Elektrotehnički fakultet, Univerzitet u Beogradu
Katedra za Signale i Sisteme

Milica Makević 0054/2018

Prepoznavanje govornika bazirano na Mel-frekvencijskim kepstralnim koeficijentima

Mentor:
prof. dr. Željko Đurović

Beograd
Septembar 2023

Rezime rada

Prepoznavanje govora ima dugu istoriju koja seže još do pedesetih godina prošlog veka. Rani sistemi su se uglavnom bazirali na jednostavnim akustičnim obeležjima koji nisu bili u mogućnosti da u potpunosti opišu govor. Jedan od najpoznatijih sistema iz ovog perioda je *Audrey* razvijen od strane Bel laboratorija. Već 1963. godine Bogert, Hili i Tuki su primetili da dalja spektralna analiza samog log-spektra može dati dobar uvid u osobine govora i definisali novi, *kepstralni*, domen [3]. Nešto kasnije, Pol Mermelštajn je na osnovu kepstra definisao Mel-frekvencijske kepstralne koeficijente. Ovi koeficijenti su ubrzo postala standardna obeležja u ovakvim sistemima. Danas je upotreba dubokog učenja sve češća praksa za potrebe prepoznavanja govora i govornika. Međutim i dalje je česta praksa da se Mel kepstralni koeficijenti koriste kao obeležja koja se prosleđuju modelu.

Ovaj rad će dati pregled relevantne teorijske pozadine Mel-frekvencijskih kepstralnih koeficijenata kao i neparametarskih metoda klasifikacije. Zatim će na bazi sa sedam različitih govornika biti testirana uspešnost ovakvog modela. Takođe će biti razmatran uticaj različitih načina za ekstrakciju koeficijenata na uspešnost klasifikacije.

Zahvalnica

Zahvaljujem se prof. dr. Željku Đuroviću na idejama, savetima i mentorstvu tokom pisanja ovog rada. Takođe se zahvaljujem svim profesorima, asistentima i saradnicima sa katedre za Signale i Sisteme od kojih sam imala priliku da učim. Znanje koje ću poneti sa sobom na kraju ovih studija je neprocenjivo.

Zahvaljujem se prijateljima i porodici na učestvovanju u ovom radu i na beskrajnom strpljenju i podršci koju ste mi ukazali. Bez vas ovaj rad ne bi mogao biti završen.

Posebno se zahvaljujem svojim roditeljima što su mi omogućili školovanje i verovali u mene do samog kraja.

Spisak slika

Sadržaj

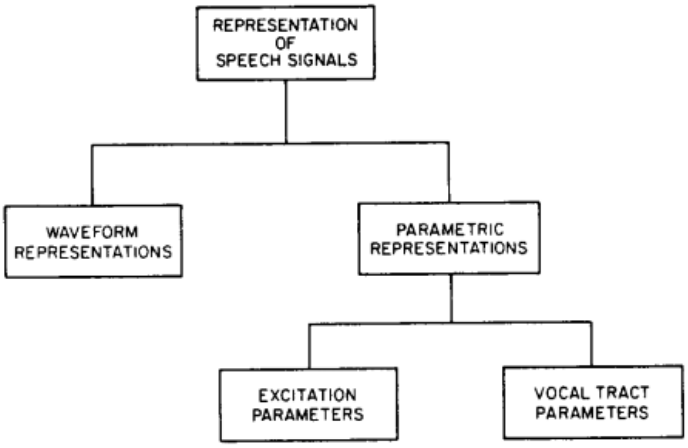
Glava 1

Uvod

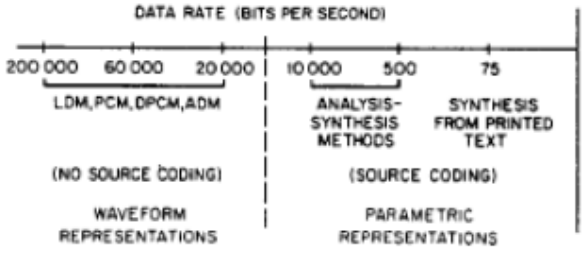
Značaj obrade govora leži u mnogobrojnim i raznovrsnim oblastima. Neke od njih su:

- Bezbednosni sistemi - Govor je biometrijsko obeležje čoveka, te se kao takvo može koristiti za potvrdu (verifikaciju) identiteta osobe.
- Forenzičke primene - Slično prethodnom, u forenzici se neretko koriste sistemi za identifikaciju govornika, kako bi se utvrdilo koja od osoba iz skupa je zapravo proizvela govor.
- Pomoć osobama sa invaliditetom - Neke od najznačajnijih primena u ovoj oblasti su konverzija govornog signala u tekst za osobe sa oštećenim sluhom, i obrnuto, konverzija teksta u govorni signal, tj sinteza govora, za osobe sa oštećenim vidom. Prepoznavanje govora je takođe od pomoći osobama sa ograničenom mobilnošću, i omogućava lakše korišćenje računara i sličnih uređaja.
- Poboljšanje kvaliteta signala - Govorni signal se prilikom prenosa značajno degradira. Obradom signala moguće je popraviti njegov kvalitet, npr. odstranjvanjem eha iz signala ili otklanjanjem šuma.
- Virtuelni asistenti - Prepoznavanje i sinteza govora su esencijalni za tehnologije kao što su *Siri*, *Alexa* i *Google Assistant*.

Prvi korak u (digitalnoj) obradi signala je uvek pogodna reprezentacija govora u digitalnom obliku. Krajnja primena obrade govora zapravo diktira šta je najefikasniji način predstavljanja signala, a neka opšta podela je data na slici 1.1. Prvi način je naravno, prestavaljanje signala u vidu samog akustičkog talasa pomoću odabiranja i kvantizacije signala. Alternativa je parametarska reprezentacija koja podrazumeva parametre vezane za eksitaciju kao i parametre vezane za sam vokalni trakt. Prednost parametarske reprezentacije je daleko veća efikasnost - oni se menjaju daleko sporije nego odbirci akustičkog talasa te je moguće diskretizaciju raditi sa daleko manjom periodom odabiranja (slika 1.2) [7]. Neki od primera parametarske reprezentacije signala su LPC koeficijenti, kepstralni koeficijenti, usrednjena energija signala, formanti itd.



Slika 1.1: Reprezentacija govornog signala [7]



Slika 1.2: Uobičajene brzine odabiranja, zavisno od vrste reprezentacije signala [7]

1.1 Metodologija rada

Ideja iza ovog rada je da predstavi korisnost Mel-frekvencijskih kepstralnih koeficijenata za potrebe prepoznavanja govornika. Razmatrana je uspešnost klasifikatora za podatke različite prirode, kao i za određene modifikacije pri ekstrakciji koeficijenata.

Sam klasifikator, kao i ekstrakcija obeležja implementirani su ručno u jeziku *Python* koristeći primarno biblioteke *NumPy* zbog efikasne implementacije potrebnih matematičkih funkcionalnosti, kao i biblioteke *Matplotlib* i *Seaborn* korišćene radi vizuelizacije rezultata. Za efikasno eksperimentisanje i logovanje rezultata korišćen je *MLflow*.

1.1.1 Formiranje baze podataka

Postoje tri baze podataka sa kojima je rađena klasifikacija. Snimanje je obavljeno pomoću softvera *Audacity* i komercijalnog mikrofona sa periodom odabiranja od $16kHz$. Snimanje je uvek bilo obavljano sa što manje pozadinske buke istim mikrofonom.

Prva baza se sastoji od devet govornika. Govornici su snimani kako izgovaraju reč *'padati'* dvadeset puta. Nakon izolovanja reči iz originalnog snimka, deo je iskorišćen za set za obučavanje a deo kao set za testiranje.

Kasnije je snimanje ponovljeno opet u nameri da se novo snimljene reči koriste kao test skup. Ideja je bila izbeći da svi odbirci dolaze sa istog snimka. Nažalost nije bilo moguće obaviti ponovljeno snimanje svih devet govornika, jer ih je samo pet bilo dostupno. Zbog ovoga su napravljene još dve baze. Druga baza je zapravo podskup prve, to jest, umesto svih devet govornika u drugoj se nalazi podaci o onih pet koji su kasnije iznova snimljeni. Treća baza sadrži isti obučavajući skup kao druga, ali test skup potiče sa ponovljenog snimanja.

Na ovaj način bilo je moguće uporediti rezultate identifikacije sa različitim brojem govornika u bazi, kao i rezultate kada su svi odbirci iz istog izvora, naspram kada su test odbirci naknadno prikupljeni.

Informacije o govornicima se mogu videti u tabeli 1.1.

redni broj govornika	godine	pol
1	20-25	ž
2	20-25	ž
3	20-25	ž
4	20-25	ž
5	20-25	m
6	20-25	m
7	55-60	m
8	25-30	m
9	50-55	ž

Tabela 1.1: Osnovni podaci o govornicima

Nakon snimanja svi audio snimci su čuvani u *.wav* formatu.

Glava 2

Teorijska pozadina

2.1 Govorni trakt

Govorni trakt (slika 2.1) se sastoji od više organa koji se, zavisno od njihove uloge, mogu podeliti na generatore i modulatore govora. Generatori dalje ubrajaju induktore - pluća, dijafragmu i traheje (ovaj deo se naziva još i subglotalnim sistemom), i fonatore - grkljan (lat. *larynx*), glasne žice i glotis. U modulatore ubrajamo resicu (lat. *velum*), farinks, jezik, zube, vilicu, nepce i usne.

Pluća, uz dijafragmu, prilikom izdisaja snadbevaju čitav trakt vazduhom koji se potom prenosi do glotisa i glasnih žica. Glotis je zapravo otvor između glasnih žica i oni zajedno predstavljaju deo grkljana. Vibriranjem glasnica prilikom prolaska vazduha javlja se akustički talas, tj. zvuk kojeg kasnije dodatno oblikuju modulatori. Prilikom izgovora nekih glasova, pored usne duplje, u formiranju učestvuje i nosna duplja tako što se resica spusti [5].

Dužina čitavog trakta se kreće od $13cm$ do $18cm$, pri čemu je on nešto kraći kod žena. Poprečni presek se menja prilikom govora i može se kretati od $0cm^2$ do $20cm^2$.

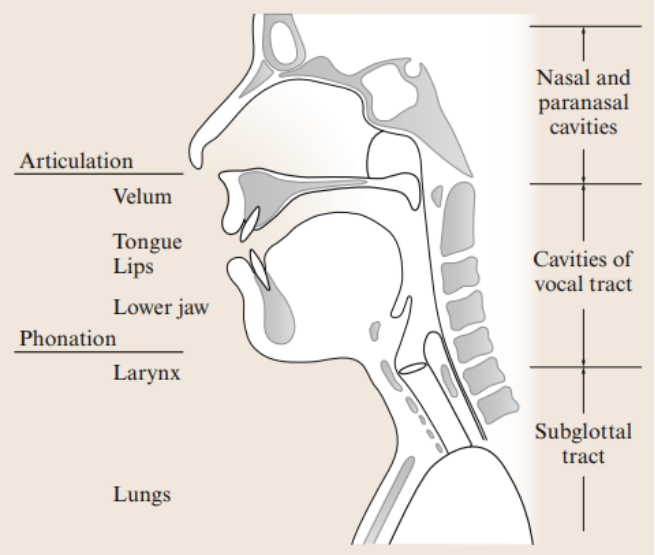
2.2 Foneme

Fonema ili glas je najmanja jezička jedinica u jeziku i njihova podela u srpskom jeziku se može videti na slici 2.2. Prilikom generisanja samoglasnika (vokala) pobuda se može smatrati kvazi-periodičnom. Eksitacija u slučaju frikativa je širokopojasni šum, dok je za plozive to impulsni signal [7].

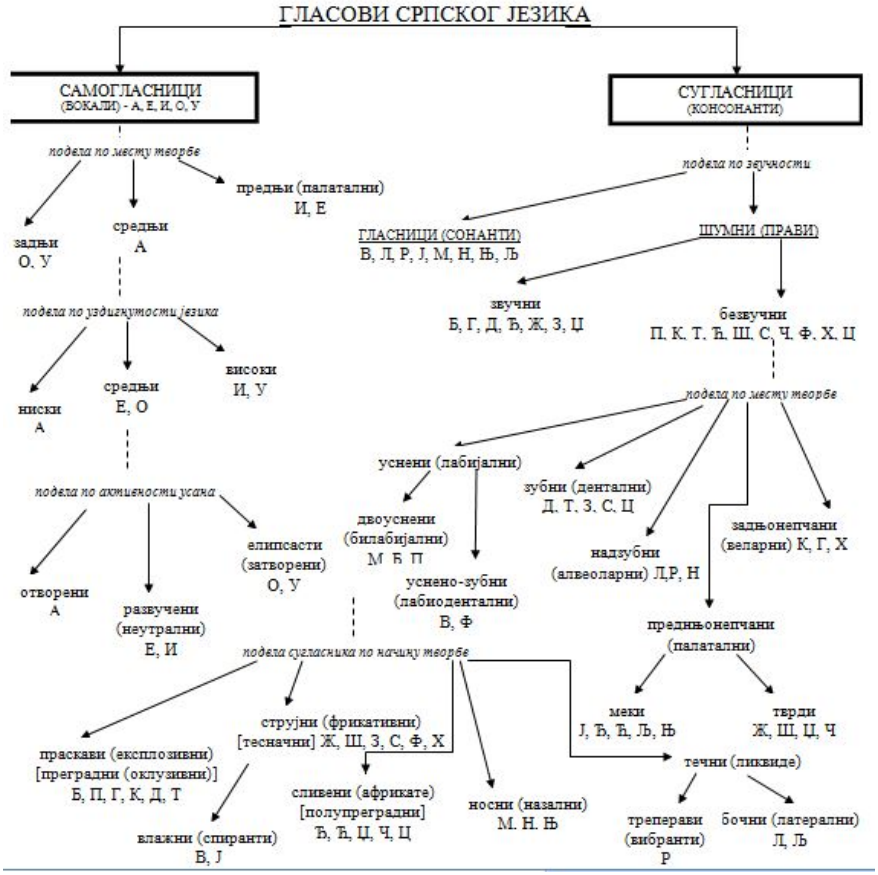
2.3 Matematički model govornog trakta

Zakoni održanja mase, održanja energije i održanja momenta, kao i zakoni termodinamike i mehanike fluida su, uz određene aproksimacije, podloga za formiranje matematičkog modela vokalnog trakta. Pri analizi se i naredni fenomeni moraju uzeti u obzir:

a. vremenski promenljiva priroda oblika vokalnog trakta,



Slika 2.1: Osnovni delovi govornog trakta



Slika 2.2: Podela glasova u srpskom jeziku

b. gubici nastali usled toplotne kondukcije i viskoznog trenja vazuha duž zidova trakta,

c. radijacija zvuka sa usana,

d. čvrstina zidova vokalnog trakta,

e. eksitacija zvuka na početku trakta,

f. učestvovanje nosne duplje.

