



Univerzitet u Beogradu - Elektrotehnički fakultet  
Katedra za signale i sisteme



# Diplomski rad

## Prepoznavanje govora pomoću kepstralnih koeficijenata

### Kandidat:

Marija Rakonjac, br. indeksa 2020/0222

### Mentor:

prof. dr Željko Đurović

Beograd, jul 2024. godine

# Sadržaj

1	UVOD	3
2	MODELIRANJE I NASTANAK GOVORNOG SIGNALA	5
2.1	Fizički model govornog aparata . . . . .	5
2.2	Propagacija govornog signala . . . . .	7
2.3	Model uniformne tube . . . . .	8
3	ELEMENTI KEPSTRALNE ANALIZE	11
3.1	Definicija kepstra i kompleksnog kepstra . . . . .	11
3.2	Kratkovremenski kepstar . . . . .	13
3.3	Računanje kepstra . . . . .	13
3.3.1	Računanje kepstra pomoću DFT-a . . . . .	13
3.3.2	Analiza z-transformacijom . . . . .	14
3.3.3	Rekurzivno računanje kompleksnog kepstra . . . . .	16
3.4	Kratkovremensko homomorfno filtriranje govora . . . . .	16
3.5	Primena u detekciji pitch periode . . . . .	18
3.6	Primena u prepoznavanju oblika . . . . .	19
3.6.1	Kompenzacija linearног filtriranja . . . . .	19
3.6.2	Mera udaljenosti lifterovanog kepstra . . . . .	20
3.6.3	Kepstralni koeficijenti . . . . .	21
3.7	Uloga kepstra . . . . .	23
4	METOD K NAJBLIŽIH SUSEDА	24
4.1	Osnovni principi kNN metode . . . . .	24
4.2	Metrike distance . . . . .	25
4.3	Odabir parametra k . . . . .	25
4.4	Prednosti i ograničenja . . . . .	27
5	OPIS SISTEMA ZA PREPOZNAVANJE GOVORA	28
5.1	Formiranje baze podataka . . . . .	28
5.2	Vizuelizacija i predprocesiranje podataka . . . . .	29
5.2.1	Spektrogram i amplitudski spektar . . . . .	29
5.2.2	Kratkovremenska energija . . . . .	29
5.2.3	Brzina prolaska kroz nulu . . . . .	31
5.2.4	Segmentacija reči . . . . .	32
5.2.5	Uklanjanje šuma . . . . .	33
5.2.6	Reč "jedan" . . . . .	33
5.2.7	Reč "kuća" . . . . .	36
5.2.8	Reč "grožđe" . . . . .	39
5.2.9	Reč "sedam" . . . . .	43
5.2.10	Reč "signal" . . . . .	45
5.2.11	Normalizacija . . . . .	48
5.2.12	Pre-emphasis filter . . . . .	49
5.3	Izdvajanje obeležja . . . . .	50
5.4	Klasifikacija oblika . . . . .	50
6	ZAKLJUČAK	54
7	LITERATURA	56

## ZAHVALNICA

Želim da izrazim svoju duboku zahvalnost profesoru Željku Đuroviću za mentorstvo tokom pisanja ovog diplomskog rada. Njegova podrška, saveti i izdvojeno vreme su dramatično doprineli uspehu mog istraživanja. Pored toga, želim da se zahvalim i za preneta znanje i stalan trud tokom celih mojih osnovnih studija.

Takođe, želim da izrazim zahvalnost svim profesorima i saradnicima Fakulteta koji su svojim znanjem, strpljenjem i posvećenošću doprineli mom obrazovanju. Njihov trud i zalaganje nisu prošli neprimećeno i duboko su uticali na moj akademski razvoj.

Posebno se zahvaljujem svima koji su mi pomogli u snimanju baze podataka za ovaj diplomski rad. Njihova pomoć i saradnja bili su od izuzetnog značaja za realizaciju ovog projekta.

Marija Rakonjac

U Beogradu, jul 2024.

## 1 UVOD

Prepoznavanje govora je tehnologija koja omogućava računarima i drugim uređajima da razumeju i interpretiraju ljudski govor. Ova tehnologija koristi napredne algoritme iz oblasti obrade signala, mašinskog učenja i veštačke inteligencije kako bi prepoznala i pretvorila govorni signal u tekst. Računar "sluša" govor i prepoznaće koje su reči izgovorene – vrši se pretvaranje govornog signala u niz reči ili tekst. Dakle, prepoznavanje govora iziskuje znanja iz mnogih naučnih disciplina poput lingvistike, akustike, elektronike, informatike i računarstva, mašinskog učenja itd.

Tehnologija prepoznavanja govora prošla je kroz nekoliko ključnih faza razvoja, počevši od sredine 20. veka pa sve do današnjih sofisticiranih sistema koji koriste veštačku inteligenciju i mašinsko učenje. Prvi značajan korak u ovoj oblasti bio je 1952. godine, kada je *Bell* laboratorija razvila sistem nazvan *Audrey*. *Audrey* je bio sposoban da prepoznaće brojeve od 0 do 9 sa velikom preciznošću, ali je mogao raditi samo sa jednim govornikom. Sledеći važan napredak postignut je 1962. godine, kada je *IBM* predstavio *Shoebox*, sistem koji je mogao prepoznati 16 reči na engleskom jeziku, uključujući brojeve i osnovne komande, što je omogućilo kontrolu računarskih funkcija putem govora. Krajem 1970-ih, skriveni Markovljevi modeli (eng. *Hidden Markov Models-HMM*) postali su popularni jer su koristili statističke modele za predstavljanje i prepoznavanje sekvenci zvukova, što je značajno poboljšalo prepoznavanje varijacija u govoru. Godine 1986. *IBM* je predstavio sistem *Tangora* koji je mogao prepoznati do 20.000 reči i bio je jedan od prvih sistema sposobnih za prepoznavanje kontinuiranog govoru, a ne samo izolovanih reči. 1990. godine *Dragon Systems* je lansirao *Dragon Dictate*, prvi komercijalno dostupan sistem za prepoznavanje govora, koji je omogućavao korisnicima da diktiraju tekst direktno računaru. Deceniju kasnije, 2011. godine, *Apple* je predstavio *Siri*, virtuelnog asistenta za *iPhone*, koji je koristio prepoznavanje govora za interakciju sa korisnicima, inspirisavši razvoj sličnih sistema kod drugih kompanija. *Google* je takođe unapredio ovu tehnologiju integrisanjem prepoznavanja govora u svoje pretraživačke alate – *Google Voice Search*, čime je omogućeno pretraživanje interneta glasovnim komandama. Razvojem savremenih tehnologija, prepoznavanje govora je postiglo značajan napredak, naročito sa razvojem dubokih neuralnih mreža koje omogućavaju modelima da bolje razumeju kompleksne akustičke i jezičke obrasce. Ovo je rezultiralo pojmom virtuelnih asistenata kao što su *Amazon Alexa*, *Google Assistant* i *Microsoft Cortana*. Sa napretkom u računarstvu, posebno sa razvojem grafičke procesorske jedinice (eng. *Graphics Processing Unit-GPU*), obrada velikih količina podataka postala je brža i efikasnija, što je ključno za trening i implementaciju modela za prepoznavanje govora. Tehnike kao što su analiza pomoću kepstralnih koeficijenata (eng. *Mel-Frequency Cepstral Coefficients*), LPC analiza (eng. *Linear Predictive Coding*) i skriveni Markovljevi modeli omogućavaju precizniju analizu i interpretaciju govornog signala.

Sa društvenog aspekta, prepoznavanje govora povećava pristupačnost tehnologiji za osobe sa invaliditetom, omogućavajući lakšu komunikaciju, posebno za slepe i slabovidne osobe, kao i za osobe sa motoričkim poteškoćama. Govorna interakcija je često brža i intuitivnija od kucanja ili korišćenja tradicionalnih interfejsa,

što poboljšava korisničko iskustvo i produktivnost. Prepoznavanje govora našlo je široku primenu u različitim industrijama. U zdravstvu, koristi se za transkripciju medicinskih beleški, omogućavajući lekarima da se fokusiraju na pacijente umesto na administrativne zadatke. U obrazovanju, sve je popularnija automatska transkripcija predavanja. U finansijskom sektoru, prepoznavanje govora se koristi za autentifikaciju korisnika i analizu telefonskih razgovora u svrhu poboljšanja usluga.

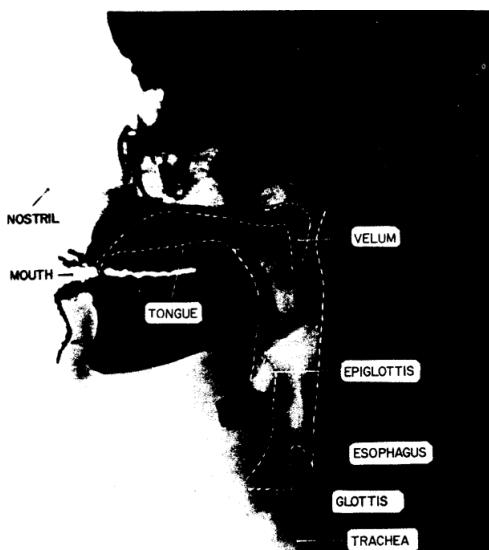
Iako je prepoznavanje govora postiglo značajan napredak, i dalje se suočava sa mnogim izazovima. Razumevanje dijalekata, brzine govora, emocionalnih intonacija, i okruženja sa šumom može značajno uticati na tačnost modela. Prikupljanje balansiranih i reprezentativnih podataka, kao i obezbeđivanje privatnosti i bezbednosti, ostaju ključni faktori za dalji razvoj i poboljšanje ove tehnologije.

Cilj ovog rada je dizajniranje sistema za prepoznavanje govora koji koristi ograničen rečnik za klasifikaciju reči. Glavni zadaci su ispitivanje efiknosti kepstralne analize za ovakav problem, kao i uticaj podele reči na segmente na tačnost klasifikacije. Ovaj rad se sastoji iz nekoliko celina. Za početak, u drugom poglavljju je objašnjeno modeliranje i nastanak govornog signala. Zatim, dat je detaljan pregled kepstralne analize koja je metoda odabrana za analizu govora u ovom radu. U četvrtom poglavljju je opisan algoritam  $k$  najbližih suseda (eng. *k-Nearest Neighbours*) koji se koristi kao klasifikator reči. Peto poglavљje opisuje sistem za prepoznavanje govora koji je implementiran, a kasnije su prikazani i njegovi postignuti rezultati. Na posletku, dat je zaključak sa predlozima za nastavak istraživanja, kao i pregled korišćene literature.

## 2 MODELIRANJE I NASTANAK GOVORNOG SIGNALA

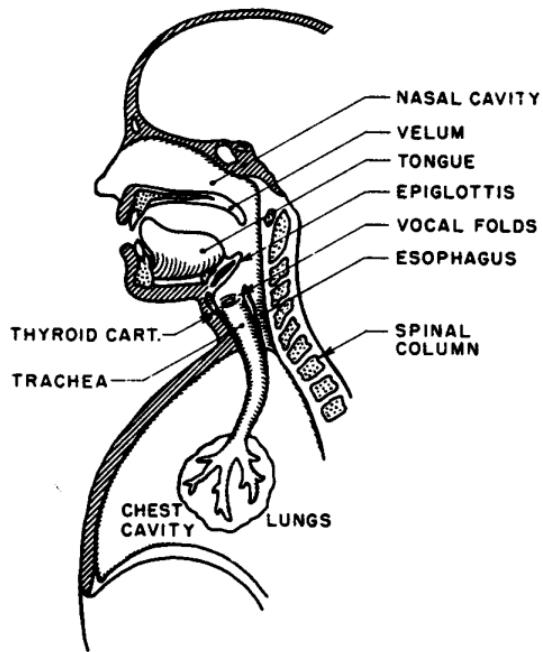
### 2.1 Fizički model govornog aparata

Vokalni trakt čoveka, uokviren isprekidanim linijama na slici 1, prostire se od usana pa sve do glasnih žica odnosno glotisa. Vokalni trakt se sastoji od farinksa (veza između jednjaka i usana) i usne šupljine. Kod prosečnog muškarca, dužina vokalnog trakta iznosi 17cm, dok je kod žena nešto kraći. Poprečni presek vokalnog trakta varira zavisno od položaja jezika, usana, veluma i vilice, ali se kreće između  $20\text{cm}^2$  i  $0\text{cm}^2$ , kada je trakt potpuno zatvoren. Nazalni trakt počinje od veluma (resice) i završava se sa nozdrvama. Kada je velum spušten, nazalni trakt je akustički povezan sa vokalnim i tada se proizvode takozvani nazalni glasovi.



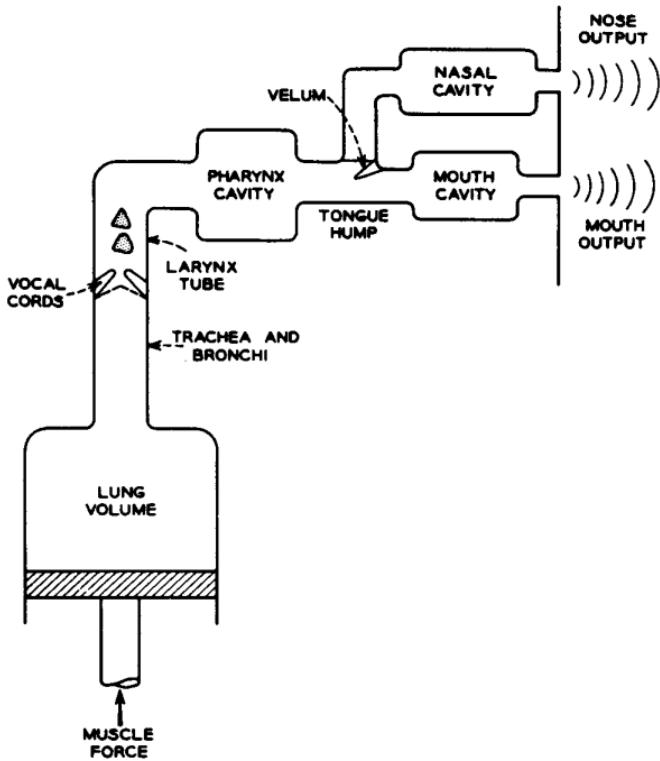
Slika 1: Vokalni trakt čoveka, preuzeto iz rada [1]

Potpuna šema čovekovog vokalnog mehanizma prikazana je na slici 2. Vazduh ulazi u pluća preko disajnih puteva. U trenucima kada vazduh napušta dušnik, zategnute glasne žice u larinksu počinju da vibriraju. Protok vazduha može se tretirati kao niz kvazi-periodičnih impulsa koji zatim prolaze kroz farinks, usnu i potencijalno nosnu duplju. Zavisno od pozicije vilice, jezika, veluma i usana, produkuju se različiti glasovi.



Slika 2: Vokalni mehanizam čoveka, preuzeto iz rada [1]

Pojednostavljena reprezentacija kompletног fizioloшког procesa prikazana je na slici 3. Ekscitaciju vrše pluća i okolni mišići koji predstavljaju izvor vazduha. Mišići zatim silom izbacuju vazduh iz pluća kroz bronhije i dušnik. Zavisno od zategnutosti glasnih žica i postojanja neke prepreke u vokalnom traktu, govorni glasovi se mogu podeliti prema ekscitaciji na tri kategorije. Prva od njih su zvučni glasovi koji nastaju kada su glasne žice potpuno zategnute, pa počinju da vibriraju relaksiranim oscilacijama pod uticajem strujanja vazduha. Ovakve oscilacije proizvode kvazi-četvrtke koje pobuđuju vokalni trakt. Zvučni glasovi su uglavnom samoglasnici. Drugu kategoriju predstavljaju frikativi ili bezvučni glasovi koji nastaju kada se vokalni trakt vrlo suzi a vazdušna struja prolazeći kroz ovako uzan prolaz dobija veliku brzinu i nastaju turbulentne struje predstavljajuće površinsku turbulenciju. Ovako turbulentne vazdušne struje predstavljaju pobudu za vokalni trakt koji je okarakterisan širokim spektralnim sadržajem. Tipičan predstavnik frikativa je glas 'š'. Treću kategoriju glasova predstavljaju plosivi, koji nastaju kada se vokalni trakt potpuno zatvori, te se ispred ove pozicije stvara komora visokog pritiska, a zatim se naglo otvoriti i pritisak naglo opadne. Ovako nastaju glasovi kao što su 'p', 'b' i 'd'.[1]



Slika 3: Fiziološki proces nastanka govora, preuzeto iz rada [1]

## 2.2 Propagacija govornog signala

Zvučni talasi su mehanički talasi koji se šire kroz različite medijume poput vode, čvrstih materijala i u slučaju govora kroz vazduh. Ovi talasi nastaju kao rezultat vibracija izvora zvuka, poput glasnih žica pri nastanku govora, koje uzrokuju oscilacije čestica medijuma. Shodno tome, nastajanje i propagaciju zvuka u vokalnom sistemu mogu opisati zakoni fizike. U cilju formiranja matematičkog modela ovog fenomena neophodno je obratiti pažnju na zakon održanja mase i energije, uz korišćenje zakona termodinamike i mehanike fluida. Matematička analiza i formulacija ovog modela je veoma složena i mora uzeti u obzir sledećih šest fenomena:

1. Vremenski promenljiva priroda vokalnog trakta – površina poprečnog preseka vokalnog trakta je promenljiva i u vremenu i po poziciji.
2. Gubici energije usled topotne kondukcije i viskoznog trenja vazdušnih struja uz zidove vokalnog trakta – molekuli vazduha gube energiju usled topotne kondukcije i viskoznog trenja.
3. Čvrstina zidova vokalnog trakta – krutost zidova vokalnog trakta je konačna.
4. Radijacija zvuka na usnama – proces emitovanja zvučnog talasa iz usana u spoljašnji prostor tokom govora.
5. Povezivanje nosne šupljine sa vokalnim traktom – prilikom izgovora određenih glasova ('m', 'n' i 'nj') nazalni trakt postaje deo vokalnog.

6. Zvučna eksitacija na početku vokalnog trakta – različiti glasovi podrazumevaju različitu pobudu.

Ukoliko se uzmu u obzir zakoni održanja mase i energije, a zanemare energetski gubici zbog topotne konduktanse i viskoznog trenja, dobijamo sledeći set jednačina:

$$\begin{aligned} -\frac{\partial p(x, t)}{\partial x} &= \rho \frac{\partial(u(x, t)/A(x, t))}{\partial t} \\ -\frac{\partial u(x, t)}{\partial x} &= \frac{1}{\rho c^2} \frac{\partial(p(x, t)A(x, t))}{\partial t} + \frac{\partial A(x, t)}{\partial t} \end{aligned}$$

gde je sa  $p = p(x, t)$  označen vazdušni pritisak u vokalnom traktu na poziciji  $x$  merno od početka vokalnog trakta, sa  $u = u(x, t)$  zapreminske vazdušni protok, sa  $\rho$  je označena gustina vazduha u vokalnom traktu, sa  $c$  brzina zvuka, i na kraju sa  $A = A(x, t)$  površina poprečnog preseka trakta na poziciji  $x$  u trenutku  $t$ . Ovaj set jednačina je sastavljen od parcijalnih jednačina čije rešenje u zatvorenoj formi nije moguće naći. Rešenje se može naći pomoću određenih numeričkih metoda, uz poznate potrebne početne uslove. Međutim, uz izvesne aproksimacije i pojednostavljenja, moguće je rešiti ove jednačine čak i u zatvorenoj formi. Jedan od pojednostavljenih modela prikazan je u nastavku.

