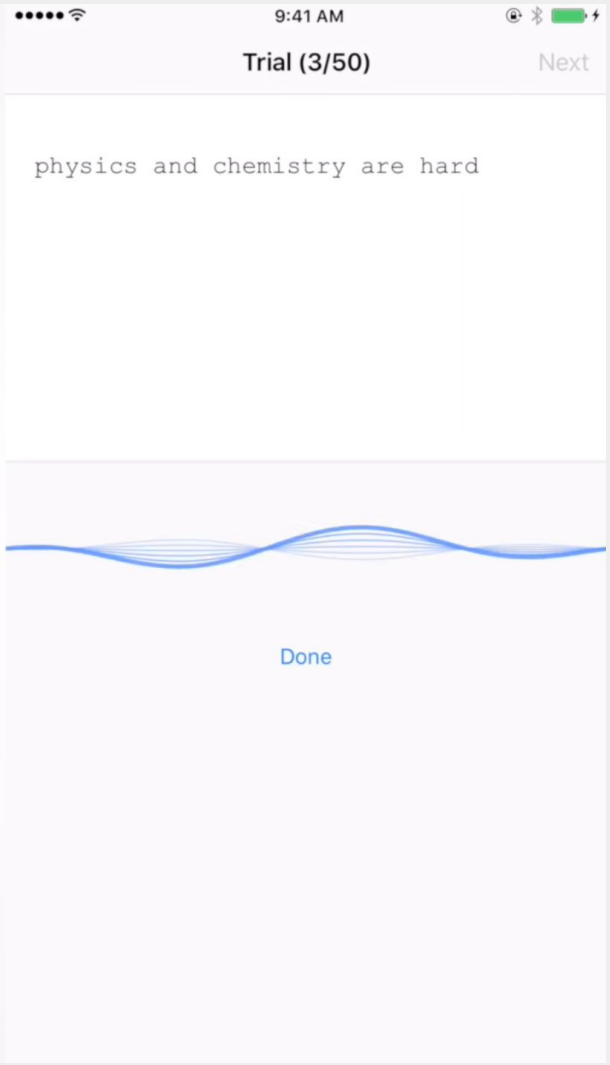
Govorni asistenti



Smart home

Stanford: „Govorne komande **3x brže** od kucanja”





Vrste prepoznavanja govora

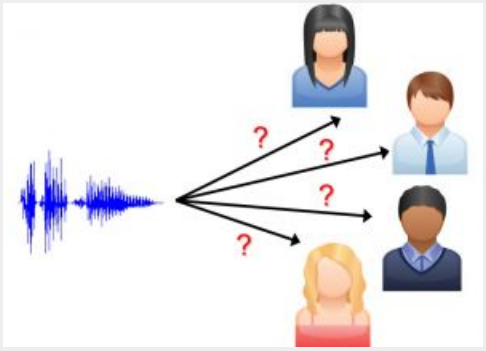
- Govor u tekst
- Detekcija ključnih reči (word spotting)
- Identifikacija govornika (Verifikacija)

Vrste prepoznavanja govora

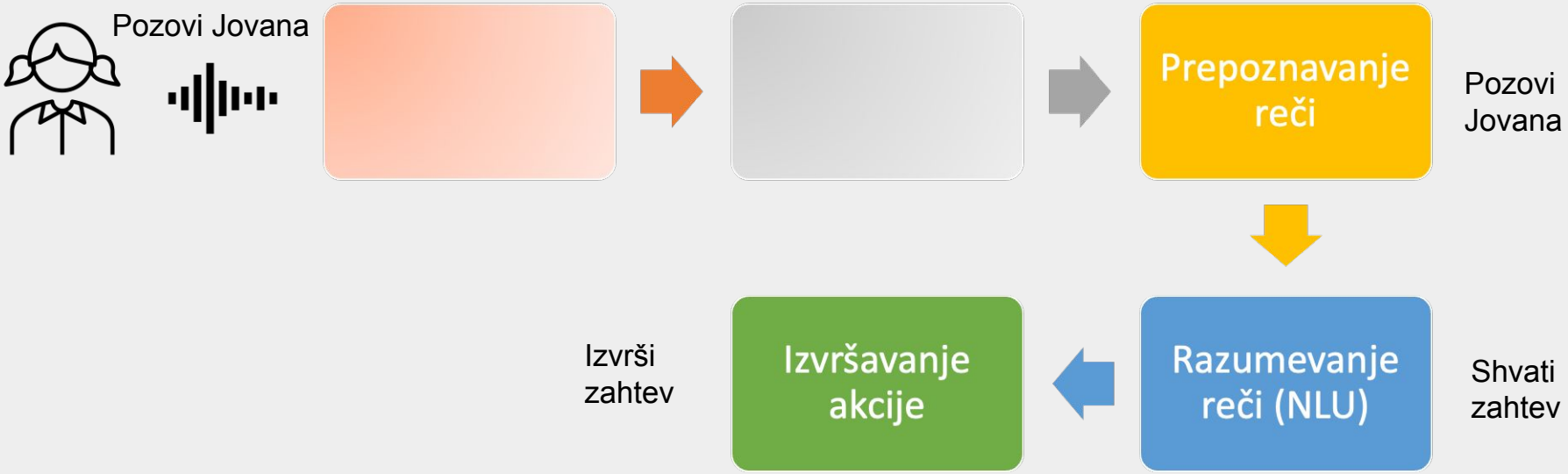
- Govor u tekst
- Detekcija ključnih reči (word spotting)
- Identifikacija govornika (Verifikacija)

Vrste prepoznavanja govora

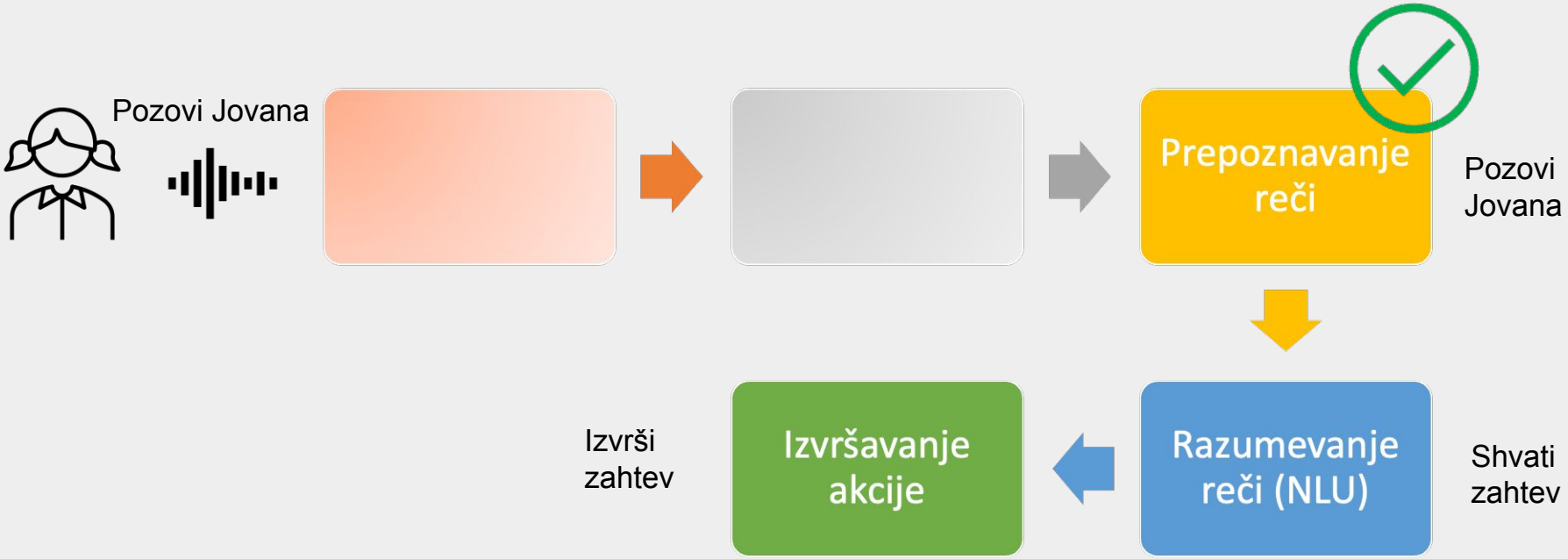
- Govor u tekst
- Detekcija ključnih reči (word spotting)
- Identifikacija govornika (Verifikacija)



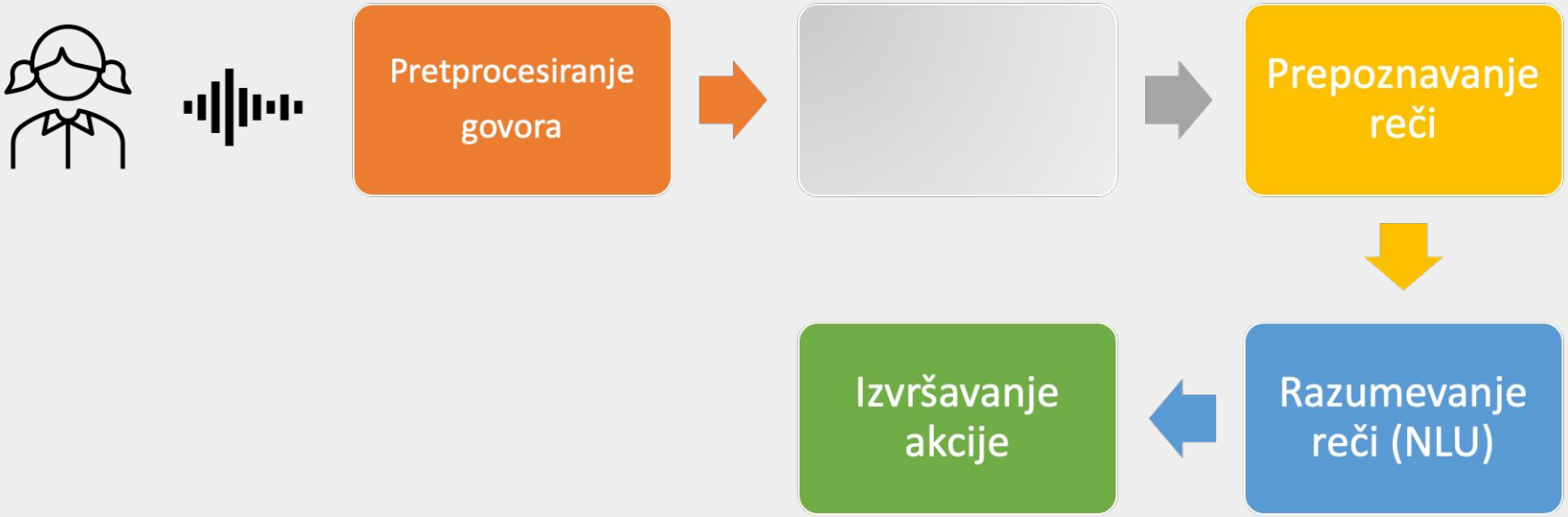
Kako rade govorni asistenti?



Kako rade govorni asistenti?

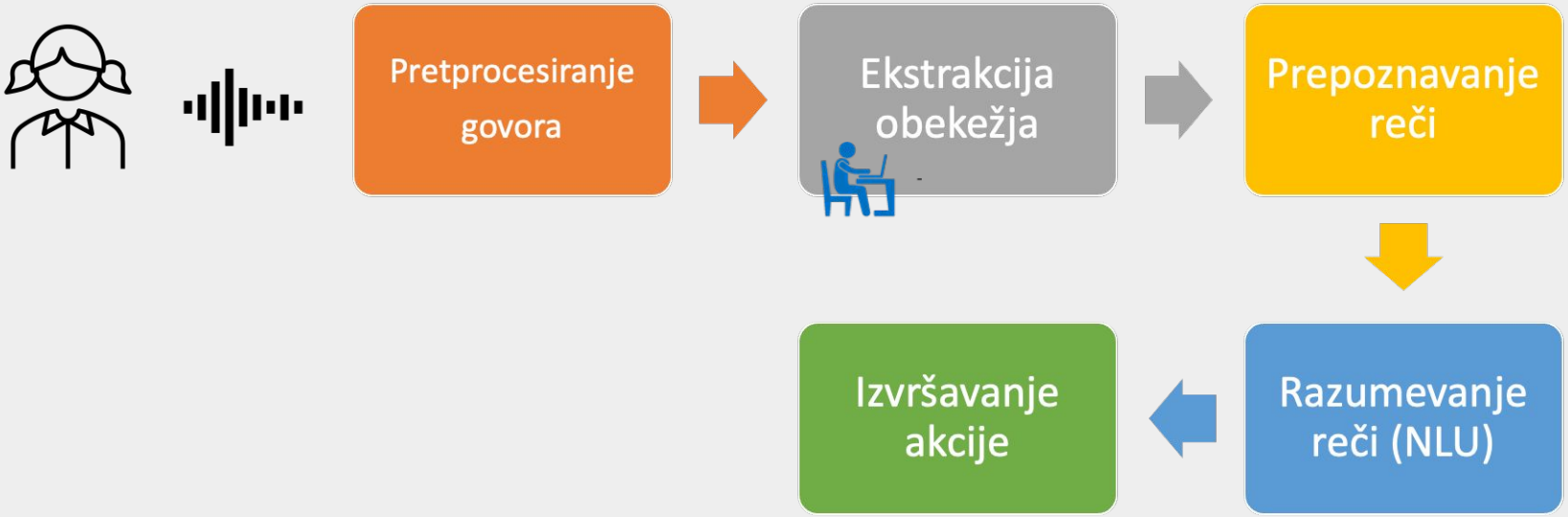


Kako rade govorni asistenti?



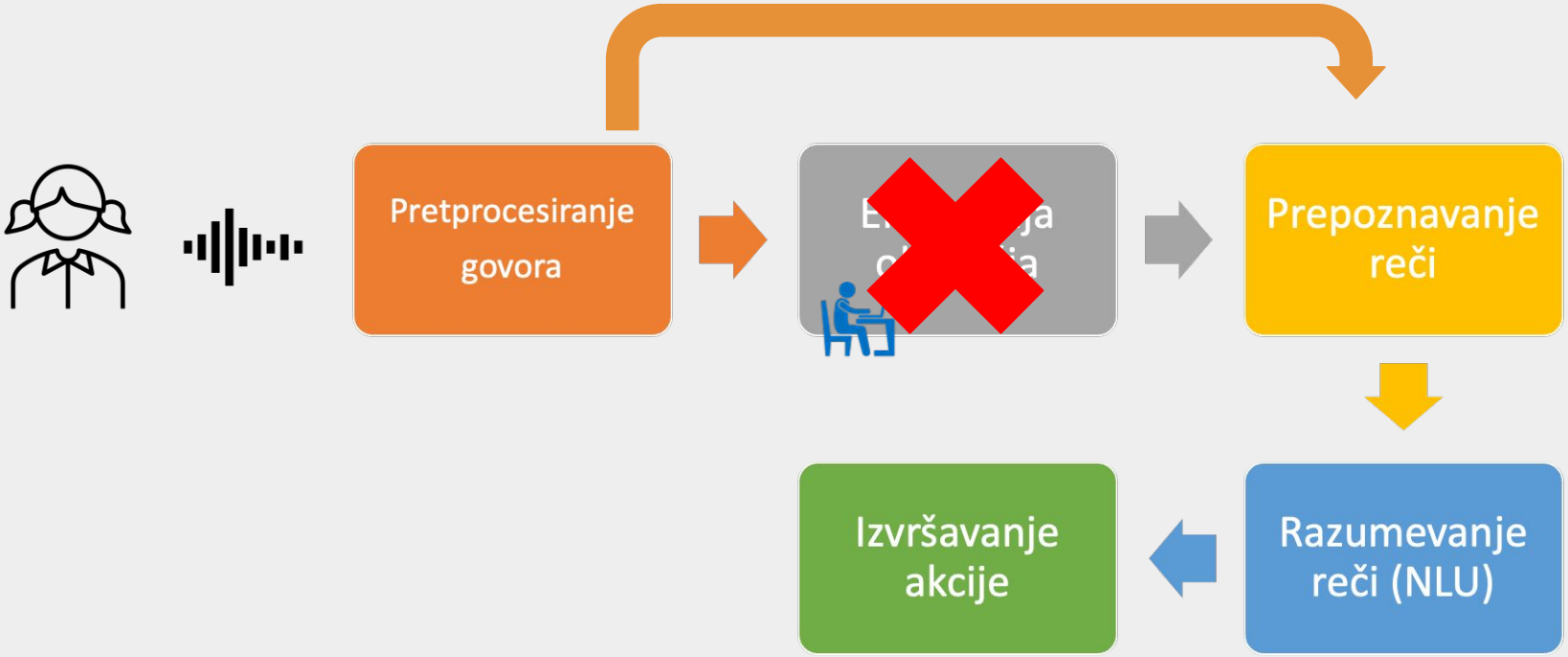
- Pretporcesiranje
  - Uklanjanje šuma, augmentacija...
  - Spektogram, Mel spektogram...

Kako rade govorni asistenti?



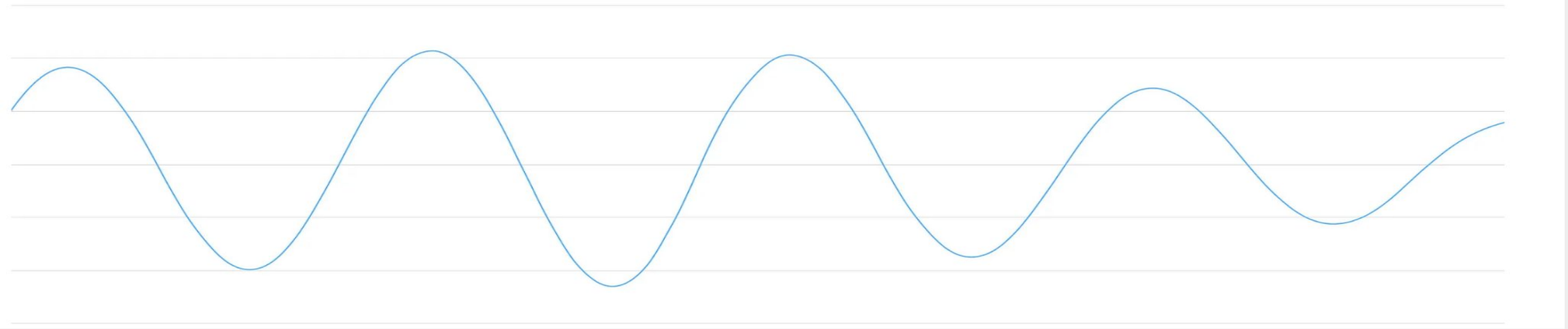
- Ekstrakcija obeležja
  - Izdvajamo domensko znanje koje smatramo korisnim
    - Standardna devijacija, snaga signala...
  - Metode dubokog učenja teže da izbegnu ovaj korak.