Za učestanosti manje od $4000Hz$ može se smatrati da je talasna dužina zvuka veća od poprečne dimenzije vokalnog trakta. Uz ovakvo ograničenje vokalni trakt je moguće predstaviti kvazi-jedno-direkcionim modelom. Ovi modeli se koriste u raznim oblastima nauke da opišu fizičke sisteme koji imaju jedan dominantan pravac. Suštinski oni pojednostavljuju problem svodeći ga na jednu dimenziju. Ako se takođe zanemare gubici usled konduktivnosti i trenja, a uzmu u obzir zakoni održanja mase, energije i momenta dolazimo do sledećeg modela [6]:

$$\begin{aligned} -\frac{\partial p}{\partial x} &= \rho \frac{\partial (\frac{u}{A})}{\partial t} \\ -\frac{\partial u}{\partial x} &= \frac{1}{\rho c^2}\frac{\partial (\rho A)}{\partial t} + \frac{\partial A}{\partial t} \end{aligned} \tag{2.1}$$

Gde je:

- c brzina zvuka,
- ρ gustina vazduha u traktu,
- $A = A(x,t)$ površina poprečnog preseka trakta na poziciji x u trenutku t,
- $p = p(x,t)$ vazdušni pritisak na poziciji x u trenutku t,
- $u = u(x,t)$ zapreminski vazdušni protok na poziciji x u trenutku t.

Rešenje u zatvorenoj formi, u opštem slučaju, nije dostupno, te se do njega dolazi numeričkim metodama. Da bi se došlo do definitivnog rešenja neophodno je poznavati vrednosti pritiska i protoka na nultoj i krajnjoj poziciji trakta. Za nultu poziciju se smatraju glasne žice, a za krajnju l-tu (gde je l dužina govornog trakta) usne. Takođe je neophodno poznavati površinu poprečnog preseka tokom vremena.

Nažalost, izuzetno je teško pratiti promenu poprečnog preseka u vremenu. Čak i kad bi on u potpunosti bio poznat, zahtevno je rešiti sistem jednačina (2.1). Srećom u praksi je sasvim dovoljno koristiti veoma jednostavne modele koje je moguće rešiti u zatvorenoj formi. Jedan od takvih modela je **model uniformne tube bez gubitaka** [7].

2.3.1 Model uniformne tube

Model uniformne tube podrazumeva da se poprečni presek ne menja tokom vremena i da je $A(x,t) = A, \forall x$. Dalje pretpostavka je da se pritisak na usnama ne menja i da je konstanta. Ovim se jednačine (2.1) svode na:

$$\begin{aligned} -\frac{\partial p}{\partial x} &= \frac{\rho}{A}\frac{\partial u}{\partial t} \\ -\frac{\partial u}{\partial x} &= \frac{A}{\rho c^2}\frac{\partial \rho}{\partial t} \end{aligned} \tag{2.2}$$

Ako predstavimo $p(x,t)$ i $u(x,t)$ kao kompleksne sinusoide

$$\begin{aligned} p(x,t) &= P(x,\Omega)e^{j\Omega t} \\ u(x,t) &= U(x,\Omega)e^{j\Omega t} \end{aligned} \tag{2.3}$$

parcijalne jednačine nam se svode na:

$$\begin{aligned} -\frac{\mathrm{d}P}{\mathrm{d}x} &= ZU \\ -\frac{\mathrm{d}U}{\mathrm{d}x} &= YP \end{aligned} \tag{2.4}$$

gde su Z i Y redom *akustička impendansa* i *akustička admitansa* i iznose

$$\begin{aligned} Z &= j\Omega\frac{\rho}{A} \\ Y &= j\Omega\frac{A}{\rho c^2} \end{aligned} \tag{2.5}$$

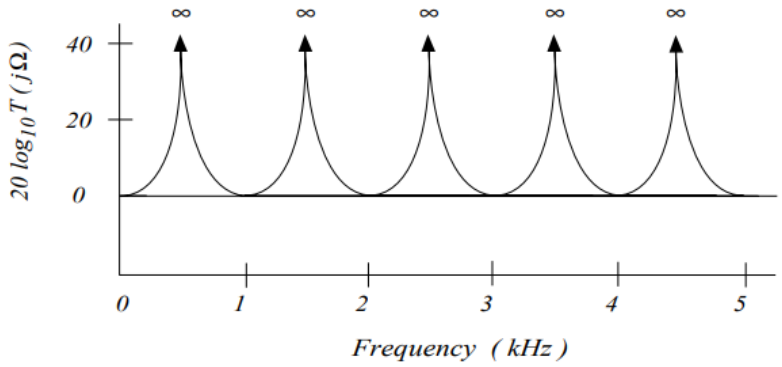
Sada nam se rešenje jednačina (2.4) javlja u formi:

$$\begin{aligned} P(x,\Omega) &= Ae^{\gamma x} + Be^{-\gamma x} \\ U(x,\Omega) &= De^{\gamma x} + Ce^{-\gamma x} \end{aligned} \tag{2.6}$$

gde je: $\gamma = \sqrt{ZY} = j\frac{\Omega}{c}$. Koeficijente nalazimo zahvaljujući graničnim uslovima

$$\begin{aligned} P(l,\Omega) &= 0 \\ U(0,\Omega) &= U_G(\Omega) \end{aligned} \tag{2.7}$$

Tako da nam se rezultat svodi na:



Slika 2.3: Frekvencijski odziv za uniformnu tubu bez gubitaka [2]

$$p(x,t) = jZ_0 \frac{\sin \Omega(l-x)c}{\cos \frac{\Omega l}{c}} U_G(\Omega) e^{j\Omega t}$$

$$u(x,t) = \frac{\cos \frac{\Omega(l-x)}{c}}{\cos \frac{\Omega l}{c}} U_G(\Omega) e^{j\Omega t} \tag{2.8}$$

gde je $Z_0 = \frac{\rho c}{A}$. Posmatrajmo sad zapreminski protok vazduha na usnama:

$$\begin{aligned} u(l,t) &= U(l,\Omega) e^{j\Omega t} \\ &= \frac{1}{\cos \frac{\Omega l}{c}} U_G(\Omega) e^{j\Omega t} \end{aligned} \tag{2.9}$$

Količnik

$$\frac{U(l,\Omega)}{U_G(\Omega)} = V_a(j\Omega) = \frac{1}{\cos \frac{\Omega l}{c}} \tag{2.10}$$

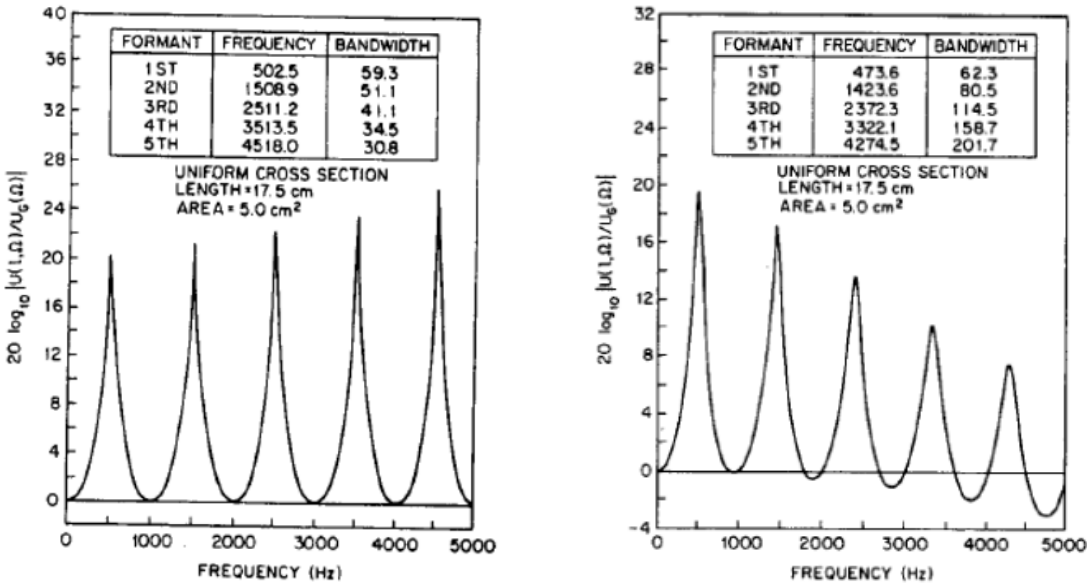
je frekvencijski odziv uniformne tube bez gubitaka. Na slici 2.3 se može videti ova funkcija za dužinu trakta $l = 17,5cm$ i brzinu zvuka $c = 35000\frac{cm}{s}$. Formanti imaju beskonačnu amplitudu i nalaze se tačno na učestanostima od $500Hz$, $1000Hz$,..

2.3.2 Uticaj gubitka energije i radijacije na usnama

Prethodno izvedeni rezultat podrazumeva da nema gubitaka u tubi, što je nerealistična situacija. Usled uticaja trenja, toplotne kondukcije i vibracija zidova trakta tokom govora gubici su neminovni. Ukoliko bi se u obzir uzeli ovi efekti

došlo bi se do frekvencijskog odziva prikazanog na slici 2.4a. Primetiti da formanti nisu više ekvidistantni (pomereni su ka višim učestanostima), kao i da im pikovi nisu pod jednakih širina. Takođe, same amplitude više nisu beskonačno velike. Uticaj ovih gubitaka je, očigledno, više izražen na nižim učestanostima.

U obzir bi takođe trebalo uzeti i radijaciju sa usana, jer je uslov koji smo inicijalno pretpostavili - $p(l,t) = 0$, nerealan. Uticaj ovog fenomena je daleko više izražen na višim učestanostima, što se može i videti na slici 2.4b [7].



(a) uz modeliranje gubitaka usred trenja, toplotne kondukcije i vibracija zidova

(b) uz modeliranje radijacije na usnama

Slika 2.4: Frekvencijski odziv vokalnog trakta [7]

2.4 Kratkovremenska analiza govora

Govorni signal nije stacionaran, tj. njegove osobine se menjaju sa vremenom. Ono što nam omogućava obradu signala i ekstrakciju obeležja u vremenskom domenu je takozvana *kratkovremenska analiza*. Naime, osobine govora se tokom vremena menjaju dovoljno sporo da na veoma kratkim delovima signal možemo smatrati stacionarnim.

Opšti oblik svih kratkovremenskih metoda se može napisati kao:

$$Q_n = \sum_{m=-\infty}^{\infty} T[x(m)]w(n-m) \tag{2.11}$$

gde je T transformaciona funkcija a $w(n)$ prozorksa funckija. Ova jednačina zapravo predstavlja diskretnu konvoluciju prozorske funkcije sa signalom, te se Q_n može smatrati i izlazom linearnog, vremenski invarijantnog (eng. *Linear Time-Invariant, LTI*) sistema.

Pri izboru prozorske funkcije postavlja se pitanje njenog oblika, kao i njene dužine. Ukoliko je prozor predugačak neće biti u mogućnosti da obuhvati promenljive osobine signala kako treba. S druge strane, ukoliko je prekratak osobine će previše oscilovati. U praksi se pokazalo najboljim da prozor bude dužine $1-3$ pitch periode

$$\begin{aligned} NT &= (1 \div 3)T_{pitch} \\ N &= (1 \div 3)\frac{T_{pitch}}{T} \\ &= (1 \div 3)\frac{F_S}{F_{pitch}} \end{aligned} \tag{2.12}$$

gde je N dužina prozora u odbircima, $T = \frac{1}{F_s}$ perioda odabiranja i $T_{pitch} = \frac{1}{F_{pitch}}$ pitch perioda.