### 2.3 Model uniformne tube

Kod modela uniformne tube usvaja se pretpostavka da se vokalni trakt ne menja tokom vremena i da je površina poprečnog preseka konstantna i iznosi  $A$ . Po red toga, usvaja se i da je pritisak na usnama konstantan i da promena postoji samo u brzini vazduha koji izlazi iz tube. Ova pojednostavljenja rezultuju takozvanom uniformnom tubom bez gubitaka i tada su parcijalne jednačine:

$$\begin{aligned} -\frac{\partial p(x, t)}{\partial x} &= \frac{\rho}{A} \frac{\partial u(x, t)}{\partial t} \\ -\frac{\partial u(x, t)}{\partial x} &= \frac{A}{\rho c^2} \frac{\partial p(x, t)}{\partial t} \end{aligned}$$

Kako bi se došlo do rešenja, usvaja se da se pritisak i zapreminske vazduha ponašaju kao kompleksne sinusoide učestanosti  $\Omega$  pri čemu amplituda ovih oscilacija zavisi i od pozicije  $x$  u vokalnom traktu. Imajući ovo u vidu, jednačine su sada:

$$\begin{aligned} p(x, t) &= P(x)e^{j\Omega t} \\ u(x, t) &= U(x)e^{j\Omega t} \end{aligned}$$

Zamenom ovih relacija u početne parcijalne jednačine dalje se dobija:

$$\begin{aligned} -\frac{\partial P(x)}{\partial x}e^{j\Omega t} &= \frac{\rho}{A}j\Omega U(x)e^{j\Omega t} \implies -\frac{\partial P(x)}{\partial x}e^{j\Omega t} = ZU \\ -\frac{\partial U(x)}{\partial x}e^{j\Omega t} &= \frac{A}{\rho c^2}j\Omega P(x)e^{j\Omega t} \implies -\frac{\partial U(x)}{\partial x}e^{j\Omega t} = YP \end{aligned}$$

pri čemu  $Z$  predstavlja akustičku impedansu po jedinici dužine i definiše se sa:

$$Z = \frac{\rho}{A} j\Omega$$

dok je  $Y$  akustička admitansa po jedinici dužine i računa se kao:

$$Y = \frac{A}{\rho c^2} j\Omega$$

Rešenje diferencijalnih jednačina se može zapisati u formi:

$$P(x) = C_1 e^{\gamma x} + C_2 e^{-\gamma x}$$

$$U(x) = C_3 e^{\gamma x} + C_4 e^{-\gamma x}$$

gde je:

$$\gamma = \sqrt{ZY} = \frac{j\Omega}{c}$$

Kako bi se odredili nepoznati koeficijenti, potrebno je uvesti granične uslove koji govore o vazdušnom pritisku i zapreminskom vazdušnom protoku na početku vokalnog trakta:

$$U(0) = U_g \implies C_3 + C_4 = U_g$$

$$P(l) = 0 \implies C_1 e^{\gamma l} + C_2 e^{-\gamma l} = 0$$

Rešavanjem jednačina dobija se konačan rezultat:

$$p(x, t) = j \frac{\rho c}{A} \cdot \frac{\sin(\Omega \frac{l-x}{c})}{\cos(\Omega l/c)} U_g e^{j\Omega t}$$

$$u(x, t) = \frac{\cos(\Omega \frac{l-x}{c})}{\cos(\Omega l/c)} U_g e^{j\Omega t}$$

Frekvencijski odziv sistema na pobudu zapreinskog protoka ima oblik:

$$V(\Omega) = \frac{U(l)}{U(0)}$$

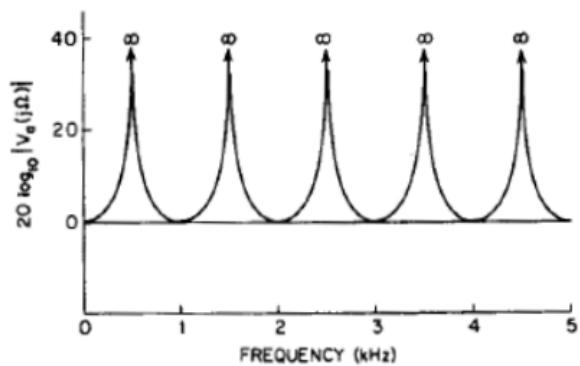
gde je  $U(l)$  zapreinski pritisak na usnama, a  $U(0)$  zapreinski pritisak na glotisu. Koristeći dobijena rešenja diferencijalnih jednačina dobija se:

$$V(\Omega) = \frac{1}{\cos(\Omega l/c)}$$

Potrebno je pronaći tačke u kojima je vrednosti imenioca jednaka nuli:

$$\frac{\Omega l}{c} = (2k - 1)\pi/2 \implies F_k = \frac{c}{4l}(2k - 1)$$

gde je  $F_k = \frac{\Omega}{2\pi}$ . Kako je  $\frac{c}{4l} \sim 500$  Hz može se videti da su prva tri formanta na učestanostima od 500, 1500 i 2500 Hz redom. Na slici 4 prikazan je frekvencijski odziv sistema na pobudu zapreinskog protoka, gde se jasno vidi da amplituda na pojedinim mestima teži beskonačnosti, a širina svakog špica teži 0, što nisu poželjne osobine.[2]



Slika 4: Frekvencijski odziv  $V(\Omega)$ , preuzeto iz rada [2]

### 3 ELEMENTI KEPSTRALNE ANALIZE

Kratkovremenska Furijeova transformacija (eng. *Short-Time Fourier Transform-STFT*) jedan je od najkorisnijih alata za analizu signala i na njemu se bazira veliki broj tehnika za obradu govora. Važan koncept koji direktno proizlazi iz kratkovremenke Furijeove transformacije je kepstar, tačnije kratkovremenski kepstar (eng. *Short-Time Cepstrum*). U ovom poglavljju opisana je upotreba kratkovremenskog kepstra kao jednog od načina za reprezentaciju govora.

#### 3.1 Definicija kepstra i kompleksnog kepstra

Kepstar je prvi put definisan od strane Bogert-a, Healy-a i Tukey-a kao inverzna Furijeova transformacija logaritamskog spektra signala. Njihova originalna definicija bila je motivisana činjenicom da logaritam Furijeovog spektra signala koji poseduje echo ima aditivnu periodičnu komponentu koja zavisi samo od amplitude i frekvencije echa, te da dalja Furijeova analiza logaritamskog spektra može biti odličan alat za detekciju tog echa. Oppenheim, Schafer i Stockham su pokazali da je kepstar povezan sa opštijim konceptom homomorfnog filtriranja signala koji su kombinovani konvolucijom, i oni su definisali kepstar diskretnog signala kao:

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |X(e^{j\omega})| e^{j\omega n} d\omega$$

gde je  $\log |X(e^{j\omega})|$  logaritam intenziteta diskretnе Furijeove transformacije (eng. *Discrete-Time Fourier Transform-DTFT*) signala, i proširili su koncept definisanjem kompleksnog kepstra kao:

$$\hat{x}[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log X(e^{j\omega}) e^{j\omega n} d\omega$$

gde je  $\log X(e^{j\omega})$  kompleksni logaritam  $X(e^{j\omega})$  definisan kao:

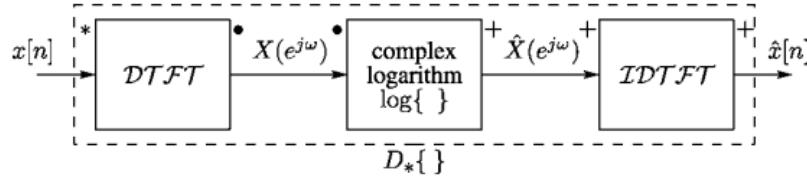
$$\hat{X}(e^{j\omega}) = \log X(e^{j\omega}) = \log |X(e^{j\omega})| + j \arg X(e^{j\omega})$$

Transformacija implicirana izrazom za kompleksni kepstar oslikana je kroz blok dijagram na slici 5. Isti dijagram predstavlja računanje kepstra ukoliko se kompleksni logaritam zameni logaritmom intenziteta DTFT-a. Pošto je od interesa realna sekvencia  $x[n]$ , iz svojstava simetrije Furijeove transformacije sledi da je kepstar parni deo kompleksnog kepstra, tj.

$$c[n] = \frac{\hat{x}[n] + \hat{x}[-n]}{2}$$

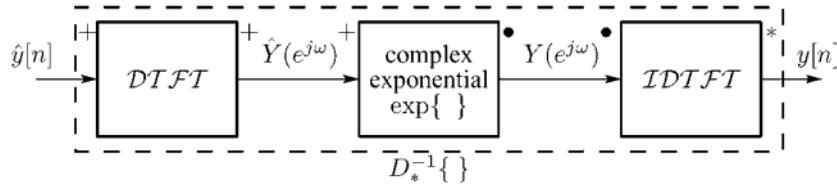
Kao što je prikazano na slici 5, računanje kompleksnog kepstra koji dolazi na ulaz sistema može se predstaviti kao  $\hat{x}[n] = D^*\{x[n]\}$ . U teoriji homomorfnih sistema  $D^*\{\cdot\}$  se naziva karakteristični sistem za homomorfnu dekonvoluciju. Ovaj sistem koristi činjenicu da konvolucija signala u vremenskog domenu rezultuje proizvodom u spektralnom domenu, ali i da se proizvod dva signala može transformisati u zbir primenom logaritamske funkcije. Ako je  $x[n] = x_1[n] * x_2[n]$ , tada sledi:

$$\hat{x}[n] = D^*\{x_1[n] * x_2[n]\} = \hat{x}_1[n] + \hat{x}_2[n]$$



Slika 5: Karakteristični sistem za homomorfnu dekonvoluciju, preuzeto iz rada [3]

Dakle, operacija izračunavanja kompleksnog kepstra transformiše konvoluciju u zbir. Ova odlika važi i za kepstar i za kompleksni kepstar, zbog čega je kepstar veoma koristan za analizu govora, budući da model za generisanje govornog signala uključuje konvoluciju ekscitacije sa impulsnim odzivom vokalnog trakta, a često je cilj razdvojiti ova dva signala. Na slici 6 prikazan je inverzni karakteristični sistem za homomorfnu dekonvoluciju kompleksnog kepstra.



Slika 6: Inverzni karakteristični sistem za homomorfnu dekonvoluciju, preuzeto iz rada [3]

Homomorfno filtriranje konvoluiranih signala se postiže formiranjem modifikovanog kompleksnog kepstra:

$$\hat{y}[n] = g[n]\hat{x}[n]$$

gde je  $g[n]$  prozor (u terminologiji Bogert-a i saradnika, "lifter") koji selektuje deo kompleksnog kepstra za inverznu obradu. Modifikovani izlazni signal  $y[n]$  se može dobiti kao izlaz na slici 6 gde je  $\hat{y}[n]$  ulaz u sistem. U prethodnoj relaciji je definisan linearni operator u konvencionalnom smislu, tj. ako je  $\hat{x}[n] = \hat{x}_1[n] + \hat{x}_2[n]$ , tada je  $\hat{y}[n] = g[n]\hat{x}_1[n] + g[n]\hat{x}_2[n]$ . Stoga će izlaz inverznog karakterističkog sistema imati oblik  $y[n] = y_1[n] * y_2[n]$ , gde je  $\hat{y}_1[n] = g[n]\hat{x}_1[n]$  kompleksni kepstar od  $y_1[n]$ , itd.

Glavni problem u definiciji i računanju kompleksnog kepstra je računanje kompleksnog logaritma, preciznije, računanje faznog ugla  $\arg X(e^{j\omega})$ , koje mora biti izvedeno tako da se očuva aditivna kombinacija faza za dva signala koja ulaze u konvoluciju.

Nezavisna promenljiva kepstra i kompleksnog kepstra je nominalno vreme. Ključna opservacija koja vodi ka konceptu kepstra je da se logaritamski spektar može tretirati kao talasni oblik koji se podvrgava daljoj Furijeovoj analizi. Imajući ovo u vidu, Bogert i saradnici su skovali reč "kepstar" transponujući neka slova u reči "spektar". Na isti način su kreirali i mnoge druge posebne termine, uključujući

”kvefrencija” (eng. *quefrency*) kao naziv za nezavisnu promenljiva kepstra i ”lifterovanje” za operaciju linearнog filtriranja logaritamskog intenziteta spektra. Danas su široko korišćeni samo termini ”kepstar”, ”kvefrencija” i ”lifterovanje”.

### 3.2 Kratkovremenski kepstar

Aplikacija dosadašnjih definicija na govor zahteva da se *DTFT* zameni sa *STFT* tj. kratkovremenskom Furijeovom transformacijom, pa se dolazi do sledeće definicije:

$$c_{\hat{n}}[m] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |X_{\hat{n}}(e^{j\hat{\omega}})| e^{j\hat{\omega}m} d\hat{\omega}$$

gde je  $X_{\hat{n}}(e^{j\hat{\omega}})$  kratkovremenska Furijeova transformacija, a kratkovremenski kompleksni kepstar se definiše zamenom  $X(e^{j\hat{\omega}})$  sa  $X_{\hat{n}}(e^{j\hat{\omega}})$  u definiciji iz prethodne podsekcije. Kratkovremenski kepstar je niz kepstara prozorovanih segmenata konačne dužine govornog talasa. Analogno, ”kepstrogram” bi bio slika dobijena iscrtavanjem amplituda kratkovremenskog kepstra kao funkcije kvefrencije  $m$  i vremena  $\hat{n}$ .

### 3.3 Računanje kepstra

Kepstar i kompleksni kepstar su prethodno definisani pomoću *DTFT*-a. Ovo je korisno prilikom definisanja pojmove, međutim ne može se koristiti u obradi uzorkovanih, realnih signala.

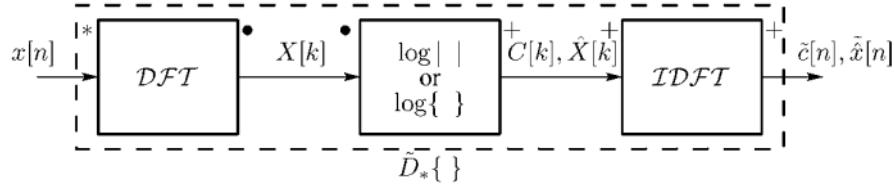
#### 3.3.1 Računanje kepstra pomoću DFT-a

Pošto je *DFT* (izračunat pomoću *FFT* algoritma) uzorkovana verzija (u frekvencijskom domenu) sekvence konačne dužine *DTFT*-a (tj.  $X[k] = X(e^{j2\pi k/N})$ ), *DFT* i inverzni *DFT* mogu biti zamenjeni sa *DTFT*-om i njegovim inverzom u ranijim definicijama, kao što je prikazano na slici 7, koja pokazuje da se kompleksni kepstar može aproksimativno izračunati korišćenjem jednačina:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j(2\pi k/N)n}$$

$$\hat{X}[k] = \log |X[k]| + j \arg\{X[k]\}$$

$$\tilde{x}[n] = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}[k] e^{j(2\pi k/N)n}$$



Slika 7: Računanje cepstra/kompleksnog cepstra pomoću  $DFT$ -a, preuzeto iz rada [3]

Svrha znaka "tilda" iznad  $\hat{x}[n]$  je da naglasi da korišćenje  $DFT$ -a umesto  $DTFT$ -a rezultira aproksimacijom zbog nastanka *aliasing*-a u vremenskom domenu. To jest:

$$\hat{x}[n] = \sum_{r=-\infty}^{\infty} \hat{x}[n + rN]$$

gde je  $\hat{x}[n]$  kompleksni kepstar. Idenična jednačina važi za kepstar  $\tilde{c}[n]$ .

Efekat *aliasing*-a u vremenskom domenu može postati zanemarljiv korišćenjem velike vrednosti  $N$ . Teži problem u izračunavanju kompleksnog cepstra je izračunavanje kompleksnog logaritma. Ugao kompleksnog broja obično se izražava u opsegu od 0 do  $2\pi$ . Da bi se pravilno izračunao kompleksni kepstar, faza uzorkovanog  $DTFT$ -a mora biti kontinualna funkcija frekvencije. Ako se faza prvo izračuna na diskretnim frekvencijama, mora se "razmotati" kako bi se konvolucije pretvorile u sabiranja. Iako je tačno razmotavanje faze izazov u izračunavanju kompleksnog cepstra, to nije problem za običan kepstar, jer faza nije potrebna. Razmotavanje faze se može izbeći korišćenjem numeričkog izračunavanja polova i nula z-transformacije.

### 3.3.2 Analiza z-transformacijom

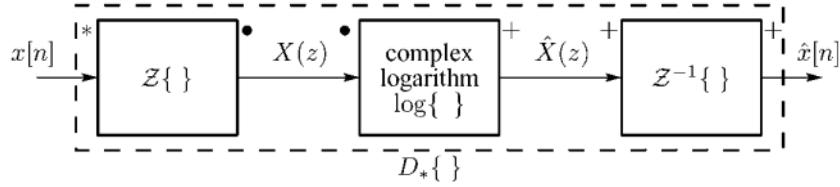
Karakteristični sistem za homomorfnu dekonvoluciju takođe može biti predstavljen dvostranom z-transformacijom, kao što je prikazano na slici 8. Prepostavlja se da ulazni signal  $x[n]$  ima racionalnu z-transformaciju oblika:

$$X(z) = X_{\max}(z)X_{\text{uc}}(z)X_{\min}(z),$$

gde je:

$$\begin{aligned} X_{\max}(z) &= z^{M_0} \prod_{k=1}^{M_0} (1 - a_k z^{-1}) = \prod_{k=1}^{M_0} (-a_k) \prod_{k=1}^{M_0} (1 - a_k^{-1} z) \\ X_{\text{uc}}(z) &= \prod_{k=1}^{M_{\text{uc}}} (1 - e^{j\theta_k} z^{-1}) \\ X_{\min}(z) &= A \frac{\prod_{k=1}^{M_i} (1 - b_k z^{-1})}{\prod_{k=1}^{N_i} (1 - c_k z^{-1})} \end{aligned}$$

Nule  $X_{\max}(z)$ , tj.  $z_k = a_k$ , su nule  $X(z)$  izvan jediničnog kruga ( $|a_k| > 1$ ). Dakle,  $X_{\max}(z)$  je deo  $X(z)$  koji nosi informaciju o maksimalnoj fazi.



Slika 8: Reprezentacija karakterističnog sistema za homomorfnu dekonvoluciju pomoću z-transformacije, preuzeto iz rada [3]

$X_{uc}(z)$  sadrži sve nule (sa uglovima  $\theta_k$ ) na jediničnom krugu.  $X_{min}(z)$  je deo koji nosi informaciju o minimalnoj fazi, gde su  $b_k$  i  $c_k$  nule i polovi, respektivno, unutar jediničnog kruga ( $|b_k| < 1$  i  $|c_k| < 1$ ). Faktor  $z^{M_0}$  implicira pomeranje za  $M_0$  uzorka uлево. Uključen je da bi se rezultati u narednim jednačinama pojednostavili.

Kompleksni kepstar signala  $x[n]$  se određuje pomoću pretpostavke da kompleksni logaritam od  $\log\{X(z)\}$  rezultira zbirom logaritama svakog od činilaca, tj.

$$\begin{aligned} \hat{X}(z) = \log \left( \prod_{k=1}^{M_0} (-a_k) \right) + \sum_{k=1}^{M_0} \log(1 - a_k^{-1} z) + \sum_{k=1}^{M_{uc}} \log(1 - e^{j\theta_k} z^{-1}) \\ + \log |A| + \sum_{k=1}^{M_i} \log(1 - b_k z^{-1}) - \sum_{k=1}^{N_i} \log(1 - c_k z^{-1}) \end{aligned}$$

Primenom razvoja u red:

$$\log(1 - a) = - \sum_{n=1}^{\infty} \frac{a^n}{n}, \quad |a| < 1$$

na svaki od članova u gore definisanoj jednačini dolazi se do:

$$\hat{x}[n] = \begin{cases} \sum_{k=1}^{M_0} \frac{a_k^n}{n}, & n < 0 \\ \log |A| + \log \left( \prod_{k=1}^{M_0} (-a_k) \right), & n = 0 \\ - \sum_{k=1}^{M_{uc}} \frac{e^{j\theta_k n}}{n} - \sum_{k=1}^{M_i} \frac{b_k^n}{n} + \sum_{k=1}^{N_i} \frac{c_k^n}{n}, & n > 0 \end{cases}$$

Poznavanjem svih polova i nula z-transformacije  $X(z)$ , prethodni izraz omogućava izračunavanje kompleksnog kepstra bez ikakve aproksimacije. Ovo važi u teorijskoj analizi gde su polovi i nule precizno određeni. Poslednji izraz je koristan kao osnova za proračun. Sve što je potrebno je na neki način dobiti z-transformaciju u obliku racionalne funkcije i pronaći nule brojioca i imenoca. Ovo je postalo izvodljivo sa povećanjem računske snage i sa novim naprecima u pronalaženju korena velikih polinoma. Jedan metod za dobijanje z-transformacije je jednostavno odabratи sekvencu konačne dužine odbiraka signala. Z-transformacija je tada jednostavno polinom sa odbircima  $x[n]$  kao koeficijentima, tj.

$$X(z) = \sum_{n=0}^M x[n] z^{-n} = A \prod_{k=1}^{M_0} (1 - a_k z^{-1}) \prod_{k=1}^{M_i} (1 - b_k z^{-1})$$

### 3.3.3 Rekurzivno računanje kompleksnog kepstra

Postoji još jedan pristup računanju kompleksnog kepstra koji je moguće primeniti samo za signale minimalne faze, tj. signale čija z-transformacija ima polove i nule unutar jediničnog kruga. Primer bi bio impulsni odziv modela vokalnog trakta čija funkcija prenosa sadrži samo polove:

$$H(z) = \frac{G}{1 - \sum_{k=1}^p \alpha_k z^{-k}} = G \prod_{k=1}^p (1 - c_k z^{-1})$$

U ovom slučaju, svi polovi  $c_k$  moraju biti unutar jediničnog kruga radi stabilnosti sistema. Iz definicije iz prethodnog poglavlja sledi da je kompleksni kepstar impulsnog odziva  $h[n]$  koji odgovara  $H(z)$ :

$$\hat{h}[n] = \begin{cases} 0 & n < 0 \\ \log |G| & n = 0 \\ \sum_{k=1}^p \frac{c_k^n}{n} & n > 0 \end{cases}$$

Može se pokazati da su impulsni odziv i njegov kompleksni kepstar povezani rekurzivnom formulom:

$$\hat{h}[n] = \begin{cases} 0 & n < 0 \\ \log G & n = 0 \\ \frac{h[n]}{h[0]} - \sum_{k=1}^{n-1} \left( \frac{k}{n} \right) \hat{h}[k] \frac{h[n-k]}{h[0]} & n \geq 1 \end{cases}$$

Može se pokazati da postoji direktna rekurzivna veza između koeficijenata polinoma imenioca u prethodno pomenutom modelu vokalnog trakta sa polovima u funkciji prenosa i kompleksnog kepstra impulsnog odziva modela filtera, tj.

$$\hat{h}[n] = \begin{cases} 0 & n < 0 \\ \log G & n = 0 \\ \alpha_n + \sum_{k=1}^{n-1} \left( \frac{k}{n} \right) \hat{h}[k] \alpha_{n-k} & n > 0 \end{cases}$$

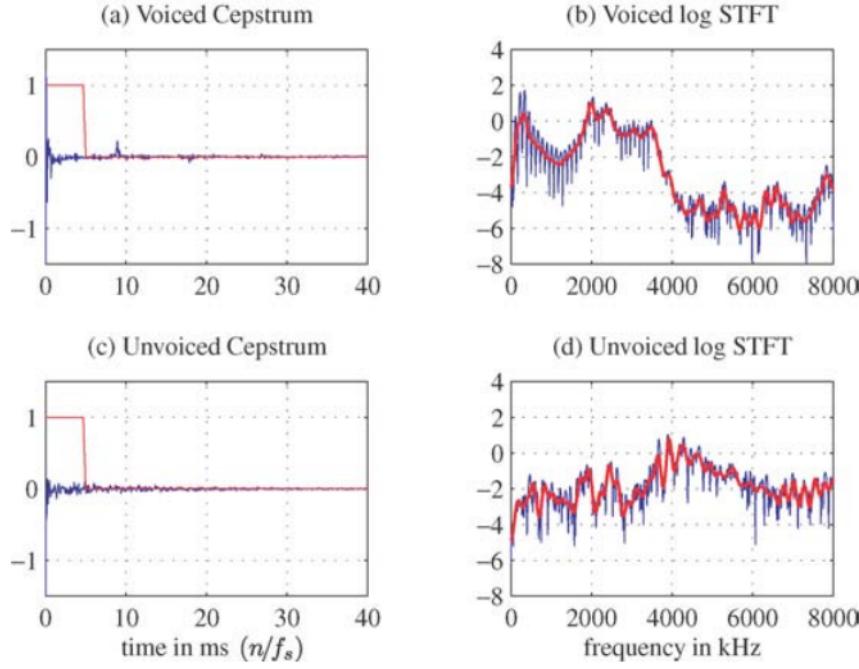
Sledi da se koeficijenti polinoma imenioca mogu dobiti iz kompleksnog kepstra koristeći relaciju:

$$\alpha_n = \hat{h}[n] - \sum_{k=1}^{n-1} \left( \frac{k}{n} \right) \hat{h}[k] \alpha_{n-k} \quad 1 \leq n \leq p$$

Iz ovoga sledi da je  $p+1$  vrednosti kompleksnog kepstra dovoljno da se u potpunosti odredi sistem modela govora definisan na početku jer se svi koeficijenti imenioca i  $G$  mogu izračunati iz  $\hat{h}[n]$  za  $n = 0, 1, \dots, p$  koristeći poslednju jednačinu. Ova činjenica je osnova za korišćenje kepstra u kodiranju govora i prepoznavanju govora kao vektorske reprezentacije svojstava vokalnog trakta.