Kako rade govorni asistenti?



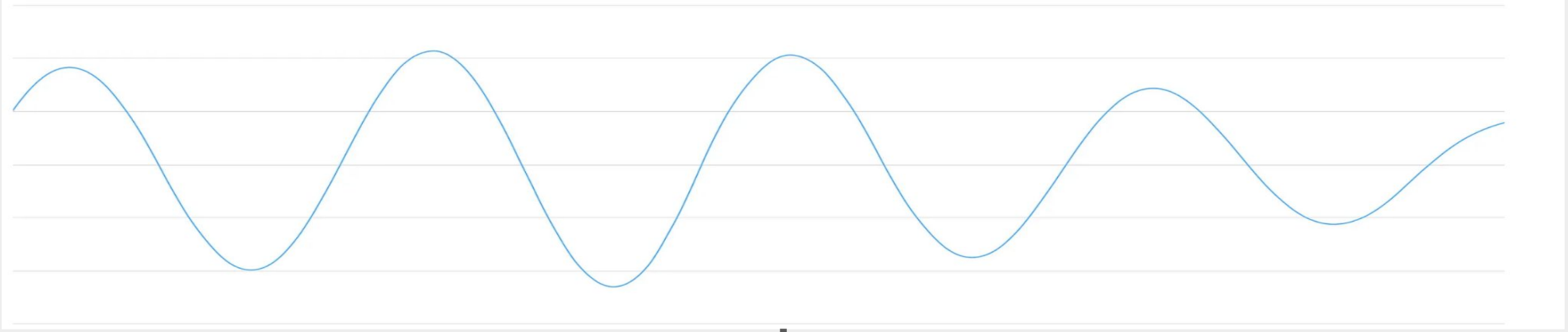
- Ekstrakcija obeležja
  - Izdvajamo domensko znanje koje smatramo korisnim
    - Standardna devijacija, snaga signala...
  - Metode dubokog učenja teže da izbegnu ovaj korak

Uzorkovanje govornog signala

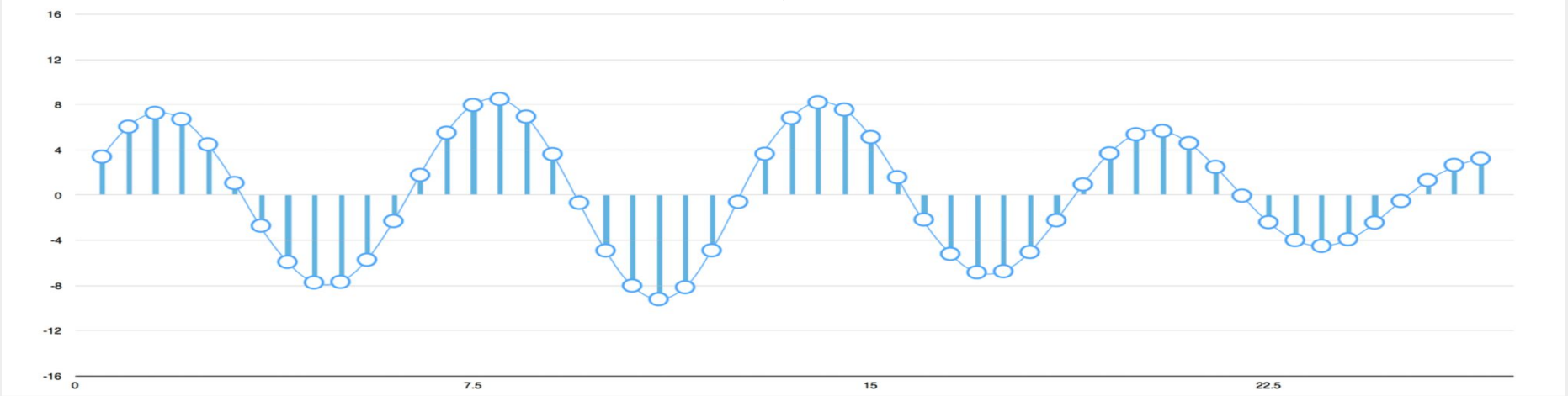


- Želimo da pretstavimo signal diskretno
- Uobičajna frekvencije odabiranja je 16kHz

Uzorkovanje govornog signala

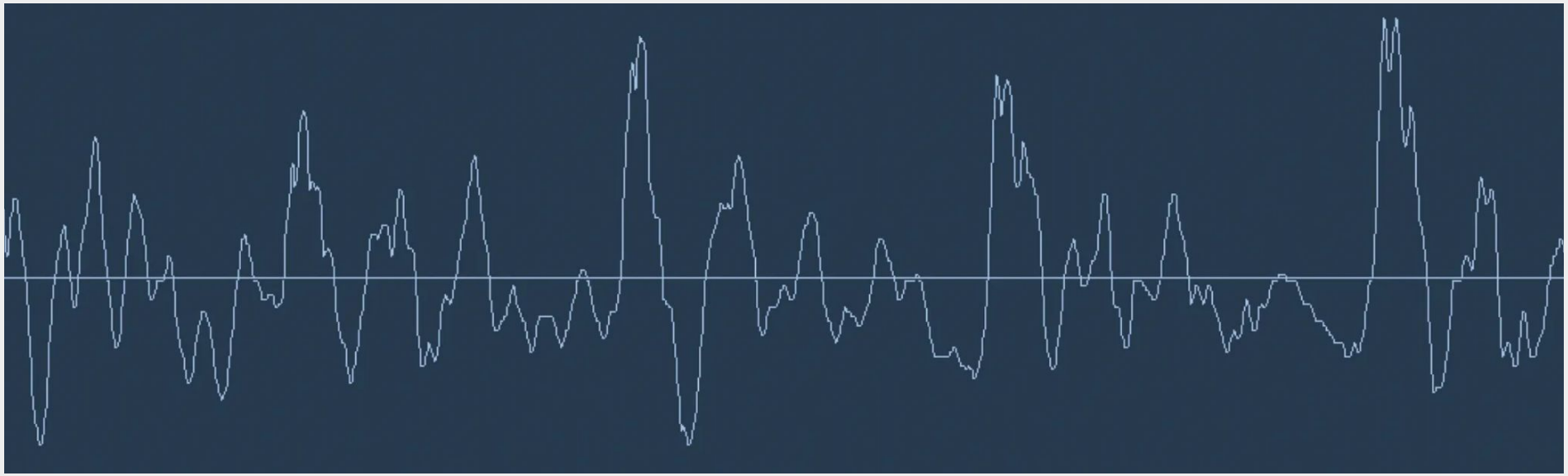


- Želimo da pretstavimo signal diskretno
- Uobičajna frekvencije odabiranja je 16kHz



Uzorkovanje govornog signala

- Govorni singal reči: „Hello”:

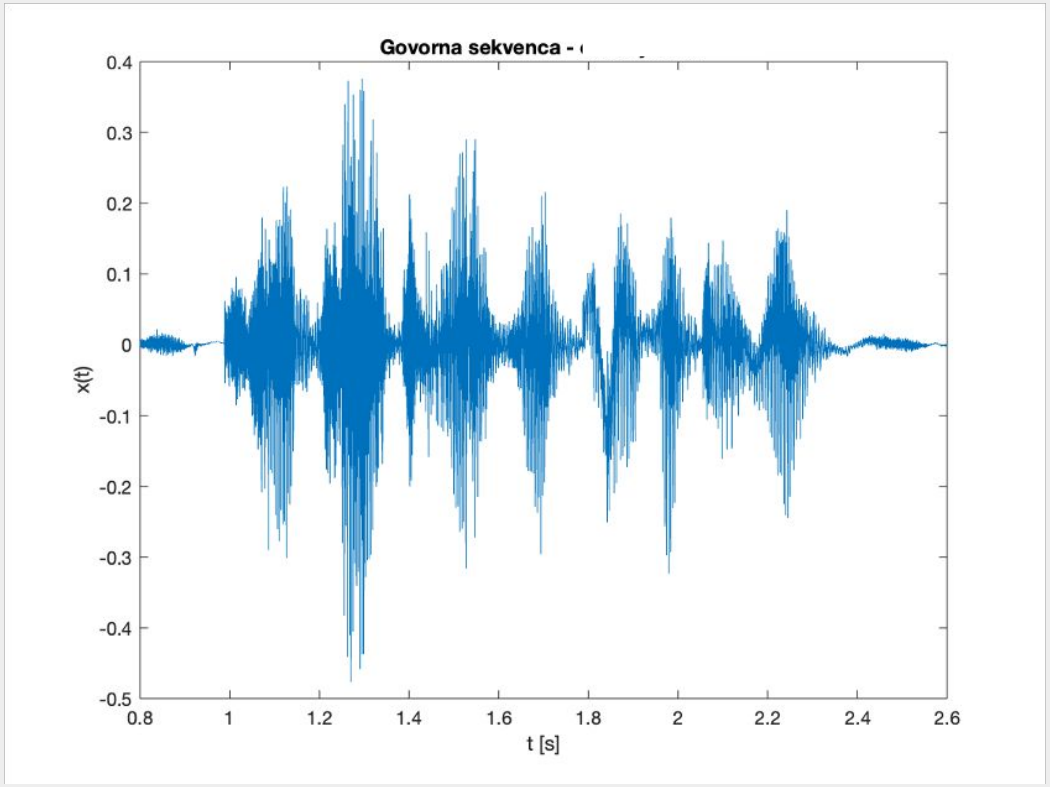


- Pretstavljamo sekvencom brojeva (vrednosti signala u trenutku odabiranja)

```
[-1274, -1252, -1160, -986, -792, -692, -614, -429, -286, -134, -57, -41, -169, -456, -450, -541, -761, -1067, -1231, -1047, -952, -645, -489, -448
, -397, -212, 193, 114, -17, -110, 128, 261, 198, 390, 461, 772, 948, 1451, 1974, 2624, 3793, 4968, 5939, 6057, 6581, 7302, 7640, 7223, 6119, 5461,
 4820, 4353, 3611, 2740, 2004, 1349, 1178, 1085, 901, 301, -262, -499, -488, -707, -1406, -1997, -2377, -2494, -2605, -2675, -2627, -2500, -2148, -
1648, -970, -364, 13, 260, 494, 788, 1011, 938, 717, 507, 323, 324, 325, 350, 103, -113, 64, 176, 93, -249, -461, -606, -909, -1159, -1307, -1544]
```

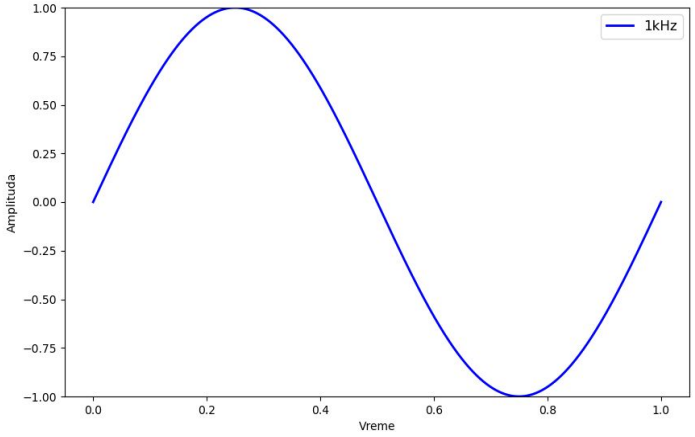
Vrmenski domen

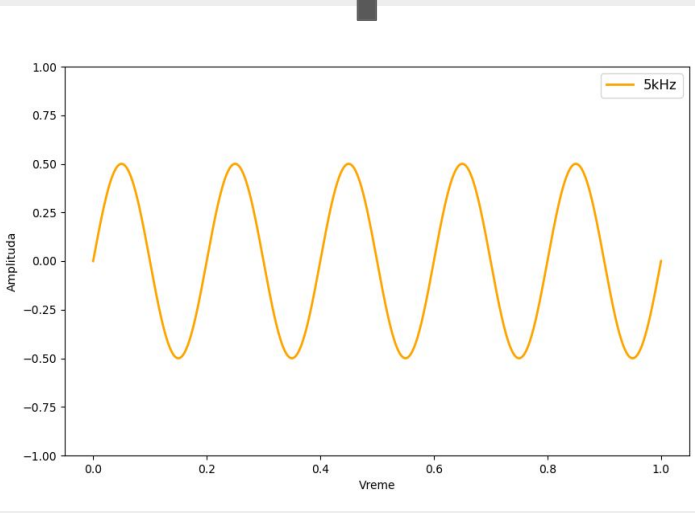
- Primer govornog signala u **vremenskom domenu**:



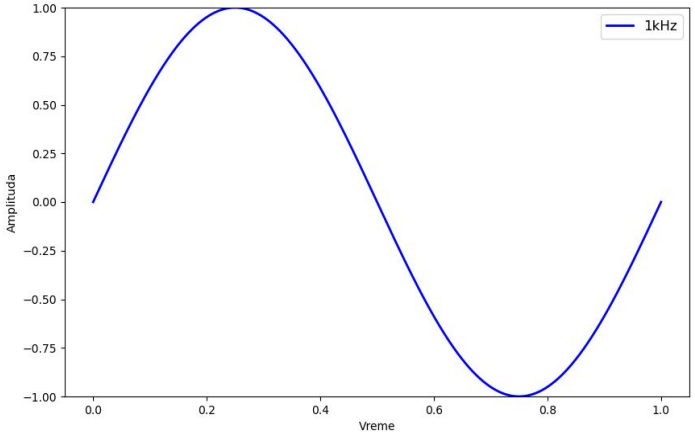
- Želimo da rastavimo ovaj signal na frekvencije koje ga čine, tj. da ga pretstavimo u **frekvencijskom domenu**.

