



Univerzitet u Beogradu - Elektrotehnički fakultet

Katedra za signale i sisteme



## **DIPLOMSKI RAD**

# **Transformacija govora na osnovu promene fundamentalne učestanosti**

### **Kandidat**

Uroš Petković, br. indeksa 2016/0186

### **Mentor**

dr Aleksandra Marjanović, docent

Beograd, *jul* 2020. godine

## PREDGOVOR

Diplomski rad baziran je na primenjivanju znanja stečenog na predmetima Obrada i prepoznavanje govora, sa šifrom *13E054OPG* pod vođstvom prof. dr Željka Đurovića i dr Aleksandre Marjanović, i Neuralne mreže, sa šifrom *13E054NM* pod vođstvom prof. dr Gorana Kvašćeva pohađanih u sedmom i osmom semestru osnovnih akademskih studija. Rad nosi naziv „Transformacija govora na osnovu promene fundamentalne učestanosti“ u čijoj izradi je učestvovao student Uroš Petković. Ovaj rad predstavlja jedno od mogućih rešenja danas jako primenjivih algoritama za transformaciju govora, baziran na promeni fundamentalne učestanosti i LPC analize i sinteze govornog signala.

Izrada diplomskog rada omogućena je uz korišćenje *CMU Artic Database* seta podataka snimljenog u studijskim uslovima na *Language Technologies Institute* na *Carnegie Mellon* univerzitetu u Pittsburgu, Pensilvaniji. Set se sastoji od 115 rečenica izgovorenih od strane muških i ženskih osoba i od velikog je značaja za ovaj diplomski rad. Praktični deo diplomskog rada i implementacija samog rešenja rađena je samostalno od strane studenta, uz nesebičnu pomoć mentora.

## REZIME RADA

U ovom diplomskom radu zadatak je bio izvršiti transformaciju govornih sekvenci izgovorenih od strane jednog pola (muške ili ženske osobe) u govornu sekvencu suprotnog pola. Postoji više načina kojim se može pristupiti rešavanju ovog problema, pri čemu je potrebno znati osnovna svojstva i osobine govornog signala, kao i to kako se njima može pristupati. Dostupni set podataka govornih sekvenci koristi se za treniranje neuralne mreže koja će uspešno prediktovati odgovarajuće LPC koeficijente traženog signala. Kao parametri neuralne mreže prosleđuju se *LSF* koeficijenti polaznih i traženih govornih sekvenci, dobijeni prethodnom konverzijom *LPC* koeficijenata radi stabilnosti, nakon čega se, posle predikcije, inverznim postupkom vraćaju ponovno u polazni oblik. Vremenskom obradom govornog signala dolazi se do procene fundamentalne učestanosti uz pomoć autokorelacione metode, nakon čega se vrši zamena iste preko *PSOLA* algoritma za obradu govornih signala. Krajnjom *LPC* sintezom dolazi se do konačnog oblika traženog govornog signala suprotnog pola. Ovakav vid obrade govornog signala je danas široko rasprostranjen i postoji više načina kako se on može rešiti, a ovo predstavlja samo jednu od mogućih implementacija.

## ZAHVALNICA

Posebna zahvalnost ističe se mentoru dr Aleksandri Marjanović na predanosti, strpljenju, pomoći i izdvojenom vremenu kada su u pitanju saveti i konsultacije u vezi rešavanja zadatog problema ovog diplomskog rada, kao i snabdevanju potrebnom literaturom za uspešno realizovanje. Takođe, velika zahvalnost se duguje i *Carnegie Mellon* univerzitetu u Pensilvaniji na dozvoli za korišćenje dostupnog seta podataka u okviru diplomskog rada.

Uroš Petković

U Beogradu, *jul* 2020. godine

## SADRŽAJ

PREDGOVOR .....	2
REZIME RADA .....	3
ZAHVALNICA .....	4
SADRŽAJ .....	5
1 UVOD .....	6
2 OBRADA GOVORNOG SIGNALA .....	10
2.1 Modeliranje govornog signala .....	10
2.2 Analiza govornog signala u vremenskom domenu .....	15
2.3 LPC analiza i sinteza govornog signala .....	17
2.4 DTW algoritam .....	18
2.5 PSOLA algoritam .....	20
3 METODOLOGIJA RADA .....	23
3 IMPLEMENTACIJA I REZULTATI .....	27
4 DISKUSIJA .....	40
5 ZAKLJUČAK .....	42
6 LITERATURA .....	43
PRILOG A .....	44

## 1 UVOD

Govorni signal je sam po sebi jedan veoma složen stohastički proces. Osnovna namena govornog signala je komunikacija, a postoji nekoliko različitih načina kako se može okarakterisati komunikaciona moć govora. Jedan od tih načina je uveden na osnovu Šenonove (*Shannon*) teorije informacija po kojoj se govor može reprezentovati porukama, odnosno informacijama koje su u njemu sadržane. Drugi pristup u karakterizaciji komunikacione moći govora jeste vezan za signal koji nosi govornu informaciju, odnosno za akustički talas. Proces nastajanja govora je izuzetno složen, pa u cilju ilustracije ove pojave, treba poći od činjenice da pre nego što nešto izgovorimo, neophodno je da se u našem mozgu stvori apstraktna slika sadržaja koji treba da bude izgovoren. Sledeći korak je da se kao posledica ovog apstraktnog sadržaja, formira niz nervnih aktivacija koje treba da pripreme naš artikulacioni govorni aparat (pokretanje jezika, usana, glasnih žica) kako bi konačni rezultat celog procesa bio akustični talas koji sadrži informaciju sadržanu u početnoj apstraktnoj ideji ili poruci. U tehničkim sistemima koji se bave govornom komunikacijom, govor se obrađuje na mnogo različitih načina, zato mora biti jasno definisano u kom obliku se govor reprezentuje, prenosi i čuva, tako da se sačuva osnovna informacija koja je sadržana u govornom signalu, bez ozbiljne degradacije sadržaja.

Govorni signal se koristi u raznim sferama tehnologije, a i ostalih prirodnih nauka, u memorisanju, zaštiti i prenosu podataka, u sistemima za sintezu govora, identifikaciju i verifikaciju govornika, prepoznavanju govora, pomoći hendikepiranim osobama, poboljšanju kvaliteta signala, LPC vokoderima, formantnoj analizi i mnogim drugim oblastima. Jednu od bitnih tema danas predstavlja transformacija govora. Transformacija govora je tehnika kojom se modifikuje glas izvornog govornika gde se on doživljava kao da ga je ciljni govornik izgovorio. Spada u opštu kategoriju modifikacije govora koja je danas od velikog interesa, sa brojnim aplikacijama koje uključuju sintezu teksta u govor, prepoznavanje govora, uređivanje glasa, emitovanje i zabavu itd. Sa razvijanjem tehnologije poslednjih godina došlo je i do razvitka različitih algoritama za transformaciju govora koji imaju široku primenu u navedenim oblastima. U ovom radu akcenat će biti postavljen na jedan od tih algoritama. Postavlja se pitanje kako doći do željenih rezultata. Između ideje i realizacije stoji dugačak put koji treba preći kako bi traženi rezultati bili zadovoljavajući. Najpre, na raspolaganju se nalazi set podataka od 115 izgovorenih rečenica od strane muške i ženske osobe. Ove rečenice treba iskoristiti za treniranje neuralne mreže. Kako bi se na datom skupu izvršilo i testiranje rezultata, dati set treba podeliti na trening skup i test skup podataka. Međutim, potrebno je naći odgovarajuća obeležja koja će efikasno hraniti neuralnu mrežu i omogućiti dobijanje adekvatnih rezultata. Ideja je da se neuralnoj mreži kao trening set ulaza i izlaza proslede koeficijenti dobijeni LPC analizom datih govornih sekvenci. Kako govorni signal ne nosi samo informacije o tome ko je govornik, već i o tome šta je izgovoreno, šta potiče od