Kada je u pitanju oblik prozora, neki od tipičnih izbora za obradu govora su pravougaona, *Hann*-ova i *Hamming*-ova prozorska funkcija (Slika 2.5):

Pravougaona prozorska funkcija:

$$h(n) = \begin{cases} 1, & 0 \leq n \leq N-1 \\ 0, & \text{inače} \end{cases} \tag{2.13}$$

Hann-ova prozorska funkcija:

$$h(n) = \begin{cases} 0.5(1 - \cos\frac{2\pi n}{N}), & 0 \leq n \leq N \\ 0, & \text{inače} \end{cases} \tag{2.14}$$

Hamming-ova prozorska funkcija

$$h(n) = \begin{cases} 0.54 - 0.46\cos\frac{2\pi n}{N-1}, & 0 \leq n \leq N-1 \\ 0, & \text{inače} \end{cases} \tag{2.15}$$

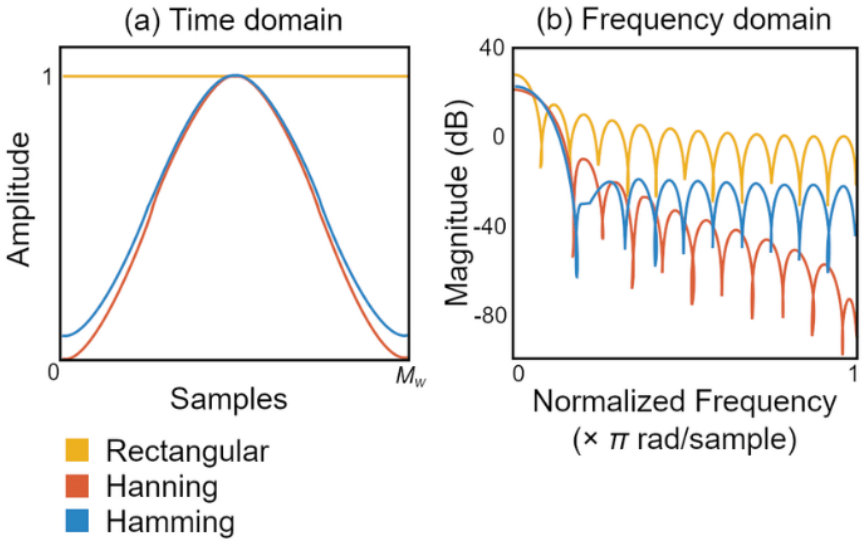
Može se primetiti da *Hann*-ova i *Hamming*-ova funkcija imaju širi glavni lob od pravougaone, ali i daleko jače slabljenje signala van propusnog opsega [7].

2.4.1 Kratkovremenska Furijeova transformacija

Jedna od najpoznatijih metoda za dvodimenzionalnu, vremensko-frekvencijsku, predstavu govornog signala je kratkovremenska Furijeova transformacija (eng. *Short-Time Fourier Transform, STFT*). Kako se osobine signala menjaju tokom vremena, ideja je da se spektar procenjuje na segmentima signala uz pomoć prozorske funkcije.

Razmatrajući diskretne signale, za neki ulaz $x[n]$, segmenti koje razmatramo su dati sa:

$$x_l[n] = w[n]x[n + lL], \quad 0 \leq n \leq N-1 \tag{2.16}$$



Slika 2.5: Poređenje pravougaone, *Hann*-ove i *Hamming*-ove prozorke funkcije [1]

gde je N dužina prozora, l indeks odgovarajućeg prozora a L veličina pomeraja. Praksa je, kako i za kratkovremensku Furijeova transformaciju tako i za ostale metode ovog tipa, da se prozori delimično preklapaju, tj. $L < N$.

Za svaki segment traži se diskretna Furijeova transformacija (eng. *discrete Fourier transform, DFT*):

$$\begin{aligned} X[k,l] &= \sum_{n=0}^{N-1} x_l[n] e^{-i2\pi n \frac{k}{K}} \\ &= \sum_{n=0}^{N-1} w[n] x[n+lL] e^{-i2\pi n \frac{k}{K}} \end{aligned} \tag{2.17}$$

gde je K broj tačaka u kojima je rađena transformacija, a k frekvencijski indeks.

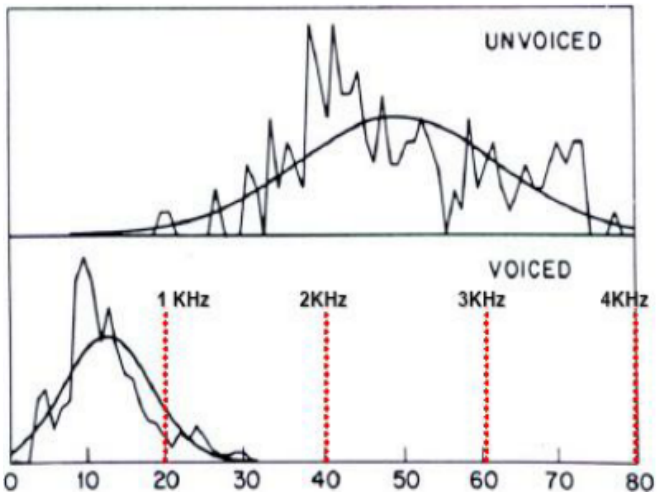
Kratkovremenska Furijeova transformacija se vizuelno prikazuje pomoću spektrograma, koji se najčešće računa kao $S[f,n] = |X_{STFT}[f,n]|^2$. Mana spektrograma je što, ako je prozor prešiirok on daje dobru frekvencijsku rezoluciju (lakše je razdvojiti bliske frekvencijske komponente) a lošu vremensku, i obrnuto. [5]

2.4.2 Kratkovremenska energija signala

Jedna od najbitnijih metoda u ovom domenu je kratkovremenska energija signala. Naime, energija diskretnog signala se definiše kao

$$E = \sum_{m=-\infty}^{\infty} x^2(m) \tag{2.18}$$

što potpuno očekivano nije preterano informativno za govorni signal. Zbog toga se definiše energija na segmentu signala kao



Slika 2.6: Raspodela kratkovremenske brzine prolaska kroz nulu za zvučni i bezvučni govor [7]

$$E_n = \sum_{m=-\infty}^{\infty} x^2(m)h(n-m) \tag{2.19}$$

gde je $h(n) = w^2(n)$.

Značaj kratkovremenske energije je u tome što se može koristiti za distinkciju zvučnog od bezvučnog dela signala. Ukoliko je SNR visok moguće je vršiti i diskriminaciju govora od tišine (eng. *speech vs. silence discrimination*).

2.4.3 Kratkovremenska brzina prolaska kroz nulu

Ideja ove metode je procena frekvencijskog sadržaja signala. Zvučni govor je mahom koncetrisan ispod $3kHz$, dok je bezvučni deo govora na višim učestanostima. Drugim rečima, ukoliko je kratkovremenska brzina prolaska kroz nulu niska možemo pretpostaviti da se radi o zvučnom govoru, i obrnuto (slika 2.6). Definiše se kao:

$$Z_n = \sum_{m=-\infty}^{\infty} |sgn[x(m)] - sgn[x(m-1)]|w(n-m) \tag{2.20}$$

gde je $sgn[x(n)] = 1$ za $x(n) \geq 0$ i $w(n) = \dfrac{1}{2N}$ za $0 \leq n \leq N-1$.

Kao i kratkovremenska energija, i ova metoda se može koristiti za distinkciju zvučnih i bezvučnih delova. Međutim, mana joj je što je izuzetno osetljiva kako na DC komponentu signala usled ADC konvertora, tako i na šum. Praksa je da se pre njene primene izvrši filtriranje signala kako bi se odstranila DC komponenta i šum na $50Hz$ [7].

2.4.4 Segmentacija signala na govor i tišinu

Čest korak u obradi govornog signala je izdvajanje delova u kojima se zapravo javlja govor. Glavni izazovi su pauze različitog trajanja između reči, kao i određeni glasovi kada se jave na početku ili kraju reči. Problematični su nazali kada se nađu na kraju, frikativi (f,h) i slabi plozivi (p,t,k) jer ih je teško razlikovati od pozadinskog šuma usled male snage.

U praksi se za rešavanje ovog problema često koristi kombinacija kratkovremenske energije i kratkovremenskog prolaska kroz nulu. Algoritam je sledeći [7]:

1. Za dati signal se odredi kratkovremenska energija i kratkovremenska brzina prolaska kroz nulu.

2. Postavi se veoma strog prag energije za koji se je sigurno da, ako energija prelazi prag, govor postoji.

3. Odredi se manje strog prag za koji se može sigurno reći da, ako energija padne ispod njega, govor ne postoji. Onda se prethodno nađene granice govora u tački 2. šire tako što se pomeraju sve dok energija ne padne ispod nižeg praga.

4. Na samom kraju rade se dodatna poboljšanja pomoću brzine prolaska kroz nulu. Posmatra se 25 prozorskih funkcija pre i nakon granica pronađenih u tački 3. Ukoliko brzina prolaska kroz nulu preseče unapred određen prag (sa histograma na slici 2.6 se mogu zaključiti prikladne vrednosti) nekoliko puta za nov početak, odnosno kraj, se uzima trenutak kada je prvi put pala ispod praga.

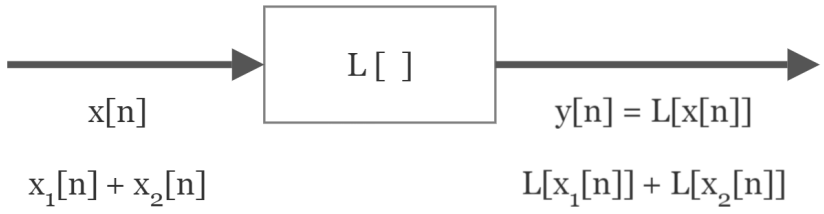
2.5 Homomorfna analiza

Kao što je već pomenuto, govor se može prikazati kao izlaz iz linearnog, vremenski promenljivog sistema čije se osobine menjaju sporo u vremenu. Što je dalje vodilo ka pretpostavci da se kratki segmenti signala mogu modelovati kao izlazi LTI sistema pobuđeni odgovarajućom eksitacijom. S obrzirom na to da je zlaz sistema zapravo konvolucija eksitacije i impulsnog odziva sistema, analiza govora se može posmatrati i iz ugla razdvajanja, tj. *dekonvolucije* tih komponenata [7].

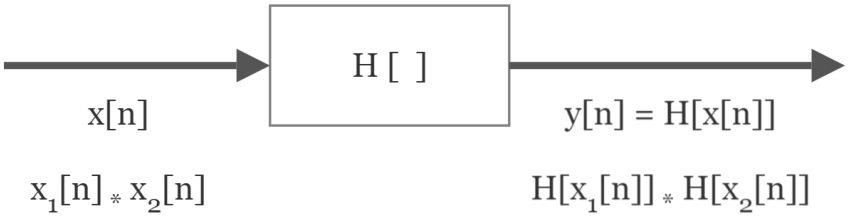
2.5.1 Homomorfni sistemi

Princip superpozicije linearnih sistema (slika 2.7), za neki linearni operator L, glasi

$$\begin{aligned} L[x(n)] &= L[x_1(n) + x_2(n)] \\ &= L[x_1(n)] + L[x_2(n)] \\ &= y_1(n) + y_2(n) \\ &= y(n) \end{aligned} \tag{2.21}$$



Slika 2.7: Sistem koji podleže principu superpozicije

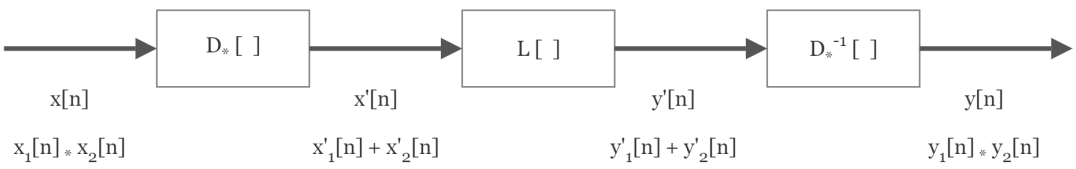


Slika 2.8: Homomorfni sistem za konvoluciju

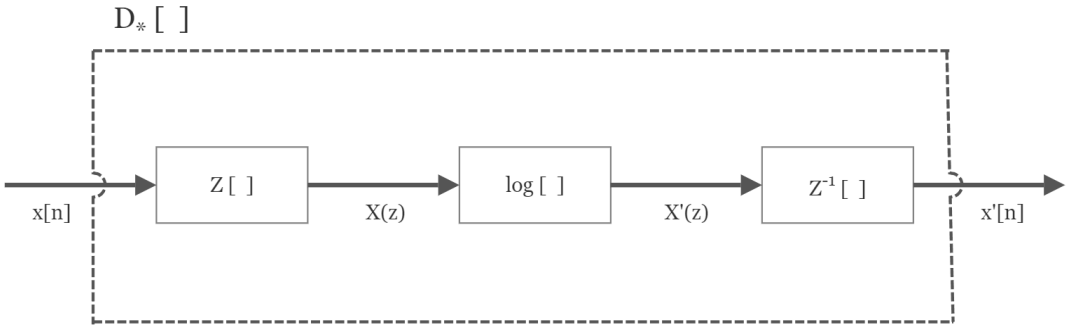
Definišimo podgrupu linearnih sistema koji podležu principu superpozicije gde je sabiranje zamenjeno konvolucijom (slika 2.8). Ovakve sisteme nazivamo *homomorfnim sistemima za konvoluciju.*

$$\begin{aligned} H[x(n)] &= H[x_1(n) * x_2(n)] \\ &= H[x_1(n)] * H[x_2(n)] \\ &= y_1(n) * y_2(n) \\ &= y(n) \end{aligned} \tag{2.22}$$

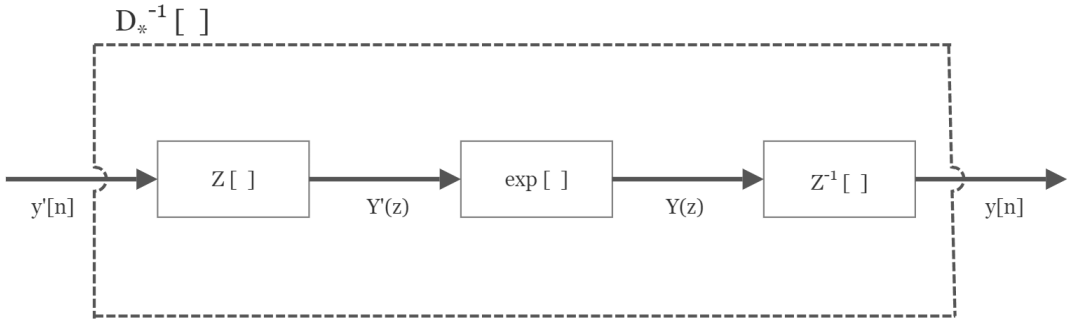
Homomorfni filter je zapravo samo homomorfni sistem kroz koji jedna komponenta prolazi nepromenjena dok nepoželjna biva potisnuta. Svaki homomorfni sistem se može prikazati u vidu tri kaskadna homomorfna sistema (slika 2.9). Prvi se naziva karakterističnim sistemom za homomorfnu dekonvoluciju (slika 2.10a). On uzima ulaze kombinovane konvolucijom i na izlazu daje odgovarajuću aditivnu kombinaciju. Drugi sistem je standardni linerani sistem koji podleže principu superpozicije iz jednačine (2.21). Treći sistem je inverzan prvom (slika 2.10b), za aditivno kombinovane ulaze daje izlaz kombinovan konvolucijom.