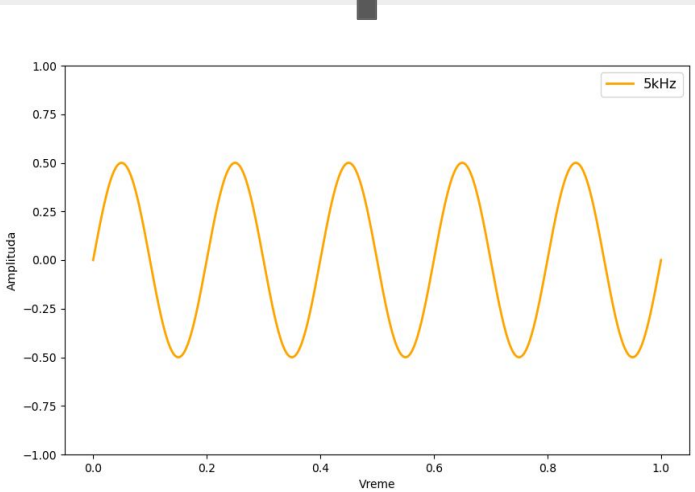
Periodični signali

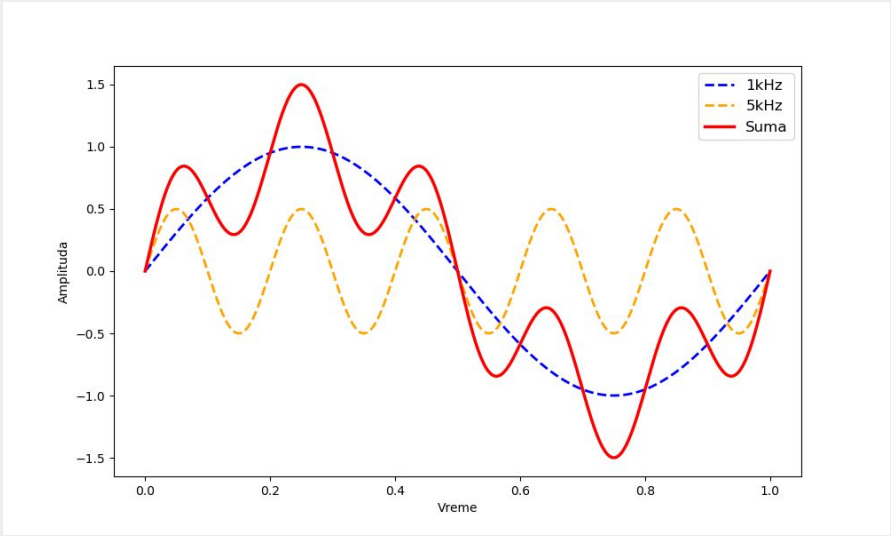




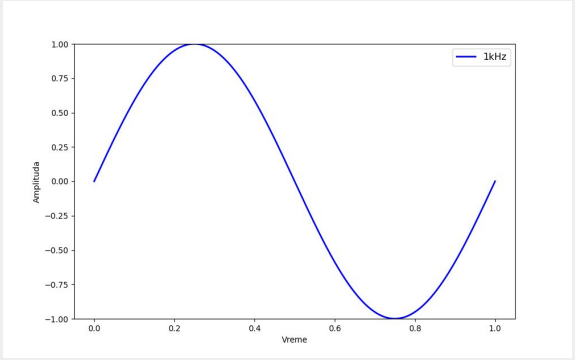
Periodični signali



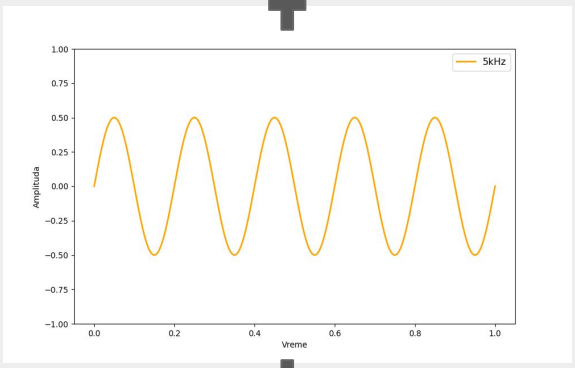




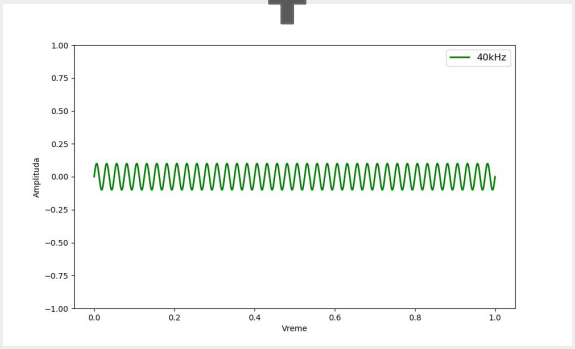
Periodični signali



+

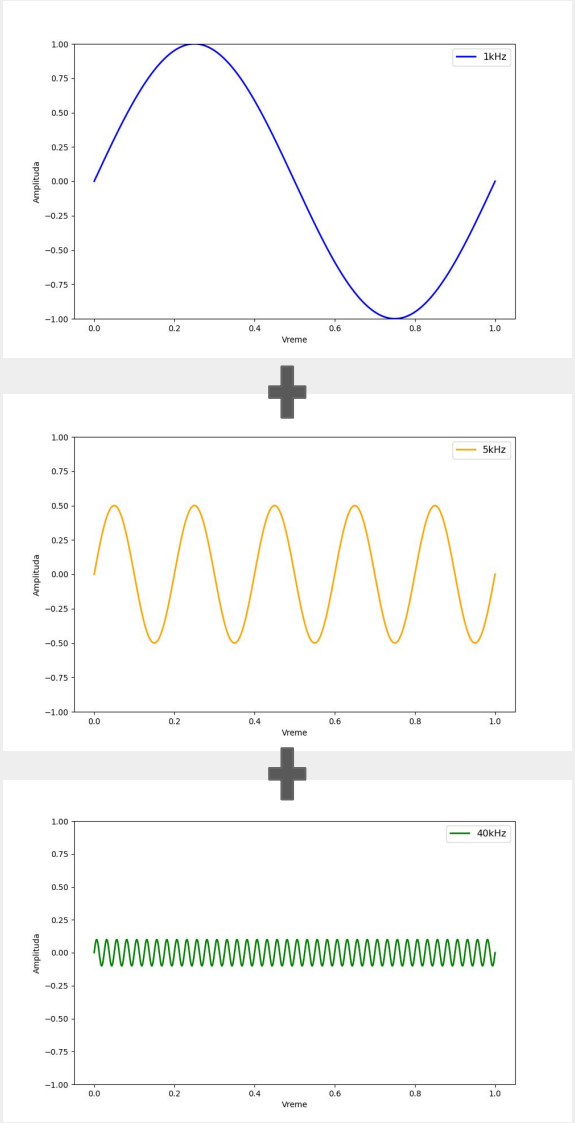


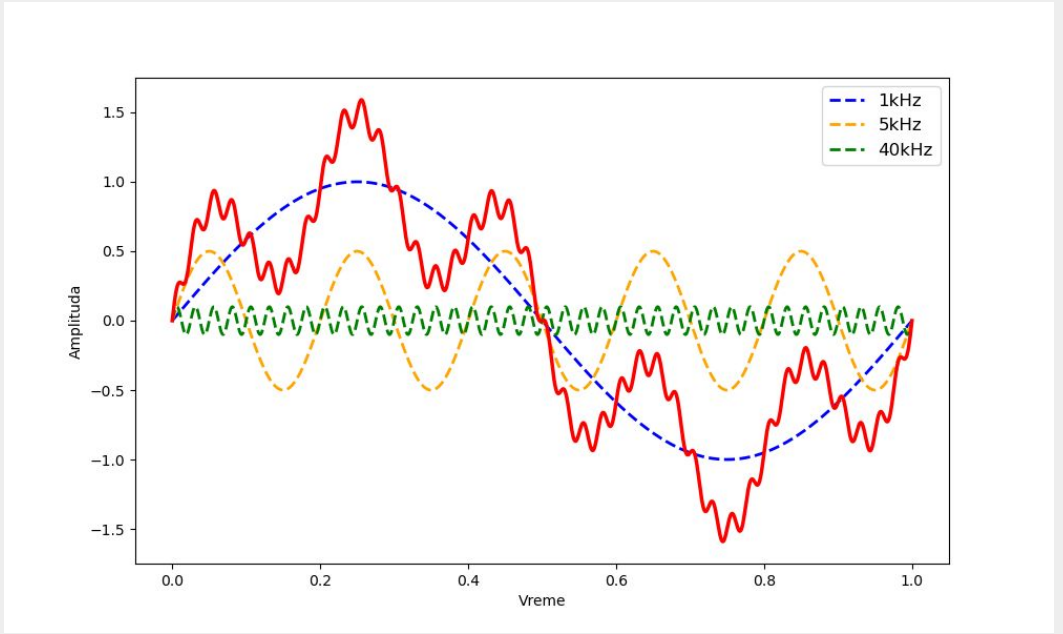
+



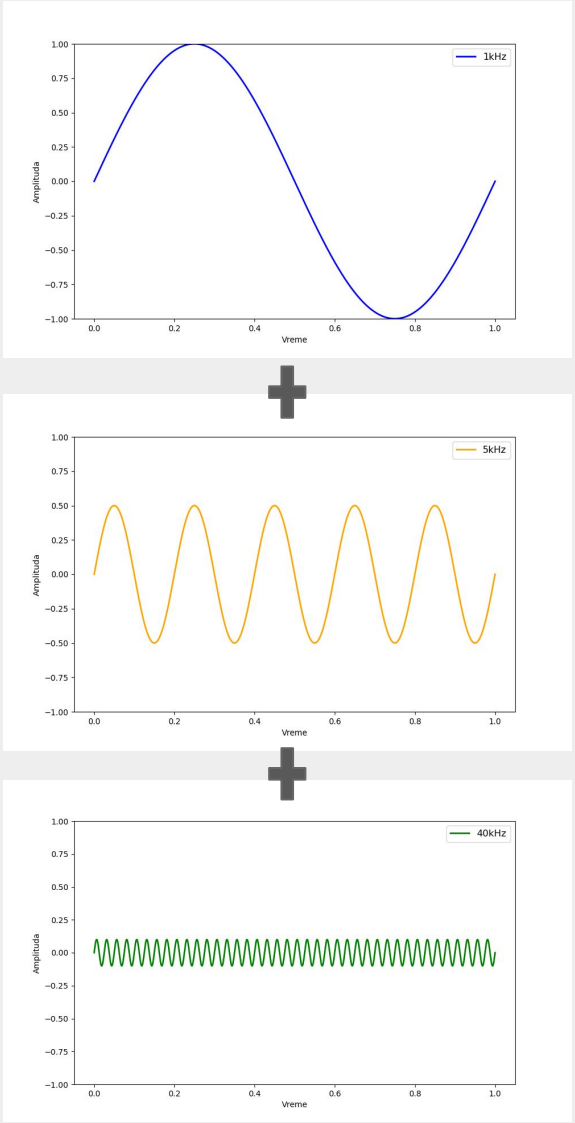
=

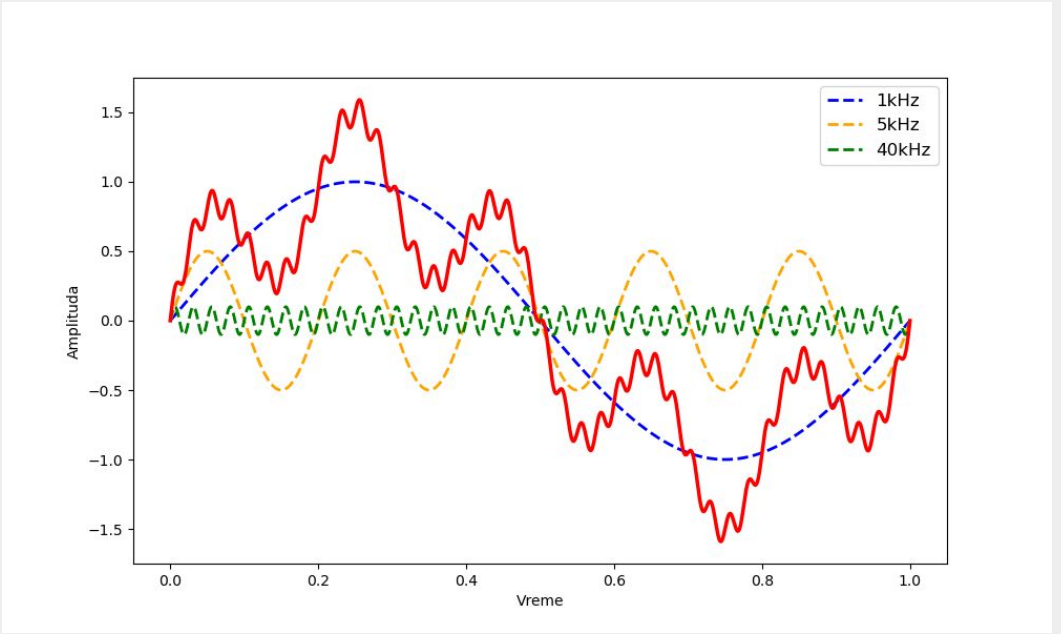
Periodični signali





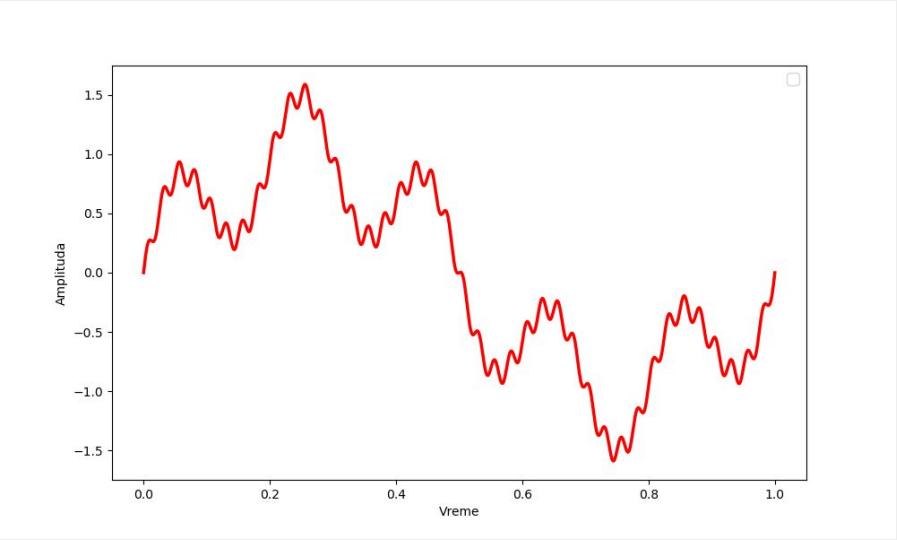
Periodični signali



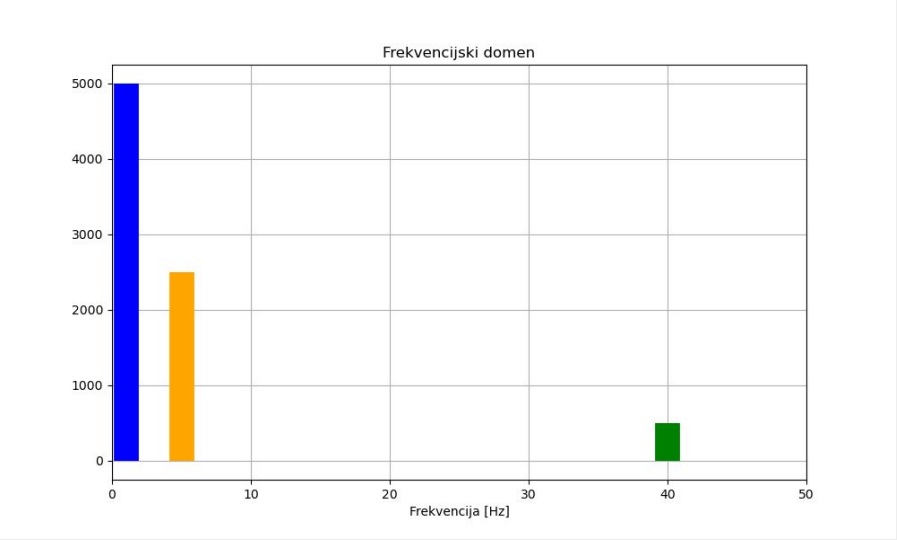


Matematički alat koji nam omogućava da periodične signale rastavimo na sumu prostih signala (sin i cos) naziva se **Furijeov red**.

Periodični signali



Vremenskom domen

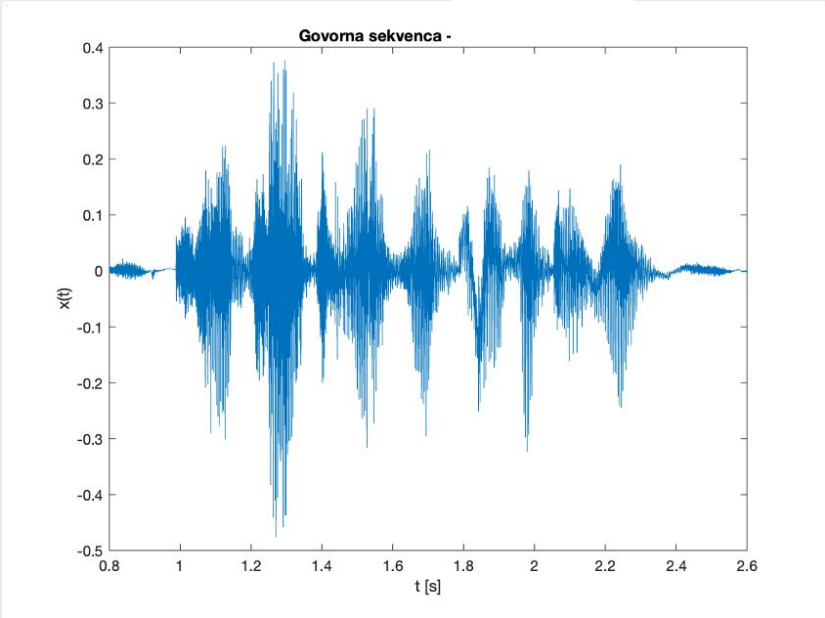


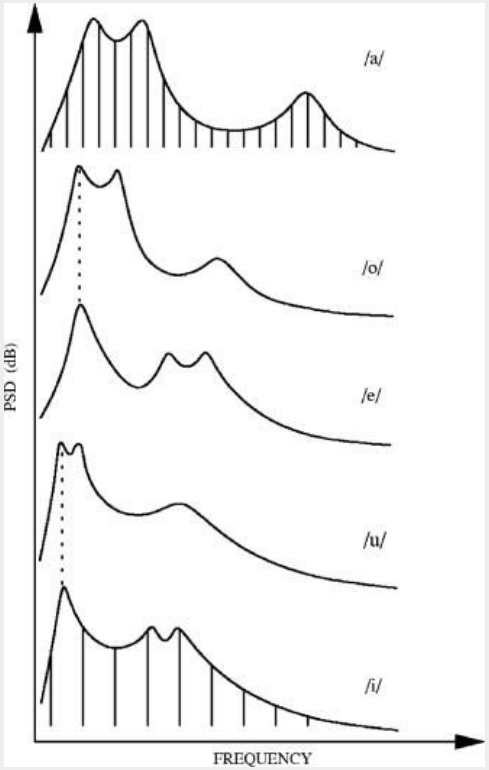
Frekvencijski domen

Neperiodični signali

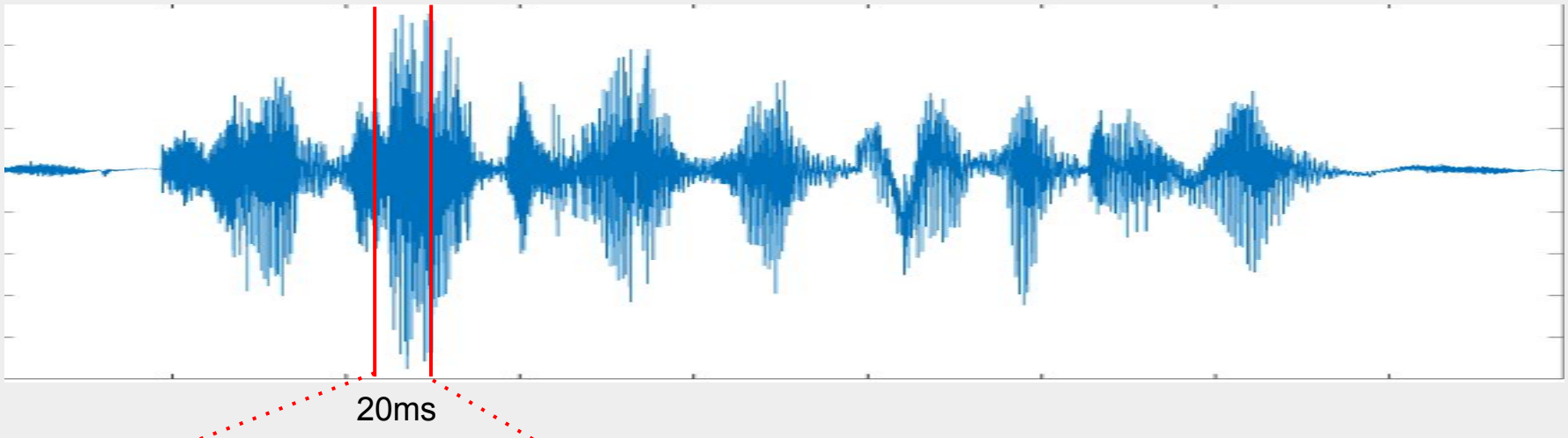
- Signali u prirodi nisu periodični, stoga njihov spektralni domen nije diskretan, tj. snaga signala je zastupljena na svim frekvencijama.