eksitacije, a šta od konstitucije i vokalnog trakta samog govornika, u ovom radu biće razmatrana 3 metoda za transformaciju govora čija će uspešnost biti poređena. Prvi od metoda predstavlja samu promenu fundamentalne učestanosti početnog govornog signala, čime se dobija izlazni govorni signal na željenoj učestanosti suprotnog pola. Mapiranje fundamentalne učestanosti vrši se uz pomoć *PSOLA* algoritma (*Pitch Synchronous OverLap Add*) gde se željena fundamentalna učestanost prediktuje na osnovu vektora fundamentalnih učestanosti trening skupa polaznih i željenih signala, a gde se dati vektori dobijaju procenom fundamentalne učestanosti svakog para govornih sekvenci uz pomoć autokorelacione metode. Na ovaj način se dobija željeni izlaz koji ima drugačiju frekvenciju oscilovanja glasnih žica, višu ako radimo prelaz sa muškog na ženski glas ili nižu ako radimo prelaz sa ženskog na muški glas, pri čemu se ostala svojstva govornog signala, poput boje glasa i ostalih osobina, ne menjaju, pa ovo rešenje i ne predstavlja jedno od najboljih rešenja. Druga dva metoda uključuju i prediktovanje LPC koeficijenata željenog signala, čime se vrši i promena pomenutih osobina, pa su se rezultati dobijeni drugim dvema metodama bolje pokazali. U samom procesu treniranja neuralne mreže učestvuju LSF koeficijenti koji predstavljaju konvertovane LPC koeficijente dobijene datom analizom radi stabilnosti, koji se nakon predikcije ponovno vraćaju u prvobitnu formu. Takođe, jednu od predobrada predstavlja *DTW* algoritam (*Dynamic Time Warping*) radi preciznosti samih prediktovanih koeficijenata. Na samom kraju, nakon prediktovanja potrebnih LPC koeficijenata, vrši se sinteza govornog signala, čime se dobijaju konačni rezultati. Kako je ova tema jako popularna i primenjiva poslednjih godina, jako je širok spektar literature koja se može pronaći za rešavanje ovog problema, kao i mnogih drugih sličnih njemu.

[1] *Mark Tse* sa kolumbijskog univerziteta u svom radu "*Voice Transformation*" bavi se rešavanjem ovog problema na sličan način. Vršiti se treniranje neuralne mreže datim parovima rečenica izgovorenih od strane osoba suprotnog pola, pri čemu se nakon prediktovanja LPC koeficijenata mapiranje željene fundamentalne učestanosti radi na eksitacionom delu signala dobijenog iz prethodne analize, pa se nakon toga LPC sintezom dobija krajnji signal. Ono što je razlika u odnosu na ovaj rad je to što se klasifikovanje na zvučni i bezvučni deo govornog signala vrši na osnovu energije eksitacionog signala i koeficijenta refleksije, poređenjem sa odgovarajućim pragom, što u potpunosti nije slučaj u ovom radu. Takođe, procena fundamentalne učestanosti bazira se na kepsstralnoj dekonvoluciji eksitacionog signala i na proceni autokorelacionom metodom, dok se ovaj rad oslanja samo na autokorelacionu procenu. Činjenica je da postoji mnogo algoritama za procenu fundamentalne učestanosti, kao što je metod paralelnog procesiranja, gorepomenuti metodi, ali svaki od njih daje približno tačne rezultate, stoga je na autoru subjektivna odluka da odredi kojim će se metodom služiti. Za mapiranje fundamentalne učestanosti korišćen je *LP-PSOLA* algoritam baziran na istim pretpostavkama. Krajnji rezultati dobijaju se LPC sintezom signala gde su dobijeni rezultati pokupili pozitivne kritike od strane slušalaca koji su ih ocenjivali.

[2] *Permanallur Ranganathan* u svom radu "*LPC based Voice Morphing*" se, takođe, bavi sličnim pristupom. Rad je baziran na transformaciji glasa na osnovu LPC analize govornog signala. Nakon analize dobijaju se koeficijenti linearnog prediktora i eksitacioni (rezidualni) signal. Nedostaci ovog rada ogledaju se u tome što se LPC koeficijenti računaju na celom signal, pri čemu se dobijaju neprecizniji rezultati. Znajući da govorni signal nije stacionaran, potrebno je raditi obradu na dovoljno malim vremenskim segmentima u kojima se signal može smatrati stacionarnim, kako bi procena bila što bolja. Takođe, ovaj rad je testiran samo na izgovoru jedne foneme, a ne cele reči ili rečenice. Još jedan od nedostataka predstavlja to što uopšte nema neuralne mreže koja će pokušati da predvidi koeficijente željenog signala, već se koeficijenti početnog signala multipliciraju nekim brojem kako bi došlo do njegove izmene pri čemu se dobija drugačiji signal, ali mogućnost dobijanja željenog signala je isključena, osim u slučaju pukog nagađanja.

[3] *Tomoki Toda et al.* su u svom radu "*Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory*" izložili jednu od novijih metoda spektralne konverzije govora. Za spektralnu konverziju između govornika koristi se GMM model (*Gaussian Mixture Model*) združene gustine verovatnoće polaznog i željenog signala. Ova konvencionalna metoda pretvara spektralne parametre frejm po frejm na osnovu minimalne srednje kvadratne greške. Iako je metod efektivan, dolazi do pogoršanja kvaliteta govora usled problema kao što je to da odgovarajući spektralni pokreti nisu uvek uzrokovani procesom konverzije na frejmovima ili da su pretvoreni spektri preterano zaglađeni statističkim modeliranjem. Kako bi se ti problem rešili, predložen je metod baziran na konverziji zasnovanoj na proceni maksimalne verodostojnosti trajektorije parametara. Za realizaciju odgovarajuće sekvence pretvorenog spektra koriste se ne samo statičke, već i dinamičke statističke karakteristike. Štaviše, efekat prekomernog smirivanja ublažen je uzimajući u obzir karakteristiku globalne varijanse konvertovanih spektara. Eksperimentalni rezultati pokazuju da se performanse konvertovanog signala mogu drastično poboljšati predloženom metodom sa strane kvaliteta govora i tačnost konverzije.