Slika 2.9: Homomorfni sistem za dekonvoluciju



(a) Karakteristični sistem za homomorfnu dekonvoluciju



(b) Inverzni karakteristični sistem za homomorfnu dekonvoluciju

Slika 2.10: Karakteristični sistem

Karakteristični sistem je definisan sa:

$$\begin{aligned} D_*[x(n)] &= D_*[x_1(n) * x_2(n)] \\ &= D_*[x_1(n)] + D_*[x_2(n)] \\ &= \hat{x}_1(n) + \hat{x}_2(n) \\ &= \hat{x}(n) \end{aligned} \tag{2.23}$$

Dok je inverzni definisan kao:

$$\begin{aligned} D_*^{-1}[\hat{y}(n)] &= D_*^{-1}[\hat{y}_1(n) + \hat{y}_2(n)] \\ &= D_*^{-1}[\hat{y}_1(n)] * D_*^{-1}[\hat{y}_2(n)] \\ &= y_1(n) * y_2(n) \\ &= y(n) \end{aligned} \tag{2.24}$$

Posebno je potrebno obratiti pažnju na logaritamsku funkciju. Ona mora biti definisana tako da je logaritam proizvoda jednak sumi logaritama:

$$\begin{aligned} \hat{X}(z) &= \log[X(z)] = \log[X_1(z)X_2(z)] \\ &= \log[X_1(z)] + \log[X_2(z)] \end{aligned} \tag{2.25}$$

Ovo je trivijalno za realne vrednosti, ali u opštem slučaju Z transformacija je kompleksna veličina, te je i logaritam kompleksni:

$$\hat{X}(e^{j\omega}) = \log|X(e^{j\omega})| + j\arg[X(e^{j\omega})] \tag{2.26}$$

Jednačina (2.25) je validna za vrednosti na jediničnom krugu $z = e^{jw}$, tj. kada se umesto Z-transformacije koristi Furijeova transformacija, što i jeste slučaj u praksi.

Izlaz karakterističnog sistema, tj inverzna Furijeova transformacija kompleksnog logaritma, je definisan sa:

$$X(e^{jw}) = \sum_{n=0}^{N-1} x(n)e^{-jwn} \tag{2.27}$$

$$\hat{X}(e^{jw}) = \log[X(e^{jw})] \tag{2.28}$$

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{X}(e^{jw})e^{jwn}d\omega \tag{2.29}$$

i naziva se kompleksnim kepstrom (eng. *complex cepstrum*). Termin 'kompleksni' dolazi iz činjenice da se koristi kompleksni logaritam, sama sekvenca je realna.

Parni deo kompleksnog kesptra, koji zapravo predstavlja inverznu Furijeovu transformaciju logaritma modula Furijeove transformacije signala, se naziva realni kepstar, ili još češće samo kepstar (eng. *cepstrum*).

$$c(n) = \mathrm{Ev}\left[\hat{x}[n]\right] = \frac{\hat{x}[n] + \hat{x}[-n]}{2} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log|X(e^{jw})|e^{jwn}d\omega \tag{2.30}$$

Naravno, u praksi se koristi aproksimacija stvarnog kepstra korišćenjem diskretne Furijeove transformacije. On se veoma efikasno računa algoritmom brze Furijeove transformacije (eng. *Fast Fourier Transform, FFT*) [7].

$$X_p(k) = \sum_{n=0}^{N-1} x(n)e^{-j\frac{2\pi}{N}kn}, \quad 0 \le k \le N-1 \tag{2.31}$$

$$c_p(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log|X_p(k)|e^{j\frac{2\pi}{N}kn}, \quad 0 \le n \le N-1 \tag{2.32}$$

S obzirom da se kepstar dobija inverznom Furijeovom transformacijom on se nalazi u vremenskom domenu. Međutim praksa je da se umesto vremena koristi pojam kefrencija (eng. *quefrency*). Slično tome, umesto filtriranja kepstar se liftrira kako bi se iz njega izvukle određene informacije.

2.5.2 Primena u govoru

Zvučni govorni segment se može predstaviti kao:

$$\begin{aligned} s[n] &= p[n] * g[n] * v[n] * r[n] \\ &= p[n] * h_z[n] \end{aligned} \tag{2.33}$$

gde je $p[n]$ pobudni signal sa periodom N_p, $g[n]$ efekat glotisa, $v[n]$ efekat vokalnog trakta, a $r[n]$ efekat radijacije na usnama. Slično tome definišemo i bezvučni govorni segment:

$$\begin{aligned} s[n] &= e[n]v[n] * r[n] \\ &= e[n] * h_b[n] \end{aligned} \tag{2.34}$$

Razlika se ogleda u činjenici da je eksitacija širokopojasni šum, a ne periodični signal, kao i u izostanku uticaja glotisa. [7]

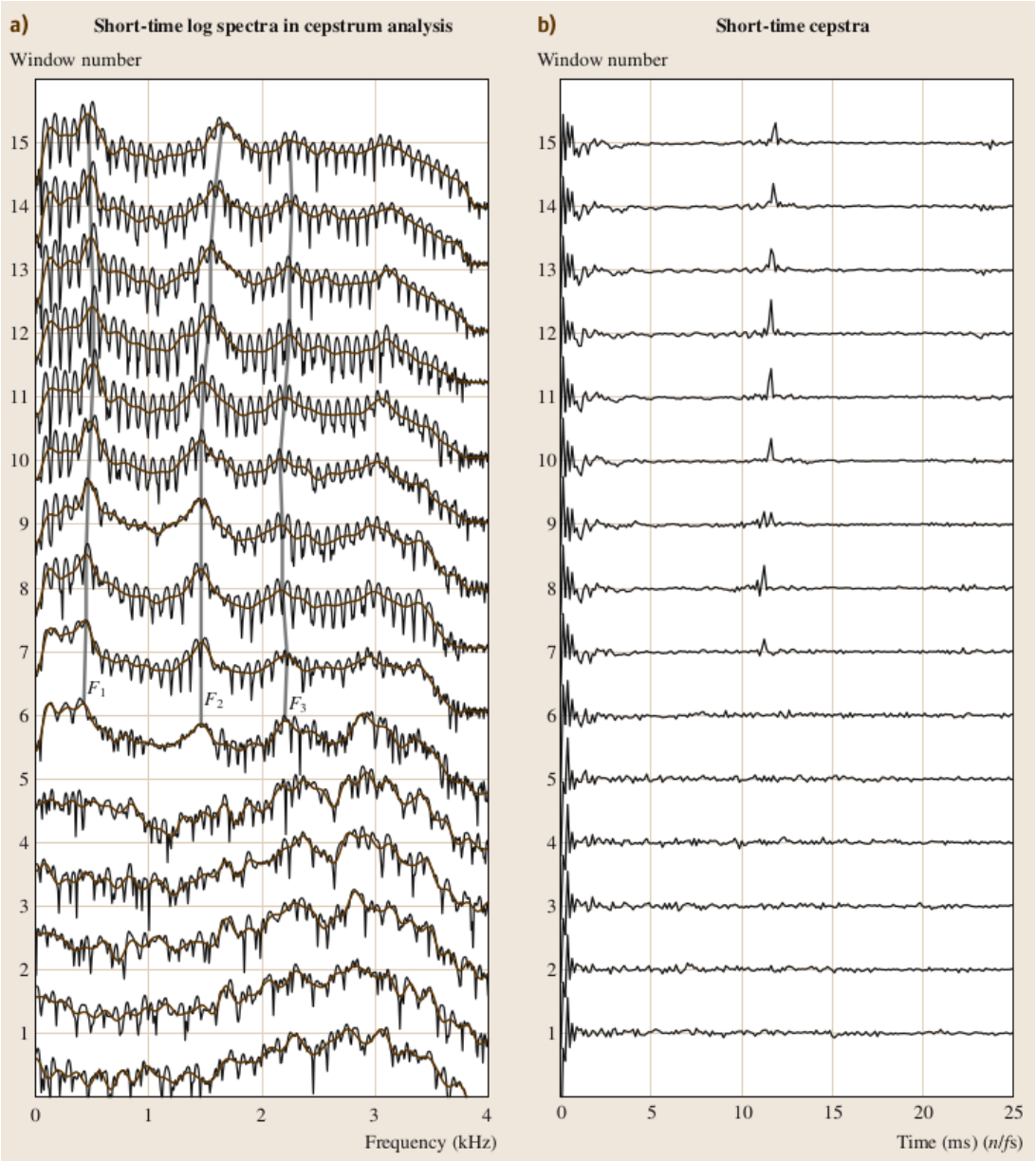
U kepstru govora moguće je razlikovati niskovremenski deo koji sadrži informacije o onome što je izgovoreno, tj eksitaciji i visokovremenski deo, koji nosi informacije o govorniku. U visokovremenskom delu moguće je, za zvučni govor, videti *pitch* periodu. Šta više kvefrencija na kojoj se nalazi njen pik zaista odgovara periodi signala. Jedan od algoritama za detekciju pitch periode se upravo sastoji od toga da se u određenom regionu kepstra, u kom se očekuje pitch perioda, traži pik, te se njegova kvefrencija uzima za vrednost periode. Sličnim postupkom je moguće raditi detekciju zvučnog i bezvučnog govora.

Nažalost ova metoda nije savršena. Na primeru 2.11 preuzetog iz [5] može se videti log-spektar i kepstar za sukcesivne prozore od $50ms$ jednog govornog signala. Prozori $0-5$ predstavljaju bezvučni govor i očekivano na njima se ne može uočiti ikakav pik u visokovremenskom delu kepstra. Prozori $6-7$ prikazuju segment koji je samo delimično zvučan, dok segmenti $8-15$ predstavljaju zvučni govor. Na segmentu 7 se može uočiti veoma mali pik, dok ga na segmentu 6 čak uopšte nema, iako je prisutan zvučni govor. Slično tome, dešava se da mnogo veći pik postoji na umnošcima kvefrencije periode, nego li na samoj kvefrenciji. Iako postoje metode koje poboljšavaju algoritam i u ovom slučaju, nije moguće sa sigurnošću reći da je segment govora bezvučan ako na visokovremnskom delu kepstra ne vidimo pitch periodu.