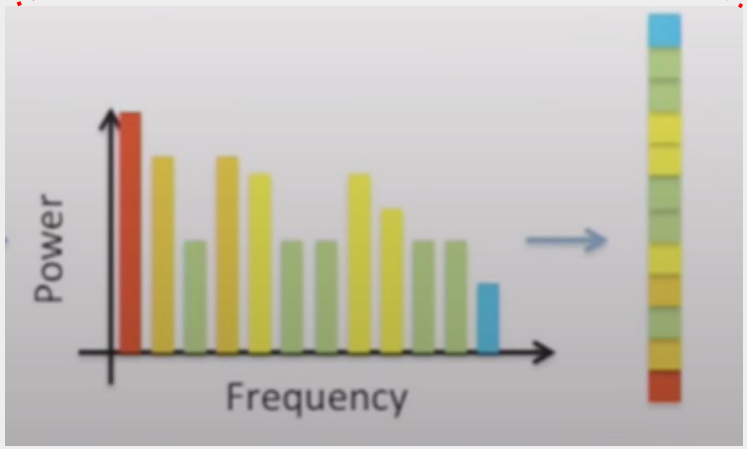
- Za spektralnu analizu neperiodičnih signala korist se **Furijeova transformacija**.





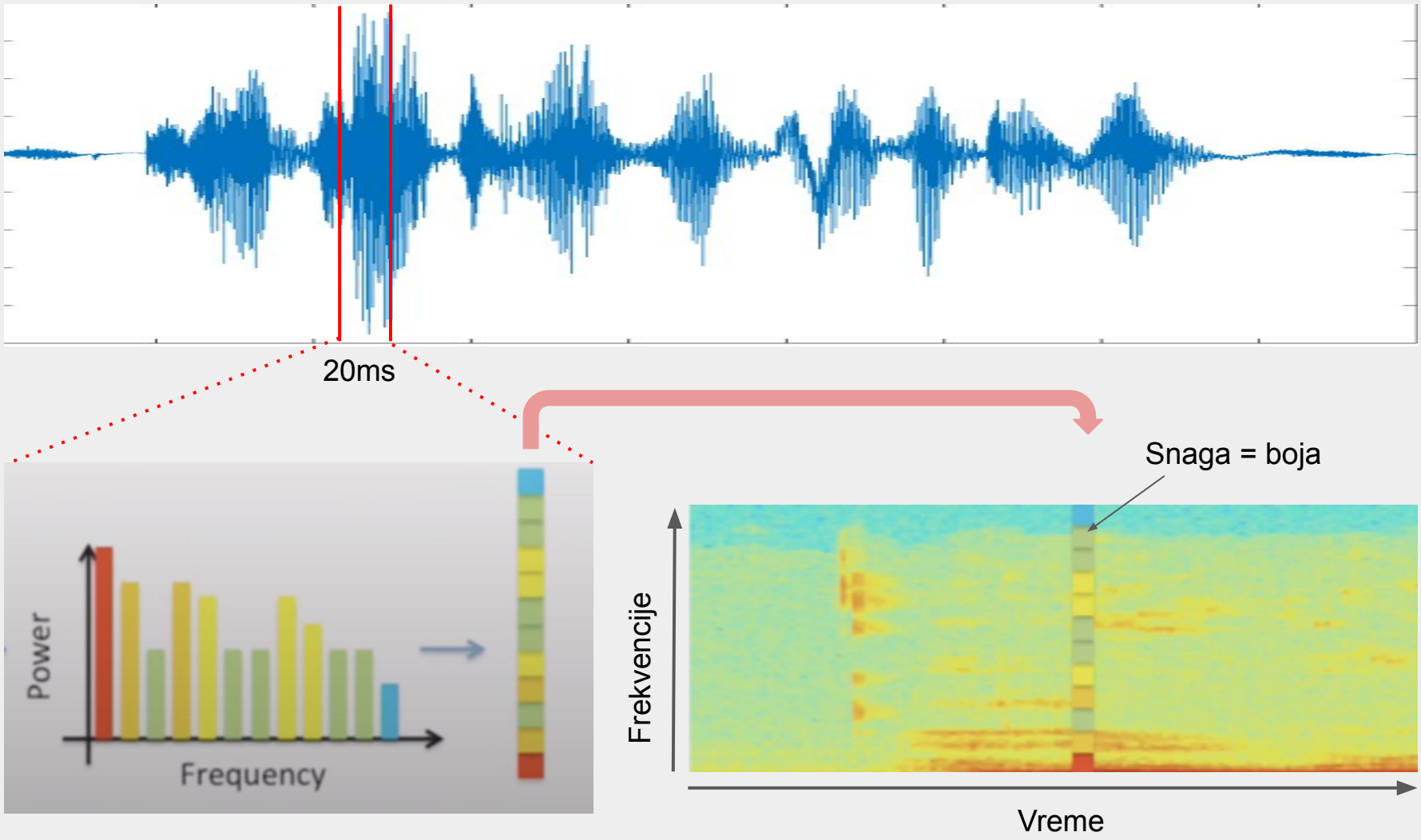
Spektogram



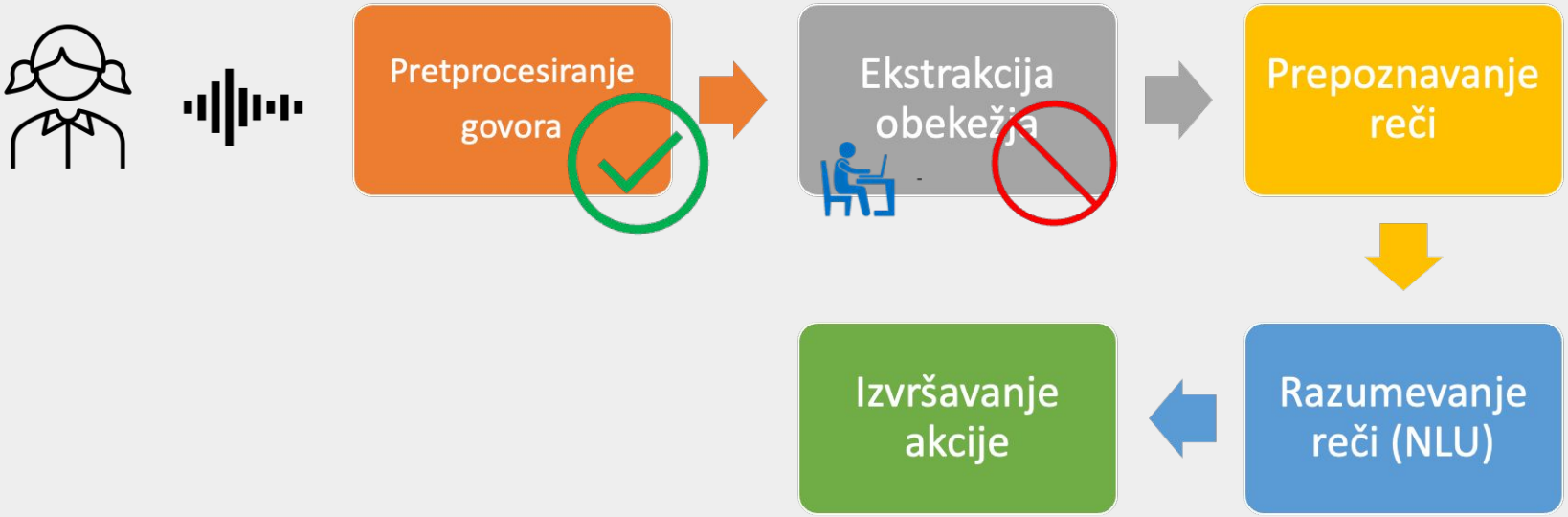


- Da li spektar govornog signala ostaje isti za svaki prozor?
- Izdelimo signal u vremenskom obliku na male prozore (npr. 20ms) i primenimo Furijeovu transformaciju na svaki.
- Snagu pojedinačnih frekvencija u tom prozoru smeštamo na spektogram.

Spektogram



Kako rade govorni asistenti?

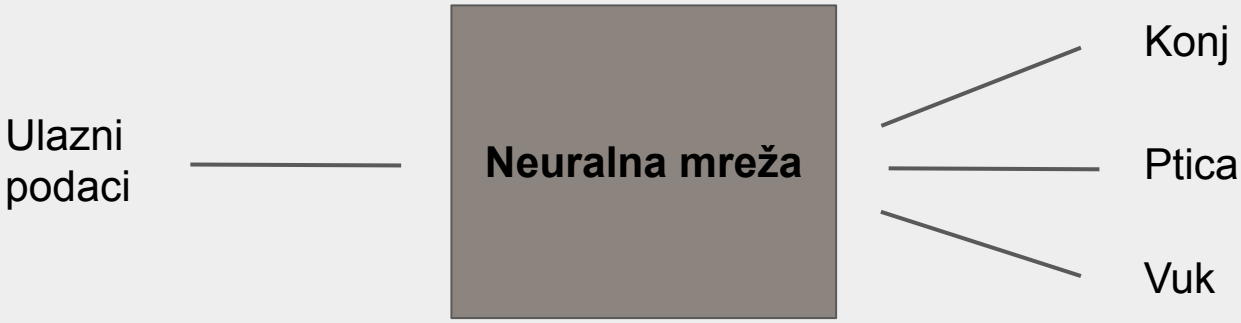


Agenda

1. Uvod u obradu govora
2. Metode dubokog učenja
3. Prepoznavanje emocija

Duboko učenje

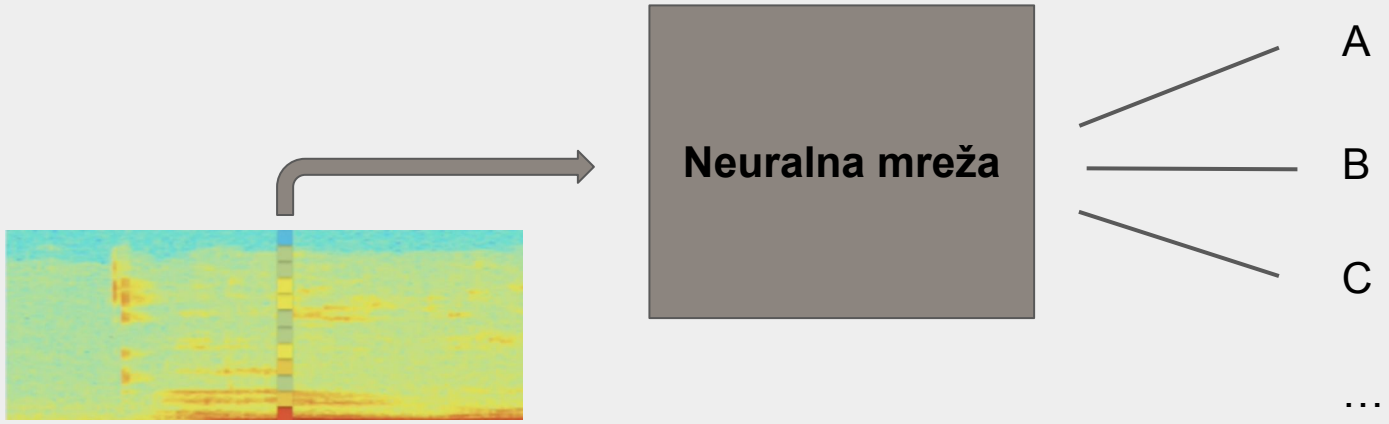
- **Duboko učenje** je grana mašinskog učenja koja ima za cilj da modeluje visoke nivoe apstrakcije u podacima koristeći se velikim brojem procesirajućih slojeva, sa ili bez kompleksih struktura i nelinearnim transformacijama.



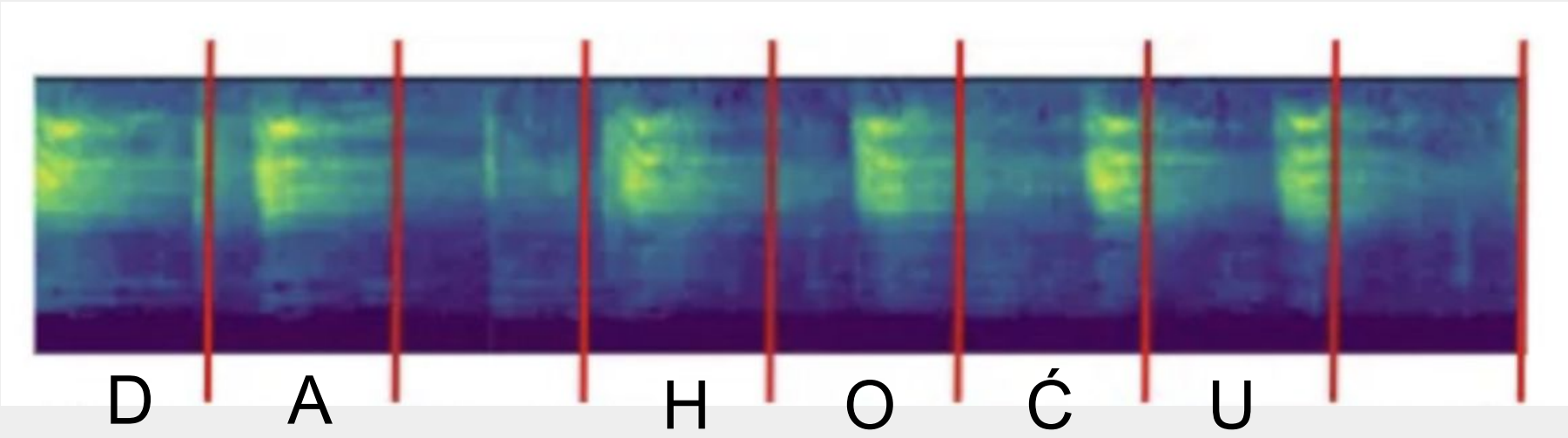
- Prepoznavanje govora, šta su
  - Ulazni podaci?
  - Klase?

Klasifikacija slova

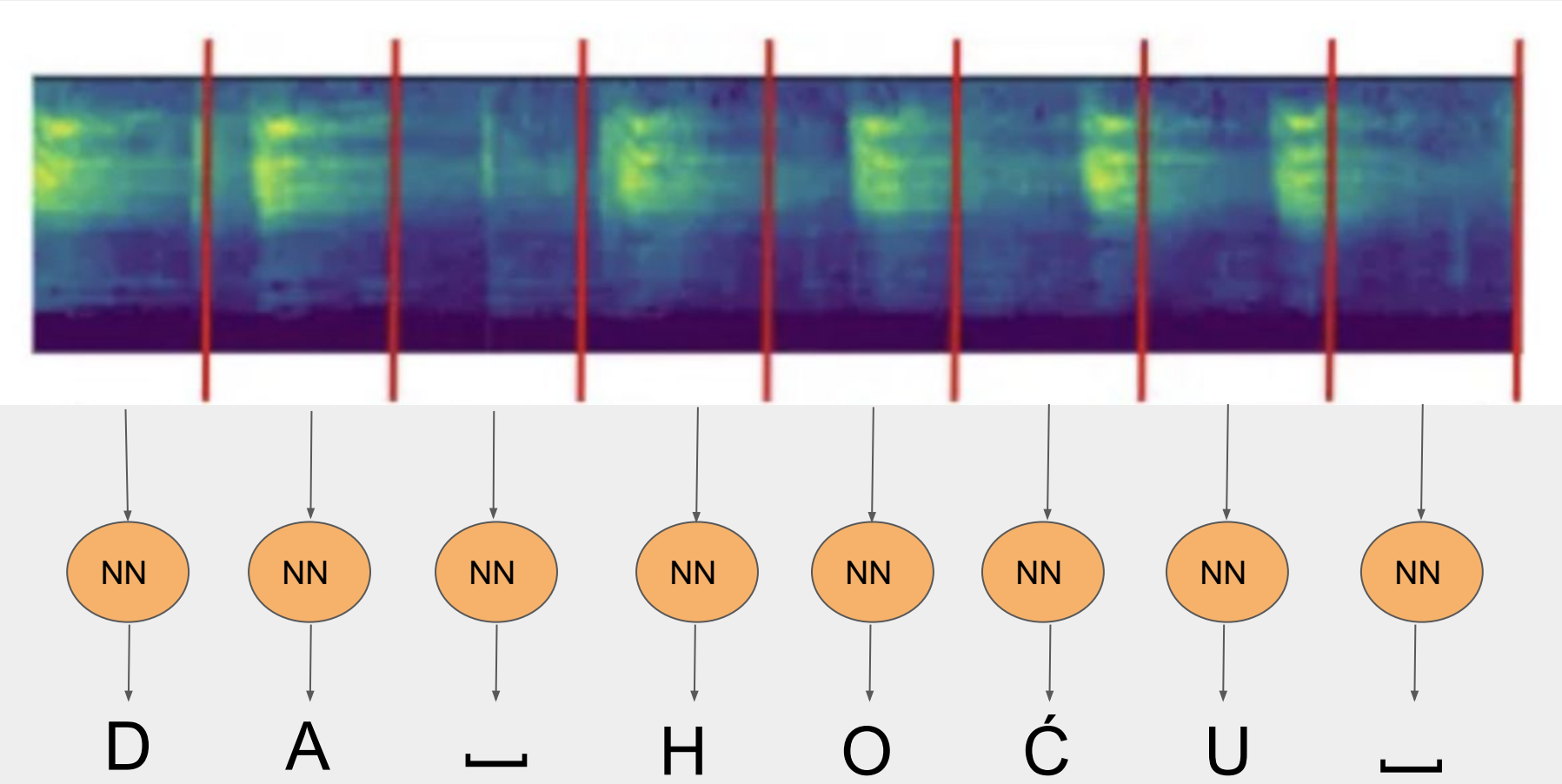
- Prepoznajemo slovo za svaki prozor spektograma