Pored ovih metoda, u literaturi se mogu naći i metodi kao što su *Spectrogram-GAN* metoda, *Phase vocoder*, metodi bazirani na sinusoidalnom modelu, *HSN (Harmonic plus noise)* modelu i mnoge druge.

Cilj ovog rada je implementirati takav algoritam koji će uspešno vršiti transformaciju govora u govor suprotnog pola, bilo to slučaj za prelazak iz muškog u ženski glas ili iz ženskog u muški glas uz što manje gubitaka u vidu kvaliteta signala i svojstava samog glasa. Primena ovakvog algoritma mogla bi se ogledati u raznim aplikacijama za konverziju govora, prepoznavanje, identifikaciju, aplikacijama za decu, generisanje glasova likova u crtanom filmu i u mnogim drugim oblastima. Ono na čemu leži osnova ovog algoritma biće opisano u narednim poglavljima. Najpre, u poglavlju *Obrada govornog signala* biće date teorijske osnove koje se tiču samog govornog signala, njegove prirode, modeliranja govornog signala, kao i tipovima njegove obrade, kao što je obrada u vremenskom domenu, na kojoj se bazira



procena fundamentalne učestanosti, *DTW* algoritam za obradu signala, *PSOLA* algoritam za mapiranje željene fundamentalne učestanosti, *LPC* analiza i sinteza signala kao i modelovanje neuralne mreže potrebne za dobro prediktovanje datih koeficijenata linearnog prediktora. Nakon toga, u poglavlju Metodologija rada prethodno definisani algoritmi obrade kombinuju se u jedan sklop koji celokupno čini srž rešenja, a čiji će pojedinačni delovi biti detaljno objašnjeni. U sekciji Implementacija i rezultati biće prikazani rezultati koji su dobijeni implementiranjem datog rešenja, kao i procene slušalaca datih dobijenih audio sekvenci. Na samom kraju, u poglavljima Diskusija i Zaključak biće izložene činjenice koje se odnose na date rezultate, kvalitet rešenja, poređenje sa postojećom literaturom, moguće nadogradnje algoritma, kao i konačni sud o ovom radu.

## 2 OBRADA GOVORNOG SIGNALA

### 2.1 Modeliranje govornog signala

U cilju primene različitih tehnika za procesiranje govornih signala neophodno je za početak razumeti osnovne mehanizme produkcije govora koji se sastoji od niza zvukova koji, kao i tranzicije između njih, služe za simboličku reprezentaciju informacija koje su sadržane u govoru. Vokalni trakt čoveka se prostire od usana, sa jedne strane, do glasnih žica sa druge strane. Sastoji se od farinksa i usne šupljine, dok se pri izdgovoru nazala vokalnom traktu priključuje i nosna šupljina. Kod prosečnog muškarca totalna dužina vokalnog trakta je oko 17cm, dok je kod žena nešto kraća. Nosna šupljina počinje od resice (*velum*) i završava nozdrvama. Kada je resica spuštena, nazalni trakt je akustički povezan sa vokalnim traktom i tada on proizvodi takozvane nazalne glasove govora. Jednostavno rečeno, govor je akustički talas koji se širi kada se vazduh pri izdisaju izbacuje iz pluća, pri čemu prolazi kroz vokalni trakt koji se na odgovarajući način artikuliše kako bi se proizveo željeni glas. Osnovni zakoni fizike mogu opisati nastajanje i propagaciju zvuka u vokalnom sistemu. U cilju formiranja matematičkog modela ovog fenomena neophodno je iskoristiti fundamentalne zakone održanja mase, momenta kao i zakon održanja energije uz korišćenje zakona termodinamike i mehanike fluida, pri čemu se određeni fenomeni prilikom modeliranja govornog signala moraju uzeti u obzir, a to su nestacionarnost i neuniformnost vokalnog trakta, gubici energije prilikom strujanja vazduha usled toplotne kondukcije i viskoznog trenja, krutost sluzokože, radijacija na usnama, ulazak nosne šupljine u sastav vokalnog trakta prilikom izgovaranja nazala i pretpostavke o tipu eksitacije koja može biti impulsne ili složenoperiodične prirode, kao i širokopojasni šum. Oslanjajući se na zakon održanja energije, mase i momenta, date su sledeće jednačine (1) i (2) :

$$-\frac{dp}{dx} = \rho \frac{d(u/A)}{dt} \quad (1)$$

$$-\frac{du}{dx} = \frac{1}{\rho c^2} \frac{d(pA)}{dt} + \frac{dA}{dt} \quad (2)$$

gde je sa  $p=p(x,t)$  označen vazdušni pritisak u vokalnom traktu na poziciji  $x$  mereno od početka vokalnog trakta (početak se smatra na glasnim žicama), sa  $u=u(x,t)$  je označen zapreminski protok vazduha, sa  $\rho$  je označena gustina vazduha u traktu, sa  $c$  brzina zvuka, dok  $A=A(x,t)$  označava površinu poprečnog preseka vokalnog trakta. Rešenje se uobičajeno traži metodama numeričke integracije, pri čemu je, u cilju rešavanja ovih jednačina, važno znati kako se menja površina poprečnog preseka tokom vremena, međutim, ako se uzmu u obzir određene pretpostavke, ovaj problem se ipak može rešiti u zatvorenoj formi. Jedna od najjednostavniji pretpostavki jeste da se govorni signal može modelovati kao uniformna tuba. U ovakvoj aproksimaciji se pretpostavlja da se vokalni trakt ne menja tokom vremena i

da je pri tome površina poprečnog preseka konstantna. Dodatno pojednostavljenje se odnosi na pretpostavku da se pritisak na krajevima usana ne menja. Ono što je, takođe, bitno je da se za pritisak i zapreminski protok vazduha smatra da se ponašaju kao kompleksne sinusoide učestanosti  $\Omega$  čiji su oblici dati formulama (3) i (4) :

$$p(x, t) = P(x, \Omega)e^{j\Omega t} \quad (3)$$

$$u(x, t) = U(x, \Omega)e^{j\Omega t} \quad (4)$$

čijom se zamenom u parcijalne jednačine (1) i (2) dobijaju jednačine (5) i (6):