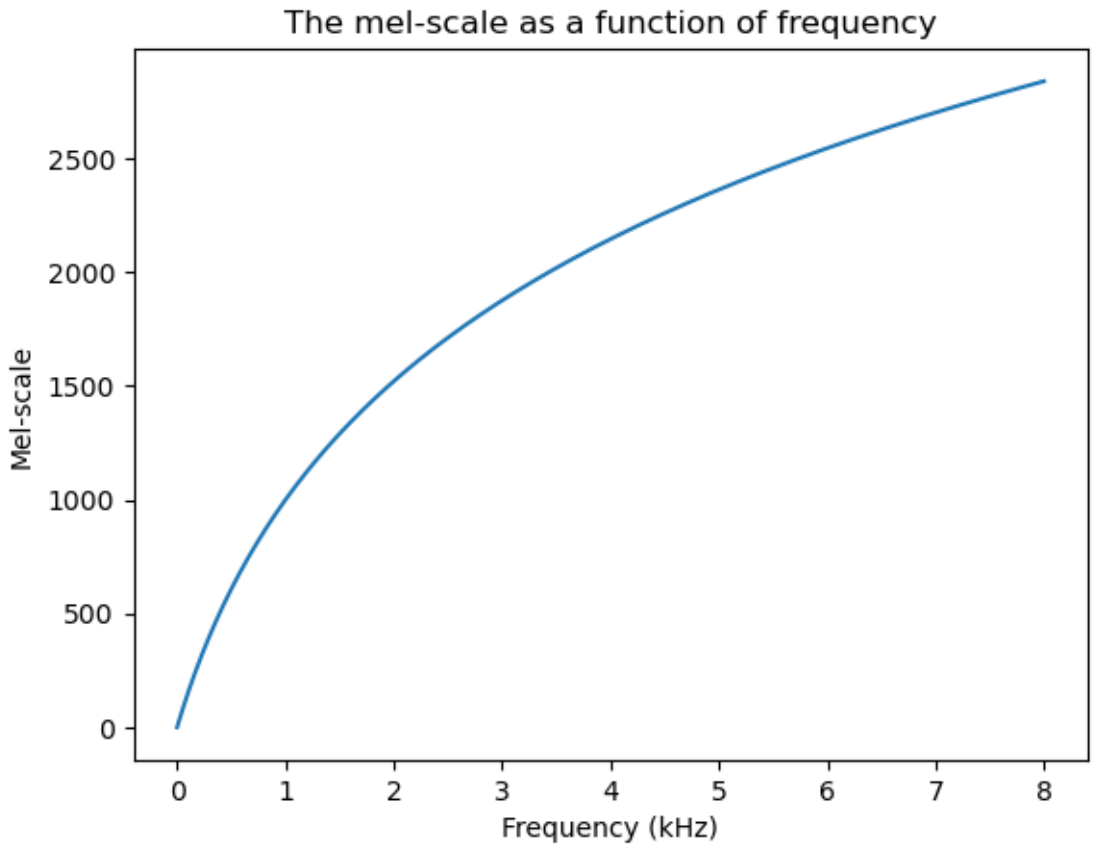
2.6 Mel-frekvencijski kepstralni koeficijenti

Iako su informacije o formantima već sadržane u kepstru, on nije preterano efikasan u smislu da nudi veliki broj koeficijenata za relativno malu količinu informacija. Jedan način da se smanji količina koeficijenata je aproksimacija spektra pomoću banke filtara. Uzmimo da su filtri trougani, ravnomerno raspoređeni do $8kHz$ i sa delimičnim preklapanjem. Suštinski filtri usrednjavaju snagu u okolini određene frekvencije i na taj način drastično smanjuju broj koeficijenata bez gubitka informacija.

Međutim, ljudi ne percipiraju frekvencije uniformno. Slično intenzitetu, ljudsko uho je mnogo osetljivije na promene nižih frekvencija te je prirodnije koristiti logaritamsku skalu. Pokazalo se da je ovakav način rezonovanja pogodan i za sisteme za prepoznavanje govornika, zbog čega se umesto standardne frekvencijske uvodi *mel skala*. Veza između mel skale i standardne frekvencijske je sledeća



Slika 2.11: Kratkovremesni spektar (levo) i kepstar (desno) segmenata govora [5]



Slika 2.12: Veza mel i frekvencijske skale

(slika 2.12):

$$m = 2595 \log_1 0(1 + \frac{f}{700}) = 1127 \ln(1 + \frac{f}{700})$$
$$f = 700(10^{\frac{m}{2595}} - 1) = 700(e^{\frac{m}{1127}} - 1) \tag{2.35}$$

Aproksimacija spektra pomoću mel skale se može videti na slici 2.13. Jasno se vidi da su više frekvencije lošije modelovane od nižih, što je u skladu sa očekivanim [4]. Sami filtri se implementiraju na sledeći način:

$$w_j(k) = \begin{cases} \frac{(\frac{k}{K})f_s - f_{c_{j-1}}}{f_{c_j} - f_{c_{j-1}}}, & l_j \leq k \leq c_j \\ f_{c_{j+1}} - (\frac{k}{K})\frac{f_s}{f_{c_{j+1}}} - f_{c_j}, & c_j < k \leq u_j \\ 0, & \text{inače} \end{cases} \tag{2.36}$$

gde su $f_{c_{j-1}}$ i $f_{c_{j+1}}$ donja i gornja granica odgovarajućeg filtra j, a l_j, c_j i u_j DFT vrednosti kojima odgovara donja, centralna i gornja granica filtra j [5].

Iako je broj koeficijenata sad daleko manji i dalje ostaje problem visoke korelisanosti među susednim koeficijentima. Za dekorelaciju koeficijenata se koristi diskretna kosinusna transformacija (eng. *Discrete Cosine Transform, DCT*). Ova metoda je veoma bliska Furijeovoj transformaciji, uz razliku da ona signal

predstavlja isključivo kao kombinaciju kosinusnih funkcija. Primarno se koristi za kompresiju u obradi slike, i predstavlja osnovu JPEG algoritma. Primenom DCT metode na logaritam mel spektra dobijamo mel-frekvencijske kepstralne koeficijente (eng. *Mel-Frequency Cepstral Coefficients, MFCC*).

$$e(j) = ln\left[\frac{1}{\sum_{k=l_j}^{u_j} w_j(k)} \sum_{k=l_j}^{u_j} |S(k)|^2 w_j(k)\right] \tag{2.37}$$

$$C(k) = \sum_{j=0}^{J} e(j) cos\left[k\left(j - \frac{1}{2}\right)\frac{\pi}{J}\right], \quad k = 1, 2, ..., K \tag{2.38}$$

gde su $S(k), \quad 0 \leq k \leq K$ koeficijenti diskretne Furijeove transformacije segmenta govornog signala. Prvi MFCC koeficijent, $C(0)$ nosi informaciju o energiji segmenta signala.

Kako je energija govora primarno skoncetrisana na nižim učestanostima može doći do problema prilikom implementacije FFT algoritma zbog lošije preciznosti na višim učestanostima. Zbog toga je praksa da se na signal inicijalno primeni tzv. *pre-emphasis* filtar (2.39) koji blago izravna spektar signala i na taj način omogući efikasniju implementaciju algoritma [4].

Konačan postupak ekstrakcije Mel-frekvencijskih kepstralnih koeficijenata je prikazan na slici 2.14.

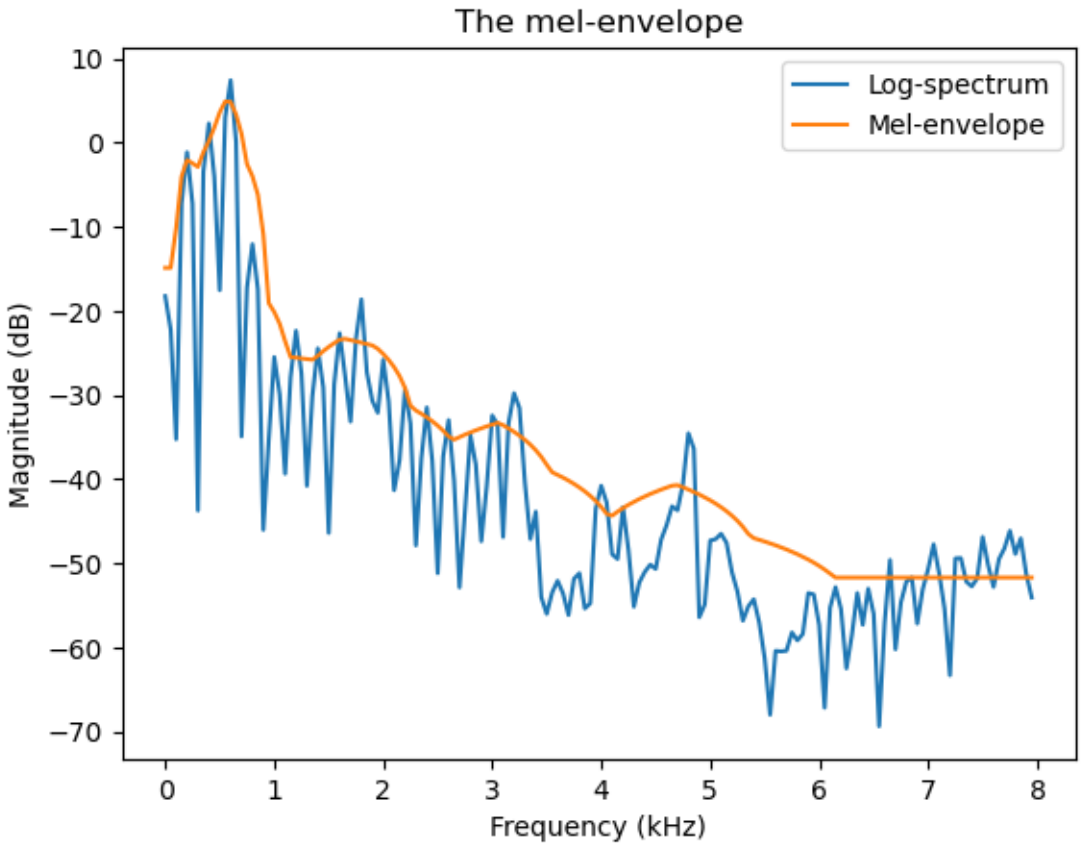
$$P(z) = 1 - 0.68z^{-1} \tag{2.39}$$

2.7 Sistemi za prepoznavanje govornika

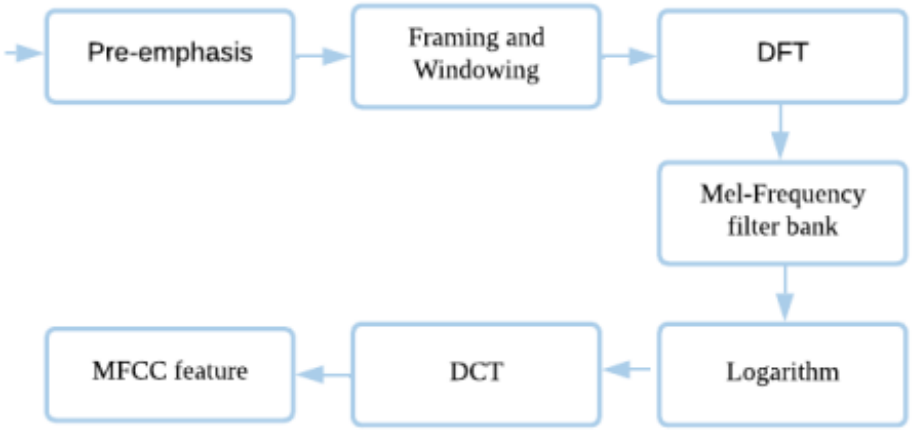
Postoje dva osnovna problema u okviru oblasti prepoznavanja govornika - *identifikacija govornika* i *verifikacija govornika*. Prilikom verifikacije govornik tvrdi svoj identitet a na modelu je da donese odluku da to prihvati ili ne. Drugim rečima, model dobijeni ulaz poredi samo sa jednim modeliranim govornikom. Identifikacija govornika podrazumeva da model ulaz poredi sa svim govornicima koje ima u bazi pre nego što donese odluku. U okviru identifikacije razlikujemo modele koji su tzv. *closed-set* i očekuju isključivo već poznate govornike i *open-set*, koji su u mogućnosti da daju *no-match* izlaz u slučaju da se pojavi govornik koji nije modeliran.

Nekad se takođe izdvaja i još jedan problem, poznat kao detekcija govornika. On je tip open-set identifikacije i podrazumeva da za nepoznat ulaz model utvrdi da li se neki od govornika iz baze nalaze na njemu. Otežavajuća okolnost je što u ovim situacijama često postoji više od jednog govornika prisutno u audio ulazu.

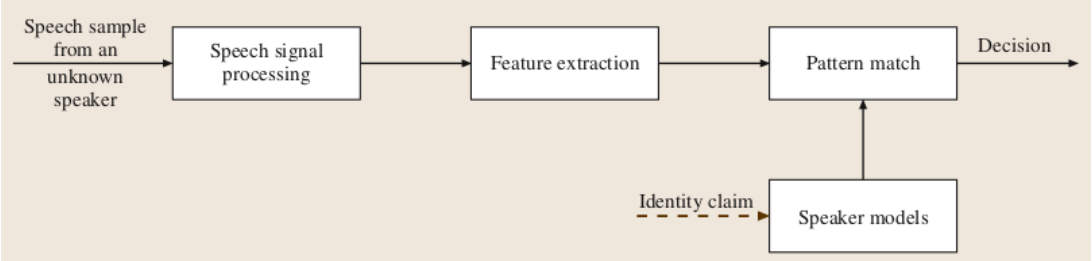
Pored toga kako donose odluke, ovi zadaci se mogu podeliti i na tekstualno zavisne i nezavisne (eng. *text dependent and text independent*). U prvim se od govornika očekuje da pruži unapred poznat tekst, dok za nezavisne to nije slučaj [5].