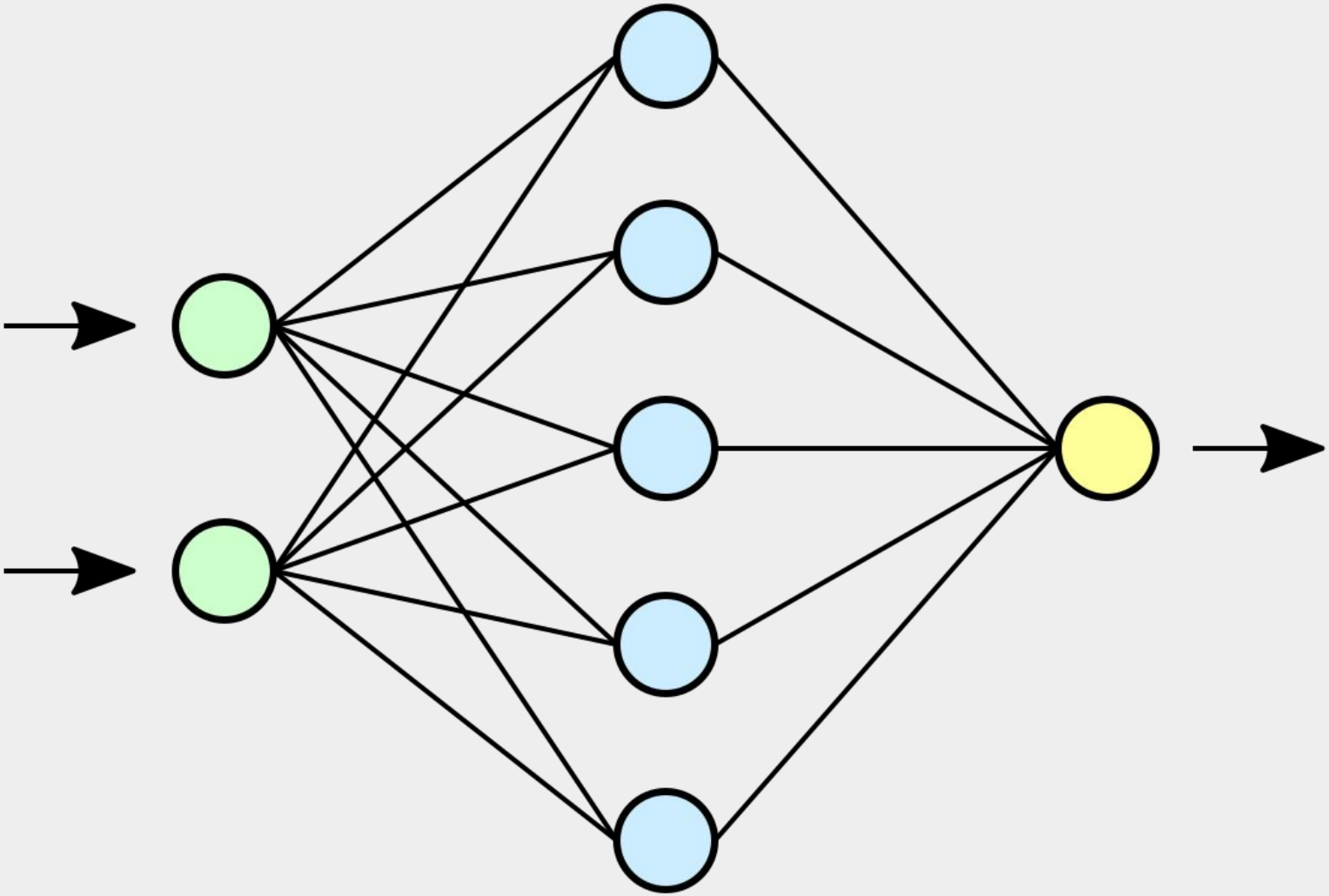
Klasifikacija slova



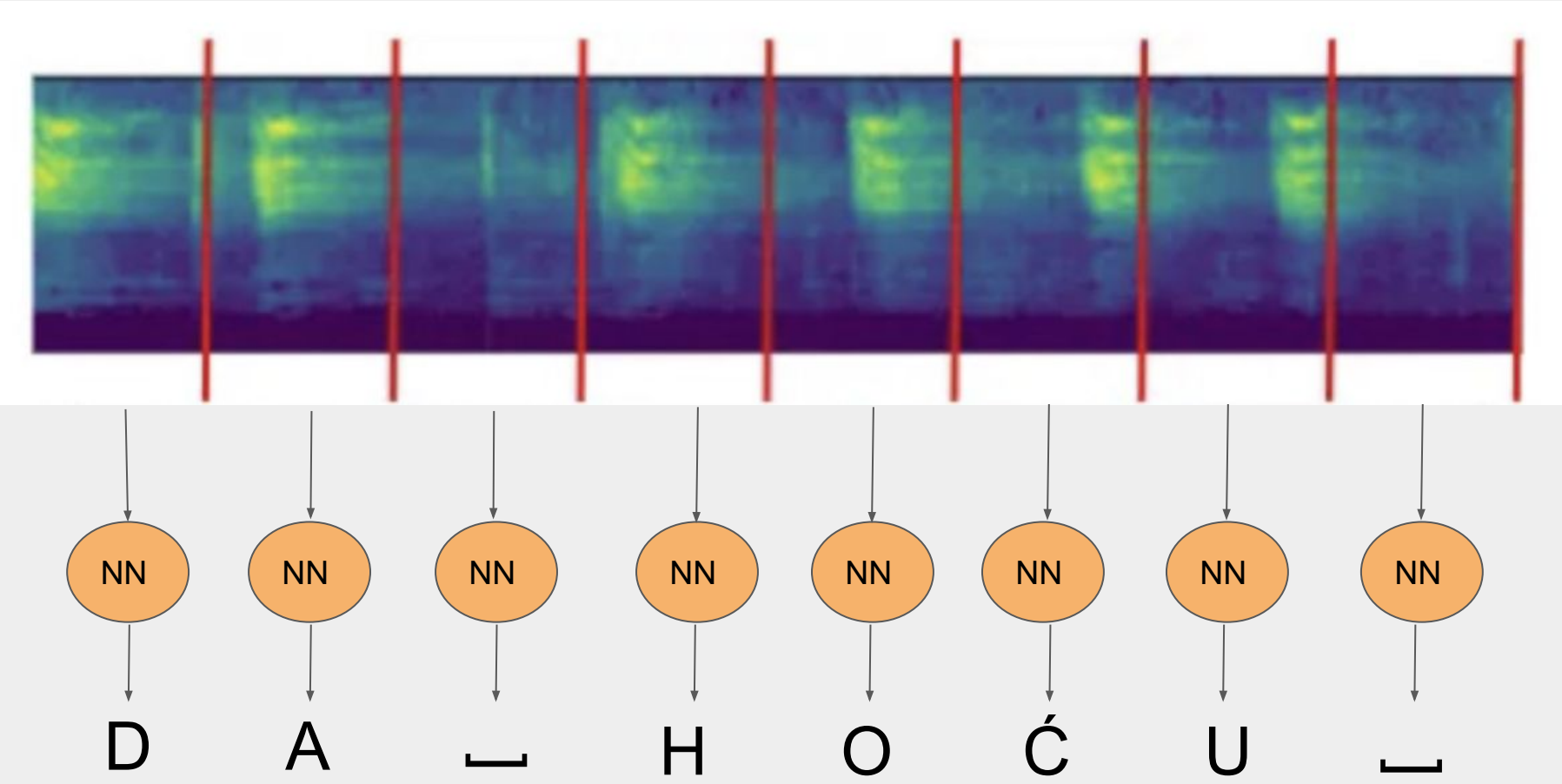
Neuralna mreža



Neuralna mreža

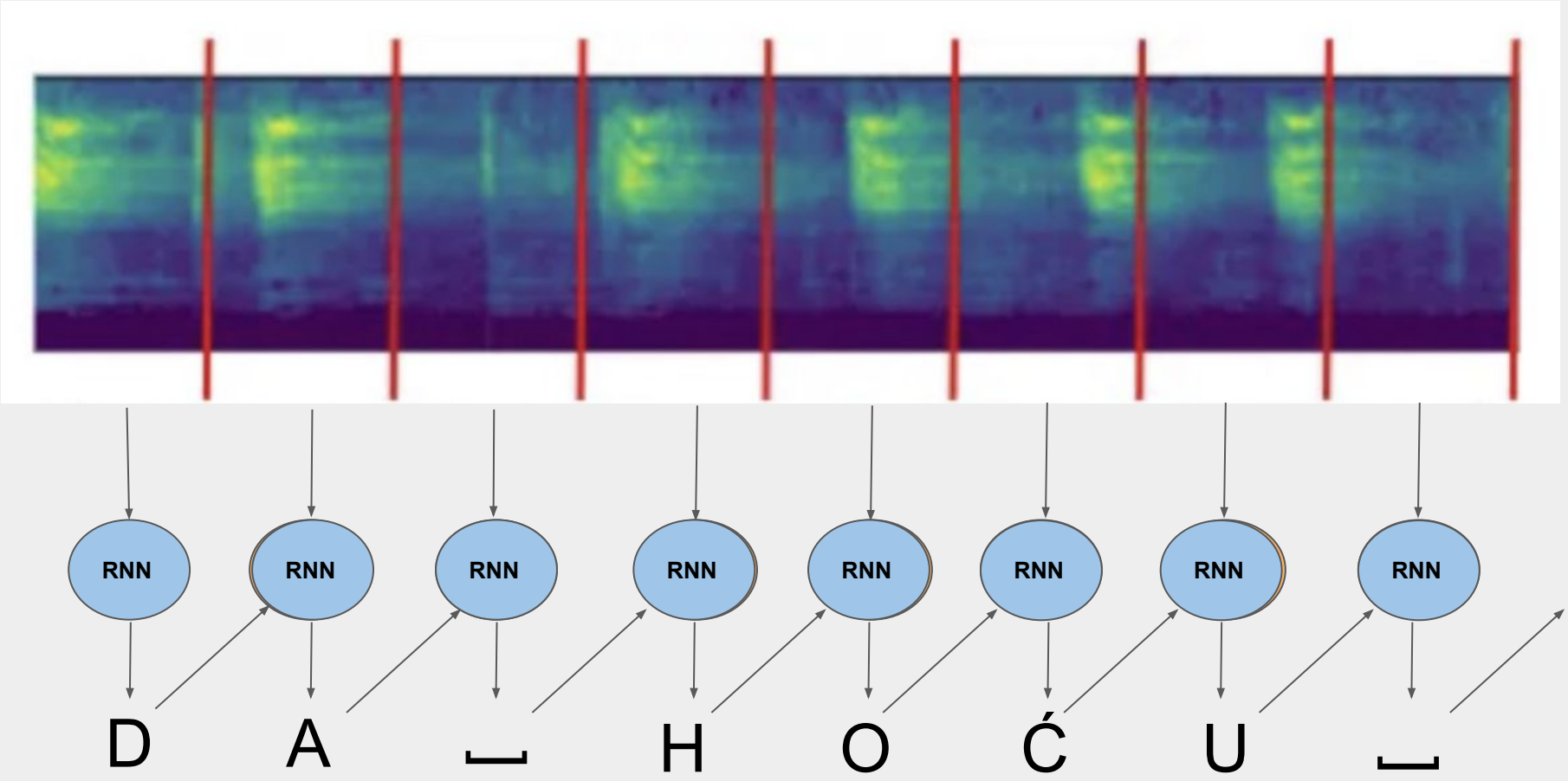


Neuralna mreža

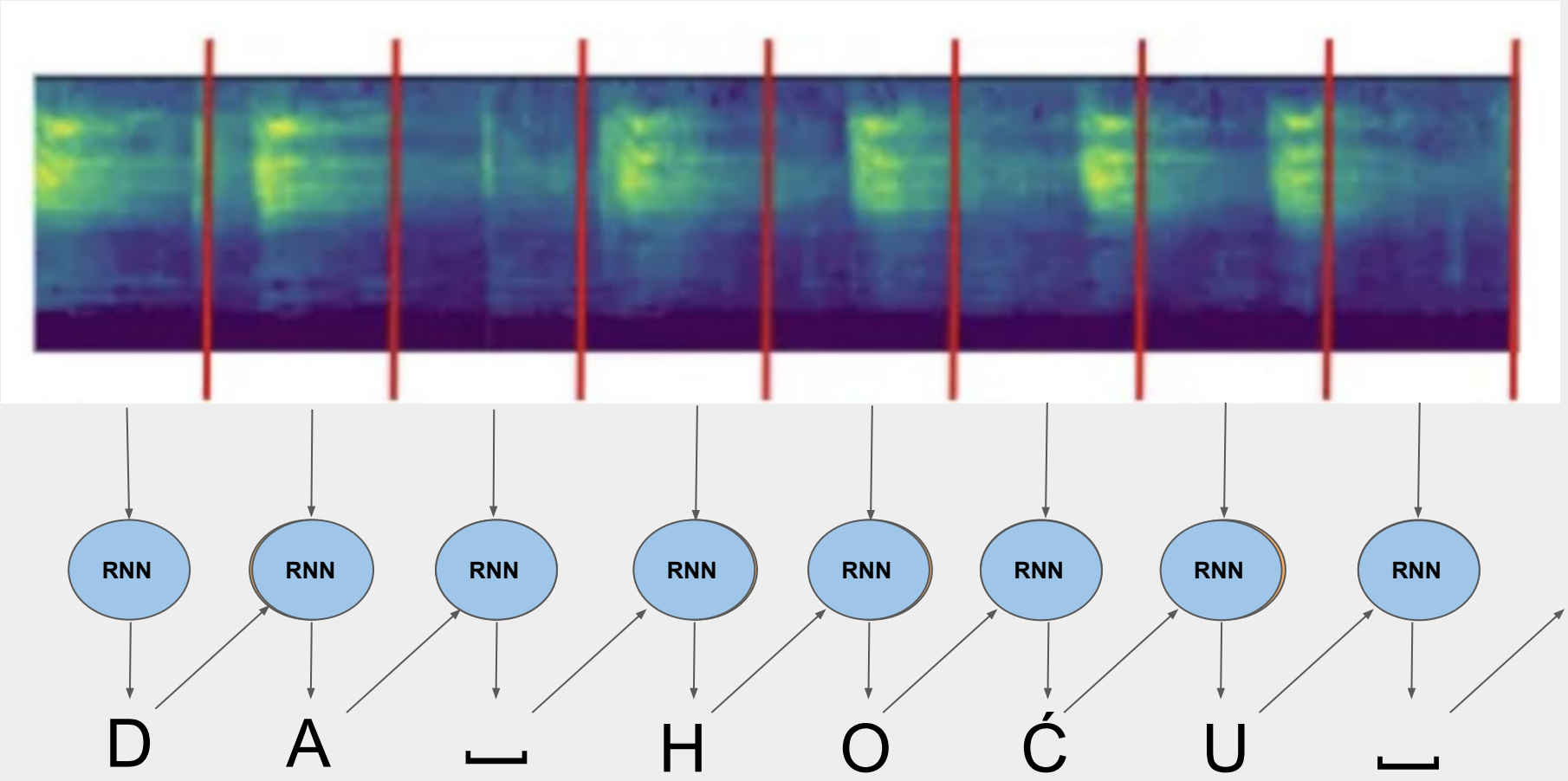


Da li koristimo sve dostupne informacije?

Rekurenta neuralna mreža



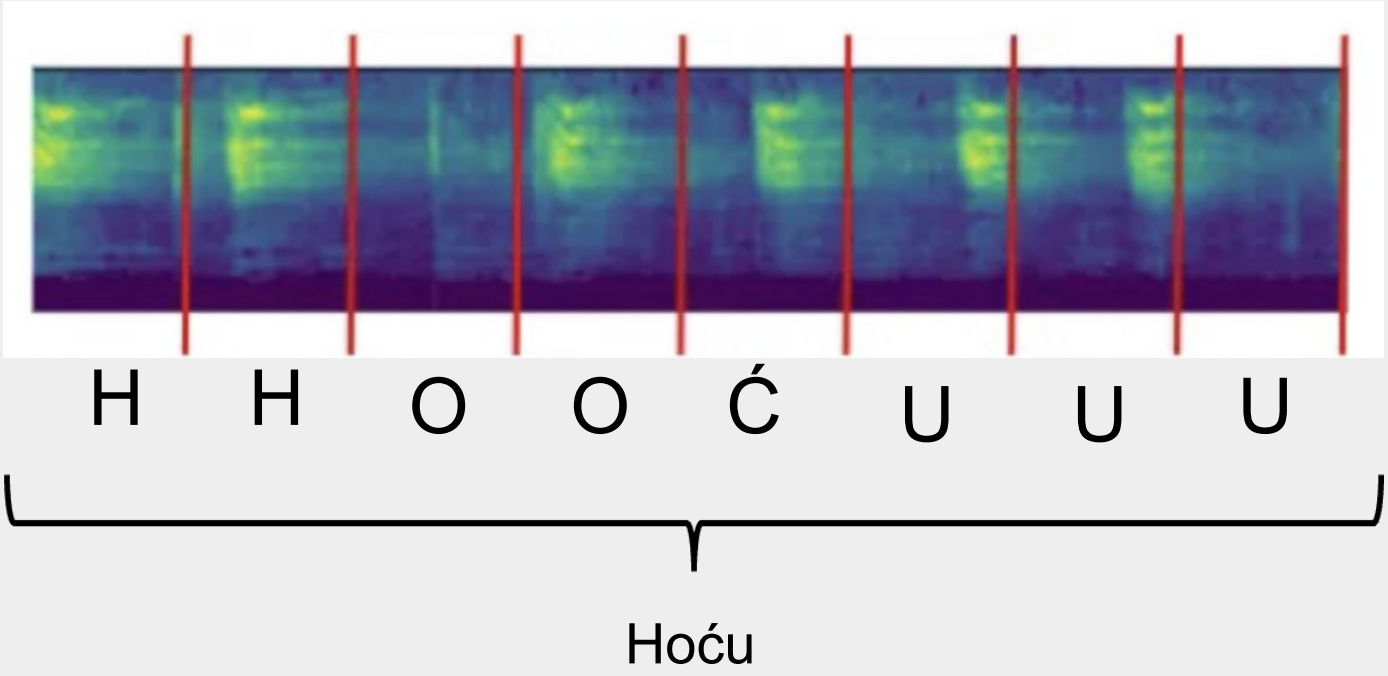
Rekurenta neuralna mreža



Svaki prozor novo slovo? Hmm...