$$-\frac{dP}{dx} = ZU \quad (5)$$

$$-\frac{dU}{dx} = YU \quad (6)$$

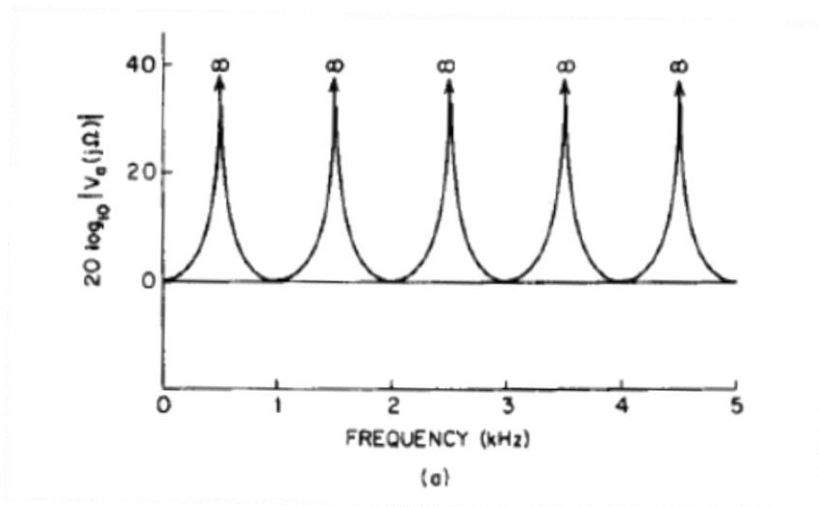
gde se  $Z$  naziva podužnom akustičnom impedansom, a  $Y$  admitansom vokalnog trakta. Rešavanjem ovih diferencijalnih jednačina i nalaženjem odgovarajućih koeficijenata dolazi se do krajnjeg rezultata koji je dat jednačinama (7) i (8) :

$$p(x, t) = jZ_0 \frac{\sin [\Omega(l-x)/c]}{\cos [\Omega l/c]} U_G(\Omega)e^{j\Omega t} \quad (7)$$

$$u(x, t) = \frac{\cos [\Omega(l-x)/c]}{\cos [\Omega l/c]} U_G(\Omega)e^{j\Omega t} \quad (8)$$

gde je  $Z_0 = \rho c/A$ . Ono što je ključno u ovoj analizi jeste zapreminski protok vazduha na vrhovima usana, odnosno frekvencijski odziv sistema na pobudu zapreminskog protoka koji je dat jednačinom (9) i prikazan na Slici 1:

$$\frac{U(l, \Omega)}{U_G(\Omega)} = V_a(j\Omega) = \frac{1}{\cos [\Omega l/c]} \quad (9)$$



Slika 1 - Frekvencijski odziv sistema za model uniformne tube; Slika preuzeta iz [4]

Lokalni maksimumi spektralne gustine snage, odnosno ovog frekvencijskog odziva predstavljaju formantne učestanosti uniformne tube. Koristeći Laplasovu transformaciju moguće je odrediti odziv vokalnog trakta i u slučaju kada se na ulazu ne nalazi prostoperiodična pobuda, već pobuda bilo kog tipa. Međutim, ovo predstavlja jedan idealan model govornog signala, koji se u realnom svetu baš i ne pojavljuje. Kvalitet modela govornog signala može se ogledati u kriterijumima kao što su  $A_1 > A_2 > A_3$  i  $B_1 < B_2 < B_3$  gde A predstavlja vrednost amplitude spektra u dB, dok B predstavlja propusni opseg datog formanta. U ovom slučaju može se primetiti da A teži beskonačnosti, dok B teži nuli. Nešto realniji slučaj jeste kada se na ovaj model doda i uticaj gubitaka energije u vokalnog trakta. Vazдушna struja gubi energiju na osnovu viskoznog trenja na zidovima vokalnog trakta, usled vibracija zidova trakta, kao i na osnovu razmene toplote sa traktom gde se kao rezultat dobijaju promenjene pozicije rezonantnih učestanosti. Može se pretpostaviti da se površina poprečnog preseka menja na sledeć i način formulom (10) :

$$A(x, t) = A_0(x, t) + \delta A(x, t) \quad (10)$$

gde je sa  $A_0(x, t)$  označena nominalna površina poprečnog preseka, a sa  $\delta A(x, t)$  perturbacija u odnosu na ovu nominalnu funkciju. Na osnovu ove promene površine poprečnog preseka, zidovi trakta ulaze u režim oscilacija koje se mogu opisati sledećom diferencijalnom jednačinom drugog reda (11) :

$$m_w \frac{d^2(\delta A)}{dt^2} + b_w \frac{d(\delta A)}{dt} + k_w (\delta A) = p(x, t) \quad (11)$$

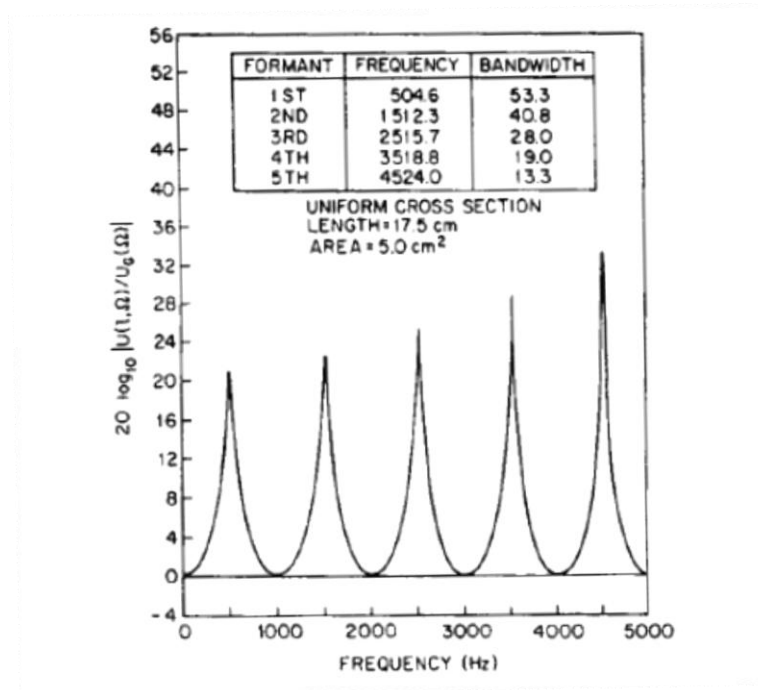
gde je  $m_w(x)$  podužna masa zidova vokalnog trakta,  $b_w(x)$  podužno prigušenje oscilacija, a  $k_w(x)$  koeficijent krutosti zidova. Ponovnim postupkom rešavanja diferencijalnih jednačina dolazi se do rezultata definisanog relacijom (12) :

$$-\frac{dU}{dx} = YP + Y_w P \quad (12)$$

gde je  $Y_w$  dato relacijom (13) :

$$Y_w(x, \Omega) = \frac{1}{j\Omega m_w(x) + b_w(x) + k_w(x)/j\Omega} \quad (13)$$

Nakon date analize, ponovo se računa frekvencijski odziv vokalnog trakta datom relacijom (9) pri čemu se dobijaju rezultati prikazani na Slici 2. Sa slike se može uočiti da više nemamo idealni slučaj, već se sada A i B koeficijenti formantnih učestanosti razlikuju. Štaviše, pozicije formantnih učestanosti su se, takođe, malo pomerile usled uvođenja ove pretpostavke. Takođe se uočava činjenica da je efekat vibracija značajnije izražen na nižim učestanostima. Ovim je dobijen model  $A_1 < A_2 < A_3$  i  $B_1 > B_2 > B_3$  što po zadatim kriterijumima ne predstavlja dobar model govornog signala, stoga će u nastavku biti uvedene dodatne pretpostavke.