## 3.4 Kratkovremensko homomorfno filtriranje govora

Slika 9 prikazuje primer kratkovremenskog kepstra govornog signala za segmente zvučnog i bezvučnog govora.



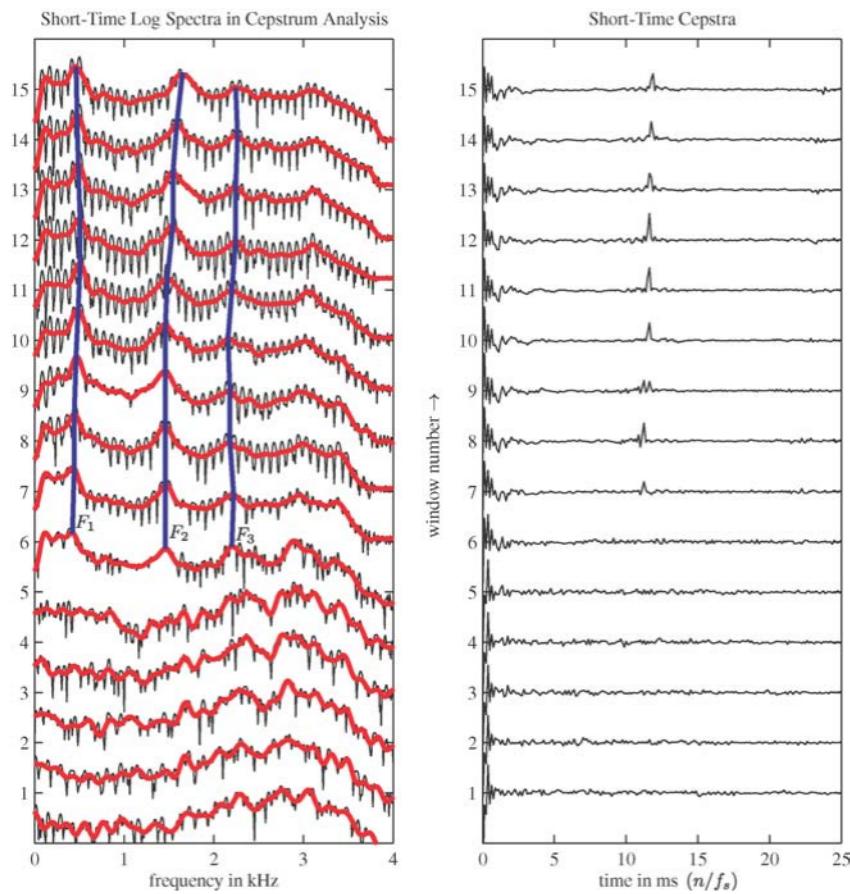
Slika 9: Kratkovremenski kepstar i odgovarajući  $STFT$  i homomorfno-usrednjeni spektar, preuzeto iz rada [3]

Očekuje se da će deo kepstra na nižim kvefrencijama imati spore varijacije u log spektru, dok će komponente na visokim kvefrencijama odgovarati bržim fluktuacijama log spektra. Logaritamske amplitude odgovarajućih kratkovremenskih spektara prikazane su desno kao slike 9(b) i 9(d). Primećuje se da spektar za segment zvučnog govora na slici 9(b) ima strukturu periodičnih talasa zbog harmonijske strukture kvazi-periodičnog zvučnog govora. Ova periodična struktura u log spektru na slici 9(b) manifestuje se u kepstralnom *peak-u* na kvefrenciji od oko 9 ms na slici 9(a). Postojanje ovog vrha u kvefrencijskom opsegu koji je očekivan za *pitch* periodu ukazuje da ovo zaista jeste zvučni govor. Takođe, kvefrencija *peak-a* je tačna procena *pitch* periode za odgovarajući govorni signal. S druge strane, čini se da brže varijacije bezzvučnih spektara izgledaju nasumično bez periodične strukture. Ovo je tipično za Furijeove transformacije kratkih segmenata nasumičnih signala. Kao rezultat, nema jasnog *peak-a* koji ukazuje na periodičnost kao u slučaju zvučnog govora.

Da bi se ilustrovao efekat lifteringa, kvefrencije iznad 5 ms su pomnožene sa 0, a kvefrencije ispod 5 ms su pomnožene sa 1.  $DFT$  rezultujućeg modifikovanog kepstra je prikazan kao glatka kriva koja je superponirana na kratkovremenskim spektrima na slikama 9(b) i 9(d), respektivno. Ovi polako varirajući log spektri jasno zadržavaju opšti spektarski oblik sa vrhovima koji odgovaraju formantnoj rezonantnoj strukturi za analizirani segment govora. Dakle, lifteringom kepstra moguće je izdvojiti informacije o doprinosima vokalnog trakta iz kratkovremenskog govornog spektra.

### 3.5 Primena u detekciji pitch periode

Kepstar je prvi put primjenjen u obradi govora za određivanje parametara po bude modela govora u diskretnom vremenu. Noll je primenio kratkovremenski kepstar za detekciju lokalne periodičnosti (zvučni govor) ili njenog odsustva (bezvručni govor). Ovo je ilustrovano na slici 10. Levo je sekvenca logaritamskih kratkovremenskih spektara (brzo varirajuće krive), a desno je odgovarajuća sekvenca kepstra izračunatih iz logaritamskih spektara sa leve strane. Uzastopni spektri i kepstri su za segmente od 50 ms dobijene pomeranjem prozora u koracima od 12.5 ms. Na osnovu diskusije iz odeljka 3.4, jasno je da za pozicije od 1 do 5 prozor uključuje samo bezvručni govor, dok je za pozicije 6 i 7 signal unutar prozora delimično zvučan i delimično bezvručan. Za pozicije od 8 do 15 prozor uključuje samo zvučni govor. Očigledno je da brze varijacije bezvručnih spektara izgledaju nasumično bez periodične strukture. S druge strane, spektri za zvučne segmente imaju strukturu periodičnih talasa zbog harmonijske strukture kvazi-periodičnog segmenta zvučnog govora. Kao što se može videti sa grafika desno, kepstralni *peak* na kvefrenciji od oko 11–12 ms ukazuje na zvučni govor, a kvefrencija *peak-a* je tačna procena osnovne periode tokom odgovarajućeg govornog intervala.



Slika 10: Kratkovremenski kepstar i odgovarajući *STFT* i homomorfno-usrednjeni spektar, preuzeto iz rada [3]

Suština algoritma za detekciju *pitch* periode koji je predložio Noll je da izračuna sekvencu kratkovremenskih kepstara i za svaki uzastopni kepstar pronađe

*peak* u očekivanom kvefrencijskom opsegu za *pitch* periodu. Prisustvo izraženog vrha implicira zvučni govor, a lokacija kvefrencijskog *peak-a* je procena za *pitch* periodu. Kao i u većini aplikacija za obradu signala zasnovanih na modelima kao što je kepstar, algoritam za detekciju tona uključuje mnoge karakteristike dizajnirane da reše slučajeve koji se ne uklapaju savršeno u osnovni model. Na primer, za segmente 6 i 7 kepstralni vrh je slab, što odgovara prelazu iz bezvučnog u zvučni govor. U drugim problematičnim slučajevima, *peak* na kvefrenciji duploj u odnosu na *pitch* može biti izraženiji od vrha na kvefrenciji *pitch* periode. Noll je primenio vremenska kontinuitetna ograničenja kako bi sprečio ovakve greške.

### 3.6 Primena u prepoznavanju oblika

Verovatno najrasprostranjenija primena kepstra u obradi govora je njegova upotreba u sistemima za prepoznavanje oblika, kao što su dizajn vektorskih kvantizatora (eng. *Vector Quantizers-VQ*) i automatski prepoznavači govora (eng. *Automatic Speech Recognizers-ASR*). U ovakvim aplikacijama, govor se predstavlja kao frejm-po-frejm sekvenca kratkovremenskih kepstara. Kao što je pokazano, kepstri se mogu izračunati ili pomoću analize z-transformacijom ili pomoću *DFT* implementacije karakterističnog sistema. U svakom slučaju, može se pretpostaviti da vektor kepstra odgovara normalizovanom po pojačanju ( $c[0] = 0$ ) impulsnom odzivu minimalno-faznog vokalnog trakta koji je definisan kompleksnim kepstrom:

$$\hat{h}[n] = \begin{cases} 2c[n] & 1 \leq n \leq n_{co} \\ 0 & n < 0 \end{cases}$$

U problemima kao što su *VQ* ili *ASR*, test obrazac  $c[n]$  (vektor kepstralnih vrednosti za  $n = 1, 2, \dots, n_{co}$ ) se upoređuje sa slično definisanim referentnim obrascem  $\bar{c}[n]$ . Takva poređenja zahtevaju meru udaljenosti. Na primer, Euklidova udaljenost primenjena na kepstar bi dala:

$$D = \sum_{n=1}^{n_{co}} |c[n] - \bar{c}[n]|^2$$

Ekvivalentno u frekvencijskom domenu:

$$D = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\log |H(e^{j\omega})| - \log |\bar{H}(e^{j\omega})||^2 d\omega$$

gde je  $\log |H(e^{j\omega})|$  logaritam amplitude *DFT* od  $h[n]$  koji odgovara kompleksnom kepstru u definiciji za  $\hat{h}[n]$  ili realni deo *DTFT* od  $\hat{h}[n]$ . Dakle, poređenja zasnovana na kepstru su vrlo povezana sa poređenjima usrednjjenih kratkovremenskih spektara. Stoga, kepstar nudi efikasnu i fleksibilnu reprezentaciju govora za probleme prepoznavanja oblika, a njegovo tumačenje kao razlike logaritamskih spektara sugerise usklađenost sa mehanizmima auditorne percepcije.

#### 3.6.1 Kompenzacija linearnog filtriranja

Prepostavlja se da imamo samo linearno filtriranu verziju govornog signala,  $y[n] = h[n] * x[n]$ , umesto  $x[n]$ . Ako je prozor od interesa za analizu dugačak u

poređenju sa dužinom  $h[n]$ , kratkovremenski kepstar jednog frejma filtriranog govornog signala  $y[n]$  će biti:

$$\hat{c}_n^{(y)}[m] = \hat{c}_n^{(x)}[m] + \hat{c}^{(h)}[m]$$

gde će  $\hat{c}^{(h)}[m]$  izgledati isto u svakom frejmu. Stoga, ako se može proceniti  $\hat{c}^{(h)}[n]$  za koji se može prepostaviti da se ne menja tokom vremena, oduzimanjem se dobija  $\hat{c}_n^{(x)}[m]$  za svaki frejm iz  $\hat{c}_n^{(y)}[m]$ , tj.  $\hat{c}_n^{(x)}[m] = \hat{c}_n^{(y)}[m] - \hat{c}^{(h)}[m]$ . Ova osobina je izuzetno važna u situacijama kada je referentni obrazac  $\bar{c}[m]$  dobijen pod različitim uslovima snimanja ili prenosa od onih koji se koriste za prikupljanje test vektora. U ovim okolnostima, test vektor se može kompenzovati za efekte linearног filtriranja pre nego što se izračunaju mere udaljenosti koriшћene za poređenje obrazaca.

Još jedan pristup uklanjanju efekata linearnih distorzija je da se uoči da je komponenta kepstra zbog distorzije ista u svakom okviru. Stoga, može se ukloniti jednostavnom operacijom razlike u obliku:

$$\Delta\hat{c}_n^{(y)}[m] = \hat{c}_n^{(y)}[m] - \hat{c}_{n-1}^{(y)}[m]$$

Jasno je da ako je  $\hat{c}_n^{(y)}[m] = \hat{c}_n^{(x)}[m] + \hat{c}^{(h)}[m]$  sa  $\hat{c}^{(h)}[m]$  koji je nezavistan od  $n$ , tada  $\Delta\hat{c}_n^{(y)}[m] = \Delta\hat{c}_n^{(x)}[m]$ , tj. efekti linearne distorzije su uklonjeni.

### 3.6.2 Mera udaljenosti lifterovanog kepстра

U primeni linearne prediktivne analize za dobijanje kepstralnih karakterističnih vektora za probleme prepoznavanja oblika, primećuje se značajna statistička varijabilnost zbog različitih faktora, uključujući poziciju prozora kratkovremenske analize, pristrasnost ka harmonijskim vrhovima i aditivni šum. Rešenje za ovaj problem je korišćenje ponderisanih mera udaljenosti u obliku:

$$D = \sum_{n=1}^{n_{co}} g^2[n] |c[n] - \bar{c}[n]|^2$$

što se može napisati kao Euklidska udaljenost lifterovanih kepstara:

$$D = \sum_{n=1}^{n_{co}} |g[n]c[n] - g[n]\bar{c}[n]|^2$$

Tohkura je otkrio da kada se uproseće kroz mnoge frejmove govora vrednosti kepstra  $c[n]$  imaju srednje vrednosti nula i varijanse reda veličine  $\frac{1}{n^2}$ . Ovo sugerise da bi se mogao koristiti  $g[n] = n$  za  $n = 1, 2, \dots, n_{co}$  kako bi se izjednačili doprinosi svakog člana u razlici kepstra.

Juang je primetio da bi se varijabilnost usled nestalnosti LPC analize mogla smanjiti korišćenjem liftera oblika:

$$g[n] = 1 + 0.5n_{co} \sin\left(\frac{\pi n}{n_{co}}\right), \quad n = 1, 2, \dots, n_{co}$$

Testovi ponderisanih mera udaljenosti pokazali su dosledna poboljšanja u zadacima automatskog prepoznavanja govora.

Itakura i Umezaki koristili su funkciju grupnog kašnjenja da izvedu drugačiju težinsku funkciju kepstra. Umesto  $g[n] = n$  za sve  $n$ , ili iznad definisanog liftera, Itakura je predložio lifter:

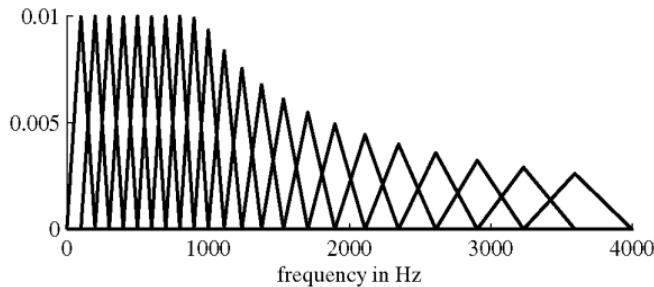
$$g[n] = n^s e^{-n^2/2\tau^2}.$$

Ovaj lifter ima veliku fleksibilnost. Na primer, ako je  $s = 0$ , ovo postaje jednostavno niskopropusno lifterovanje kepstra. Itakura i Umezaki su testirali meru udaljenosti spektra grupnog kašnjenja u sistemu za automatsko prepoznavanje govora. Otkrili su da je za čiste test uzorke razlika u stopi prepoznavanja bila mala za različite vrednosti  $s$  kada je  $\tau \approx 5$ , iako su performanse opadale sa povećanjem  $s$  za veće vrednosti  $\tau$ . Ovo je pripisano činjenici da za veće vrednosti  $s$  spektar grupnog kašnjenja počinje da ima veoma izošrene vrhove i time je osetljiviji na male razlike u lokacija formanta. Međutim, u test uslovima sa aditivnim belim šumom i sa distorzijama linearног filtriranja, stope prepoznavanja su se značajno poboljšale sa  $\tau = 5$  i povećanjem vrednosti parametra  $s$ .

### 3.6.3 Kepstralni koeficijenti

Kao što je rečeno, ponderisane mere udaljenosti kepstra imaju direktnu ekvivalentnu interpretaciju u frekvencijskom domenu. Ovo je značajno za modele za ljudsku percepciju zvuka, koji se zasnivaju na frekvencijskoj analizi koja se dešava u unutrašnjem uhu. Imajući ovo u vidu, Davis i Mermelstein formulisali su novi tip kepstralne reprezentacije koja je postala široko korišćena i poznata kao kepstralni koeficijenti (eng. *Mel-Frequency Cepstral Coefficients-MFCC*).

Osnovna ideja je da se izvrši frekvencijska analiza zasnovana na skupu filtera sa približno kritičnim opsegom filtera i širinama opsega. Za opseg širine 4 kHz, koristi se otprilike 20 filtera. U većini implementacija, prvo se primeni kratkovremenska Furijeova analiza, što rezultira  $DFT$ -om  $X_{\hat{n}}[k]$  za vreme analize  $\hat{n}$ . Zatim se vrednosti  $DFT$ -a grupišu zajedno u kritične opsege i ponderišu trougaonom težinskom funkcijom kao što je prikazano na slici 11.



Slika 11: Težinska funkcija za skup filtera kepstralnih koeficijenata, preuzeto iz rada [3]

Širine opsega na slici 11 su konstantne za središnje frekvencije ispod 1 kHz, a zatim se eksponencijalno povećavaju do polovine frekvencije uzorkovanja od 4 kHz, što rezultira ukupno 22 "filtera". Mel-frekvencijski spektar u vremenu analize  $\hat{n}$  definisan je za  $r = 1, 2, \dots, R$  kao:

$$MF_{\hat{n}}[r] = \frac{1}{A_r} \sum_{k=L_r}^{U_r} |V_r[k]X_{\hat{n}}[k]|^2$$

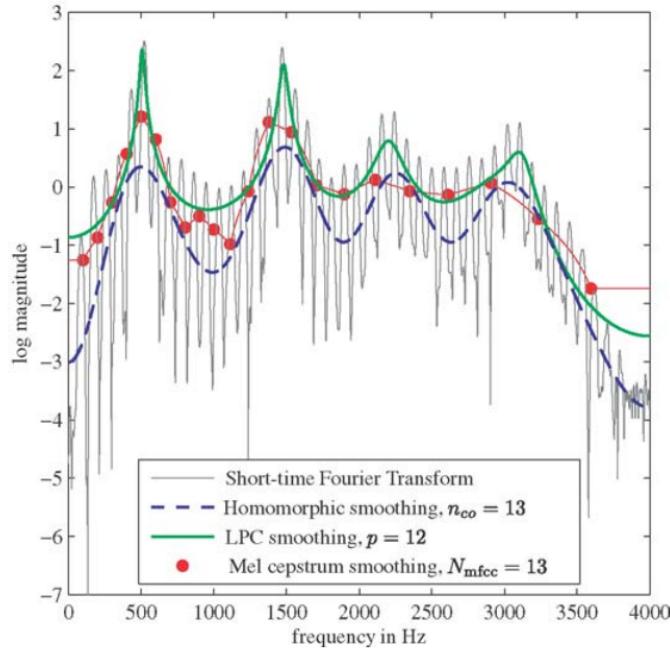
gde je  $V_r[k]$  trougaona težinska funkcija za  $r$ -ti filter koja se proteže od  $DFT$  indeksa  $L_r$  do  $U_r$ , gde:

$$A_r = \sum_{k=L_r}^{U_r} |V_r[k]|^2$$

predstavlja normalizacioni faktor za  $r$ -ti mel-filter. Ova normalizacija je ugrađena u težinske funkcije sa slike 11, a ona je potrebna kako bi savršeno ravan ulazni Furijeov spektar proizveo ravan mel-spektar. Za svaki okvir, diskretna kosinusna transformacija logaritma amplitude izlaza filtera se računa da bi se formirala funkcija  $mfcc_{\hat{n}}[m]$ :

$$mfcc_{\hat{n}}[m] = \frac{1}{R} \sum_{r=1}^R \log(MF_{\hat{n}}[r]) \cos \left( \frac{2\pi}{R} \left( r + \frac{1}{2} \right) m \right)$$

Tipično,  $mfcc_{\hat{n}}[m]$  se procenjuje za broj koeficijenata  $N_{mfcc}$ , koji je manji od broja mel-filtera, npr.  $N_{mfcc} = 13$  i  $R = 22$ . Slika 12 prikazuje rezultat  $MFCC$  analize govora u poređenju sa kratkovremenskim Furijeovim spektrom,  $LPC$  spektrom i homomorfno usrednjjenim spektrom. Velike tačke predstavljaju vrednosti  $\log(MF_{\hat{n}}[r])$  a linija interpolirana između njih predstavlja spektar rekonstruisan na originalnim  $DFT$  frekvencijama. Svi ovi spektri su različiti, ali zajedničko im je da imaju *peak-ove* na rezonancijama formanata. Na višim frekvencijama, rekonstruisani mel-spektar naravno ima više izravnavanja zbog strukture skupa filtera.