Alignment problem (poravnanje)

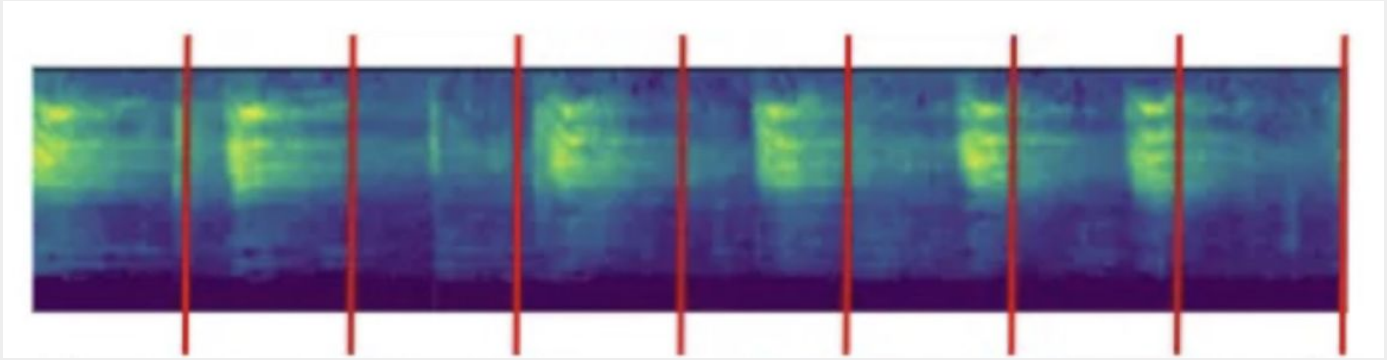
- Više uzastopnih vremenskih prozora odgovaraju istom slovu:



- Labeliranje pojedinačnih prozora je skupo, neefikasno i sklono greškama!
- Možemo li samo da sažmemo slova?

Alignment problem (poravnanje)

- Dva ista slova za redom:



Helo ✖

Treniranje mreže

- Uvodimo *blank* karakter. Ovo nije isto što i razmak!
- Prvo sažmemo, pa uklonimo *blank*.

Potencijalni izlazi iz modela:

H _ E _ _ L _ L L _ O -> HELLO V1

H H _ E _ L _ L _ O O -> HELLO V2

H _ E E _ L L L _ O -> HELO

H _ A _ L _ L _ O O O -> HALLO

Treniranje mreže

- Uvodimo *blank* karakter. Ovo nije isto što i razmak!
- Prvo sažmemo, pa uklonimo *blank*.

Potencijalni izlazi iz modela:

H _ E _ _ L _ L L _ O -> HELLO V1

H H _ E _ L _ L _ O O -> HELLO V2

H _ E E _ L L L _ O -> HELO

H _ A _ L _ L _ O O O -> HALLO

Kako ćemo da definišemo funkciju greške?

CTC funkcija greške

Potencijalni izlazi iz modela:

H _ E _ _ L _ L L _ O -> HELLO V1

H H _ E _ L _ L _ O O -> HELLO V2

H _ E E _ L L L L _ O -> HELO

H _ A _ L _ L _ O O O -> HALLO

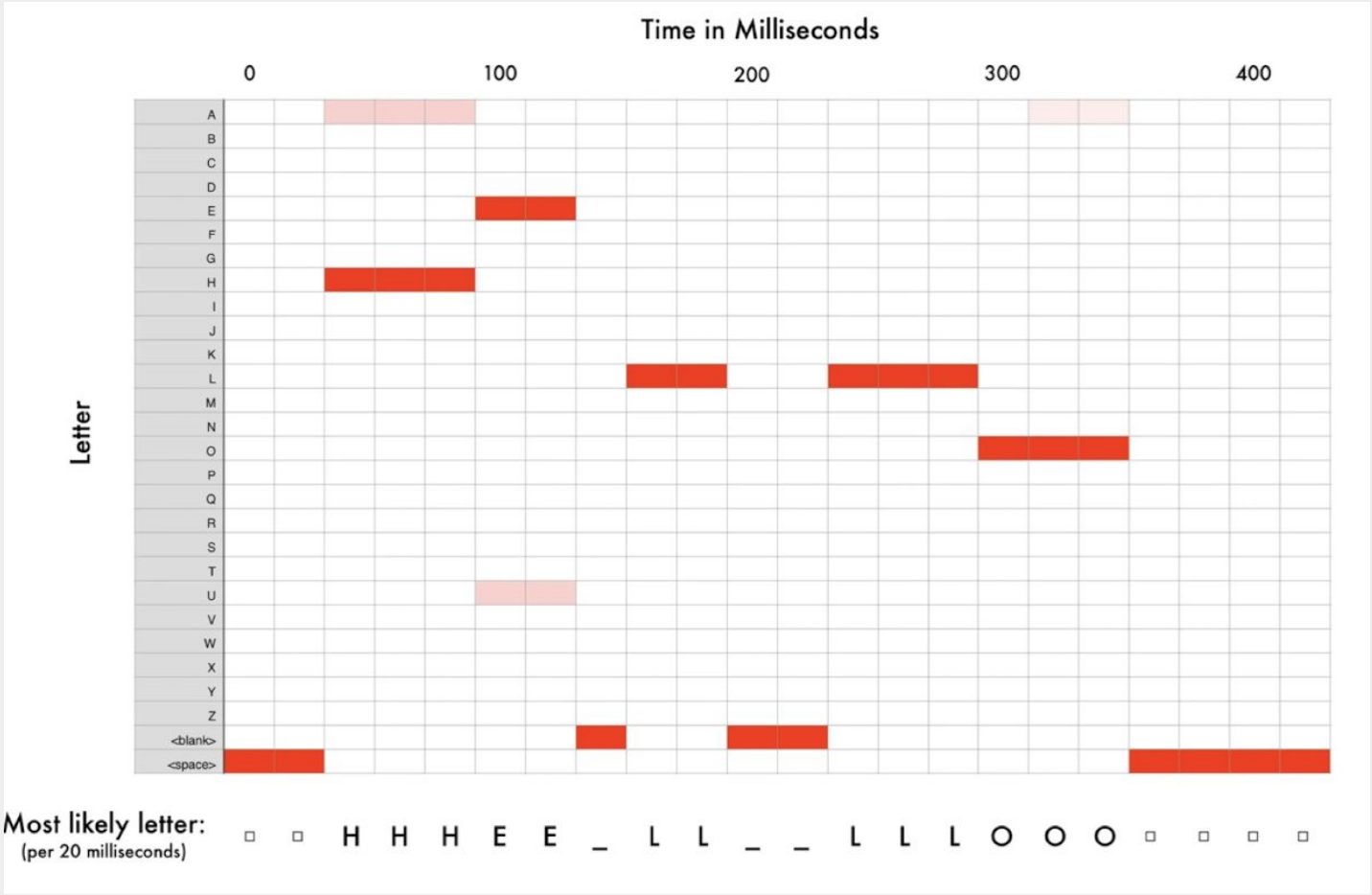
CTC (Connectionist temporal classification) **funkcija greške**: max(Pm("HELLO"))

Pm("HELLO") = P(V1) + P(V2) + P(V3) + ...

gde je: P(V1) = P(H)*P(_)*P(E)*P(_)*P(_)...

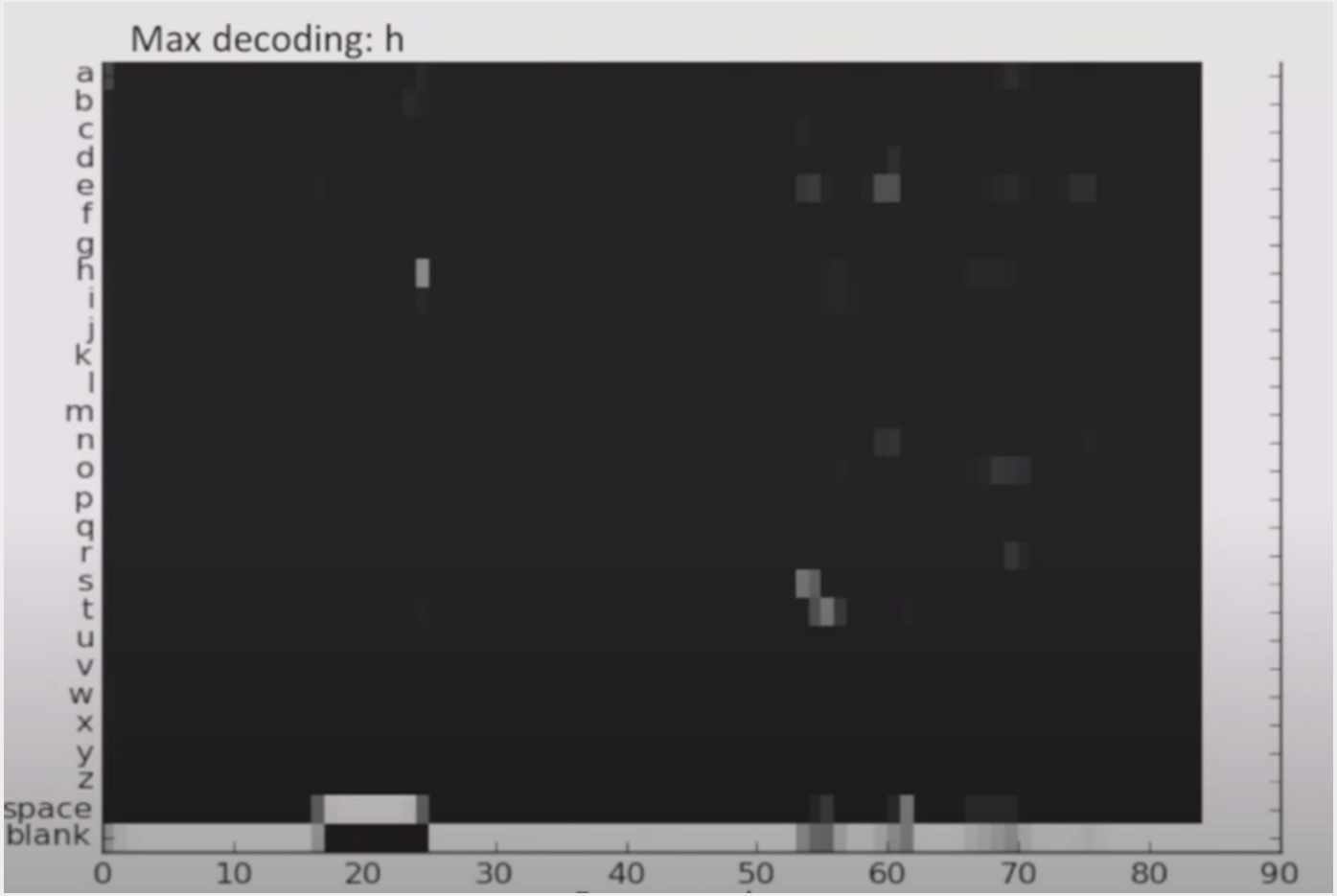
Dekodovanje - **Max Decoding**

- Uvek biraj slovo sa najvećom verovatnoćom u tom vremenskom prozoru.



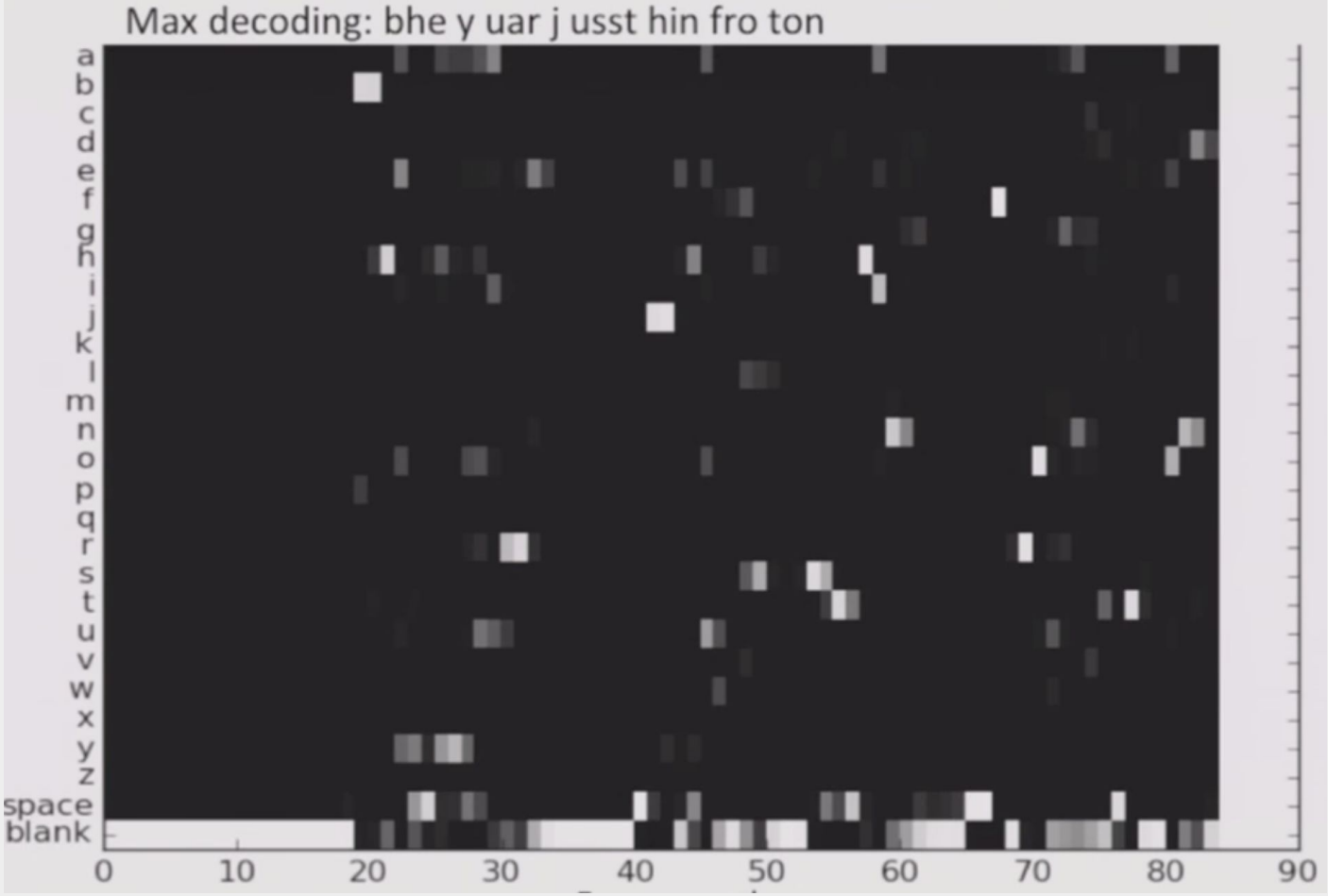
Dekodovanje - **Max Decoding**

- Nakon 300 iteracija



Dekodovanje - **Max Decoding**

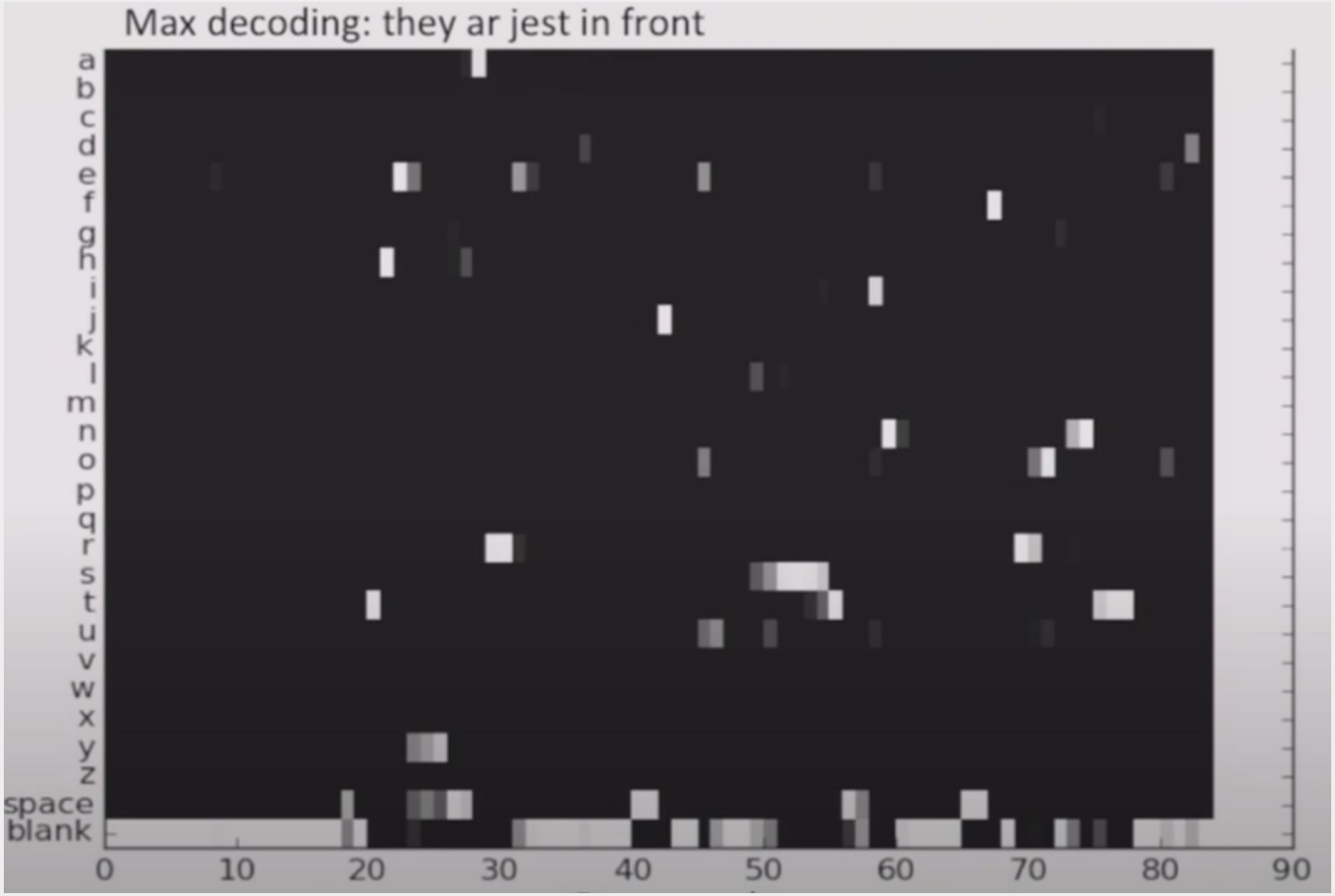
- Nakon 1500 iteracija



Dekodovanje - **Max Decoding**

- Nakon 5500 iteracija

Izgovoreno: “There just in front”



Integrisanje jezičkih modela

- Dobre sekvence su često odbačene prerano. Sistemi za prepoznavanje govora nemaju dovoljno bogate podatke za trening. (Vlastite imenice, retke reči...)