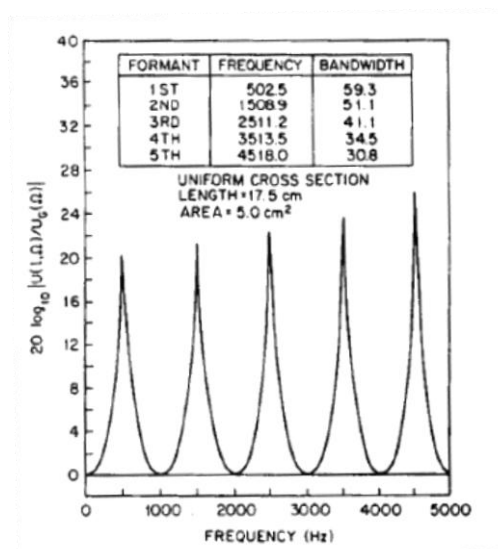
Slika 2 - Frekvencijski odziv uz modeliranje gubitaka usled vibracija; Slika preuzeta iz [4]

Uticaj viskoznog trenja i razmene toplote manje je izražen nego uticaj vibracija, i ogleda se u tome što se menjaju izrazi za  $Z$  i  $Y$  koji su sada dati relacijama (14) i (15) :

$$Z(x, \Omega) = \frac{S(x)}{[A_0(x)]^2} \sqrt{\rho \pi \mu / 2} + j\Omega \frac{\rho}{A_0(x)} \quad (14)$$

$$Y(x, \Omega) = \frac{S(x)(\eta - 1)}{\rho c^2} \sqrt{\frac{\Omega \lambda}{2c_p \rho}} + j\Omega \frac{A_0(x)}{\rho c^2} \quad (15)$$

gde je na Slici 3 prikazan rezultat frekvencijskog odziva u ovom slučaju :

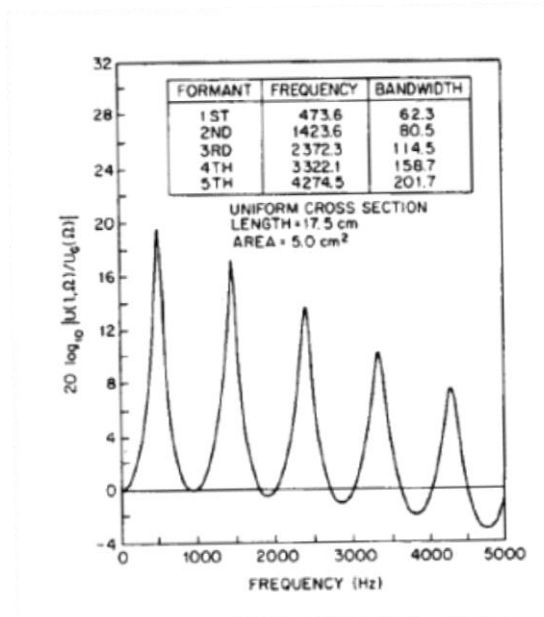


Slika 3 – Frekvencijski odziv uz postojanje viskoznog trenja i razmene toplote; Slika preuzeta iz [4]

Na samom kraju, jedna od pretpostavki koja će biti najbitnija za izradu ovog rada jeste modelovanje radijacije na usnama. U prethodnoj analizi razmatrani su gubici energije koji se dešavaju usled transmisije vazdušnog talasa kroz vokalni trakt, međutim, usvojen je granični uslov koji je zapravo vrlo nerealan. Dovoljno jednostavna pretpostavka koja uzima u obzir otvor usana i rezonatorska svojstva usne duplje jeste da se amplituda pritiska na vrhovima usana modelira kroz takozvanu radijacionu impedansu  $Z_L(\Omega) = \frac{j\Omega L_r R_r}{R_r + j\Omega L_r}$  gde su  $L_r$  i  $R_r$  jednaki  $L_r = \frac{8a^2}{3\pi c}$  i  $R_r = \frac{128}{9\pi^2}$  i nazivaju se radijaciona induktansa i radijaciona otpornost respektivno gde je sa  $a$  označen radijus otvora usana a  $c$  brzina zvuka, pa se pritisak može modelovati relacijom (16) :

$$P(l, \Omega) = Z_L(\Omega) U_l(l, \Omega) \quad (16)$$

Uvođenjem i ove pretpostavke model govornog signala postaje još složeniji. Upoređujući dobijeni rezultat sa prethodnim uočavaju se značajne razlike. Prvo se uočava proširenje rezonantnih učestanosti i najveći uticaj se ogleda na višim učestanosti. Širina prvog formanta je dominantno određena gubicima na zidovima, dok se širina drugog i trećeg formanta određuje uticajem oba efekta (i gubicima i radijacijom). Na Slici 4 prikazan je frekvencijski odziv koji uključuje i ovu pretpostavku, gde se može zaključiti da je ovaj model još bolji.



Slika 4 – Frekvencijski odziv usled modelovanja radijacije na usnama; Slika preuzeta iz [4]

Potrebno je reći da postoje tri tipična mehanizma eksitacije koja se mogu pojaviti, a to je kvaziperiodičan signal u slučaju generisanja samoglasnika i polusamoglasnika, širokopojasni šum pri izgovaranju frikativa, pri čemu dolazi do sužavanja vokalnog trakta gde dolazi do turbulentnog kretanja, zašto se i ovaj tip eksitacije i modeluje kao šum, i treći, impulsni

signal, karakterističan za plosive, kod kojih se vokalni trakt na pojedinim mestima potpuno zatvori, pa se naglim otvaranjem pritisak relaksira, što se može modelovati Dirakovim impulsima.