Slika 12: Poređenje metoda za usrednjavanje spektra, preuzeto iz rada [3]

### 3.7 Uloga kepstra

Kao što je diskutovano u ovom poglavlju, kepstar je ušao u oblast obrade govora kao osnova za detekciju *pitch* frekvencije. Ostaje jedan od najefikasnijih pokazatelja *pitch* frekvencije glasa koji su ikada osmišljeni. Kako su vokalni trakt i komponente ekscitacije dobro razdvojeni u kepstru, bilo je prirodno takođe razmotriti tehnike za procenu sistema vokalnog trakta. Dok tehnike separacije zasnovane na kepstru mogu biti veoma efikasne, metode linearne prediktivne analize su se pokazale kao efikasnije iz više razloga.[3]

## 4 METOD K NAJBLIŽIH SUSEDА

$K$  najbližih suseda (eng. *k Nearest Neighbours-kNN*) je neparametarska metoda koja se koristi u problemima klasifikacije oblika i regresije. Evelyn Fix i Joseph Hodges su razvili ovaj algoritam 1951. godine, a kasnije ga je proširio Thomas Cover.

### 4.1 Osnovni principi kNN metode

Metod  $k$  najbližih suseda je metod procene funkcije gustine verovatnoće koji se zasniva na prilagođavanju zapreminе prostora oko tačke  $X$  tako da obuhvati tačno  $k$  odbiraka iz serije podataka. Umesto da se fiksira zapremina  $v$  i dozvoli da broj uzoraka  $k$  varira, kod  $kNN$  pristupa fiksira se broj uzoraka  $k$  i prilagođava zapremina  $v$  tako da obuhvati tačno  $k$  najbližih suseda tačke  $X$ . Na taj način, zapremina  $v(X)$  i oblast  $L(X)$  postaju slučajne promenljive koje zavise od pozicije tačke  $X$ . Procena funkcije gustine verovatnoće metodom  $kNN$  se može izraziti kao:

$$\hat{f}(X) = \frac{k^{-1}}{Nv(X)}$$

Da bi procena funkcije gustine verovatnoće bila nepomerena i konzistentna, parametar  $k$  treba da zadovolji sledeće uslove:

$$\lim_{N \rightarrow \infty} k = \infty$$

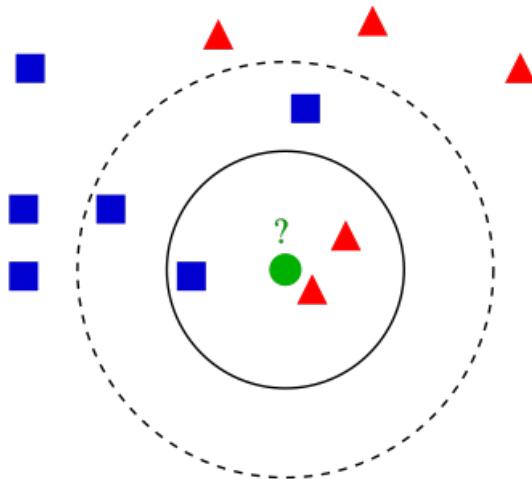
$$\lim_{N \rightarrow \infty} \frac{k}{N} = 0$$

Poštujući ove uslove, parametar  $k$  se može izabrati za zadati broj raspoloživih uzoraka  $N$  na različite načine, a uobičajeno se uzima:

$$k = \lfloor N^\alpha \rfloor, \quad \alpha \in (0, 1)$$

gde je sa  $\lfloor \cdot \rfloor$  označena funkcija celog dela.[4]

Na slici 13 oslikan je rad ovog algoritma na jednom primeru.



Slika 13: Prikaz rada  $kNN$  algoritma, preuzeto sa [5]

## 4.2 Metrike distance

Kao što je rečeno, algoritam  $kNN$  identificuje najbliže tačke za neku tačku upita. Kako bi se odredilo koje su to najbliže tačke, koriste se neke od sledećih metrika za distancu:

- **Euklidska distanca** je ništa drugo nego kartezijska udaljenost između dve tačke koje se nalaze u ravni/hiper ravni. Euklidska distanca se takođe može vizualizovati kao dužina prave linije koja spaja dve tačke od interesa, i definisana je sledećom formulom:

$$d(x, X_i) = \sqrt{\sum_{j=1}^d (x_j - X_{ij})^2}$$

- **Menhetn distanca** meri dužinu pređenog puta kada se tačka kreće samo pravolinijskim putanjama koje su paralelne sa osama koordinatnog sistema (kao po ulicama grada koje formiraju mrežu pravolinijskih blokova, kao u Njujorku – tačnije Menhetnu). Ova metrika se računa sumiranjem apsolutnih razlika između koordinata tačaka u n-dimenzija:

$$d(x, y) = \sum_{i=1}^n |x_i - y_i|$$

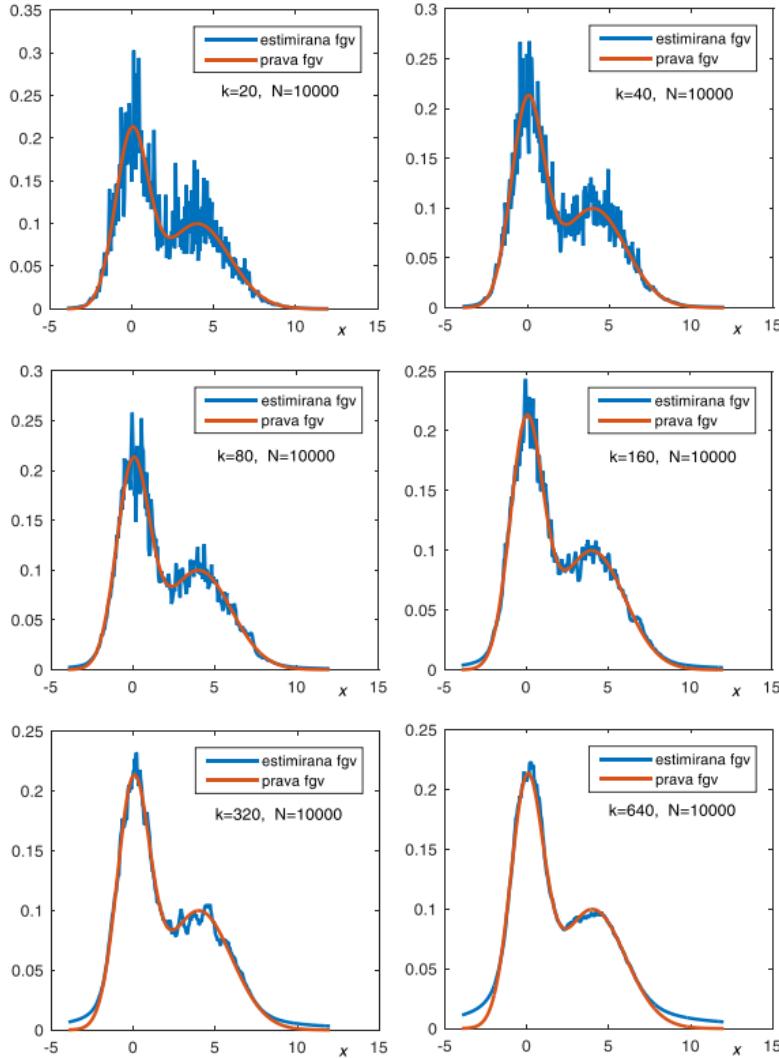
- **Minkovski distanca** je uopštena forma i Euklidske i Menhetn distance. Ona je definisana izrazom:

$$d(x, y) = \left( \sum_{i=1}^n (x_i - y_i)^p \right)^{\frac{1}{p}}$$

Ukoliko je  $p = 2$ , formula postaje identična kao ona za Euklidsku distancu, a ukoliko je  $p = 1$ , tada se dobija formula za Manhattan distancu.[5]

## 4.3 Odabir parametra $k$

Jasno je da odabir parametra  $k$  u velikoj meri utiče na kvalitet procene funkcije gustine verovatnoće. Na slici 14 prikazan je primer bimodalne Gausovske raspodele i procene funkcije gustine verovatnoće zavisno od izbora parametra  $k$ .

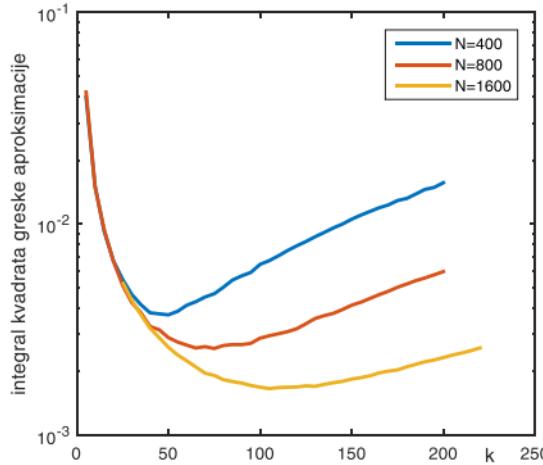


Slika 14: Procena funkcije gustine verovatnoće metodom  $kNN$  za različite vrednosti parametra  $k$ , preuzeto iz rada [4]

Na prva dva grafika gde je parametar  $k$  nedovoljno veliki primetno je da procene funkcije gustine verovatnoće nisu najbolje, pogotovo za visoke vrednosti prave funkcije gustine verovatnoće. U ovim oblastima su uočljivi veliki pikovi zato što se oni nalaze u tačkama u kojima su locirani raspoloživi odbirci. Znajući ovo loša procena je onda očekivana, jer kada se procenjuje funkcija gustine verovatnoće u tački u kojoj se već nalazi jedan od raspoloživih odbiraka, nema potrebe više tražiti  $k$  najbližih suseda već  $k - 1$ , a to znači da će okruženje  $v(X)$  biti manje, pa će i procena postati neopravданo velika. Ovaj efekat je slabo izražen u delovima gde je funkcija gustine verovatnoće niska, jer je broj skoncentrisanih odbiraka manji, a samim tim i procena opravdanija.

Ukoliko se vrednost parametra  $k$  poveća, procena funkcije gustine verovatnoće postaje kvalitetnija i 'mirnija'. Međutim, u oblastima niskih vrednosti procena će biti pomerena, jer u tim slučajevima oblast  $v(X)$  se širi i više se ne može usvojiti pretpostavka da je oblast dovoljno mala, te funkcija konstantna.

Dakle, odabir parametra  $k$  veoma utiče na performanse algoritma. Zato se postavlja pitanje, šta predstavlja dobar izbor parametra  $k$  za zadat broj odbiraka  $N$ . Na slici 15 je prikazana zavisnost integrala kvadrata greške aproksimacije od parametra  $k$ , za različit broj odbiraka  $N$ .



Slika 15: Zavisnost integrala kvadrata greške aproksimacije od parametra  $k$  broja raspoloživih odbiraka  $N$ , preuzeto iz rada [4]

Sa dijagrama se može zaključiti da je za mali izbor parametra  $k$  nebitno koliki je broj raspoloživih odbiraka, kvalitet aproksimacije funkcije gustine verovatnoće se neće značajno menjati. S druge strane, primetno je da sa povećanjem broja odbiraka  $N$  raste i parametar  $k$ , što je i logično. Takođe, optimalna vrednost kriterijumske funkcije je sve manja sa povećanjem broja odbiraka. Postavlja se pitanje da li postoji funkcionalna veza između optimalnog parametra  $k_{\text{opt}}$  i broja odbiraka  $N$ . Ona zaista postoji i glasi:

$$k_{\text{opt}} = CN^\alpha$$

gde vrednosti  $C$  i  $\alpha$  zavise od tipa raspodele.[4]

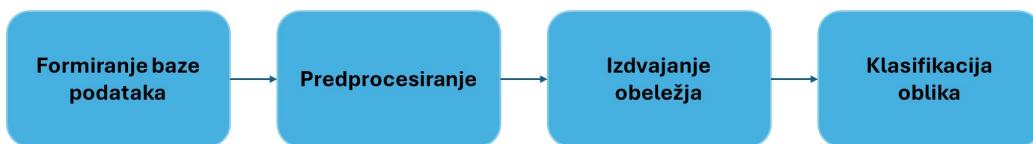
#### 4.4 Prednosti i ograničenja

*KNN* algoritam je vrlo jednostavan za implementaciju. Kako se svi podaci čuvaju u memoriji, dodavanje novih tačaka ne predstavlja nikakav problem i algoritam se lako adaptira na nove podatke. Takođe, *kNN* zahteva malo hiperparametara – potrebno je odrediti samo adekvatno  $k$  i metriku za distancu.

Međutim, *kNN* ima i svoja ogranišenja. Algoritam se ne skalira dobro i često se naziva "lenjim algoritmom". To znači da zahteva značajnu računarsku snagu i memoriju za skladištenje podataka, što ga čini vremenski zahtevnim i resursno iscrpljujućim. Pored toga, podložan je prokletstvu dimenzionalnosti, što znači da se teško nosi sa pravilnom klasifikacijom podataka kada je dimenzionalnost isuviše visoka, što je poznato kao fenomen *peaking-a*. Ovo takođe čini algoritam sklonim preobučavanju.

## 5 OPIS SISTEMA ZA PREPOZNAVANJE GOVORA

U ovom poglavlju je predstavljeno projektovanje kompletnog sistema za prepoznavanje govora. Blok šema ovakvog sistema prikazana je na slici 16.



Slika 16: Blok šema sistema za prepoznavanje govora

Prvi korak u projektovanju ovakvog sistema je formiranje adekvatne baze podataka u vidu audio snimaka. Zatim, potrebno je odraditi predprocesiranje snimaka i pripremiti ih za dalju obradu. Treći korak je primena kepstralne analize, odnosno izvlačenje kepstralnih koeficijenata kao obeležja za klasifikaciju. Kao klasifikator ko-rišćen je algoritam  $k$  najbližih suseda. Detaljan opis svakog dela sistema dat je u nastavku.

### 5.1 Formiranje baze podataka

Za projektovanje sistema za prepoznavanje govora potrebno je formirati adekvatnu bazu podataka. Baza podataka je kreirana tako što su govornici snimili svoje izgovore određenih reči. Izabrane su reči: jedan, kuća, grožđe, sedam i signal. Svaki govornik je izgovarao reči po 15 puta. Snimljeni podaci su zatim organizovani u odgovarajuće foldere za svakog ispitanika. Na primer, snimci reči govornika "Petar Petrović" su smešteni u folder pod nazivom "Petar Petrović", a unutar foldera reči raspoređene u pet različitih podfoldera, svaki za određenu reč. Tako se u svakom folderu nalazi po 75 audio snimaka, odnosno ukupno 2250 snimaka.

Govornici su koristili svoje telefone za snimanje govora. Svaki govornik je upotrebo ugrađeni mikrofon svog telefona i aplikaciju za snimanje zvuka. Snimci su prikupljeni u kontrolisanom okruženju kako bi se osigurala konzistentnost i kvalitet zvuka. Svaki govornik sniman je odvojeno kako ne bi došlo do imitacije izgovora. Snimci su originalno sačuvani u  $m4a$  formatu.

U kreiranju baze učestvovalo je 30 ljudi, od toga 14 osoba ženskog pola i 16 osoba muškog pola. Govornici su većinski mlade osobe starosti od 20 do 25 godina. Kako bi se postigla varijacija u uzrastu, učestvovala su i deca, tri devojčice uzrasta 9 do 13 godina i dva dečaka uzrasta 11 do 13 godina. Takođe, učestvovale su i dve osobe uzrasta 45+, muškog i ženskog pola.

Baza podataka kreirana je u saradnji sa koleginicama sa smera, Nikoliom Ilić (2020/0162) i Kristinom Petrić (2020/0186).

## 5.2 Vizuelizacija i predprocesiranje podataka

U cilju pronašlaska adekvatnih obeležja za klasifikaciju, potrebno je vizualizovati signale i uočiti određene karakteristike. U narednim poglavljima su detaljno opisane metode i karakteristike signala koje su od značaja za obradu i klasifikaciju, kao što su spektrogram, amplitudski spektar, kratkovremenska energija i brzina prolaska kroz nulu.

Za svaku od snimljenih reči prikazana su po tri snimka, uključujući snimak ženskog govornika, muškog govornika i deteta. Ovi snimci će omogućiti analizu varijacija u izgovoru i pomoći u boljem razumevanju karakteristika govora različitih govornika.

Za svaku reč prikazana je i segmentacija jednog primera, koja je deo predobrade signala. Na raju poglavljia detaljno su opisani i ostali elementi predobrade snimljenih govornih signala, koji uključuju uklanjanje šuma filtriranjem, normalizaciju i primenu *pre-emphasis* filtera.

### 5.2.1 Spektrogram i amplitudski spektar

Spektrogram je vizuelna reprezentacija spektralnog sadržaja zvuka kao funkcije vremena. Osnovni cilj spektrograma je prikazati kako se frekvencijski sastav zvuka menja tokom vremena. Predstavlja trodimenzionalni graf, gde su dve ose vremenska (horizontalna) i frekvencijska (vertikalna), dok se treća dimenzija (amplituda) obično prikazuje kao intenzitet boje ili nivo sivila.

Amplitudska frekvencijska karakteristika prikazuje amplitudu komponenti signala kao funkciju frekvencije. U suštini, pokazuje koliko energije signal sadrži na svakoj frekvenciji. To se postiže korišćenjem Furijeove transformacije, koja razlaže signal na njegove frekvencijske komponente.

Furijeova transformacija signala  $x(t)$  definisana je kao:

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi ft} dt$$

Amplitudska frekvencijska karakteristika signala  $x(t)$  može se dobiti kao:

$$|X(f)| = \sqrt{\Re\{X(f)\}^2 + \Im\{X(f)\}^2}$$

gde su  $\Re\{X(f)\}$  i  $\Im\{X(f)\}$  realni i imaginarni deo Furijeove transformacije  $X(f)$ .

### 5.2.2 Kratkovremenska energija

Jedna od značajnih karakteristika govornog signala jeste kratkovremenska energija, skraćeno KVE (eng. *Short-Time Signal Energy*). Ukupna energija se definiše sledećim izrazom:

$$E = \sum_{k=-\infty}^{\infty} x^2[k]$$

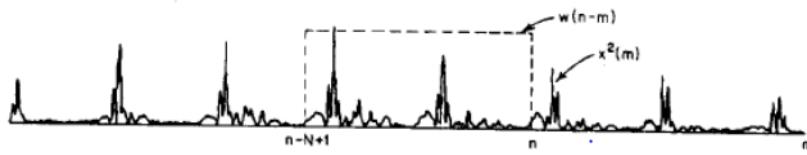
Međutim, ovako definisana veličina ne nosi nikakvu informaciju o pojavi zvučnih i bezvučnih signala, kao i o njihovoj vremenski promenljivoj prirodi. Odatle potreba da se definiše kratkovremenska energija, koja se računa kao zbir kvadrata N poslednjih odbiraka iz intervala  $(n - N + 1)$  do n-tog odbirka:

$$E_n = \sum_{k=n-N+1}^n x^2[k]$$

Ovo je identično uvođenju prozorske funkcije koja ima vrednost 1 za opseg  $0 \leq n \leq N - 1$ , a 0 za ostale vrednosti, pa se izraz može napisati i na sledeći način:

$$E_n \sum_{k=-\infty}^{\infty} x^2[k]w[n-k]$$

gde  $w$  predstavlja prozorsknu funkciju. Na slici 17 prikazan je tipičan izgled kratkovremenke energije.



Slika 17: Tipičan oblik kratkovremenske energije, preuzeto iz rada [2]

Za prozorsknu funkciju se najčešće bira Hammingova:  $w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi}{N-1}\right)$  za  $0 \leq n \leq N - 1$ , odnosno 0 za ostale vrednosti, ili pravougaona prozorska funkcija:  $w[n] = 1$  za  $0 \leq n \leq N - 1$ , odnosno 0 za ostale vrednosti. Što se tiče dužine prozora N, važno je da ona ne bude suviše mala, recimo reda dužine pitch periode ili manje, jer to dovodi do značajnog oscilovanja kratkovremenske energije zavisno od pojedinih oscilacija u vremenskom obliku signala. Takođe, N ne treba biti ni suviše veliko, reda veličine nekoliko pitch perioda, jer će za posledicu kratkovremenska energija biti isuviše spora i neće nositi informacije o značajnim promenama u obliku signala. Uobičajeno je da se uzima da je dužina prozorskne funkcije 1 do 3 pitch periode, pa dobijamo sledeće relacije:

$$NT = (1 \div 3) \Rightarrow N = (1 \div 3) \frac{T_{\text{pitch}}}{T} = (1 \div 3) \frac{F_s}{F_{\text{pitch}}}$$

Uzimajući u obzir da se pitch frekvencija ljudskog glasa kreće između 120 Hz (kod muškaraca) i 300 Hz (kod dece), i da je tipična frekvencija odabiranja 10 kHz, znamo da je dobar izbor za dužinu prozorskne funkcije  $N \in [100, 200]$ .

Kratkovremenska energija se često koristi za identifikaciju govornih delova u signalima. Na primer, segmenti signala sa visokom energijom mogu odgovarati govorima, dok segmenti sa niskom energijom odgovaraju pauzama. Grafički prikaz kratkovremenske energije može se koristiti za vizuelnu analizu i segmentaciju signala. Takođe, ova karakteristika se može koristiti i za analizu emocionalnog stanja govornika, jer energija može reflektovati intenzitet emocija u govoru.