- Možemo da iskoristimo moć jezičkih modela (NLU), trenirani su na više podatak i mogu bolje da procene verovatnoću retkih sekvenci.

Koristimo jezički model P_{LM}:

$$P(C_t) = P_m(C_t)*P_{LM}(C_t|C_{1:t-1})$$

Do sada: **LO**
Next: O? $P(LO\mathbf{V}) = P_m(\mathbf{V})*P_{LM}(\mathbf{V}|LO)$
Next: M? $P(LO\mathbf{M}) = P_m(\mathbf{M})*P_{LM}(\mathbf{M}|LO)$

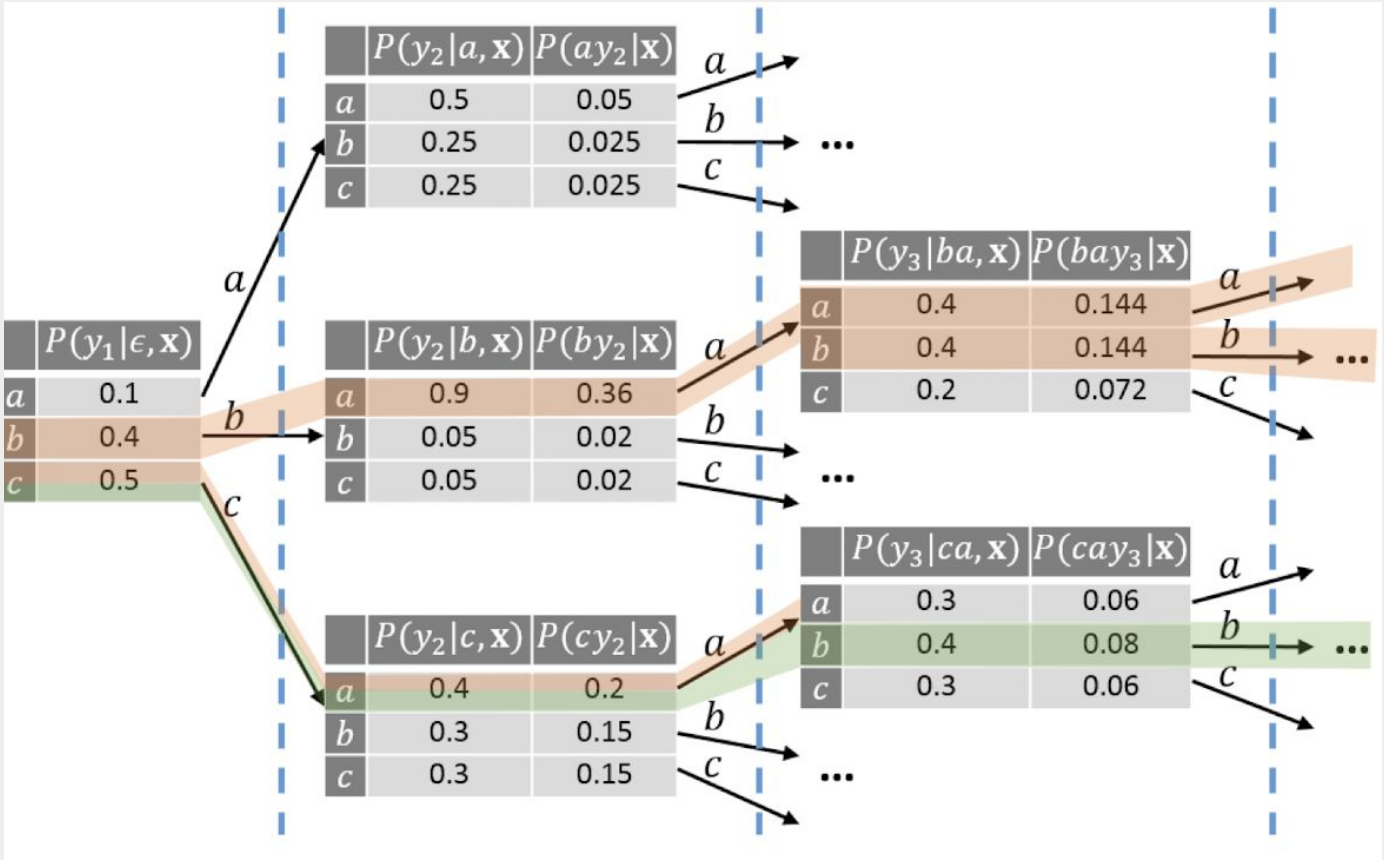
Do sada: **LA**
Next: O? $P(LA\mathbf{V}) = P_m(\mathbf{V})*P_{LM}(\mathbf{V}|LA)$
Next: M? $P(LA\mathbf{M}) = P_m(\mathbf{M})*P_{LM}(\mathbf{M}|LA)$

Previše kombinacija -> Beam search

Dekodovanje - Pretragom po snopu (**Beam search**)

Koristimo jezički model P_{LM}:

$$P(C_t) = P_m(C_t) * P_{LM}(C_t | C_{1:t-1})$$



Izazovi - Raznovrstnost podataka

- Najčešće se koristi „čitani govor'' iz izvora kao što su *Street Journal* (novine) i *LibriSpeech* (audio knjige).
- **Kvalitet podatak** je **izuzetno važan** kod metoda dubokog učenja, neadekvatni podaci dovešće do niske performanse modela u realnoj upotrebi.
- Ovi primeri ne odgovoraju većini svkodnevnih situacija ili specifičnih primena.
- Lombard efekat.
- Akcenti, pol, zamuckivanja, godište, šumovi...

Augmentacija podataka

- Veštačko formiranje skupa podataka veće raznovrsnost.
- Povećanje/smanjenje brzine, dodavanje šuma...
- Augmentacijom podataka možemo povećati skup podataka nekoliko puta.
- Nekad je jednostavnije uticati na podatke nego na robustnost modela.

Klasifikacija na foneme

- Često se umesto prepoznavanja slova, prepoznaju **fonemi**.
- Fonemi su eksperimentalno određene zvučne jedinice jezika.

Dekompozicija na:

foneme *slova*

Hello -> [**'HH' 'AH' 'L' 'OW'**] Hello -> ['H' 'E' 'L' 'L' 'O']

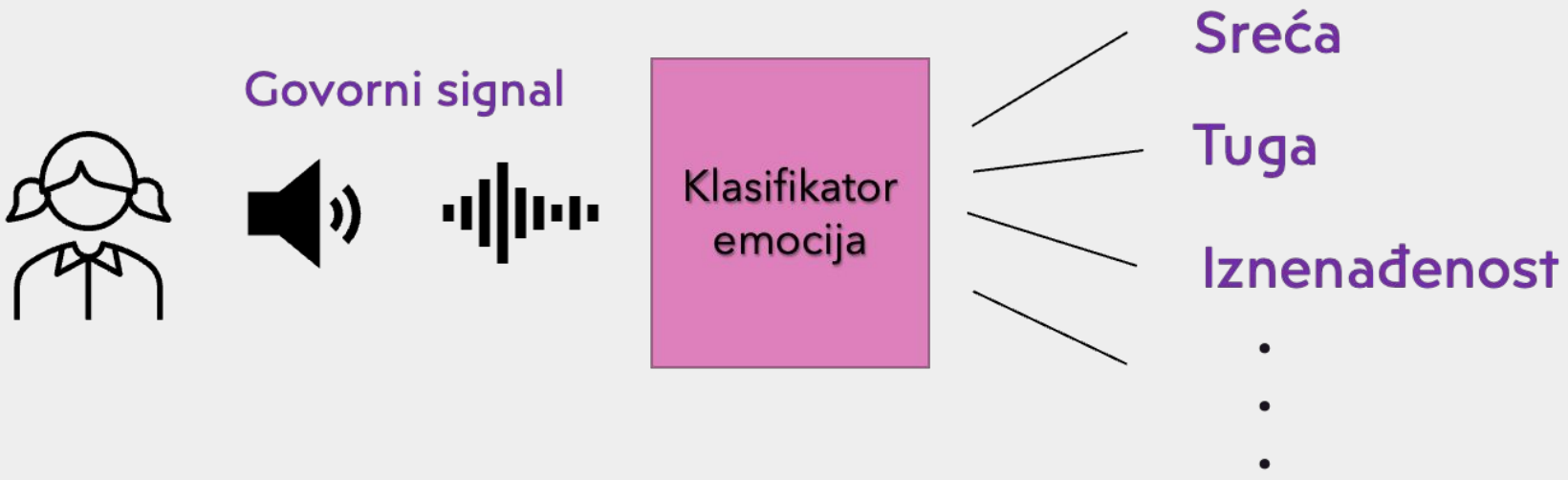
	Phone Label	Example		Phone Label	Example		Phone Label	Example
1	iy	b*ee*t	22	ch	*ch*oke	43	en	butt*on*
2	ih	b*i*t	23	b	*b*ee	44	eng	Wash*ing*ton
3	eh	b*e*t	24	d	*d*ay	45	l	*l*ay
4	ey	b*ai*t	25	g	*g*ay	46	r	*r*ay
5	ae	bat	26	p	*p*ea	47	w	*w*ay
6	aa	b*o*b	27	t	*t*ea	48	y	*y*acht
7	aw	b*ou*t	28	k	*k*ey	49	hh	*h*ay

Agenda

1. Uvod u obradu govora
2. Metode dubokog učenja
3. Prepoznavanje emocija

Klasifikacija emocija

- Klasifikacioni problem.
- Prepoznavanje emocija može biti subjektivno.
- Primene
  - Govorna podrška (analiza i unapređenje)
  - Automatski intervju asistenti
  - Razumevanje budućih akcija
  - Terapije i mentalno zdravlje



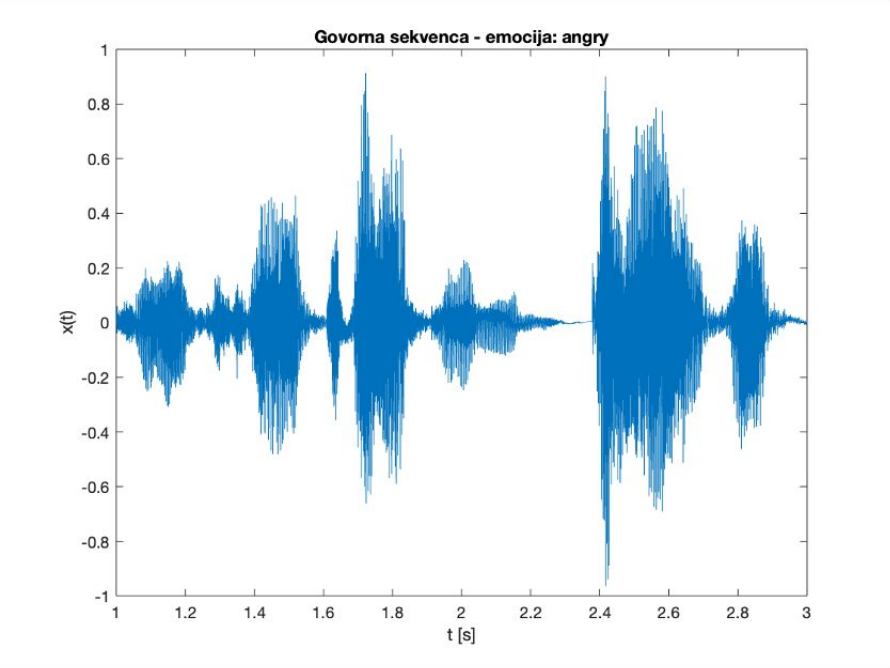
Podaci

- RAVDESS baza podataka
- 24 govornika iz istog govornog područja - glumci
- 2 rečenice
- 8 emocija

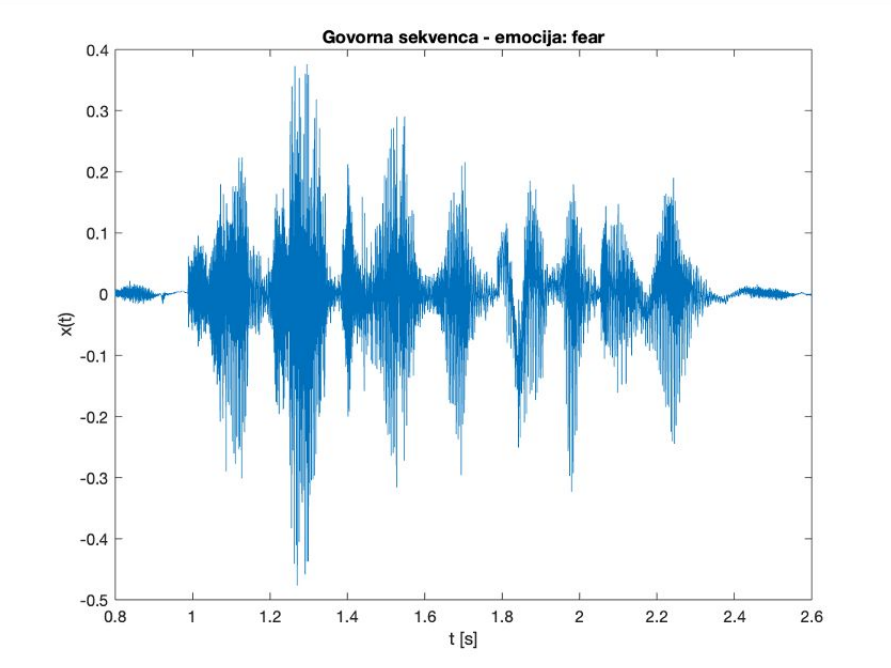
Sad Angry Disgust Fear Suprise Calm Neutral Happy