## 2.2 Analiza govornog signala u vremenskom domenu

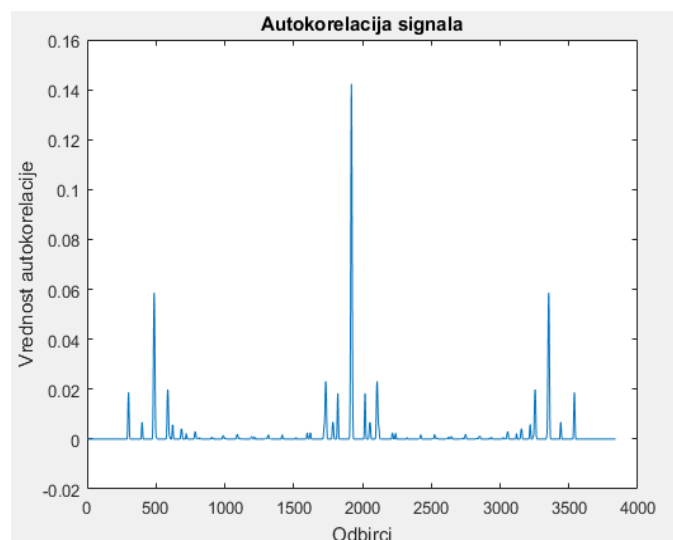
Govorni signal može se obrađivati u više domena, pri čemu obrada u svakom domenu ima specifičnu ulogu i od velikog je značaja za dalje posmatranje govornog signala. Jedan od domena analize govornog signala jeste vremenski domen. Čini se da je najprirodniji način obrade govornog signala u vremenskom domenu, međutim, ovakav pristup sa sobom nosi određeni broj problema. Kada bi se posmatrao govorni signal u vremenskom domenu, moglo bi se uočiti da se prepoznaju delovi koji odgovaraju zvučnom, kao i delovi koji odgovaraju bezvučnom delu signala. To se lako može uočiti na osnovu amplitude govornog signala, međutim, samo posmatranje amplitude nije dovoljno. Kako se u govornom signalu nalazi i dosta šuma, koji je uzrok lošeg kvaliteta mikrofona i raznih drugih smetnji, signal izgleda kao jedan stohastički proces. Može se zaključiti da govorni signal nije stacionaran, pa kako bi analiza signala bila preciznija, potrebno je posmatrati signal na malim vremenskim trenucima koji se nazivaju frejmovi, na kojima se može pretpostaviti da je govorni signal stacionaran. Obrada u vremenskom domenu primenjena je u oblasti segmentacije govornog signala, prepoznavanju govora, određivanja fundamentalne učestanosti i mnogim drugim oblastima obrade. Ono što je glavna tema ovog dela rada jeste dati teorijsku osnovu potrebnu za shvatanje određivanja fundamentalne učestanosti u vremenskom domenu. Postoji više načina kojima se efikasno može odrediti fundamentalna učestanost govornika, kao što je paralelno procesiranje, iako je ova metoda jako osetljiva na parametre, kepsralnom analizom, autokorelacionom metodom itd. U ovom radu biće dat akcenat na određivanju fundamentalne učestanosti pomoću autokorelacione metode.

Jedan od najpreciznijih metoda za određivanje fundamentalne učestanosti jeste metod koji se bazira na proceni autokorelacione funkcije. Autokorelaciona funkcija je jako sadržana informacijama i periodična je funkcija, pa se tako lako može odrediti periodičnost koja potiče od glasnica. Procena fundamentalne učestanosti bazira se na posmatranju kratkovremenskih delova signala, takozvanih frejmova, sa određenim stepenom preklapanja. Ideja je proći kroz sve frejmove signala, odrediti fundamentalnu učestanost za svaki frejm i na kraju proceniti konačnu učestanost. Procenjena vrednost fundamentalne učestanosti nalazi se na mestu pika procene autokorelacione funkcije. Najpre se svaki frejm signala propušta kroz filter čime se propuštaju samo niske komponente učestanosti, znajući da fundamentalna učestanost u proseku ne prelazi vrednost iznad 320 Hz. Nakon toga, na datom segmentu nalazi se maksimum u prvoj trećini i u trećoj trećini segmenta. Od tako dobijene dve vrednosti bira se manja vrednost koja će služiti za klipovanje signala. Klipovanje signala predstavlja odsecanje

nekim delova signala u zavisnosti od postavljenog praga koji se u ovom slučaju postavlja na 60% manjeg maksimuma. Postoji više načina za klipovanje signala, ali ovaj rad baziran je na centralnom klipovanju signala gde se delovi signala koji su manji od praga  $CL$  po apsolutnoj vrednosti zanemaruju, dok se od ostalih delova dati prag oduzima, kako za negativni, tako i za pozitivni deo. Ovaj deo predobrade koristi se kako bi se otklonili brojni pikovi koji potiču od nekih drugih izvora. Nakon klipovanja signala računa se energija signala kao kvadrirana vrednost i određuje prag koji će odlučivati da li je u pitanju zvučni ili bezzvučni deo signala. Potrebno je izračunati autokorelaciju tako dobijenog signala. Kako je često nemoguće simulirati idealan slučaj, na osnovu broja odbiraka datog signala može se izvršiti adekvatna procena autokorelacione funkcije data sledećom relacijom (17) :

$$\widehat{R}_X(k) = \frac{1}{2N+1} \sum_{n=-N}^N x(n)x(n+k) \quad (17)$$

Kako je poznato da se fundamentalna učestanost kreće otprilike između 60 Hz i 320 Hz, uzimajući u obzir dati opseg, kao i informaciju o tome da li je u pitanju zvučni ili bezzvučni deo signala, data procena fundamentalne učestanosti uzima se kao validna ili ne. Posmatrajući grafik dobijene procene autokorelacione funkcije, nalazi se pozicija maksimuma koja direktno daje informaciju o tome kolika je fundamentalna učestanost govornika. Kako bi procena učestanosti bila robusnija, prilikom samog određivanja za svaki frejm, posmatra se prozor od tri procene, dve prethodne i trenutne, gde se kao konačna procena bira medijana ove tri procene. Na samom kraju, računanjem srednje vrednosti svake validne procene, dobija se konačna procena fundamentalne učestanosti za dati govorni signal. Na Slici 5 može se videti jedna od procena autokorelacione funkcije koja se može iskoristiti za određivanje fundamentalne učestanosti.



Slika 5 – Procena autokorelacione funkcije na klipovanom signalu radi određivanja fundamentalne učestanosti



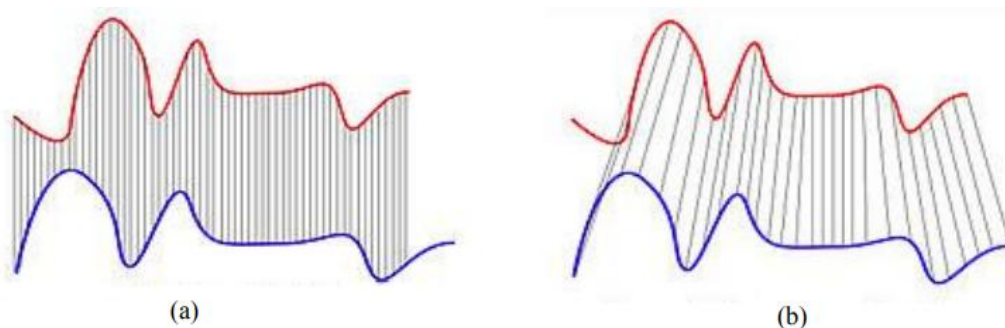
## 2.3 LPC analiza i sinteza govornog signala