### 5.2.3 Brzina prolaska kroz nulu

Još jedan parametar koji dobro oslikava sadržaj snimljenog signala je brzina prolaska signala kroz nulu (eng. *Zero-Crossing Rate-ZCR*). Ova veličina je zapravo mera frekvencijskog sadržaja signala, pa ukoliko je signal bogat visokim učestanostima, ova mera će biti veća i obrnuto. Brzina prolaska  $Z$  definiše se kao:

$$Z = \frac{N_0}{N}$$

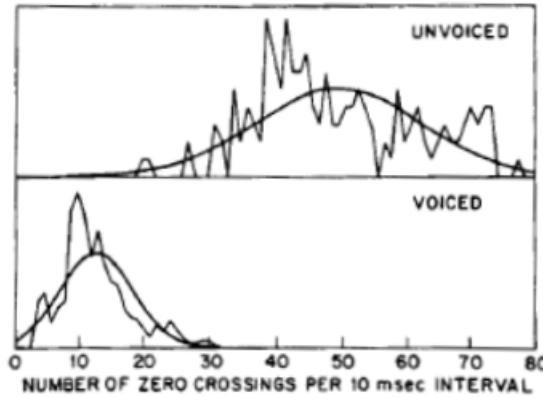
gde je  $N$  broj posmatranih odbiraka, a  $N_0$  broj odbiraka koji nisu istog znaka kao njihovi prethodnici. Na primer, ukoliko posmatramo prostoperiodičnu komponentu učestanosti  $F_0$  koja je odabirana sa učestanošću  $F_s$ , tokom jedne periode  $T_0$  imaćemo  $N = \frac{T_0}{T_s} = \frac{F_0}{F_s}$ , pri čemu će signal dva puta promeniti znak, pa je brzina prolaska kroz nulu  $Z$ :

$$N = \frac{2}{N} = \frac{2F_0}{F_s}$$

Međutim, govorni signal sadrži frekvencijske komponente iz širokog spektra učestanosti, pa sama vrednost mere  $Z$  nije tako precizna. Zbog ovoga se koristi kratkovremenska forma ova mere:

$$Z_n = \sum_{k=-\infty}^{\infty} \frac{1}{2} |\operatorname{sgn}(x[k]) - \operatorname{sgn}(x[k-1])| w[n-k]$$

gde je  $\operatorname{sgn}$  funkcija znaka, a  $w[n]$  prozorska funkcija definisana kao:  $w[n] = \frac{1}{2N}$  za  $0 \leq n \leq N-1$ , a u suprotnom 0. Kada ovu meru primenjujemo na gorovne signale polazi se od pretpostavke da je zvučni deo signala na učestanostima oko 3 kHz, dok je bezvučni deo i pozadinski šum na višim učestanostima. Zbog ovoga možemo očekivati da je mera  $Z$  za zvučni signal značajno niža u odnosu na bezvučni. Na slici 18 je prikazan histogram parametra  $Z$  za zvučni i bezvučni signal kao i njihove odgovaraće Gausovske funkcije gustine verovatnoće.



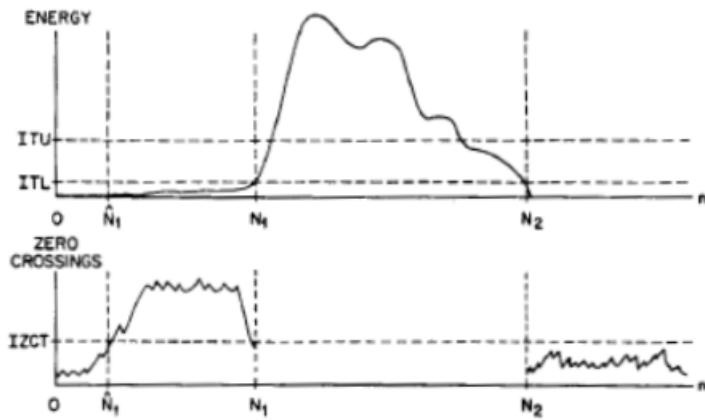
Slika 18: FGV brzine prolaska kroz nulu za zvučne i bezvučne segmente govornog signala, preuzeto iz rada [2]

Za bezvučni signal, srednja vrednost parametra  $Z$  je 49 prolaska kroz nulu na periodu od 10ms, dok je za zvučni samo 14. Iako je razlika očigledna, varijanse su velike i

funkcije gustine verovatnoće se preklapaju, pa  $Z$  ne može biti jedini parametar po kome se vrši segmentacija na zvučni i bezvučni deo govornog signala.

#### 5.2.4 Segmentacija reči

Jedan od glavnih problema za rešavanje prilikom obrade govornog signala je izdvajanje korisnog, govornog dela od raznih vrsta šumova koji su prisutni. Ono što dodatno otežava ovaj problem je i postojanje reči koje počinju ili se završavaju frikativima ('f', 'h') ili slabim plozivima ('p', 't', 'k'), ili reči koje se završavaju nazalima ('m', 'n', 'nj'). Snaga ovih glasova ne leži u očekivanom opsegu, i pri tome je prilično mala, pa se često teško izdvaja od pozadinskog šuma. Algoritam koji rešava ovaj problem koristi kombinaciju kratkovremenske energije i kratkovremenske brzine prolaska kroz nulu signala i prikazan je na slici 19.



Slika 19: Primer kratkovremenske energije i kratkovremenske brzine prolaska kroz nulu u cilju određivanja početka i kraja govornog signala, preuzeto iz rada [4]

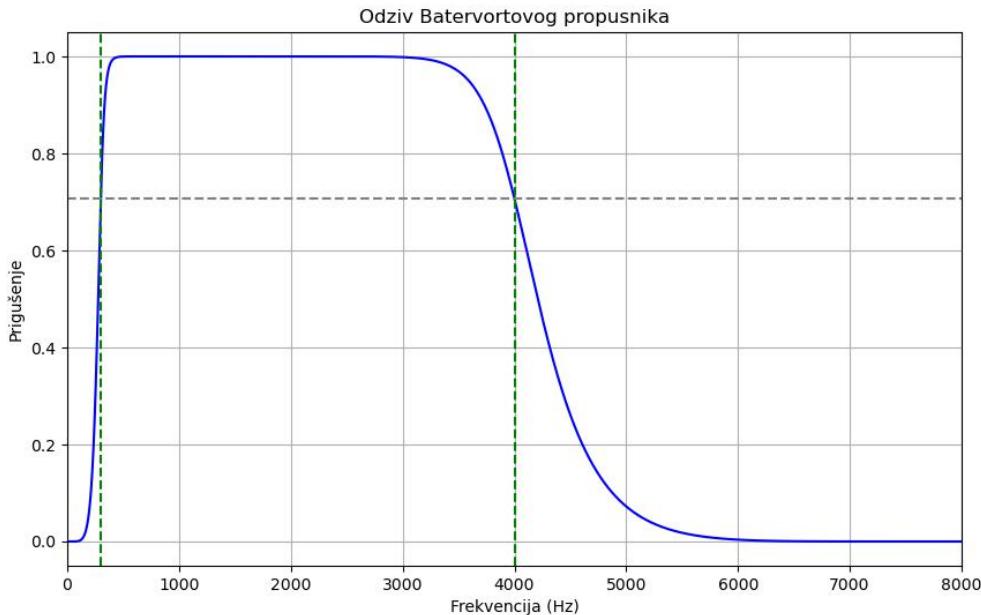
Prvo što određujemo je *ITU* (eng. *Intensity Threshold Upper*) koji predstavlja gornji prag za koji ćemo biti sigurni da ako je energija signala veća od njegove vrednosti, znamo da je u tom intervalu sigurno govorni signal. Međutim, s obzirom da je prag uglavnom dovoljno veliki, znamo i da je govor započeo i pre nego što je KVE postala veća od *ITU*, i traje i nakon što je KVE opala ispod *ITU*. Pored gornjeg praga biramo i donji prag *ITL* (eng. *Intensity Threshold Lower*), koji će biti manje konzervativan i pomoći nam da približnije odredimo početak i kraj govornog signala. Od trenutka kada je KVE postala veća od *ITU* iterativnim postupkom se pomeramo ulevo u vremenu i određujemo trenutak  $N_1$  kada KVE prvi put postane manja od *ITL*-a i na sličan način, od trenutka kada je KVE postala manja od *ITU* iterativnim postupkom se pomeramo udesno u vremenu i određujemo trenutak  $N_2$  kada je KVE postala manja od *ITL*-a. Ova dva trenutka predstavljaju naš inicijalni početak i kraj reči. U cilju finog podešavanja inicijalnog početka reči  $N_1$  posmatra se interval signala kratkovremenske brzine prolaska kroz nulu (eng. *Short-Time Zero-Crossing Rate-STZCR*) 25 frejmova (prozorskih funkcija) levo od tačke  $N_1$ . Ukoliko u tom intervalu signal STZCR više od tri puta preseče usvojeni prag (između 20 i 30) tada se  $N_1$  koriguje u  $\hat{N}_1$ , u tačku gde je prag prvi put presečen. Analogno se sprovodi postupak za prag  $N_2$ , s tim što je  $\hat{N}_2$  tačka poslednjeg preseka.[2]

### 5.2.5 Uklanjanje šuma

Za uklanjanje šuma iz govornih signala korišćen je Batervortov filter (eng. *Butterworth filter*). Batervortov filter ima maksimalno ravnu odzivnu karakteristiku, tj. nema talasanja u propusnom opsegu. Ovo je posebno važno u obradi govornih signala, jer omogućava minimalno izobličenje signala unutar propusnog opsega. Ovakav filter je definisan pomoću dva parametra: redom filtera odnosno brojem polova, i frekvencijom koja određuje propusni opseg.

U ovom radu korišćen je Batervortov filter sa opsegom od 300 do 4000 Hz, što je optimalan opseg za ovakav problem prepoznavanja reči. Govorni signali sadrže većinu svojih energetskih komponenti u ovom frekvencijskom domenu. Takođe, frekvencije ispod 300 Hz često sadrže informacije o osnovnom tonu i glasu, ali nisu od značaja za prepoznavanje reči. Sa druge strane, frekvencije iznad 4000 Hz sadrže manje bitne informacije za razumljivost govora, a mogu doprineti šumu i povećanju složenosti obrade.[6]

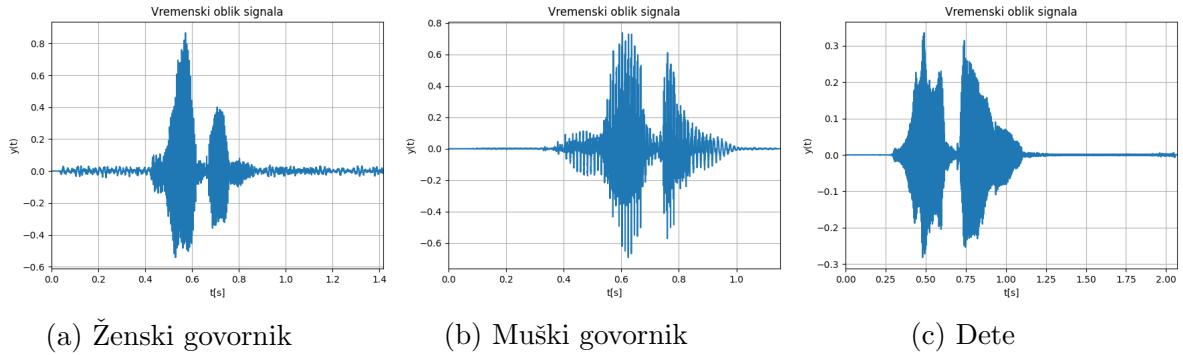
Na slici 20 prikazan je Batervortov filter, sa propusnim opsegom od 300 do 4000 Hz, reda 6 i frekvencije odabiranja 16 kHz. Na grafiku se vidi da je u propusnom opsegu karakteristika maksimalno zaravnjena, dok se signal van ovog opsega značajno prigušuje. Zelene isprekidane linije označavaju granične frekvencije filtera.



Slika 20: Odziv Batervortovog filtera

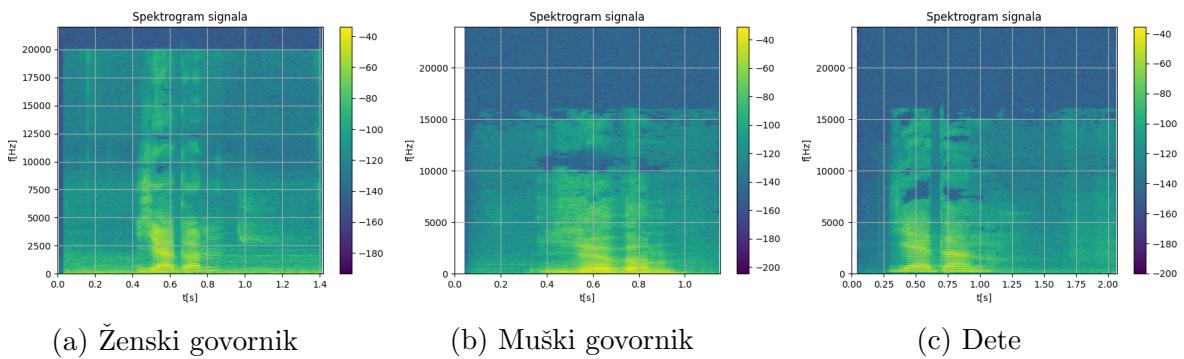
### 5.2.6 Reč "jedan"

Na slici 21 prikazani su vremenski oblici signala za reč "jedan", koje su redom izgovarali ženski govornik, muški govornik i dete.



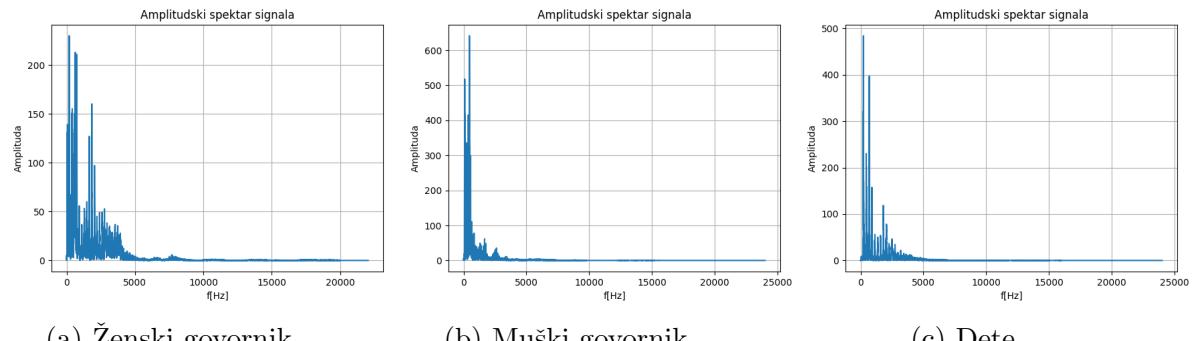
Slika 21: Vremenski oblici signala različitih govornika

Na slici 22 prikazani su spektrogrami signala za reč "jedan", koje su redom izgovarali ženski govornik, muški govornik i dete.



Slika 22: Spektrogrami signala različitih govornika

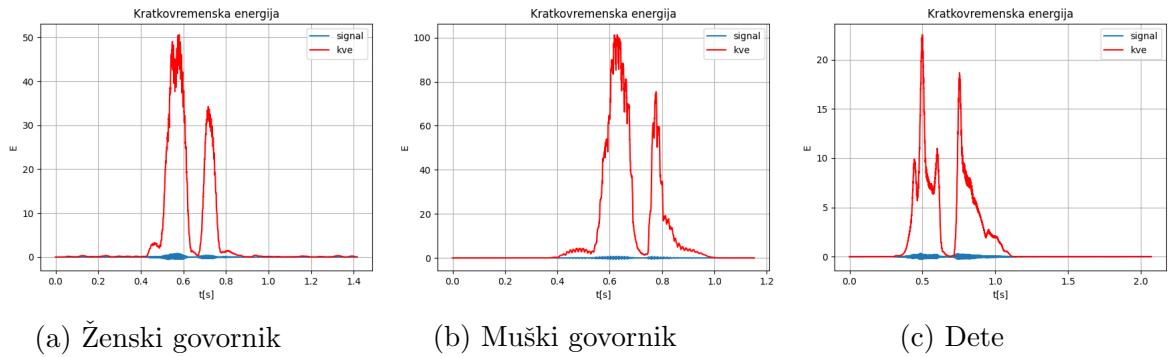
Na slici 23 prikazane su amplitudske frekvencijske karakteristike signala za reč "jedan", koje su redom izgovarali ženski govornik, muški govornik i dete.



Slika 23: Amplitudske frekvencijske karakteristike signala različitih govornika

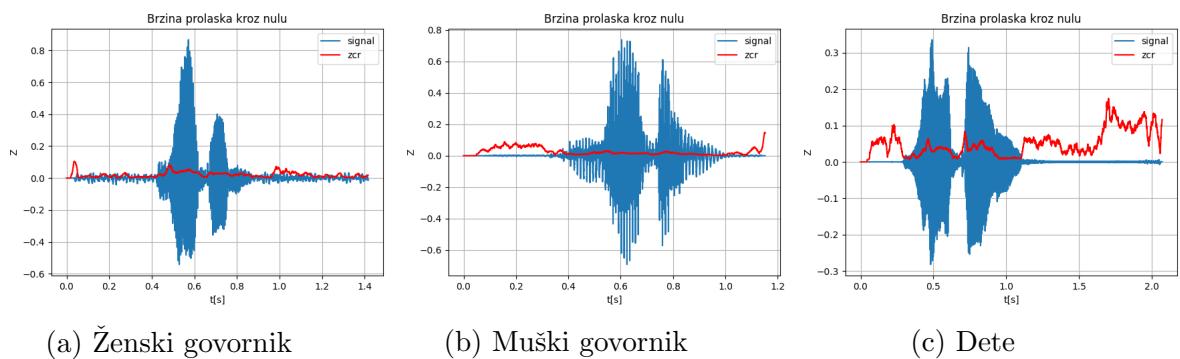
Za segmentaciju slika potrebno je odrediti i kratkovremenu energiju, kao i brzinu prolaska kroz nulu, te su i ove karakteristike prikazane na narednim graficima.

Na slici 24 prikazane su kratkovremenske energije signala za reč "jedan", koje su redom izgovarali ženski govornik, muški govornik i dete.



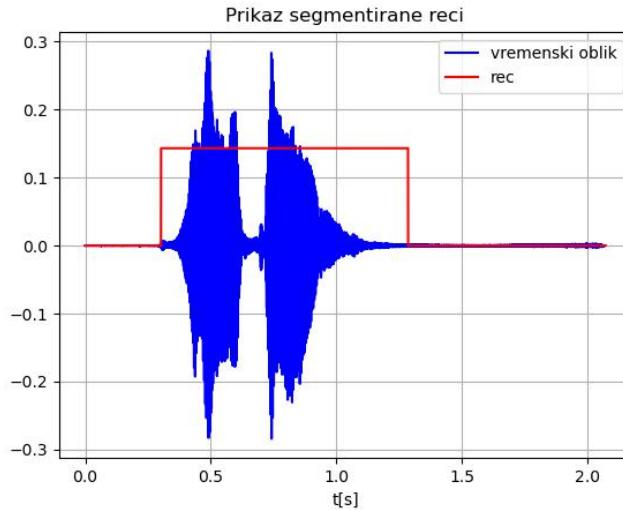
Slika 24: Kratkovremenske energije signala različitih govornika

Na slici 25 prikazane su brzine prolaska kroz nulu signala za reč "jedan", koje su redom izgovarali ženski govornik, muški govornik i dete.



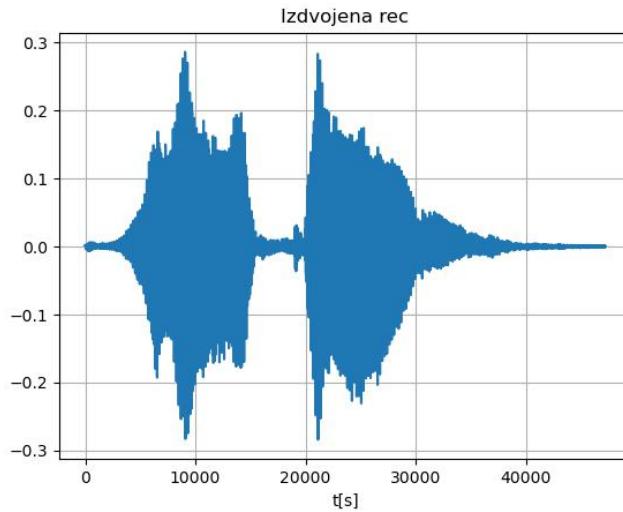
Slika 25: Brzine prolaska kroz nulu signala različitih govornika

Segmentacija reči sprovodi se gorepomenutim algoritmom koji koristi kratkovremensku energiju i brzinu prolaska kroz nulu. Na slici 26 prikazan je primer govornog signala za reč "jedan" nakon primene Batervortovog filtera, koji se koristi za otklanjanje šuma. Na istom grafiku crvenom bojom je iscrtan pravougaonik, koji predstavlja granice za segmentaciju reči. Gornji prag odnosno  $ITU$  je postavljen na 5% od maksimalne energije, dok je donji prag odnosno  $ITL$  postavljen na 0.003% od maksimalne energije.



Slika 26: Primer segmentacije reči "jedan"

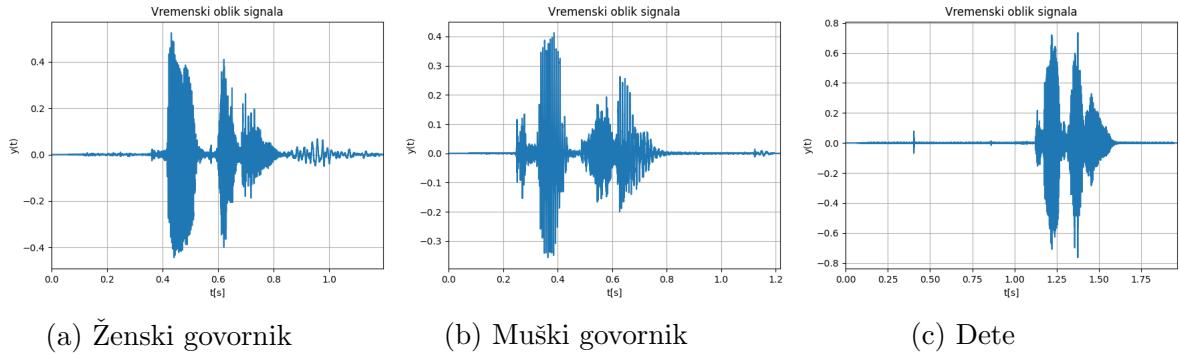
Kako je na ovom primeru šum lepo otklonjen, određivanje granica za segmentaciju nije bio težak proces. Na slici 27 prikazan je segmentisan govorni signal. Signal ima dve izražene amplitude, koje predstavljaju dva glavna formanta reči. Nakon drugog formanta, signal postepeno opada do nule, što ukazuje na kraj izgovora reči. Oblik signala je simetričan, sa izraženim vrhovima i dolinama, koji su karakteristični za glasne delove reči, dok tiši delovi imaju manje amplitude.