Pretprocesiranje i vremenski domen

- Odsecanje tišine postavljanjem praga snage

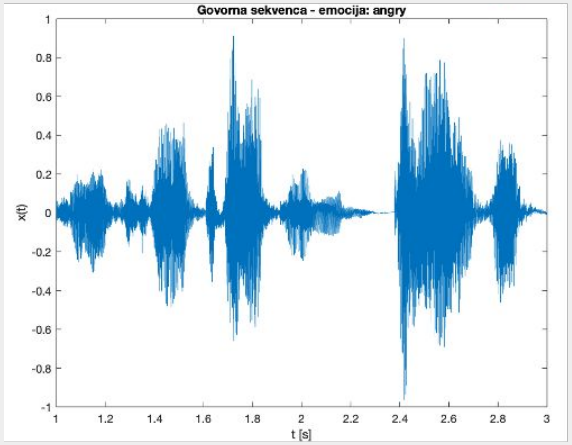


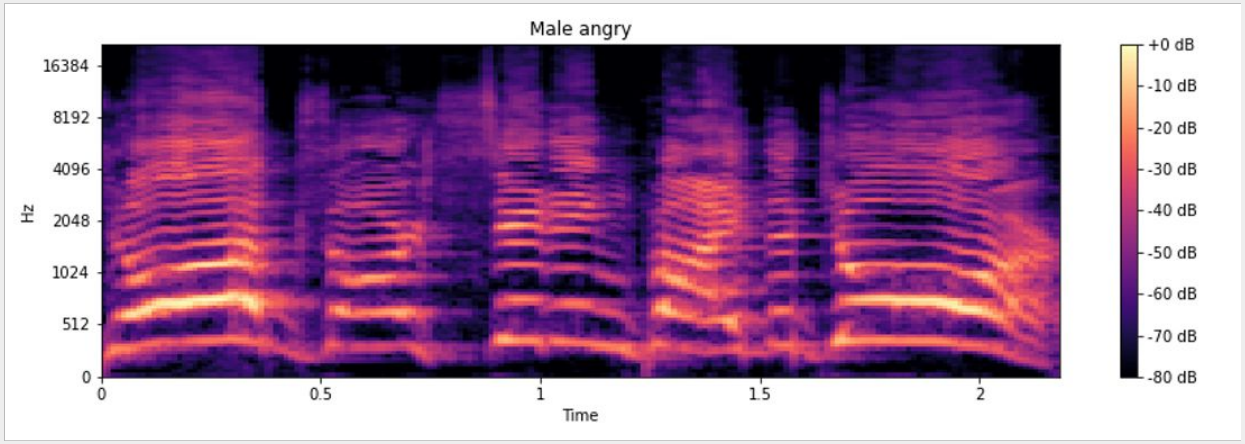
Ljutnja



Strah

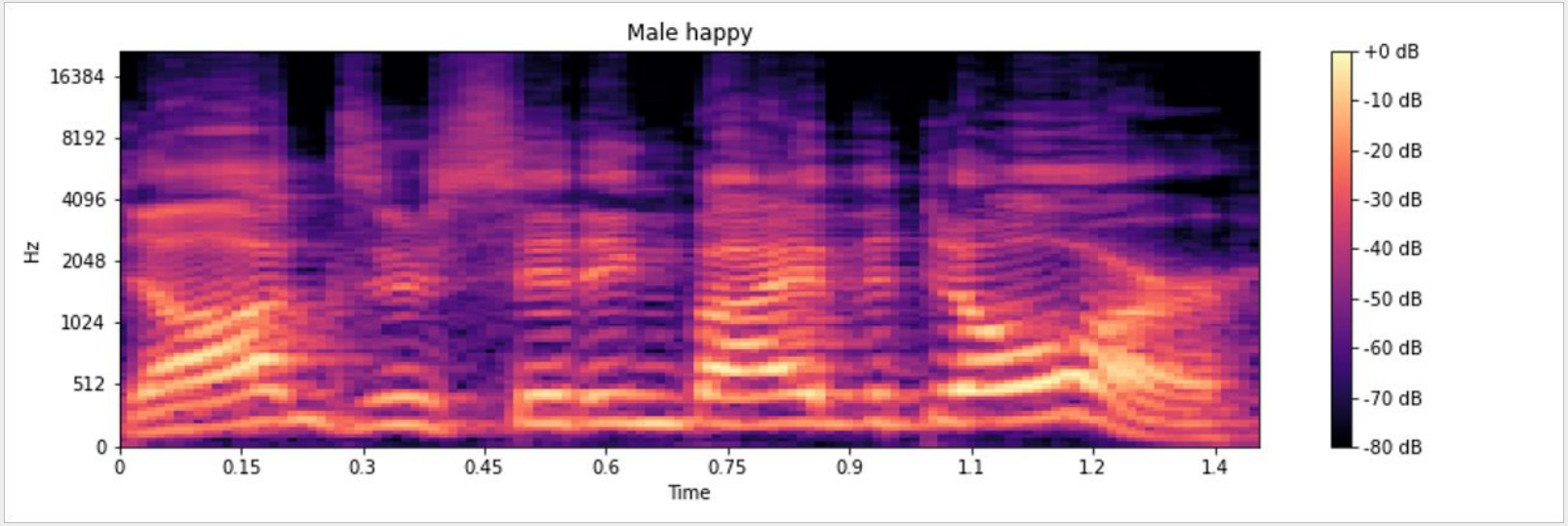
Frenkvencijski domen



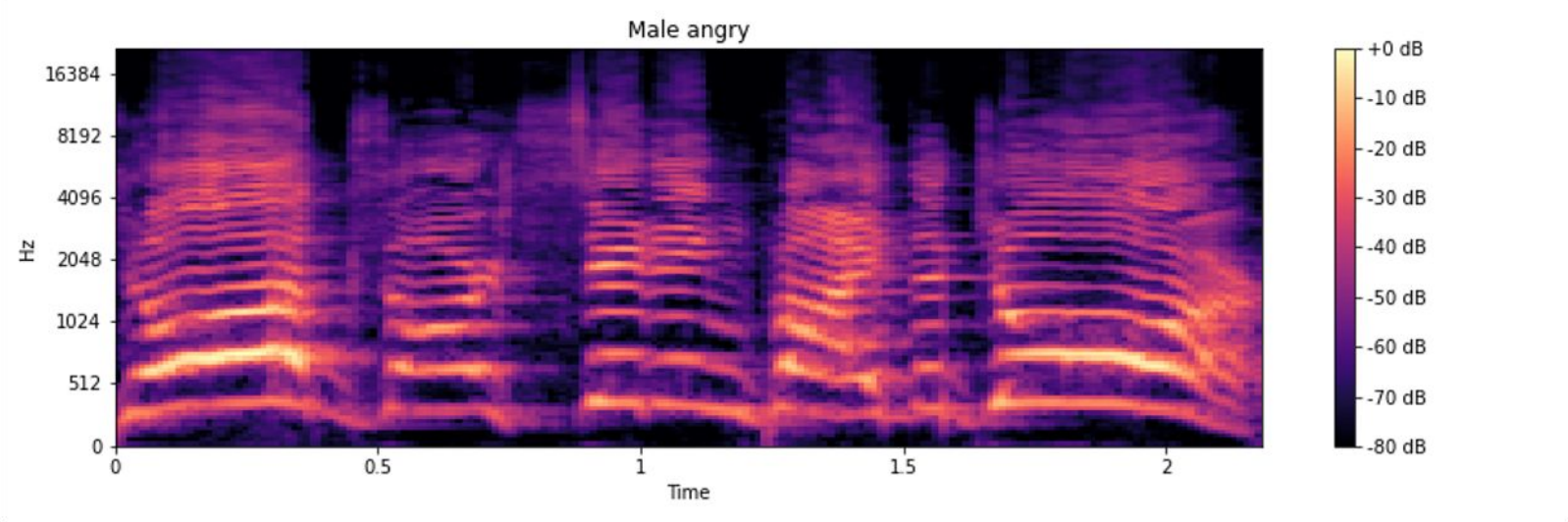


- Mel skala - ljudsko uho teže razlikuje tonove na višim frekvencijama.
- Eksperimentalna frekvencijska skala (mel skala)

Spektogram emocija



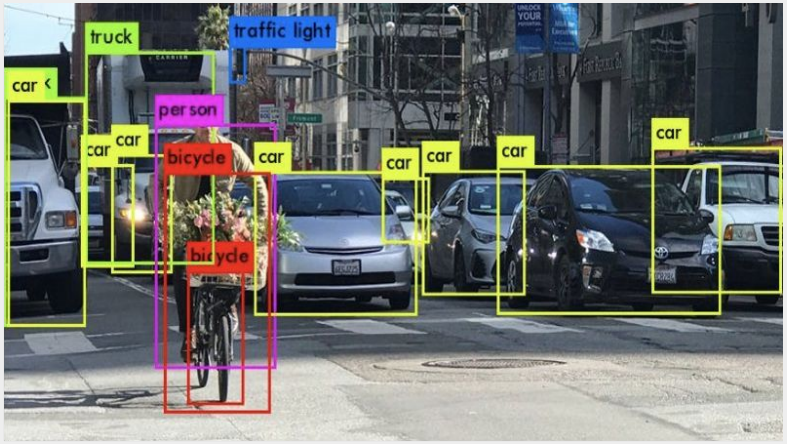
Sreća



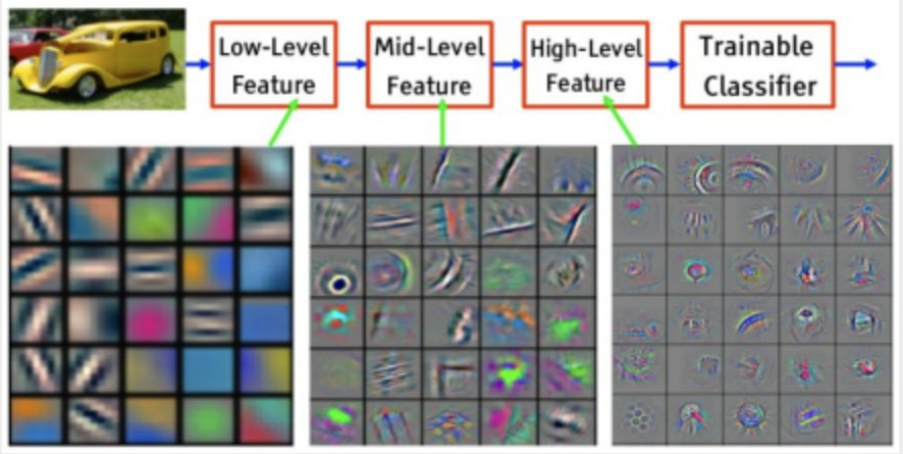
Ljutnja

Konvoluciona mreža za klasifikaciju

- Specijalni tip neuralnih mreža specijalizovan za izvlačenje podataka iz mrežastih struktura koa što su slike



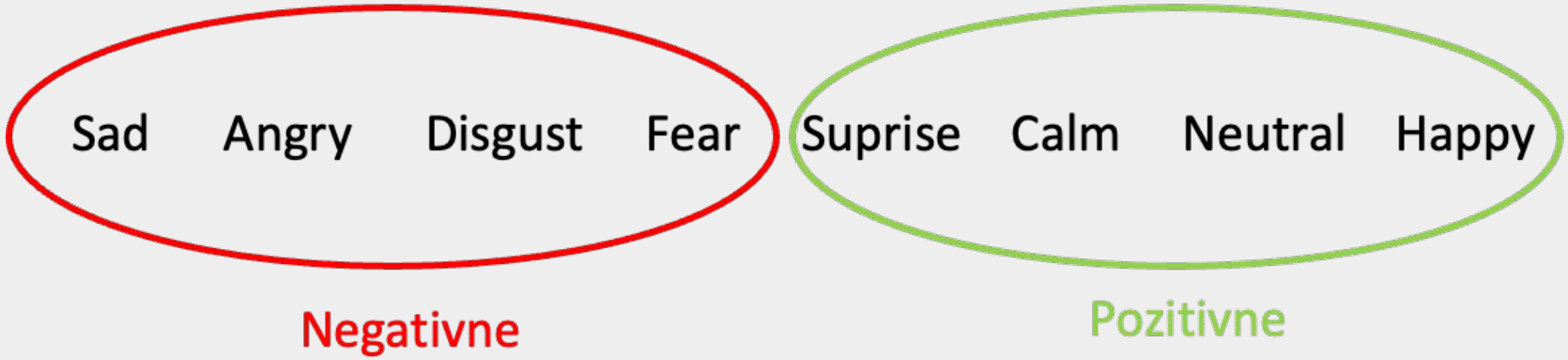
CNN za detekciju objekata



Specijalizacija za konkretna obeležja kroz slojeve

Rezultati

- Binarna klasifikacija



- Tačnost: **84%**

Rezultati

- Po tipu - 8 klasa

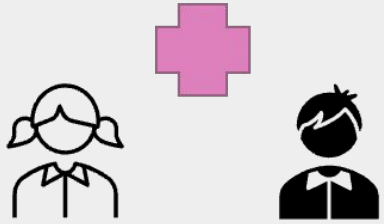
Sad Angry Disgust Fear Suprise Calm Neutral Happy

- Tačnost: **52%**, Top 3: **88%**

- Po tipu i polu - 16 klasa

Sad Angry Disgust Fear Suprise Calm Neutral Happy

- Tačnost: **50%**, Top 3: **78%**



Zanimljivi primeri

?

Sad Angry Disgust Fear Suprise Calm Neutral Happy

Zanimljivi primeri

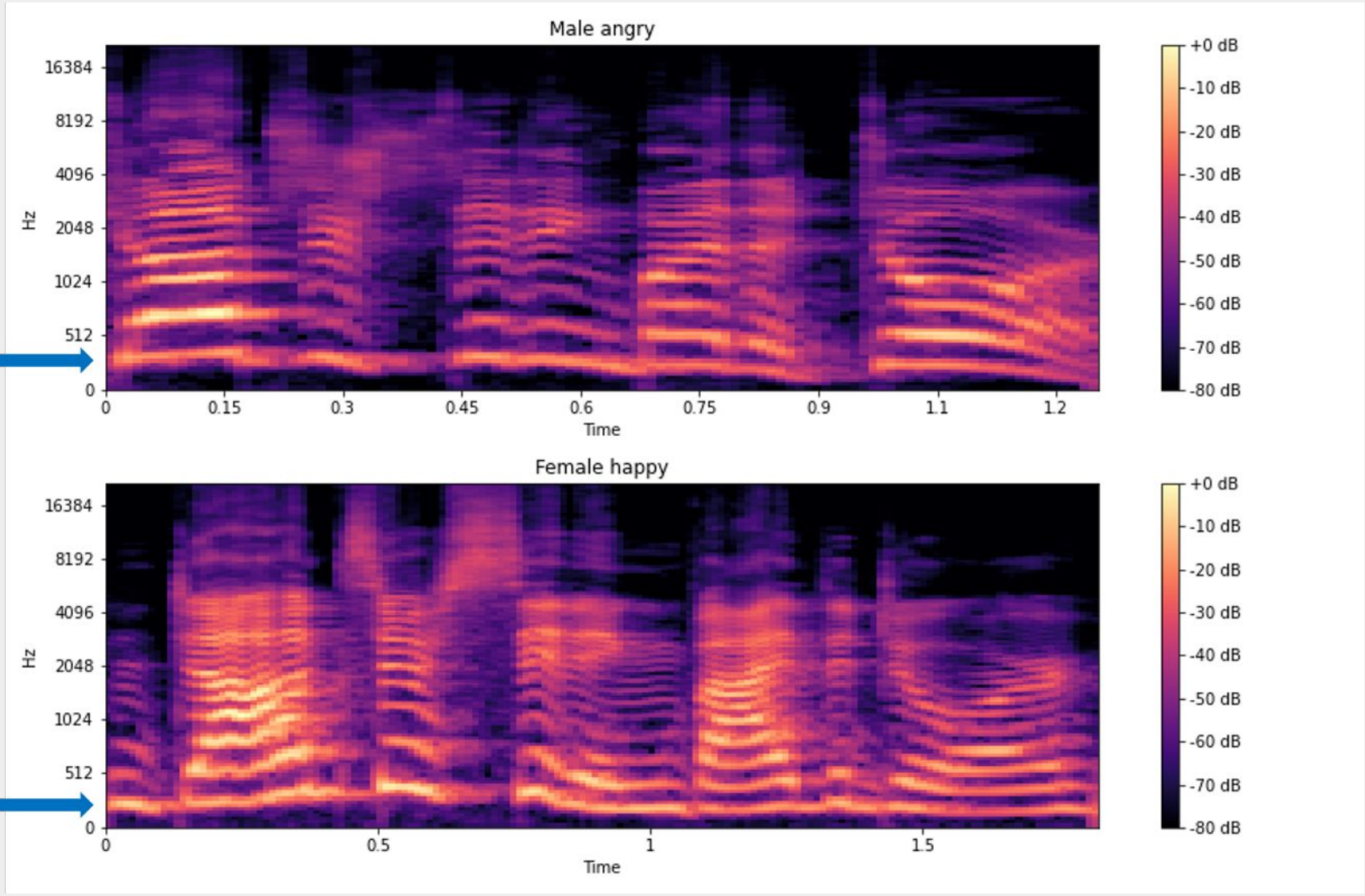
Labela	Model
Iznenadjena	Srećna

Zgrožen / Ljut Zgrožen / Uplašen ...

Anketa slušalaca

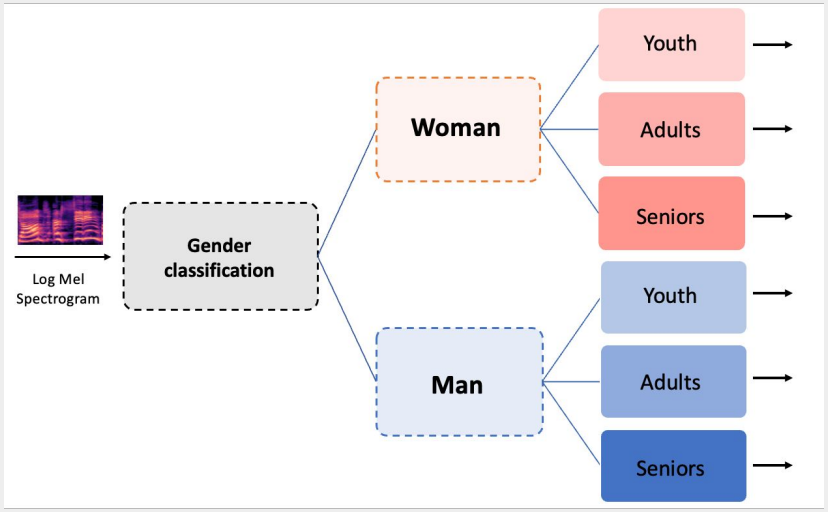
		Actor intended emotion							
		Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprise
Rater chosen emotion	Neutral/Calm	86.6	69.9	14.25	17.12	4.03	4.5	4.36	7.03
	Happy	0.63	17.27	68.44	1.48	0.23	0.59	0.59	6.56
	Sad	4.65	6.06	2.29	60.85	1.02	6.58	8.65	0.76
	Angry	3.82	1.02	1.79	2.9	81.32	4.79	6.48	2.78
	Fearful	0.63	0.66	1.67	9.64	1.39	70.71	2.31	2.22
	Disgust	1.15	1.46	0.78	3.09	8.37	1.81	69.77	3.28
	Surprise	0.28	0.33	7.88	0.69	1.2	7.76	4.13	72.29
	None	2.26	3.3	2.9	4.24	2.45	3.26	3.72	5.07

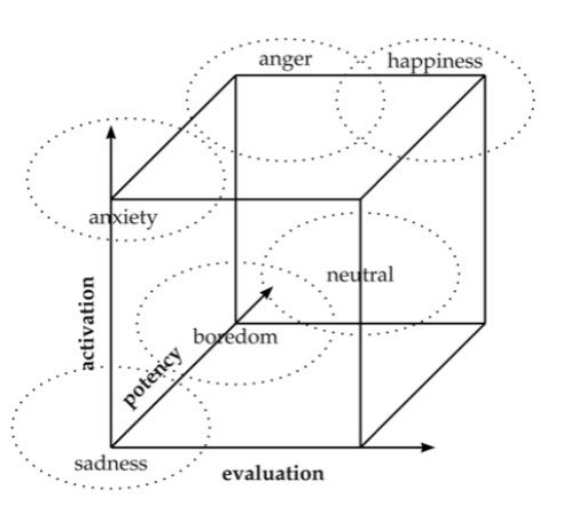
Ljut muškarac = Srećna žena?



Problemi i unapređenja

- Velika neponovljivost i nekonzistentnost labela među ispitacnicima.
- Dodavanje infromacija o izgovorenom tekstu ili facijalnim ekspresijama.
- Pojedinačni modeli za različite polove i starosna doba.
- Pretstavljane emocija u prostoru aktivacija-valentnost-dominantnost.





Hvala na pažnji!

Pitanja?

Naslov slajda

$$f_N(t) = \sum_{n=-N}^{N} c_n e^{in\omega t}$$

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

Naslov slajda

- Zdravo, ovo je neki tekst
  - On može imati i uvučene teze
- ... a i redovne teze
- ... šta god ti duša ište