Slika 2.13: Ovojnica spektra, dobijena mel bankom filtara



Slika 2.14: Ekstrakcija Mel-frekvencijskih kepstralnih koeficijenata [1]



Slika 2.15: Šematski prikaz sistema za prepoznavanje govornika [5]

Glava 3

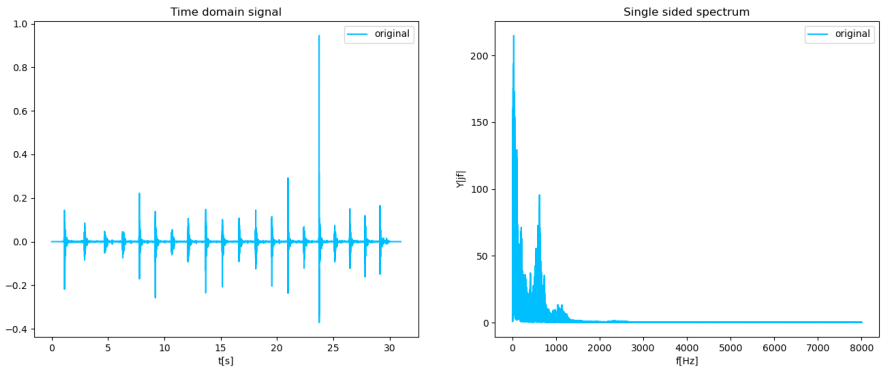
Rezultati

3.1 Pretprocesiranje podataka

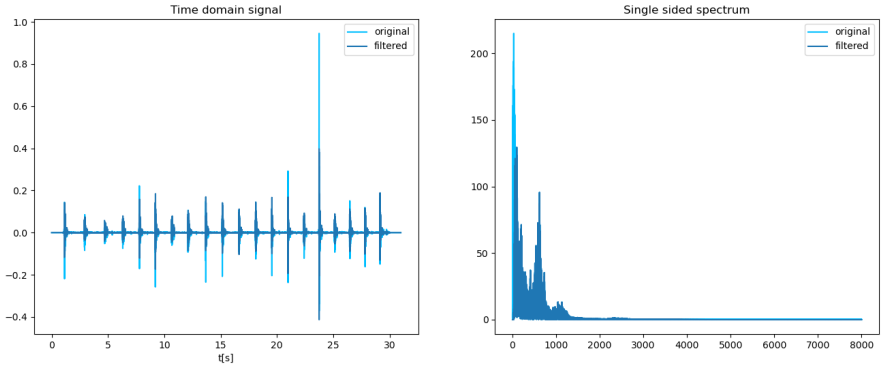
Prvi korak u obradi je podrazumevao izdvajanje pojedinačnih reči iz signala. Nakon učitavanja (slika 3.1) signal je filtriran kako bi se uklonili uticaji DC komponente i šuma na 50Hz (slika 3.2). Filtriranje je rađeno kako bi se poboljšao kvalitet kratkovremenske analize. Segmentacija na reči je rađena algoritmom opisanim u sekciji 2.4.4.

Za prozorsku funkciju je izabrana pravougana, a uticaj dužine prozora se može videti na slici 3.4. Kao što je i bilo očekivano za veoma kratke prozore javljaju se oscilacije, dok veoma dugi ne uspevaju da uhvate sve promene. Za dalji rad izabran je prozor od $20ms$. Kratkovremenska energija i kratkovrmenska brzina prolaska kroz nulu za izabrani prozor su prikazani na slici 3.3. Nažalost ispostavlja se da kratkovremenska brzina prolaska kroz nulu uopšte nije informativna. Ova metoda je veoma osetljiva na smetnje, tako da uprkos filtraciji, verovatno je i dalje ostao prisutan pozadinski šum.

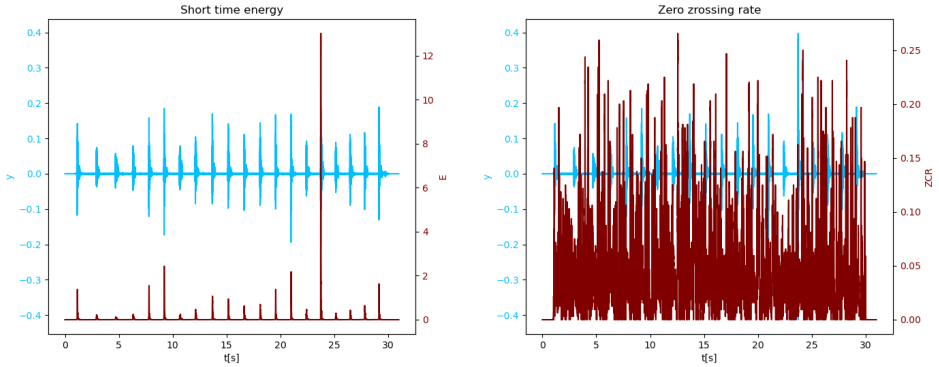
Nakon odbacivanja kratkovremenskog prolaska kroz nulu segmentacija je rađena samo pomoću kratkovremenske energije koja je daleko otpornija metoda. Rezultati segmentacije se mogu videti na slici 3.5. Izdvojene reči su prikazane na slici 3.6



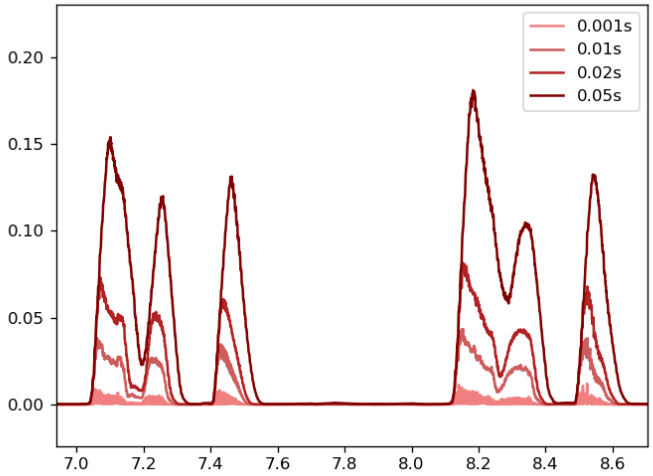
Slika 3.1: Inicijalni signal jednog od govornika iz baze



Slika 3.2: Filtriranje DC komponente i šuma na $50Hz$



Slika 3.3: Kratkovremenska energija i kratkovremenska brzina prolaska kroz nulu



Slika 3.4: Detalj prikaza kratkovremenske energije signala sračunate za različite dužine prozora

3.2 Izdvajanje obeležja

Nakon segmentacije signala, reči su korišćene kao zasebne celine iz kojih su procenjivani Mel-frekvencijski kepstralni koeficijenti. Inicijalno je na svaki signal primenjen pre-emphasis filtar (slika 3.7).

Za potrebe klasifikacije eksperimentisano je sa izdvajanjem Mel-frekvencijskih kepstralnih obeležja kako na celom signalu tako i na segmentima različitih dužina. Određivanjem koeficijenata na velikom segmentu očekuju se lošiji rezultati zbog nestacionarnosti signala. Bez obzira na veličinu segmenta, proces je uvek isti. Prvo se signal prozoruje Hanovom prozorskom funkcijom (na slici 3.8 prikazano je prozorovanje za segment od $30ms$) i na segmentima se pronađe Furijeova transformacija. Pronađeni spektar se provlači kroz mel banku filtara (slika 3.11) i na kraju se od logaritma mel-spektra primenom diskretne kosinusne transformacije dobijaju koeficijenti.

Na slici 3.10 se može videti kratkovremenska Furijeova transformacija tražena na segmentima prozorovanim Hanovom funkcijom. Segmenti su veličine od $30ms$ sa $20ms$ preklapanja. Nakon provlačenja istog kroz banku filtara, primene logaritma i kosinusne transformacije dobili su se koeficijenti na slici 3.12.

Dodatno je za segment sa slike 3.8 pronađen kepstar, te se na slici 3.9 može jasno uočiti pitch perioda u visokovremenskom segmentu.

3.3 Klasifikacija

Klasifikacija je rađena metodom k najbližih suseda (eng. *k Nearest Neighbors, kNN*). To je neparametarska metoda supervizorskog tipa. Faza obučavanja je veoma jednostavna, svodi se na prosto učitavanje podataka u klasifikacioni prostor. U ovom slučaju to je prostor sa četrnaest dimenzija, zbog toga što se za obeležja koristi četrnaest Mel-frekvencijskih kepstralnih koeficijenata.

Faza predikcije je nešto složenija. Ona podrazumeva da se za svaki nadolazeći odbirak iz test skupa nađe udaljenost od svakog odbirka iz obučavajućeg skupa. Merenje udaljenosti između odbiraka je rađeno Euklidskom i Menhetn distancom. Nakon pronalaska distanci uzima se k najbližih i, na osnovu toga kojoj klasi pripada najveći broj suseda, odbirak se klasifikuje.

$$\begin{aligned} d(p,q) &= \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+...+(p_n-q_n)^2} \\ &= ||p-q|| \qquad \text{Euklidska distanca} \end{aligned} \tag{3.1}$$

$$\begin{aligned} d(p,q) &= \sum_{i=1}^{n} |p_i - q_i| \\ &= ||p-q||_T \qquad \text{Menhetn distanca} \end{aligned} \tag{3.2}$$

Broj suseda je eksperimentalno određivan. Klasifikacija je rađena za $k \in [3,5,7,9,13,21,29,43,65]$. U eksperimentima gde se signal delio na dva ili tri dela, ili gde se uopšte nije delio broj suseda nije prelazio 7.

Inicijalno je klasifikacija rađena za skup podataka koji je dolazio sa istog snimka. Tačnije, svaki govornik je bio snimljen kako dvadeset puta izgovara istu reč. Iz originalnog snimka su reči bile izdvojene, te je petnaest korišćeno za obučavanje a pet za testiranje. Ovo nije dobra praksa jer se u svim primerima javlja ista pozadinska buka, na istoj je udaljenosti držan mikrofon, ista je habitualna pitch perioda itd. Rezultati klasifikacije u ovom slučaju su dati na slici 3.13.

Kasnije je snimanje ponovljeno opet, sa istim mikrofonom i sa istom izgovorenom rečju, te je taj snimak iskorišćen kao skup za testiranje. Sada je skup za obučavanje isti, određen iz inicijalnog snimka, a u skupu za testiranje se nalazi pet novo snimljenih reči. Opet, pouzdaniji rezultati bi se dobili u slučaju da su svi ulazi snimani posebno, ali se i u ovoj situaciji vidi realističniji (lošiji) rezultat (slika 3.15). Nažalost nije bilo moguće snimiti opet svih devet govornika, već je ponovo snimljeno samo pet. Kako ne bi bilo korektno porediti rezultate u ova dva slučaja jer kvalitet identifikacije umnogome zavisi od broja govornika u bazi, napravljen je još jedan eksperiment, sa istim podacima kao u 3.13 ali sa samo pet govornika (slika 3.14).

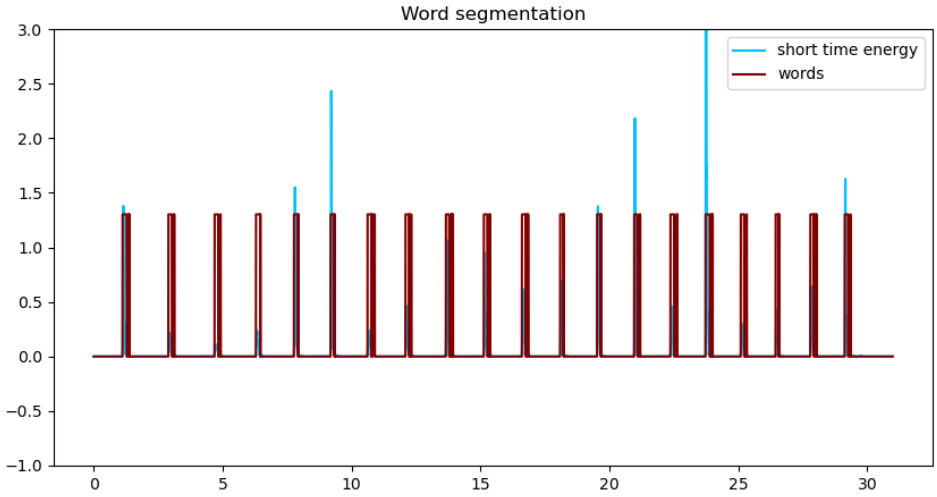
Na grafovima 3.13, 3.14 i 3.15 se mogu videti rezultati za Menhetn (levo) i euklidsku (desno) distancu. Svaki podgraf na x-osi ima prikazan broj suseda za koji je rađen specifični eksperiment. Na y-osi je pikazan F1 skor. Broj segmenata na koji je podeljen ulazni signal pri izdvajanju obeležja je prikazan različitim bojama i to tako da:

- 1 - Signal nije deljen uopšte
- 2 - Signal je deljen na dva jednaka dela
- 3 - Signal je deljen na tri jednaka dela
- (30,20) - Signal je deljen na segmente od $30ms$ sa $20ms$ preklapanja.

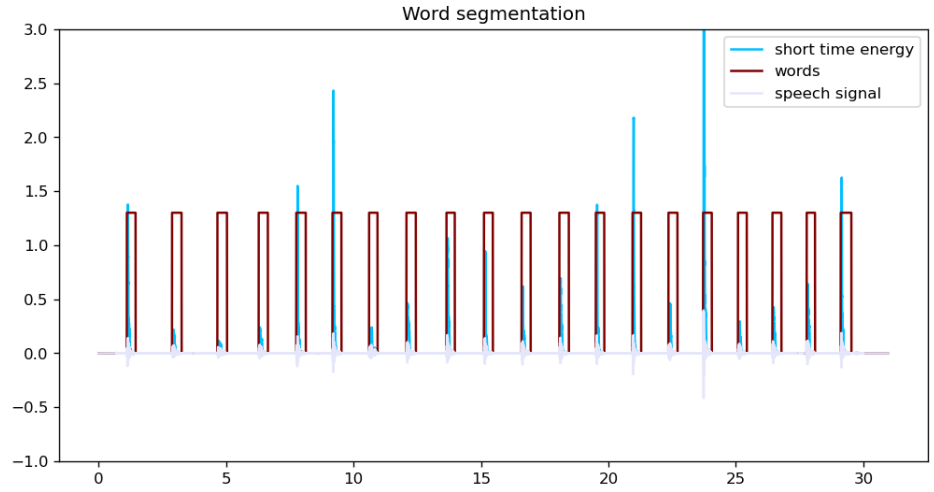
Sa grafika se može videti da su rezultati uvek najbolji kada se obeležja određuju na celom signalu i kada se klasifikacija radi sa malim brojem suseda. Razlika u uspešnosti je neznatna za euklidsku i Menhetn distancu, uz nešto bolje performanse za Menhetn. S obzirom da je problem prepoznavanje govornika, a ne govora, nestacionarnost signala ne predstavlja toliki problem, te deljenje signala na segmente ne doprinosi kvalitetu klasifikacije. Šta više, pri deljenju su se javljali i segmenti koji nisu zvučni, a oni ne nose korisne informacije o govorniku.