Slika 27: Primer reči "jedan" nakon segmentacije

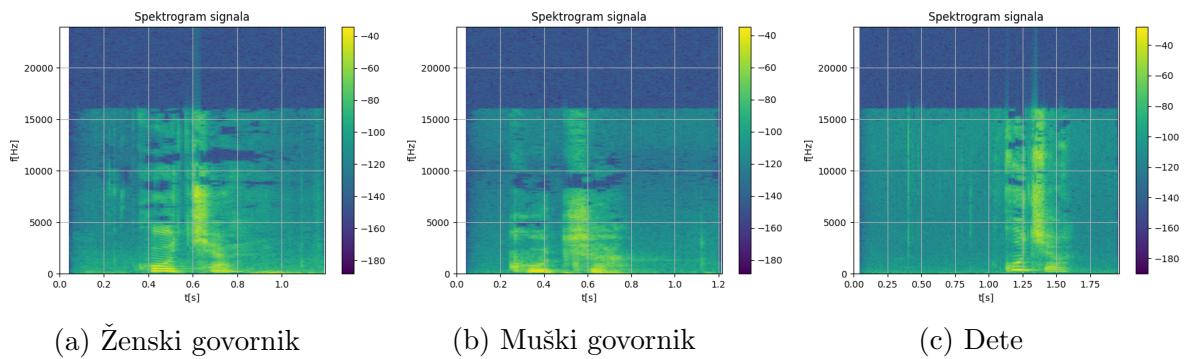
### 5.2.7 Reč "kuća"

Na slici 28 prikazani su vremenski oblici signala za reč "kuća", koje su redom izgovarali ženski govornik, muški govornik i dete.



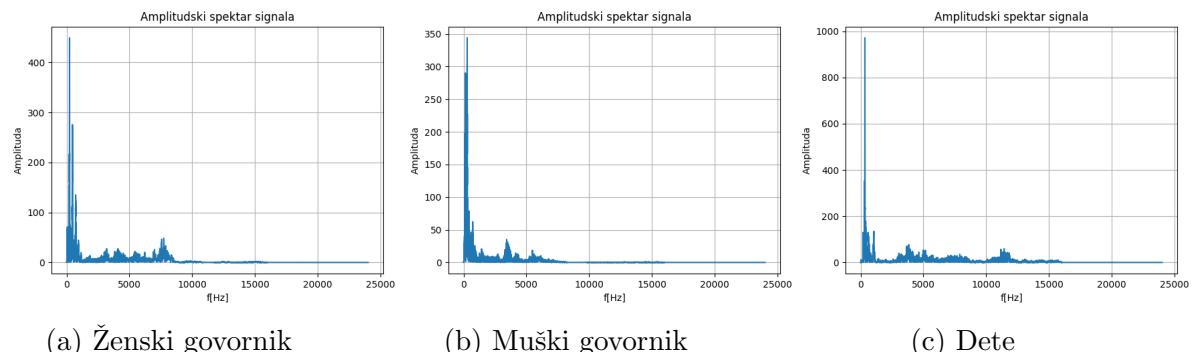
Slika 28: Vremenski oblici signala različitih govornika

Na slici 29 prikazani su spektrogrami signala za reč "kuća", koje su redom izgovarali ženski govornik, muški govornik i dete.



Slika 29: Spektrogrami signala različitih govornika

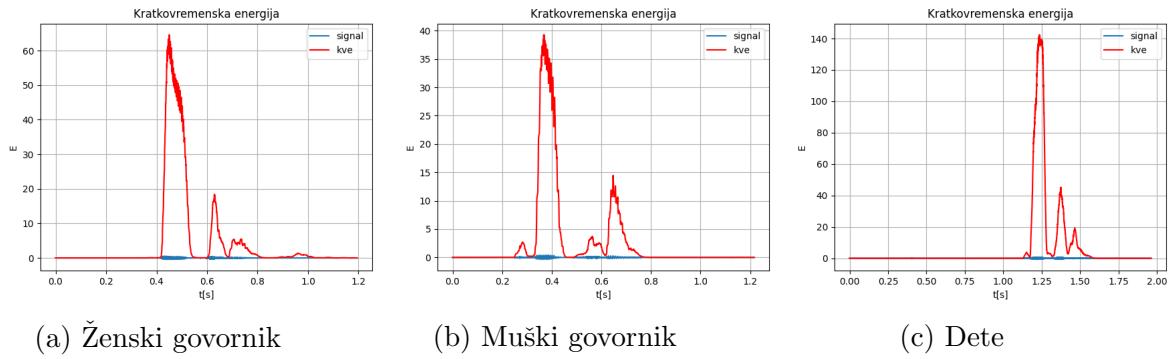
Na slici 30 prikazane su amplitudske frekvencijske karakteristike signala za reč "kuća", koje su redom izgovarali ženski govornik, muški govornik i dete.



Slika 30: Amplitudske frekvencijske karakteristike signala različitih govornika

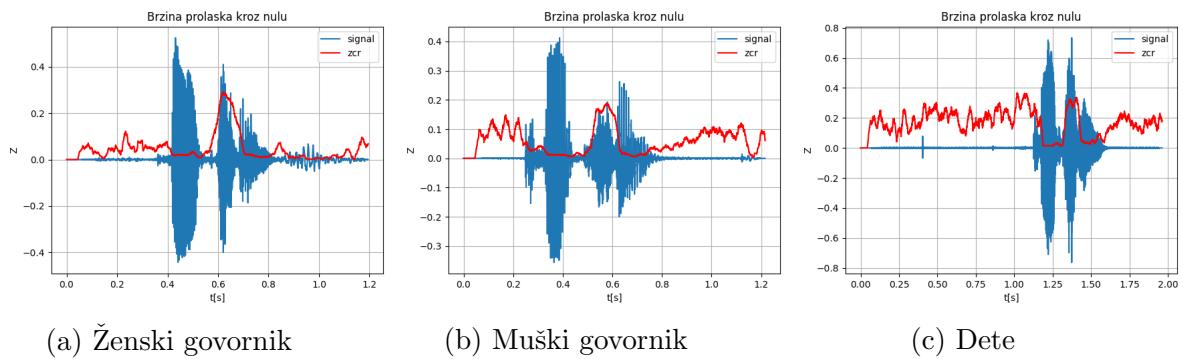
Za segmentaciju slika potrebno je odrediti i kratkovremenu energiju, kao i brzinu prolaska kroz nulu, te su i ove karakteristike prikazane na narednim graficima.

Na slici 31 prikazane su kratkovremenske energije signala za reč "kuća", koje su redom izgovarali ženski govornik, muški govornik i dete.



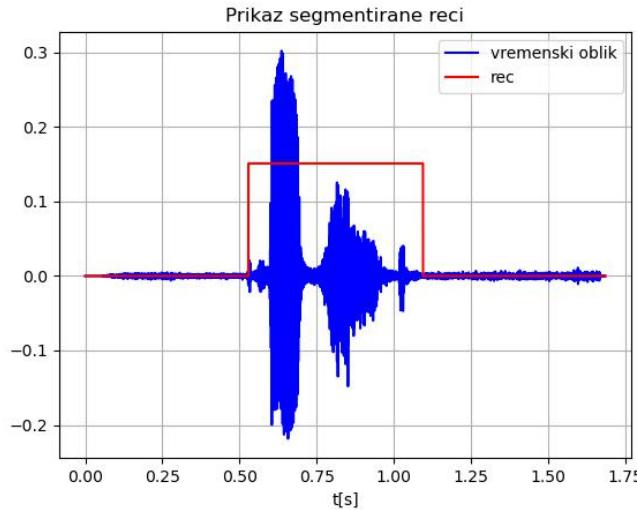
Slika 31: Kratkovremenske energije signala različitih govornika

Na slici 32 prikazane su brzine prolaska kroz nulu signala za reč "kuća", koje su redom izgovarali ženski govornik, muški govornik i dete.



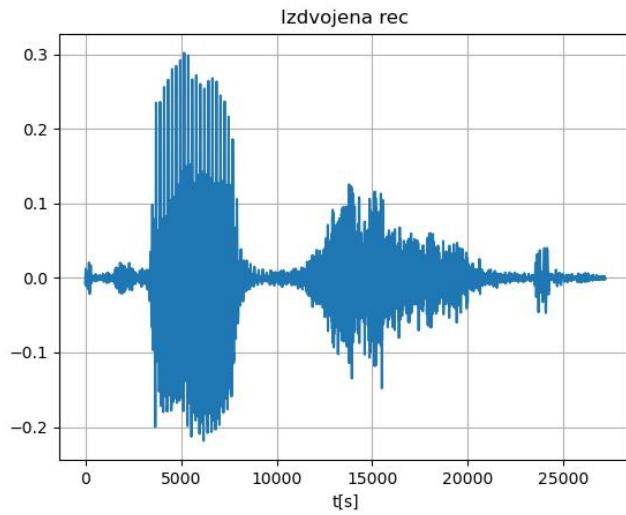
Slika 32: Brzine prolaska kroz nulu signala različitih govornika

Segmentacija reči sprovodi se gorepomenutim algoritmom koji koristi kratkovremensku energiju i brzinu prolaska kroz nulu. Na slici 33 prikazan je primer govornog signala za reč "kuća" nakon primene Batervortovog filtera, koji se koristi za otklanjanje šuma. Na istom grafiku crvenom bojom je iscrtan pravougaonik, koji predstavlja granice za segmentaciju reči. Gornji prag odnosno  $ITU$  je postavljen na 5% od maksimalne energije, dok je donji prag odnosno  $ITL$  postavljen na 0.02% od maksimalne energije.



Slika 33: Primer segmentacije reči "kuća"

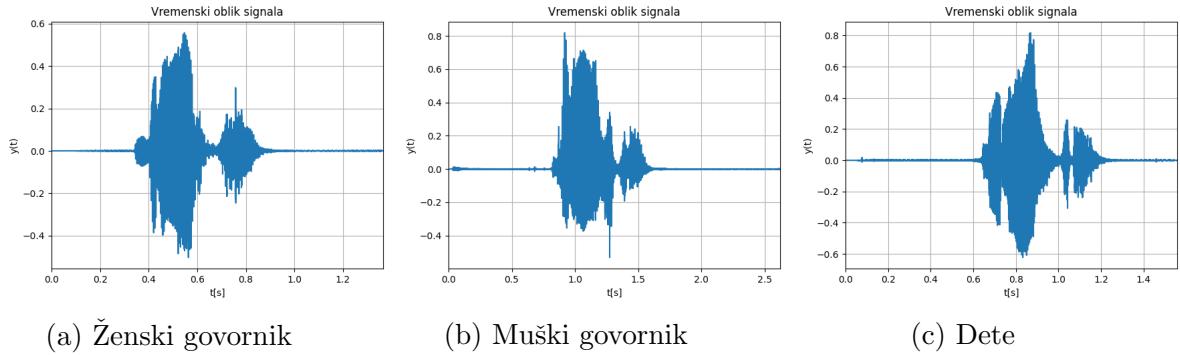
Primećuje se da je donja granica 10 puta veća u odnosu na donju granicu iz prethodnog primera, što je posledica činjenice da je signal za reč "kuća" zašumljeniji, te su amplitude veće i potrebno je primeniti viši prag kako bi se reč adekvatno segmentisala. Na slici 34 prikazan je primer segmentisane reči. Sa grafika se vidi da je prvi deo signala intenzivan sa visokim amplitudama, i on pripada glasnom izgovaranju sloga "ku". Nakon prvog vrha, amplituda se smanjuje, ali ostaje umereno visoka, i ovo predstavlja prelaz iz prvog sloga u drugi slog "ća". Zatim signal postepeno opada, što obeležava kraj reči.



Slika 34: Primer reči "kuća" nakon segmentacije

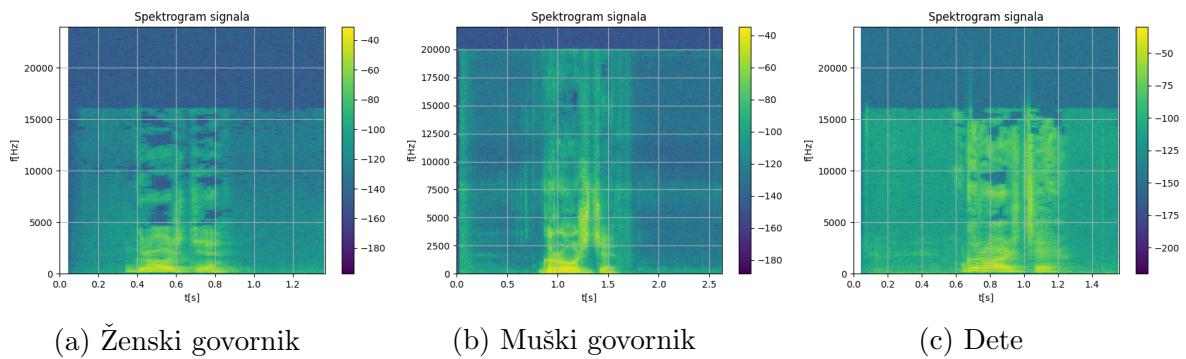
#### 5.2.8 Reč "grožđe"

Na slici 35 prikazani su vremenski oblici signala za reč "grožđe", koje su redom izgovarali ženski govornik, muški govornik i dete.



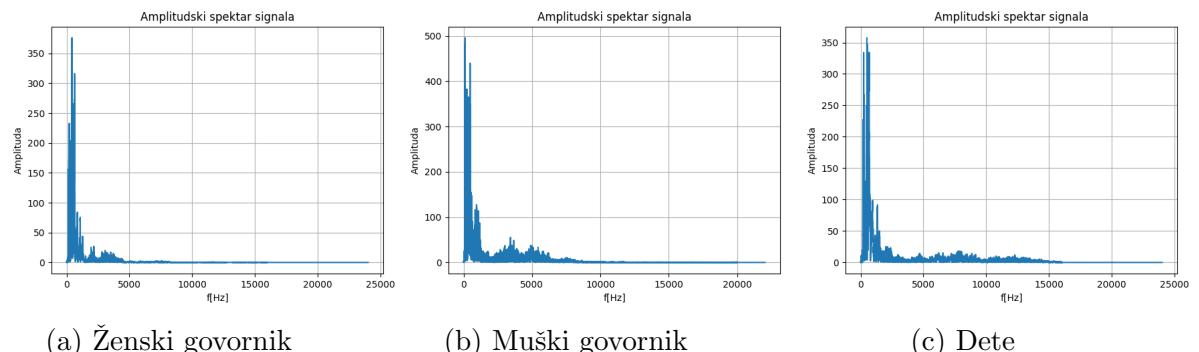
Slika 35: Vremenski oblici signala različitih govornika

Na slici 36 prikazani su spektrogrami signala za reč "grožđe", koje su redom izgovarali ženski govornik, muški govornik i dete.



Slika 36: Spektrogrami signala različitih govornika

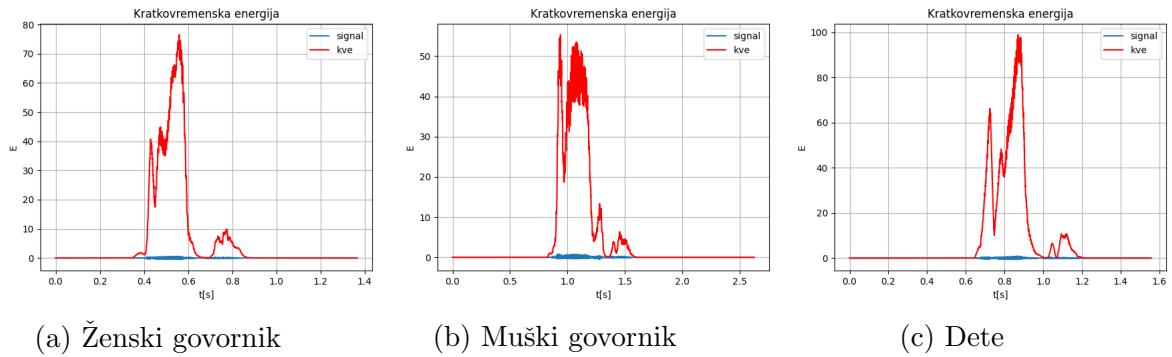
Na slici 37 prikazane su amplitudske frekvencijske karakteristike signala za reč "grožđe", koje su redom izgovarali ženski govornik, muški govornik i dete.



Slika 37: Amplitudske frekvencijske karakteristike signala različitih govornika

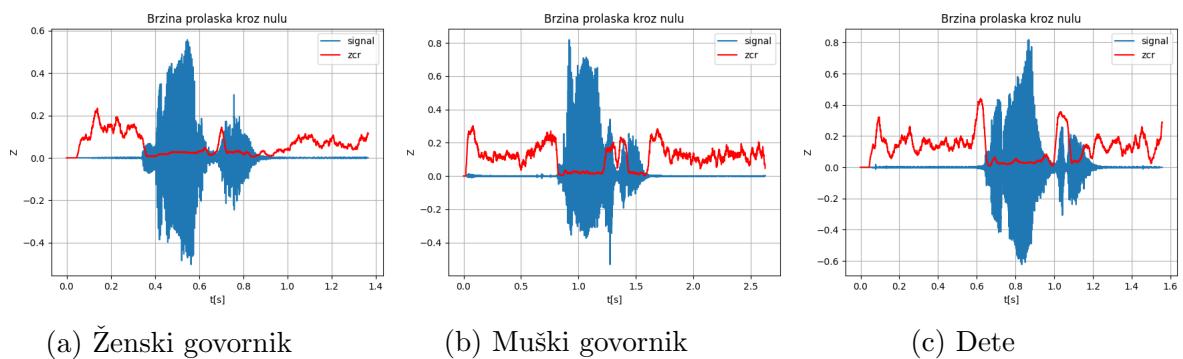
Za segmentaciju slika potrebno je odrediti i kratkovremensku energiju, kao i brzinu prolaska kroz nulu, te su i ove karakteristike prikazane na narednim graficima.

Na slici 38 prikazane su kratkovremenske energije signala za reč "grožđe", koje su redom izgovarali ženski govornik, muški govornik i dete.



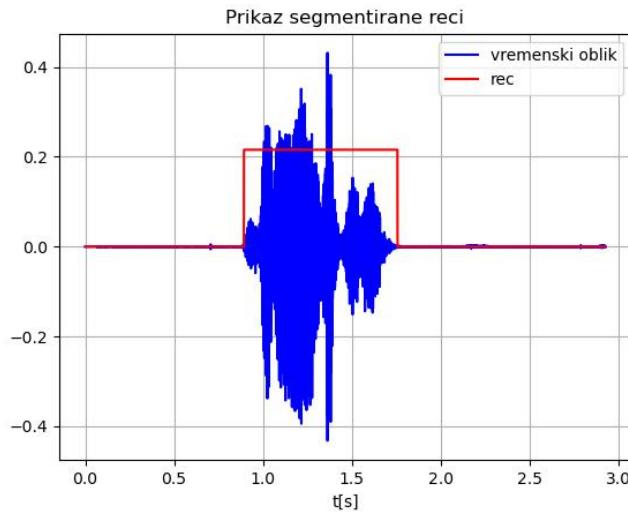
Slika 38: Kratkovremenske energije signala različitih govornika

Na slici 39 prikazane su brzine prolaska kroz nulu signala za reč "grožđe", koje su redom izgovarali ženski govornik, muški govornik i dete.



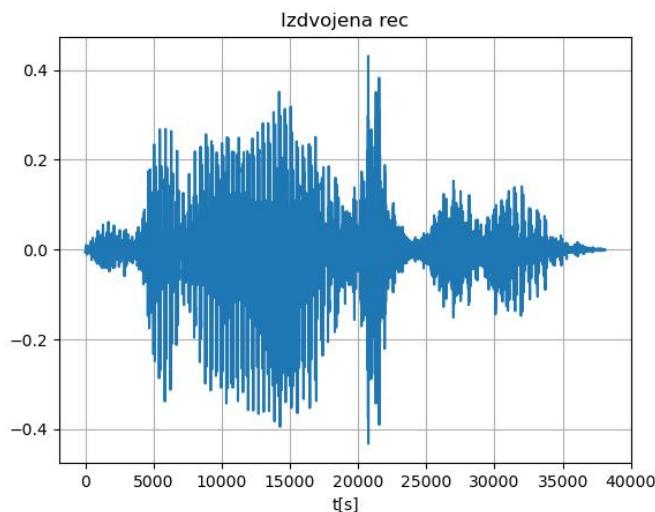
Slika 39: Brzine prolaska kroz nulu signala različitih govornika

Segmentacija reči sprovodi se gorepomenutim algoritmom koji koristi kratkovremensku energiju i brzinu prolaska kroz nulu. Na slici 40 prikazan je primer govornog signala za reč "grožđe" nakon primene Batervortovog filtera, koji se koristi za otklanjanje šuma. Na istom grafiku crvenom bojom je iscrtan pravougaonik, koji predstavlja granice za segmentaciju reči. Gornji prag odnosno *ITU* je postavljen na 5% od maksimalne energije, dok je donji prag odnosno *ITL* postavljen na 0.005% od maksimalne energije.