LPC analiza predstavlja jednu od jako primenjenih grana obrade govornog signala. Ona polazi od pretpostavke da se signal može opisati linearnim AR modelom, tj. modelom koji ima samo polove, odnosno pretpostavlja se da se on može predstaviti kao neka linearna kombinacija prethodnika. Ovako uzete pretpostavke pokazale su se dobro kada je u pitanju većina glasova, odnosno kod onih glasova sa izraženim rezonantnim učestanostima i slabo izraženim antirezonantnim učestanostima, dok su za glasove sa izraženim antirezonantnim učestanostima rezultati lošiji, kao što su glasovi „n” i „m”. Još jedna od pretpostavki je da se vokalni trakt pobuđuje slučajnim procesom, šumom, za bezvučni govor, i povorkom impulsa za zvučni govor. Rezultat LPC analize je  $p$  konjugovano-kompleksnih parova polova, a svaki par polova je određen sa dva parametra, pa je za određivanje ovih polova potrebno  $2p$  koeficijenata linearnog prediktora. Kada su u pitanju samoglasnici, očekuje se jedan formant na oko 1000 Hz, a pošto obično posmatramo signal do 4 kHz, može se zaključiti da ćemo imati 4-5 formanata. Odatle se može izvesti zaključak da je 10 LPC koeficijenata najčešće dovoljno za analizu, međutim, zbog odstupanja realnih uslova od pretpostavljenih, kao što je radijacija na usnama ili neidealna pobuda, dodatne informacije se moraju uzeti u obzir, pa se pokazuje da je za većinu primena dovoljno uzeti u obzir 14-16 LPC koeficijenata. Kada je u pitanju rešavanje problema u ovom diplomskom radu, posmatrani su rezultati za broj LPC koeficijenata od 16, 20 i 24. Jasno je da što više koeficijenata imamo, to će analiza biti bolja, ali i zahtevnija. Kao što je već poznato da govorni signal nije stacionaran signal, a LPC analizu je moguće uspešno rešavati samo na stacionarnim signalima, potrebno je posmatrati govorni signal u kratkim vremenskim intervalima, takozvanim frejmovima dovoljno malim tako da se može smatrati da je signal stacionaran, kao što je već pomenuto, gde dati frejmovi mogu biti sa određenim stepenom preklapanja ili ne. Vršiti se prozorovanje govornog signala gde se najčešće koristi Hamingov prozor, a dati rezultati u mnogome zavise od izbora tipa prozora, kao i dužine i stepena preklapanja prozora. Oblast primene ove analize je jako velika, koristi se u formantnoj analizi, LPC vokoderima, sintezi govornog signala, određivanju fundamentalne učestanosti, prepoznavanju govora. Jedan od tipičnih algoritama u prepoznavanju govora uz pomoć LPC analize podrazumeva prvobitno segmentisanje signala, odabir prozora i računanje koeficijenata nakon čega se uvidom u koeficijente dolazi do zaključka da različite reči imaju različite vremenske oblike LPC koeficijenata, pa se kao takvi mogu koristiti u obučavanjima neuralnih mreža i klasifikaciji. Metoda za računanje koeficijenata linearnog prediktora ima mnogo. Oni se mogu odrediti bilo kojom AR metodom spektralne estimacije kao što je autokorelacioni metod, kovarijacioni metod, modifikovani kovarijacioni metod, Burgov metod, metod maksimalne verodostojnosti, tehnika inverznih matrica, *lattice* metoda i mnoge druge. Danas, takođe, postoje već ugrađene funkcije koje se bave ovom estimacijom i dodatno olakšavaju priču kada je u pitanju LPC analiza signala. U programskom jeziku MATLAB postoji ugrađena funkcija *lpc()* koja uspešno prediktuje

koeficijente linearnog prediktora, vraćajući kao izlaz *alpha* koeficijente i *gain* pojačanja. Uz ovako izračunate izlaze date funkcije filtriranjem sa odgovarajućim parametrima dobija se rezidualni (eksitacioni) deo signala za svaki frejm koji se kasnije može koristiti za transformaciju govora i određivanje fundamentalne učestanosti. LPC sinteza signala predstavlja inverzan problem analize. Kada se nakon prediktovanja neuralne mreže dobijaju željeni LPC koeficijenti, pristupa se sintezi inverznim postupkom koristeći se *IIR LP Synthesis* filtrom pri čemu se dobija rekonstruisani novi signal koji se nadalje može koristiti. Takođe, treba napomenuti da se pre treniranja neuralne mreže korišćeni LPC koeficijenti konvertuju u LSF koeficijente radi stabilnosti, te da se nakon dobijanja izlaza ponovno vraćaju u prvobitnu formu, o čemu će kasnije više biti reči.

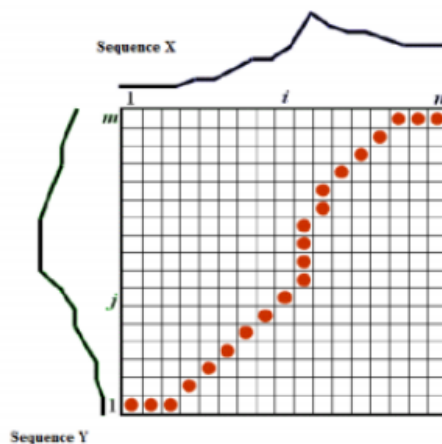
## 2.4 DTW algoritam

DTW algoritam ( *Dynamic Time Warping* ) je algoritam koji izračunava optimalnu putanju savijanja između dva podatka, tako da su izlaz vrednosti iskrivljenosti putanje i udaljenost između dva podatka. Dve iste reči izgovorene od strane istog govornika mogu biti drugačije interpretirane, kao na primer reč luk koja u zavisnosti od odabira akcenta predstavlja različite reči. DTW algoritam rešava ovaj problem tako što ispravno poravnjava reči i izračunava minimalnu distancu između dve reči. Različito vreme usklađivanja govora je osnovni problem za merenje udaljenosti u prepoznavanju govora jer male promene rezultiraju pogrešnom identifikacijom. DTW algoritam je efikasna metoda za rešavanje problema vremenskog poravnanja, stoga je ovaj algoritam realnije koristiti za merenje sličnosti obrazaca, odnosno njihovo podudaranje. Obradeni podaci se uvek nalaze u vremenskoj zoni, tako da se sekvenca koju posmatramo menja u vremenu. Ilustracija ovog problema može se videti na Slici 6, koja prikazuje originalno poravnanje i poravnanje uz korišćenje DTW algoritma.



Slika 6 – (a) Originalno poravnanje dve sekvence; (b) Poravnanje uz DTW algoritam; Slika preuzeta iz [5]

DTW algoritam namenjen je za poravnanje dve sekvence vektora okretanjem vremenske ose više puta sve dok se ne nađe optimalno podudaranje između dve sekvence. Ovaj algoritam deluje kao linearno mapiranje ose radi poravnavanja dva signala gde su u ovom slučaju ta dva vektora odbirci govornog signala polaznog i željenog glasa. Pretpostavimo da imamo dve sekvence vektora u  $n$ -dimenzionlnom stanju:  $X = [x_1 x_2 \dots x_n]$  i  $Y = [y_1 y_2 \dots y_n]$  gde se ilustracija ovog primera može predstaviti Slikom 7. Dve sekvence su poravnate na bočnim stranama, jedna iznad, a jedna sa leve strane i obe sekvence počinju u donjem levom uglu.