Kada uporedimo rezultate 3.14 i 3.15 vidno su bolji rezultati kada su i test i obučavajući skup iz istog izvora, što potvrđuje pretpostavku da ovaj vid formiranja baze nije bio dobra praksa. Kada se uporede rezultati 3.13 i 3.14 opet su daleko bolji rezultati za identifikaciju sa pet naspram sa devet govornika. Ovo je u skladu sa očekivanim, jer problem identifikacije postaje složeniji sa brojem govornika.

Rezultati sa slike 3.13 su dodatno analalizirani na konfuzionoj matrici (slika 3.16). Najviše grešaka je pravljeno za govornike istog pola, što je i očekivano. Izuzetak su ženski govornici 2 i 4 kod kojih su određeni odbirci loše klasifikovani kao da pripadaju muškim govornicima.

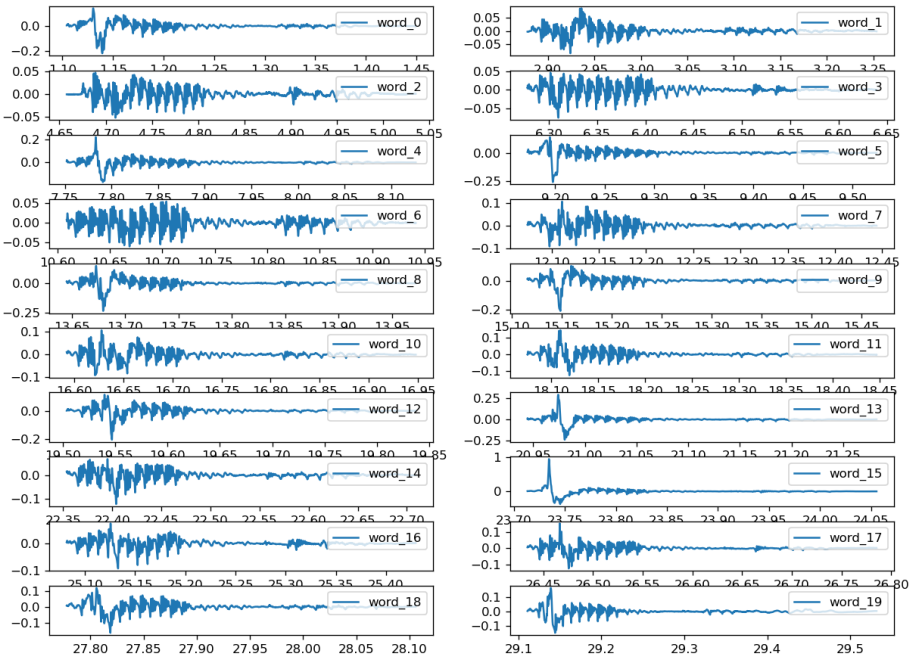


(a)

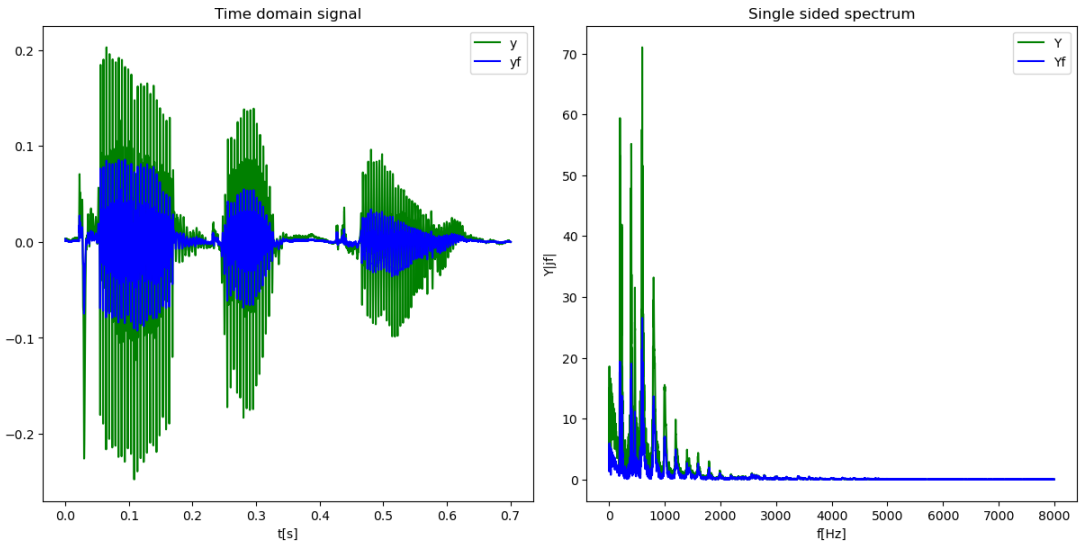


(b)

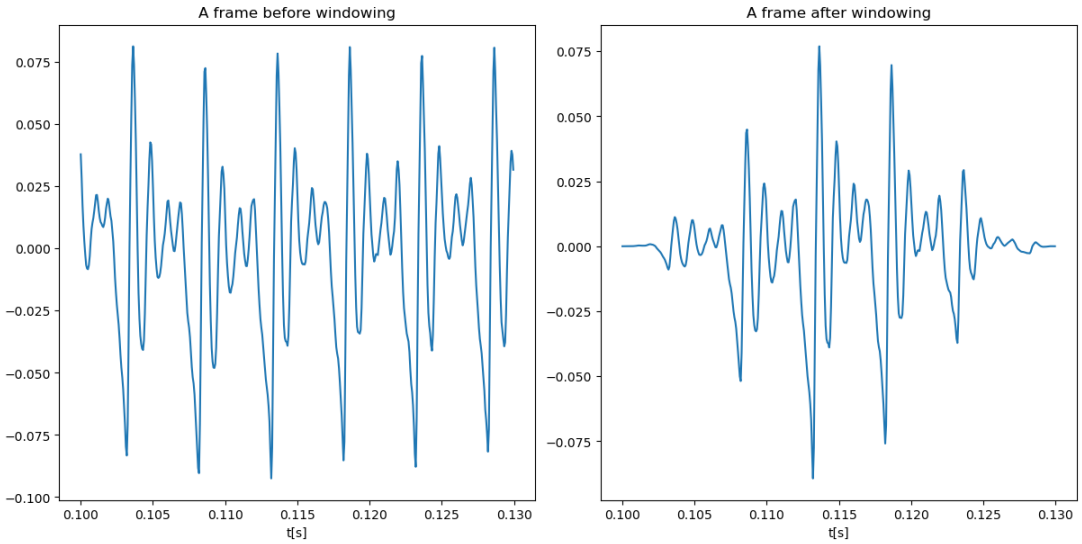
Slika 3.5: Segmentacija pomoću gornjeg praga energije (a), i njeno poboljšanje pomoću donjeg praga (b)



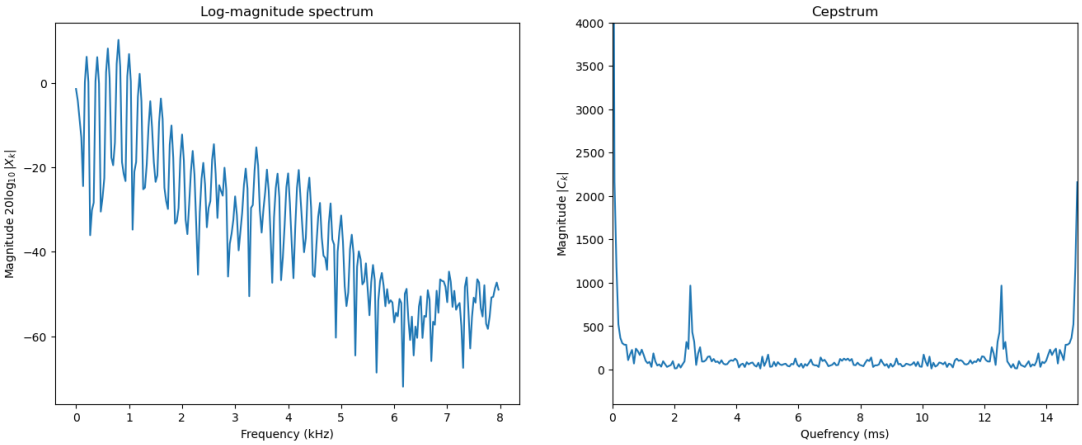
Slika 3.6: Izolovane reči jednog od govornika iz baze



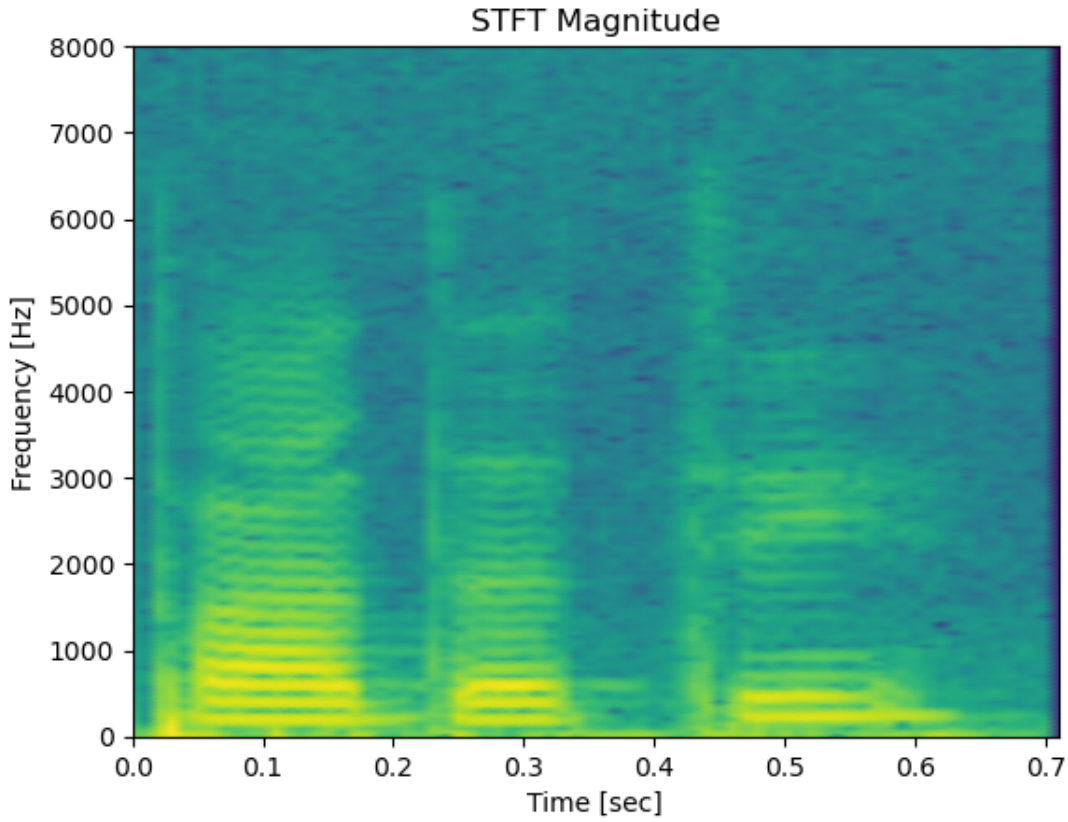
Slika 3.7: Signal nakon primene pre-emphasis filtra



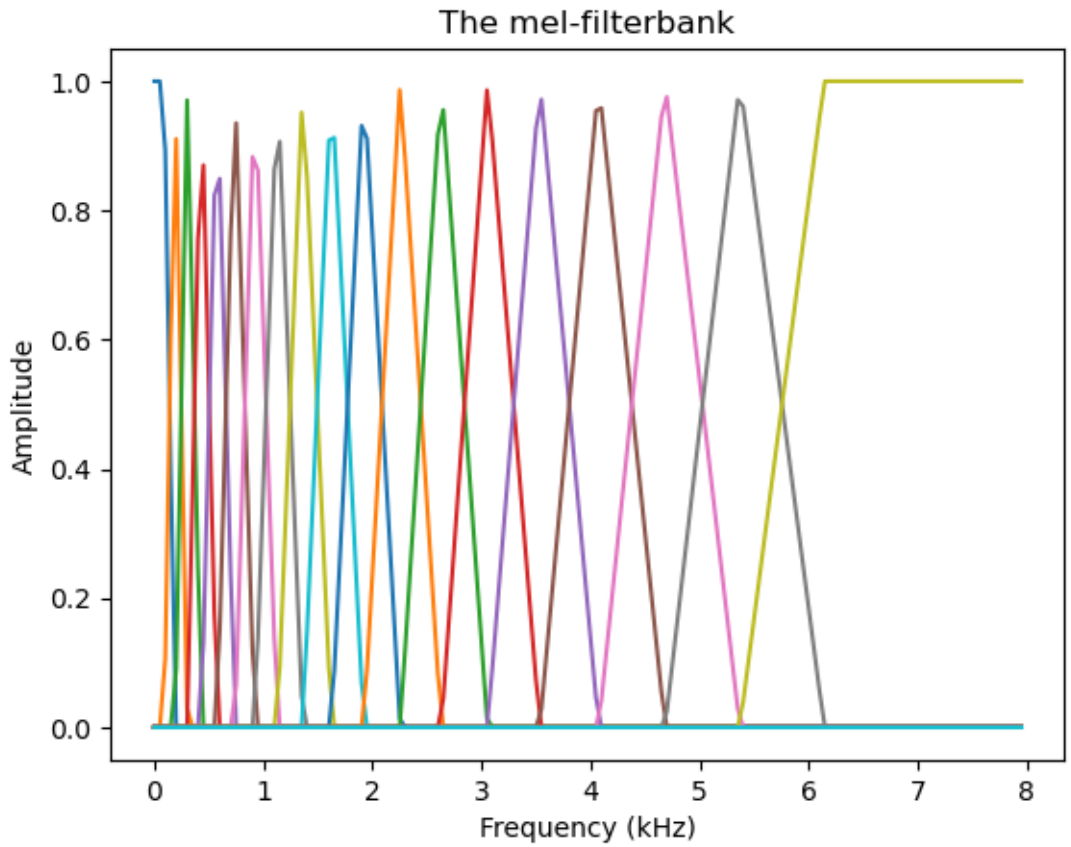
Slika 3.8: Originalni segment signala (levo) i segment nakon prozorovanja (desno)