Slika 40: Primer segmentacije reči "grožđe"

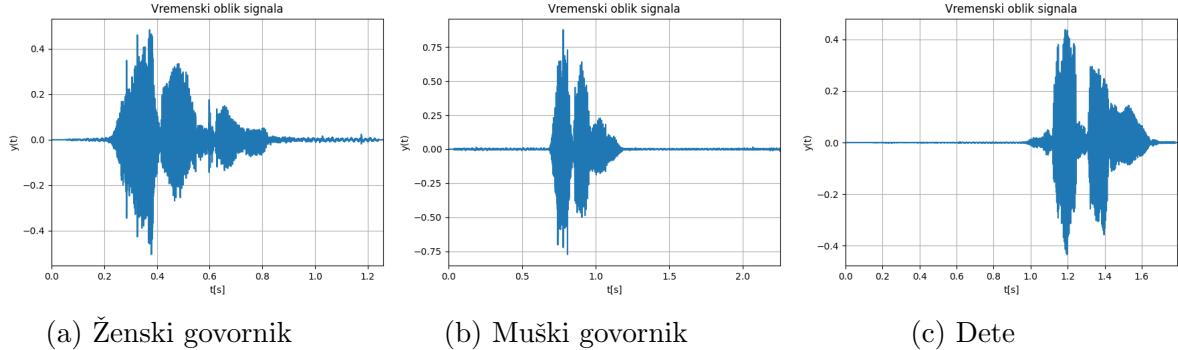
Signal je dobro filtriran i relativno "čist", na šta ukazuje potrebna mala donja granica. Na slici 41 prikazana je segmentisani primer reči "grožđe". Signal se sastoji od nekoliko karakterističnih delova: početni deo, sa povećanom amplitudom, odgovara prvom slogu "gro", sa intenzivnim oscilacijama koje označavaju artikulaciju ovog sloga. Srednji deo signala, pokazuje manje intenzivne oscilacije, verovatno zbog prelaska na drugi deo reči "žđ". Na kraju, signal pokazuje ponovni porast amplitudine, što označava završni deo reči "đe", sa jasnim vrhovima koji ukazuju na naglašavanje zvuka. Ovakav obrazac oscilacija u signalu je karakterističan za složene reči sa više naglašenih delova.



Slika 41: Primer reči "grožđe" nakon segmentacije

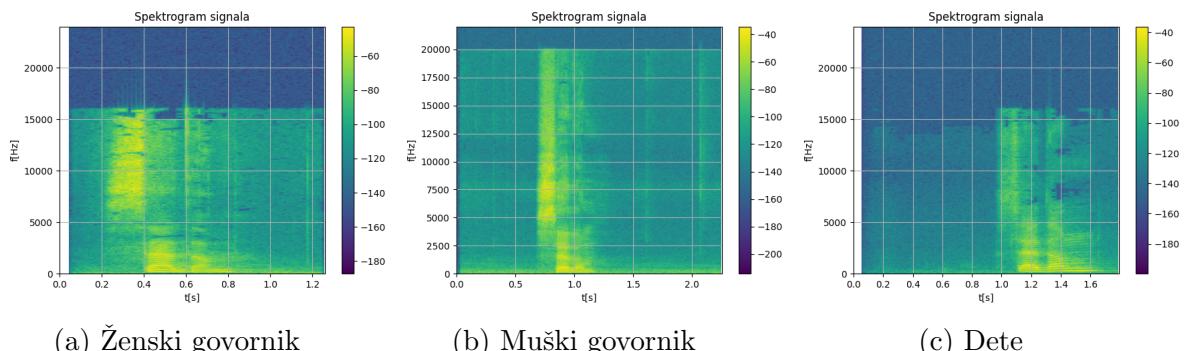
### 5.2.9 Reč "sedam"

Na slici 42 prikazani su vremenski oblici signala za reč "sedam", koje su redom izgovarali ženski govornik, muški govornik i dete.



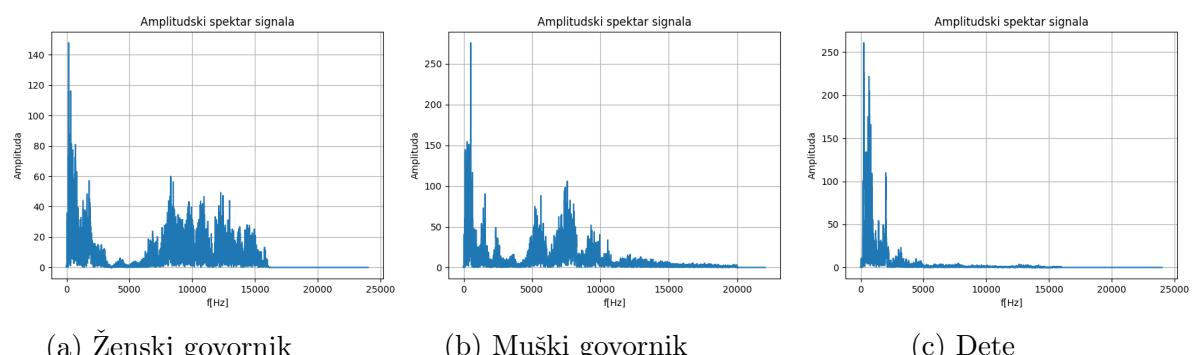
Slika 42: Vremenski oblici signala različitih govornika

Na slici 43 prikazani su spektrogrami signala za reč "sedam", koje su redom izgovarali ženski govornik, muški govornik i dete.



Slika 43: Spektrogrami signala različitih govornika

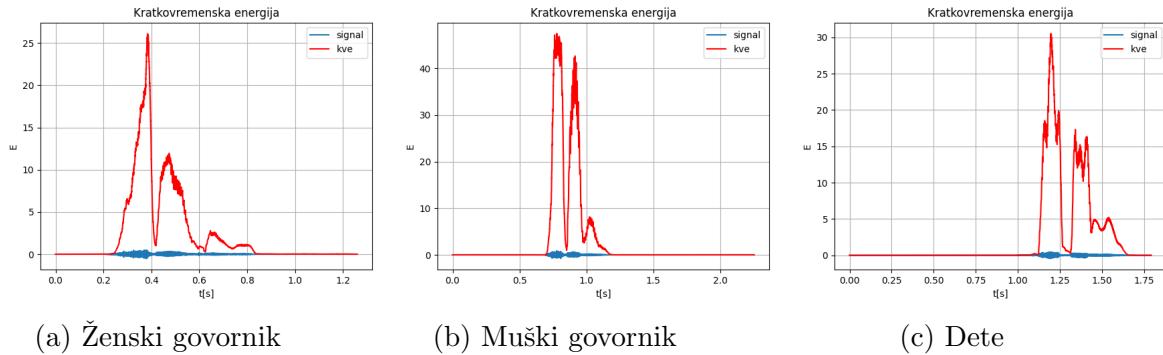
Na slici 44 prikazane su amplitudske frekvencijske karakteristike signala za reč "sedam", koje su redom izgovarali ženski govornik, muški govornik i dete.



Slika 44: Amplitudske frekvencijske karakteristike signala različitih govornika

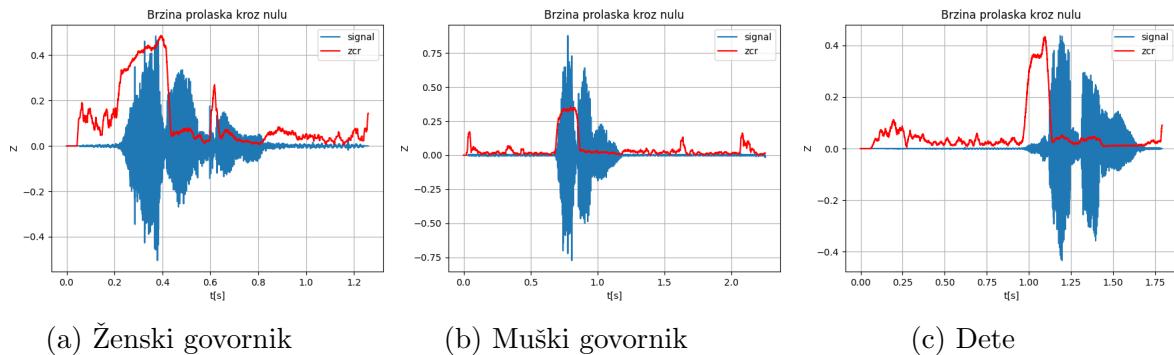
Za segmentaciju slika potrebno je odrediti i kratkovremensku energiju, kao i brzinu prolaska kroz nulu, te su i ove karakteristike prikazane na narednim graficima.

Na slici 45 prikazane su kratkovremenske energije signala za reč "sedam", koje su redom izgovarali ženski govornik, muški govornik i dete.



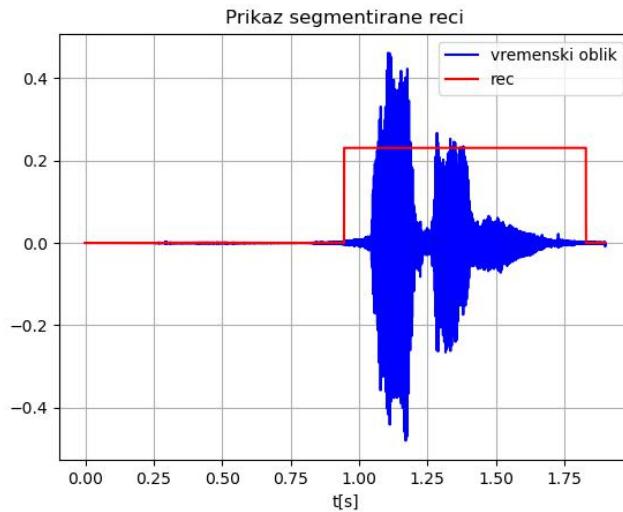
Slika 45: Kratkovremenske energije signala različitih govornika

Na slici 46 prikazane su brzine prolaska kroz nulu signala za reč "sedam", koje su redom izgovarali ženski govornik, muški govornik i dete.



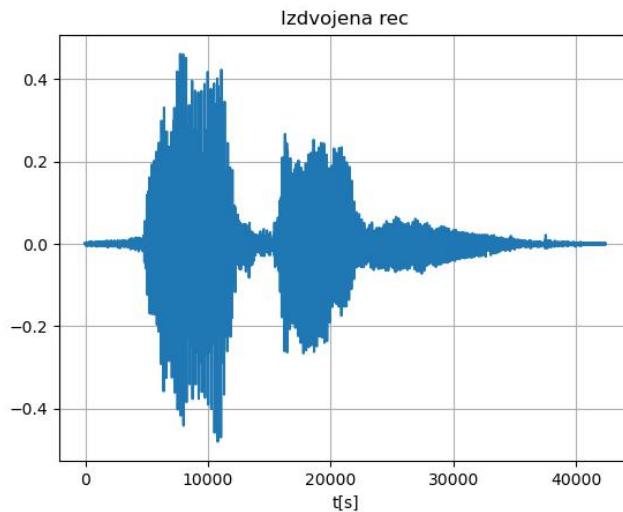
Slika 46: Brzine prolaska kroz nulu signala različitih govornika

Segmentacija reči sprovodi se gorepomenutim algoritmom koji koristi kratkovremensku energiju i brzinu prolaska kroz nulu. Na slici 47 prikazan je primer govornog signala za reč "sedam" nakon primene Batervortovog filtera, koji se koristi za otklanjanje šuma. Na istom grafiku crvenom bojom je iscrtan pravougaonik, koji predstavlja granice za segmentaciju reči. Gornji prag odnosno  $ITU$  je postavljen na 5% od maksimalne energije, dok je donji prag odnosno  $ITL$  postavljen na 0.005% od maksimalne energije.



Slika 47: Primer segmentacije reči "sedam"

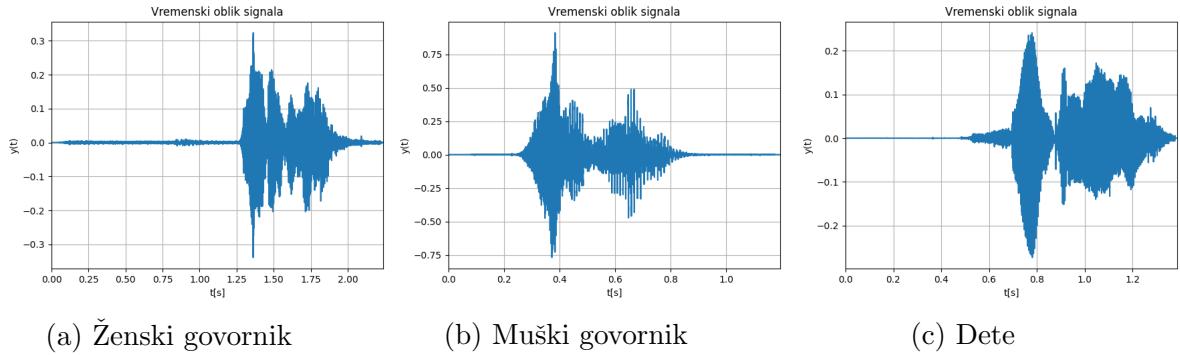
Kao i u nekim ranijim primerima, ovaj signal je prilično dobor filtriran i "čist", pa je donja granica niska. Na slici 48 prikazan je segmentisan govorni signal. Početak signala je okarakterisan brzim dostizanjem visoke amplitude, što ukazuje na izgovor prvog sloga "se". Zatim se amplituda smanjuje, međutim signal održava značajnu vrednost, što predstavlja prelaz na drugi slog "dam". Amplituda ponovo raste pri izgovoru ovog sloga i na kraju postepeno opada, što označava kraj izgovorene reči.



Slika 48: Primer reči "sedam" nakon segmentacije

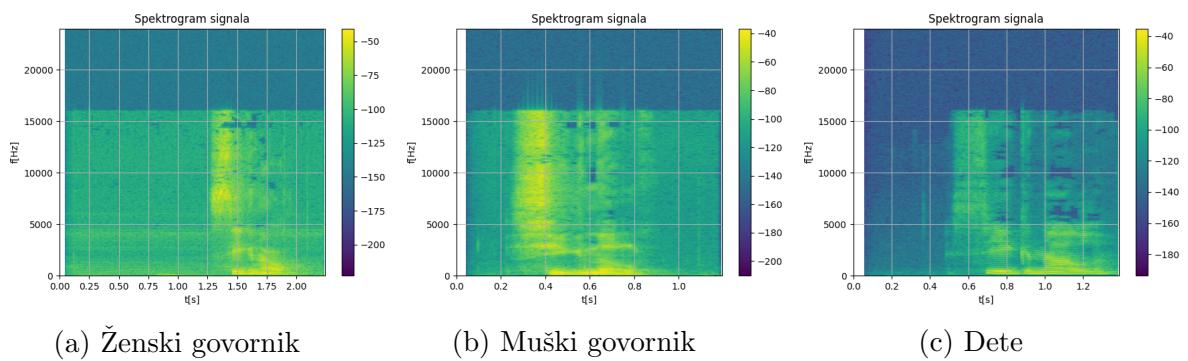
#### 5.2.10 Reč "signal"

Na slici 49 prikazani su vremenski oblici signala za reč "signal", koje su redom izgovarali ženski govornik, muški govornik i dete.



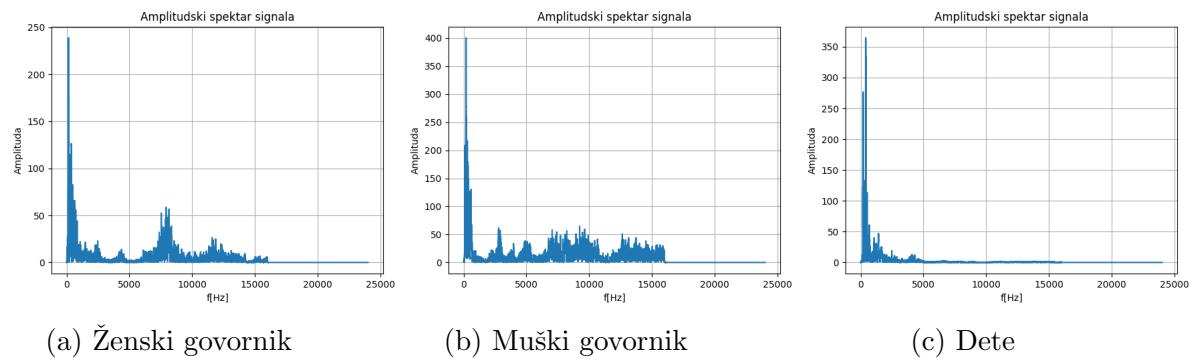
Slika 49: Vremenski oblici signala različitih govornika

Na slici 50 prikazani su spektrogrami signala za reč "signal", koje su redom izgovarali ženski govornik, muški govornik i dete.



Slika 50: Spektrogrami signala različitih govornika

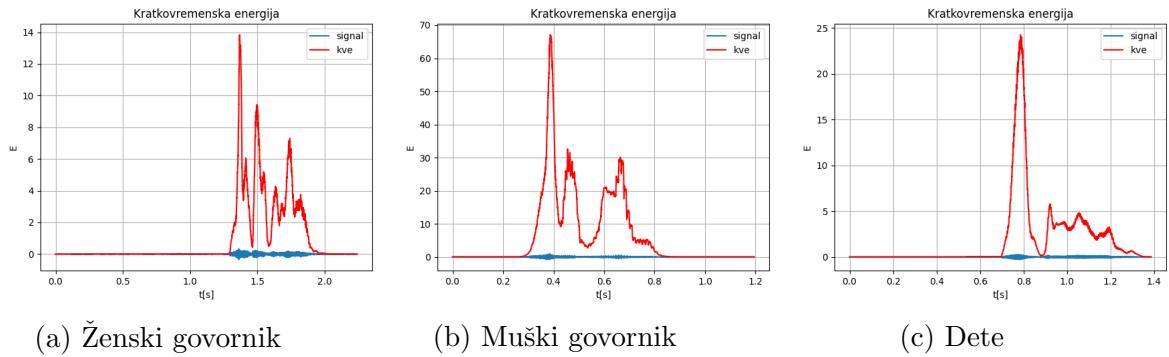
Na slici 51 prikazane su amplitudske frekvencijske karakteristike signala za reč "signal", koje su redom izgovarali ženski govornik, muški govornik i dete.



Slika 51: Amplitudske frekvencijske karakteristike signala različitih govornika

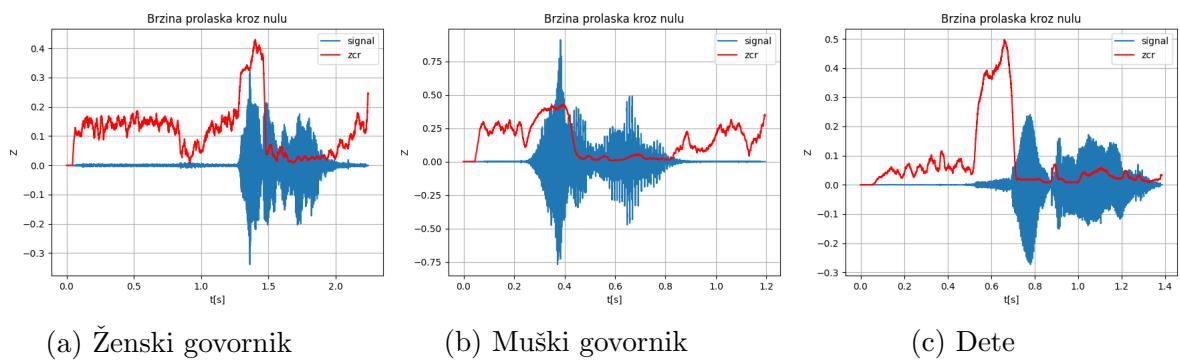
Za segmentaciju slika potrebno je odrediti i kratkovremensku energiju, kao i brzinu prolaska kroz nulu, te su i ove karakteristike prikazane na narednim graficima.

Na slici 52 prikazane su kratkovremenske energije signala za reč "signal", koje su redom izgovarali ženski govornik, muški govornik i dete.



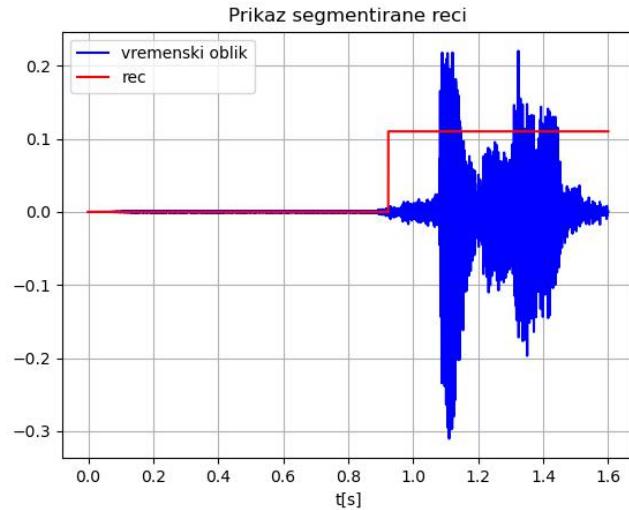
Slika 52: Kratkovremenske energije signala različitih govornika

Na slici 53 prikazane su brzine prolaska kroz nulu signala za reč "signal", koje su redom izgovarali ženski govornik, muški govornik i dete.



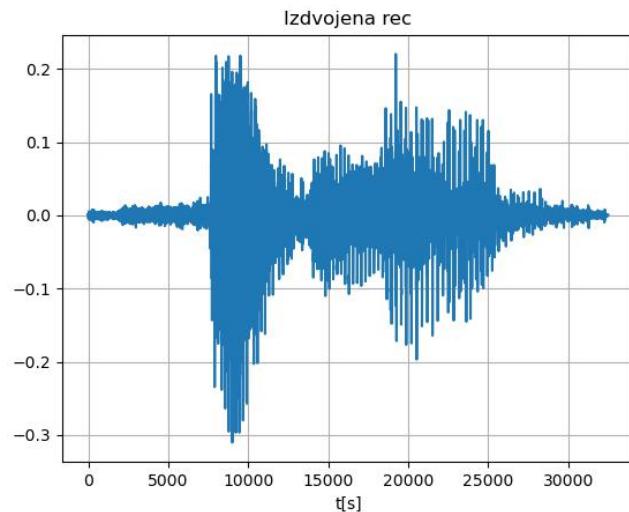
Slika 53: Brzine prolaska kroz nulu signala različitih govornika

Segmentacija reči sprovodi se gorepomenutim algoritmom koji koristi kratkovremensku energiju i brzinu prolaska kroz nulu. Na slici 54 prikazan je primer govornog signala za reč "signal" nakon primene Batervortovog filtera, koji se koristi za otklanjanje šuma. Na istom grafiku crvenom bojom je iscrtan pravougaonik, koji predstavlja granice za segmentaciju reči. Gornji prag odnosno  $ITU$  je postavljen na 5% od maksimalne energije, dok je donji prag odnosno  $ITL$  postavljen na 0.005% od maksimalne energije.