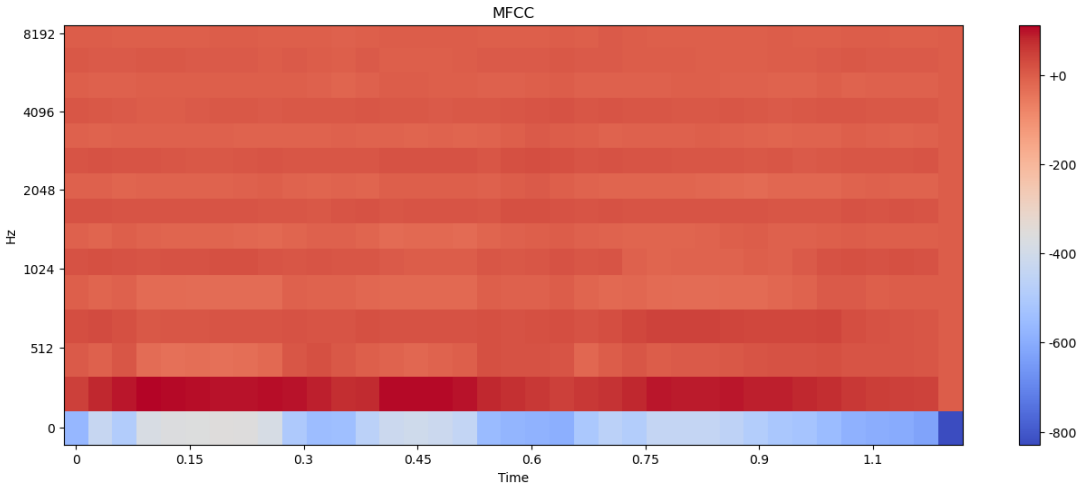
Slika 3.9: Logaritam spektra i kepstar zvučnog segmenta



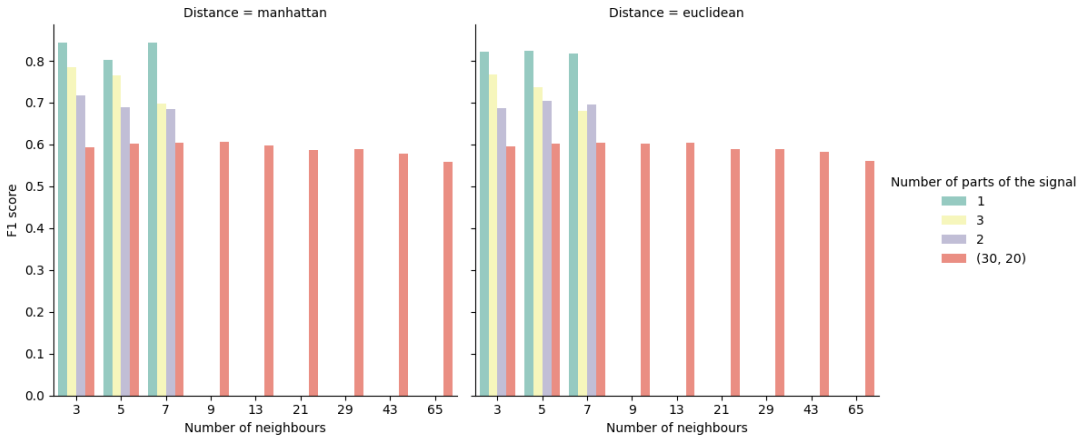
Slika 3.10: Kratkovremenska Furijeova transformacija vizuelizovana na decibelskoj skali



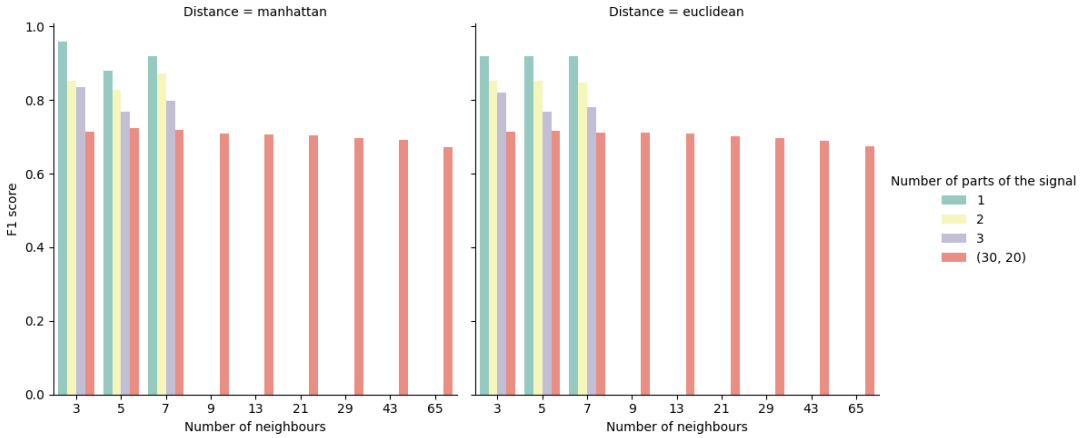
Slika 3.11: Mel banka sa 15 filtara



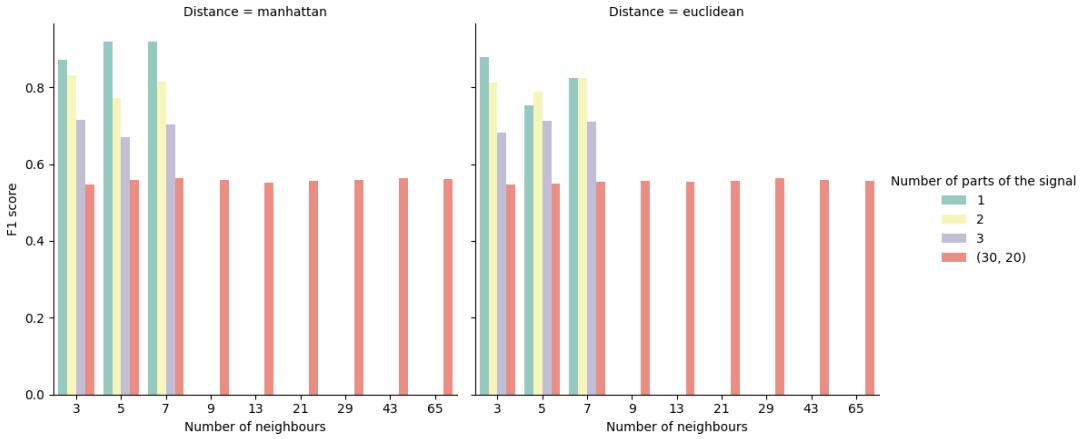
Slika 3.12: Mel-frekvencijski kepstralni koeficijenti za jedan segment



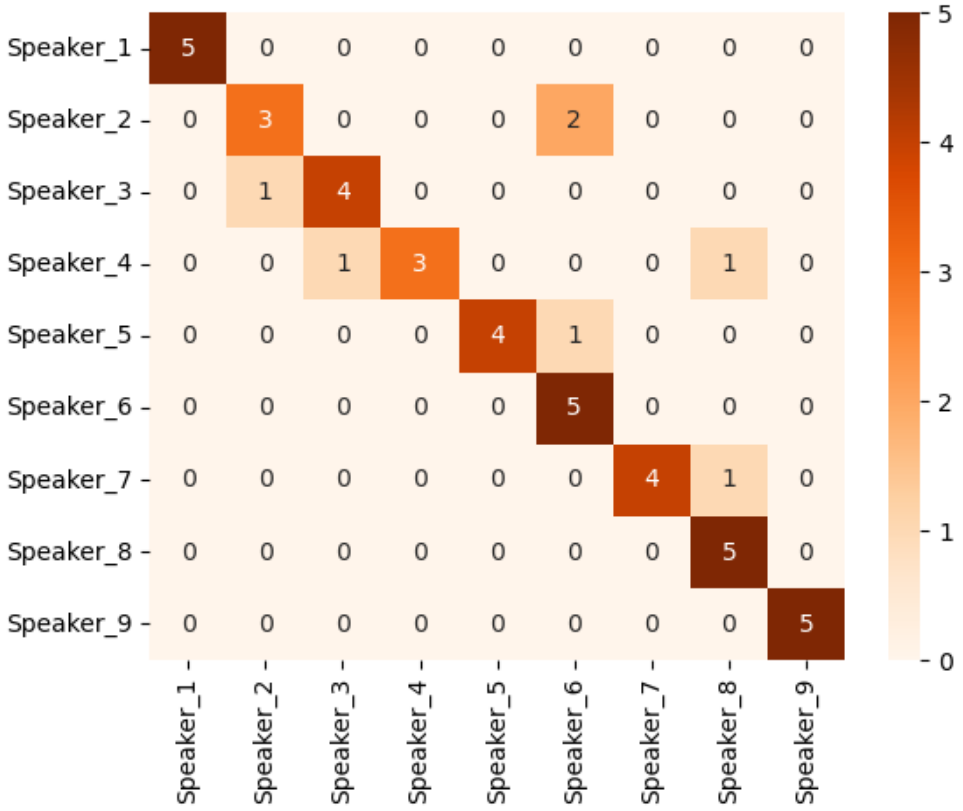
Slika 3.13: Rezultati - skup za obučavanje i skup za testiranje iz istog izvora, identifikacija 9 govornika



Slika 3.14: Rezultati - skup za obučavanje i skup za testiranje iz istog izvora, identifikacija 5 govornika



Slika 3.15: Rezultati - skup za obučavanje i skup za testiranje iz različitog izvora, identifikacija 5 govornika



Slika 3.16: Konfuziona matrica za slučaj identifikacije 9 govornika

Glava 4

Zaključak

Ovaj rad je pružio uvid u jedan od načina projektovanja sistema za prepoznavanje govornika, posebno u slučaju kada su hardverski resursi za treniranje zahtevnih modela, kao što su neuralne mreže, ograničeni. Posebno je istaknut značaj same obrade signala i ekstracije obeležja za potrebe uspešne klasifikacije.

Dalji rad bi prvenstveno podrazumevao opširnuju bazu podataka. Ovo bi otvorilo mogućnosti za korišćenje različitih modela koji mogu biti zahtevniji po pitanju veličine obučavajućeg skupa. Neki od mogućih izbora su metoda potpornih vektora (eng. *Support Vector Machine, SVM*), pomešani Gausovski modeli (eng. *Gaussian Mixture Models, GMM*) i skriveni Markovljevi modeli (eng. *Hidden Markov Models, HMM*).

Što se same ekstracije obeležja tiče, ovaj rad se ograničio na tekstualno zavisne modele te je korišćenje svih Mel-frekvencijskih kepstralnih koeficijenata bio razuman izbor. Za tekstualno nezavisne modele samo bi visokovremenski deo kepstra bio informativan. Potencijalni kandidati za dodatna obeležja bi prvenstveno bili delta Mel-frekvencijski kepstralni koficijenti, koji predstavljaju izvode (prvi i drugi) standardnih koeficijenata. Njihova korisnost leži u mogućnosti da obuhvate dinamiku govora.

Literatura

[1] Zrar Kh. Abdul i Abdulbasit K. Al-Talabani. "Mel Frequency Cepstral Coefficient and its Applications: A Review". U: *IEEE Access* 10 (2022.), str. 122136–122158. DOI: 10.1109/ACCESS.2022.3223444.

[2] Automatic Speech Recognition, MIT course, Spring 2003.

[3] J.W. Tukey B.P. Bogert M.J.R. Healy. "The Quefrency Alanysis [sic] of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking". U: *Proceedings of the Symposium on Time Series Analysis*. 1963.

[4] Tom Bäckström i dr. *Introduction to Speech Processing*. 2. izdanje. 2022. DOI: 10.5281/zenodo.6821775. URL: https://speechprocessingbook.aalto.fi.

[5] Yiteng Huang Jacob Benesty M. Mohan Sondhi. *Springer Handbook of Speech Processing*. Springer Berlin, Heidelberg, 2008.

[6] Michael Rodney Portnoff. *A Quasi-One-Directional Digital Simulation for the Time-Varying Vocal Tract*. 1973.

[7] Lawrence Rabiner i Ronald W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall, 1978.