Slika 54: Primer segmentacije reči "signal"

Donja granica je ponovo relativno niska, što ukazuje na lepo filtriran signal. U ovom primeru izgovor reči je pri samom kraju snimka, i pravougaoni prozor za segmentaciju to tačno uspeva da isprati. Na slici 55 prikazana je segmentisana reč. Signal počinje niskom applitudom koja se povećava, što odgovara izgovoru sloga "sig", a nakon toga, applituda se smanjuje, ali signal ostaje aktivan, sa izraženim oscilacijama što odgovara prelazu na drugi slog "nal". Intenzitet signala zatim postepeno opada ukazujući na završetak izgovora reči.



Slika 55: Primer reči "signal" nakon segmentacije

### 5.2.11 Normalizacija

Normalizacija signala je važan korak u obradi signala kako bi se osigurala konzistentnost i pouzdanost rezultata. Cilj je transformacija podataka u standar-dizovan oblik, čime se smanjuje mogućnost da algoritam bude pristrasan prema

većim vrednostima. Takođe, normalizacija smanjuje varijabilnost između trening i test podataka, i time omogućava modelu da se bolje generalizuje na nove, neviđene podatke. Normalizacija po varijansi je metod normalizacije koji je korišćen u ovom radu. Ovaj pristup transformiše signale tako da svi imaju istu varijansu.

Normalizacija signala po varijansi se vrši prema sledećoj formuli:

$$x_{\text{norm}}[n] = \frac{x[n]}{\sigma_x}$$

gde je  $x[n]$  originalni signal, a  $\sigma_x$  standardna devijacija signala  $x[n]$ . Standardna devijacija signala  $x[n]$  se računa kao:

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{n=0}^{N-1} (x[n] - \mu_x)^2}$$

gde je  $N$  broj uzoraka signala, a  $\mu_x$  srednja vrednost signala  $x[n]$ , koja se računa kao:

$$\mu_x = \frac{1}{N} \sum_{n=0}^{N-1} x[n]$$

Primena normalizacije po varijansi ima nekoliko prednosti:

- Uklanjanje uticaja amplitude signala, što omogućava pravedno poređenje signala sa različitim amplitudama.
- Poboljšanje robusnosti algoritama za obradu signala, jer se svi signali dovode na istu skalu.
- Smanjenje efekta šuma i nepoželjnih varijacija u signalu, što može dovesti do preciznijih rezultata obrade.

### 5.2.12 Pre-emphasis filter

*Pre-emphasis* filter je linearni digitalni filter koji se koristi za poboljšanje kvaliteta govornog signala. Primarna svrha *pre-emphasis* filtera je da pojača komponente visokih frekvencija govornog signala u odnosu na niže frekvencije. Ovo pomaže da se kompenzuje prirodni pad energije govora na višim frekvencijama i poboljšava odnos signal-šuma (eng. *Signal to Noise Ratio-SNR*) za te frekvencije.

Standardni *pre-emphasis* filter se obično implementira kao FIR filter propusnik višokih učestanosti prvog reda. Matematička reprezentacija ovog filtera je data sa:

$$H(z) = 1 - \alpha z^{-1}$$

gde je  $\alpha$  koeficijent *pre-emphasis* filtera, obično u opsegu  $0.9 \leq \alpha \leq 1.0$

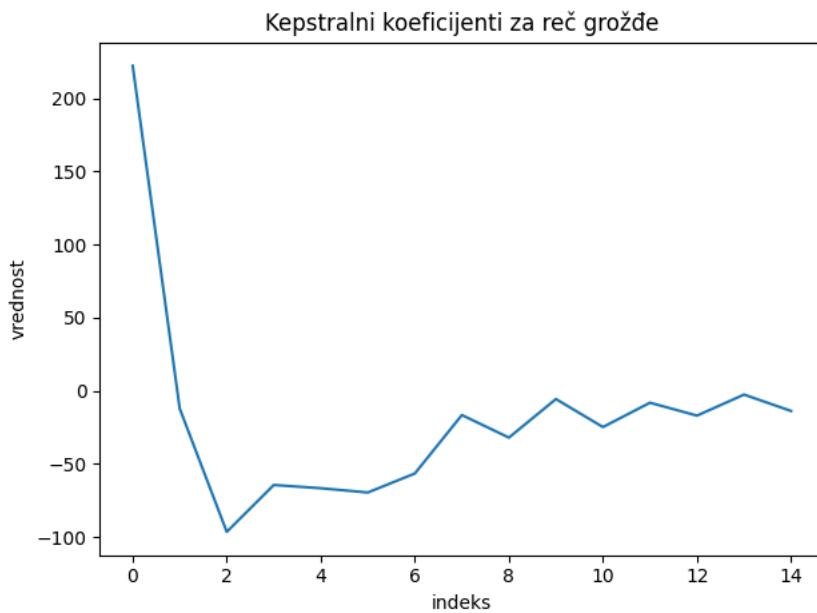
Svrha ovog filtera je da poveća amplitudu visokofrekventnih komponenti dok niskofrekventne komponente ostavlja relativno nepromenjenim. Ovo pomaže u smanjenju efekta spektralnog nagiba i olakšava kepstralnoj analizi da uhvati važne karakteristike govornog signala.[7]

### 5.3 Izdvajanje obeležja

Za ekstrakciju obeležja korišćena je kepstralna analiza, detaljno objašnjena u sekciji 3. Red kepstralne analize, odnosno broj kepstralnih koeficijenata direktno utiče na performanse modela prepoznavanja govora. Manji broj koeficijenata može dovesti do gubitka važnih informacija, što može smanjiti tačnost modela. Veći broj koeficijenata može dodati suvišne informacije i povećati složenost modela, što može rezultovati preobučavanjem (eng. *overfitting*). Tipičan broj kepstralnih koeficijenata je između 10 i 20, a u ovom radu je usvojen broj 15.

Polazi se od prepostavke da je govorni signal stacionaran, pa se za izračunavanje kepstralnih koeficijenata ne koristi prozorovanje. U daljem radu ispituje se kakvi se rezultati dobijaju kada se za izračunavanje koeficijenata koristi cela reč, zatim ukoliko se signal podeli na dva dela i odvojeno izračunaju koeficijenti za svaki deo, i isto ponovi za podelu na tri dela.

Na slici 56 prikazan je jedan primer izračunatih kepstralnih koeficijenata za reč grožđe.



Slika 56: Primer kepstralnih koeficijenata za reč "grožđe"

### 5.4 Klasifikacija oblika

Sledeći korak je izvršiti klasifikaciju oblika na osnovu obeležja koja su izračunata kepstralnom analizom. Klasifikacija se vrši korišćenjem algoritma  $k$  najbližih suseda. Baza podataka sa svim govornim signalima je podeljena na trening i test skup, gde su govornim signala dece stavljeni u test skup, kako bi sistem bio testiran na veće varijacije u izgovoru koje su prisutne kod dece. Tokom treniranja klasifikatora korišćena je pretraga hiperparametara sa petostrukom kros-validacijom (eng. *Grid Search*), kako bi se pronašli optimalni parametri za dati problem. Postavlja se mreža koja pretražuje sledeće hiperparametre:

- broj suseda (*n\_neighbors*): testira vrednosti 3, 5, 7 i 9
- težinsku funkciju (*weights*): testira slučaj kada svi susedi imaju istu težinu, bez obzira na udaljenost od tačke koja se klasificuje, i kada se daje prednost susedima koji su bliži trenutnoj tački
- metriku (*metric*): testira Euklidsku i Menhetn distancu

Najpre se za klasifikaciju koristi cela dužina govornih signala, i za svaki signal izračunavaju odgovarajući kepstralni koeficijenti. Na osnovu ovih koeficijenata formira se *kNN* klasifikator. Za nepodeljene gorovne signale, čija se koristi cela dužina, dobijaju se sledeći rezultati:

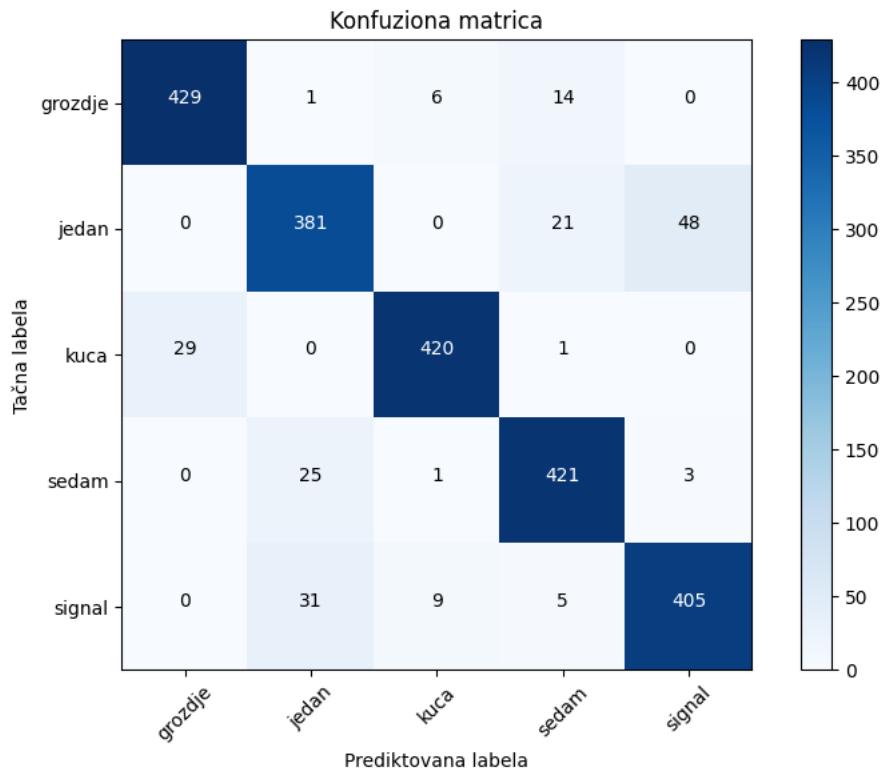
Optimalan broj suseda: 3

Optimalna težinska funkcija: funkcija težina distance

Optimalna metrika: Euklidska distanca

Tačnost klasifikacije: 91%

Na slici 57 prikazana je konfuziona matrica za ovu klasifikaciju.



Slika 57: Konfuziona matrica za nepodeljene signale

U narednom delu, metodologija će biti promenjena tako što će se svaki govorni signal podeliti na dva dela. Nakon podele, za svaki segment se izračunava 15 kepstralnih koeficijenata. Na kraju se kepstralni koeficijenti oba segmenta spajaju i formira se konačni skup obeležja. Na taj način, za svaku reč ćemo imati ukupno 30 obeležja. Ovakvom metodologijom dobijaju se sledeći rezultati:

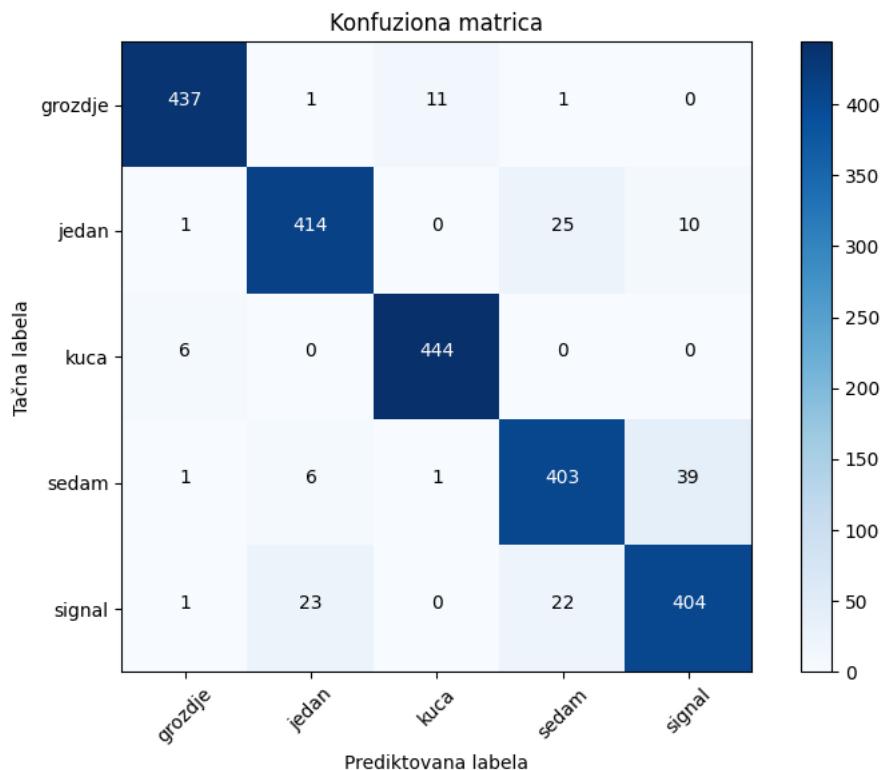
Optimalan broj suseda: 9

Optimalna težinska funkcija: funkcija težina distance

Optimalna metrika: Euklidska distanca

Tačnost klasifikacije: 93%

Na slici 58 prikazana je konfuziona matrica za ovu klasifikaciju.



Slika 58: Konfuziona matrica za signale podeljene na dva dela

Na posletku, isti postupak je primenjen za signale koji su podeljeni na tri dela i dobijeni su sledeći rezultati:

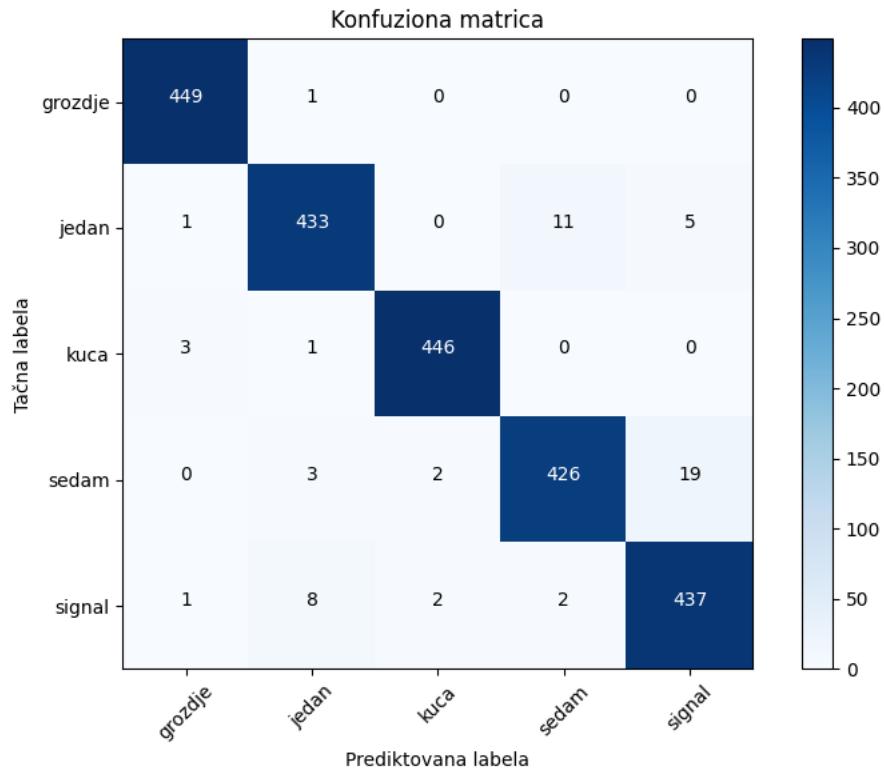
Optimalan broj suseda: 7

Optimalna težinska funkcija: funkcija težina distance

Optimalna metrika: Menhetn distanca

Tačnost klasifikacije: 97%

Na slici 59 prikazana je konfuziona matrica za ovu klasifikaciju.



Slika 59: Konfuziona matrica za signale podeljene na tri dela

Dakle, povećanje tačnosti modela pokazuje da je podela signala na segmente i kombinovanje kepstralnih koeficijenata iz različitih delova signala efikasan pristup za obogaćivanje skupa obeležja i povećanje performansi klasifikacionog modela. Ovakav pristup omogućava preciznije razlikovanje različitih govornih uzoraka i doprinosi boljoj generalizaciji modela na nove podatke.

## 6 ZAKLJUČAK

U ovom radu istraženo je prepoznavanje govora pomoću kepstralnih koeficijenata. Fokus je bio na analizi tačnosti prepoznavanja pet osnovnih reči: "jedan", "sedam", "grožđe", "signal" i "kuća". Primena Batervortovog i *pre-emphasis* filtera, normalizacija signala, *resampling-a* na 8 kHz i računanje kepstralnih koeficijenata omogućila je preciznu ekstrakciju relevantnih karakteristika govornih signala. Do- stignuta je tačnost od čak 97%.

Rezultati su pokazali da preemphasis filter značajno doprinosi poboljšanju kvaliteta signala uklanjanjem niskofrekventne komponente. Normalizacija signala i *resampling* su osigurali konzistentnost i uporedivost među različitim signalima. Kepstralni koefficijenti su se pokazali kao efikasna metoda za prepoznavanje govora, omogućavajući visok nivo tačnosti u klasifikaciji.

Metoda *kNN* (eng. *k-Nearest Neighbours*) se pokazala efikasnom za zadatak klasifikacije govornih signala. Ova metoda je koristila udaljenost između kepstralnih koeficijenata za prepoznavanje reči, omogućavajući visok nivo tačnosti uz relativno jednostavnu implementaciju. Ipak, tačnost modela zavisi od pravilnog izbora broja suseda *k* i karakteristika signala.

Jedna od ključnih tačaka ovog istraživanja bila je demonstracija nestacionarnosti govornog signala. Primena prozorovanja, odnosno podela reči na dva ili tri dela, te pronalaženje kepstralnih koeficijenata za te segmente, dala je znatno bolje rezultate. Ovaj pristup omogućava bolju detekciju lokalnih karakteristika signala koje su presudne za preciznu klasifikaciju. Tačnost klasifikacije nepodeljenih signala iznosila je 91%, dok je podela signala na dva dela podigla tačnost na 93%, a na tri čak na 97%. Rezultati su pokazali da prozorovanje znatno unapređuje performanse sistema za prepoznavanje govora, čineći ga otpornijim na varijabilnosti unutar govornog signala.

Ukupna dobijena tačnost klasifikacije je zadovoljavajuća, međutim zanimljivo je razmotriti kako tačnost klasifikacija varira među različitim rečima. Reč "signal" se pokazala kao najproblematičnija zbog velike varijabilnosti u izgovoru među govornicima, što je rezultiralo nižom tačnošću klasifikacije. Fonetska sličnost reči "jedan" i "sedam" takođe je izazvala konfuziju unutar modela, što je dovelo do pogrešnih klasifikacija i lošije tačnosti.

Dalje istraživanje može uključivati primenu alternativnih tehniki ekstrakcije obeležja, kao što su *LPC* (eng. *Linear Predictive Coding*) i *PLP* (eng. *Perceptual Linear Predictive*) koeficijenti. Takođe, korisno je ispitati primenu drugih metoda klasifikacije, kao što su *Random Forest* i *Decision Trees*, kako bi se utvrdilo koje metode pružaju najbolje rezultate za prepoznavanje govora. Implementacija i poređenje performansi dubokih neuralnih mreža, kao što su konvolucione neuralne mreže (eng. *Convolutional Neural Networks*) i rekurentne neuralne mreže (eng. *Recurrent Neural Networks*), može pružiti dublji uvid u efikasnost ovih metoda za prepoznavanje govora u odnosu na *kNN*.

Zaključno, ovo istraživanje je pokazalo značaj izbora metode ekstrakcije obelježja i tehnike klasifikacije za prepoznavanje govora. Dalje unapređenje i prilagođavanje modela mogu doprineti razvoju preciznijih i robusnijih sistema za prepoznavanje govora, što je od ključnog značaja za primenu u realnim uslovima.

## 7 LITERATURA

- [1] Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Prentice Hall.
- [2] Đurović, Ž. (2012). *Autorizovane beleške sa predavanja na predmetu Obrada i prepoznavanje govora 13E054OPG*. Univerzitet u Beogradu - Elektrotehnički fakultet.
- [3] Rabiner, L. R., & Schafer, R. W. (2007). *Introduction to digital speech processing (Foundations and Trends in Signal Processing)*. Now Publishers. Retrieved from <http://libgen.lc>
- [4] Đurović, Ž. (2012). *Autorizovane beleške sa predavanja na predmetu Prepoznavanje oblika 13E054PO*. Univerzitet u Beogradu - Elektrotehnički fakultet.
- [5] GeeksforGeeks. (n.d.). K-Nearest Neighbours. Retrieved July 17, 2024, from <https://www.geeksforgeeks.org/k-nearest-neighbours/>
- [6] Xu, Z., Santurkar, S., Tsipras, D., Yu, F., Kalai, A. T., & Madry, A. (2020). Intriguing properties of adversarial training at scale. *arXiv preprint arXiv:2002.03130*. Retrieved from <https://arxiv.org/pdf/2002.03130>
- [7] Yang, S., Zhang, Z., & Yamagishi, J. (2024). Pre-emphasis filters for speech enhancement. *arXiv preprint arXiv:2401.09315*. Retrieved from <https://arxiv.org/pdf/2401.09315>