Slika 8 – Scatter dijagram podudaranja sekvenci; Slika preuzeta iz [5]

Za svaku ćeliju računa se distanca upoređujući odgovarajuće elemente datih sekvenci. Udaljenost između dve tačke računa se preko Euklidove distance koja je data relacijom (18):

$$Dist(x, y) = |x - y| = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2]^{1/2} \quad (18)$$

Najbolje podudaranje, odnosno poravnanje između ova dva niza je put kroz mrežu (*grid*) koji minimizira ukupnu udaljenost između njih, a koja se naziva globalna udaljenost. Globalna udaljenost izračunava se pronalaženjem i prolaskom kroz sve moguće rute mreže, svaki put računajući ukupnu distancu. Za svaku sekvencu dovoljno dugu, broj mogućih puteva kroz mrežu će biti veliki. Funkcija optimalne vrednosti definisana je kao  $D(i, j)$  kao DTW udaljenost između  $t(i, m)$  i  $r(j, n)$  sa preslikavanjem putanje od  $(i, j)$  do  $(m, n)$  datom relacijom (19), gde su  $t$  i  $r$  posmatrane sekvence :

$$D(i, j) = |t(i) - r(j)| + \min \begin{cases} D(i + 1, j) \\ D(i + 1, j + 1) \\ D(i, j + 1) \end{cases} \quad (19)$$

sa inicijalnim uslovima  $D(n, m) = |t(m) - r(n)|$ .

Što je manja udaljenost definisana, tada će zvuk biti sličniji početnom, drugim rečima, zvuk se može prepoznati. Vreme izgovaranja i jačina signala utiču na rastojanje koje nastaje. Kada je u pitanju primena ovog algoritma u diplomskom radu, on se primenjuje na izračunate LPC koeficijente polaznog i željenog signala prethodno konvertovane u LSF signale. Za izračunate koeficijente svakog frejma računa se euklidsko rastojanje na koje se dodaje izabrani odgovarajući minimum po formuli definisanoj relacijom (19). Vršiti se poravnavanje signala odabirom dobijenih odgovarajućih indeksa tako da posmatrani signali najviše liče jedan na drugi.

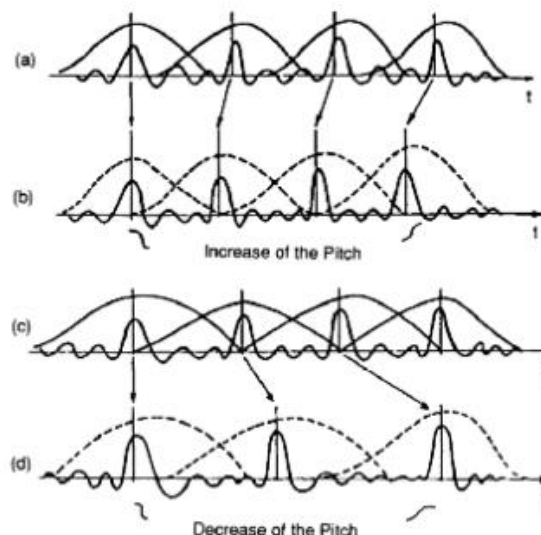
Kao što je rečeno, dinamičko vremensko savijanje (*DTW*) je algoritam za merenje sličnosti između dve vremenske sekvence koje mogu da variraju u vremenu i brzini. Na primer, sličnost obrazaca može da se detektuje korišćenjem ovog algoritma, čak i ako je jedna osoba išla brže od ostalih, ili ako je bilo ubrzanja i usporavanja tokom jednog posmatranja. *DTW* je primenjen na vremenske sekvence video, audio i grafičkih podataka, uopšteno, bilo kakvi podaci koji mogu biti pretvoreni u linearan niz, mogu biti analizirani *DTW* algoritmom. Jedna od bitnih primena je poznata aplikacija za automatsko prepoznavanje govora, gde je problem izboriti se sa različitim brzinama govora.

U principu, *DTW* je metod koji izračunava optimalno podudaranje između dve date sekvence (npr. vremenske serije) sa određenim ograničenjima. Sekvence su "izvijene" nelinearno na vremensku dimenziju kako bi se odredila mera njihove sličnosti nezavisno od nekih nelinearnih varijacija u vremenskoj dimenziji. Ovaj metod sekvenci se često koristi u klasifikaciji vremenskih serija. Iako *DTW* algoritam meri daljinu kao količinu između dve date sekvence, to ne garantuje trougao nejednakosti za održavanje, pri čemu je složenost ovog algoritma srazmerna  $O(n^2)$ .

## 2.5 PSOLA algoritam

PSOLA algoritam ( *Pitch Synchronous Overlap Add* ) predstavlja jedan od najpopularnijih metoda za promenu fundamentalne učestanosti danas. Algoritam je prvobitno otkriven u Francuskoj, u *Telecom*-u gde je nosio naziv *CNET*. On zapravo i nije algoritam sinteze, ali omogućava da se prethodno snimljeni uzorci govora glatko sjedine i pruže dobru kontrolu visine i trajanja, pa se koristi u nekim komercijalnim sistemima sinteze kao što su *ProVerbe* i *HADIFIX*. Postoji nekoliko verzija PSOLA algoritma, ali sve one funkcionišu u osnovi na isti način. Verzija u vremenskom domenu *TD-PSOLA*, najčešće se koristi zbog svoje računске efikasnosti. Osnovni algoritam sastoji se iz tri koraka. Korak analize, u kojem se originalni signal deli na odvojene, ali često preklapajuće kratkovremenske analizirajuće signale, modifikacija tih signala i dobijanje signala za sintezu i, konačno, korak sinteze gde se ovi segmenti rekombinuju koristeći *Overlap Add* metodu. Kratkovremenski signali dobijaju se iz

digitalnog govornog signala prozorovanjem uz korišćenje odgovarajućeg tipa prozora koji je najčešće *Hanning window*, koji je centriran oko uzastopnih delova signala koji se nazivaju "pitch marks", odnosno oko pikova amplitude originalnog signala. Te se oznake postavljaju sinhronom brzinom tona na izraženim delovima signala i stalnom brzinom na deonicama bez glasa. Korišćena dužina prozora proporcionalna je lokalnoj fundamentalnoj učestanosti, što je obično od 2 do 4 takve učestanosti. Dužina prozora ne bi trebalo biti ni prevelika ni premala, kako bi procena bila dobra. Manipulacija osnovne frekvencije postiže se promenom vremenskih intervala. Izmena trajanja postiže se ponavljanjem ili izostavljanjem segmenata govora što u principu podrazumeva modifikaciju trajanja, a u zavisnosti od toga da li je ton željenog govornika viši ili niži. Na kraju, preostali manji segmenti se kombinuju preklapanjem i dodavanjem. Rezultat je signal istog spektra kao original, ali sa drugačijom osnovnom frekvencijom. Tako se menja visina zvuka, ali ostale vokalne osobine ostaju iste. Primer ove obrade može se videti na Slici 9.



Slika 9 – Modifikacija fundamentalne učestanosti zvučnog dela signala na primeru spuštanja date učestanosti; Slika preuzeta iz [6]

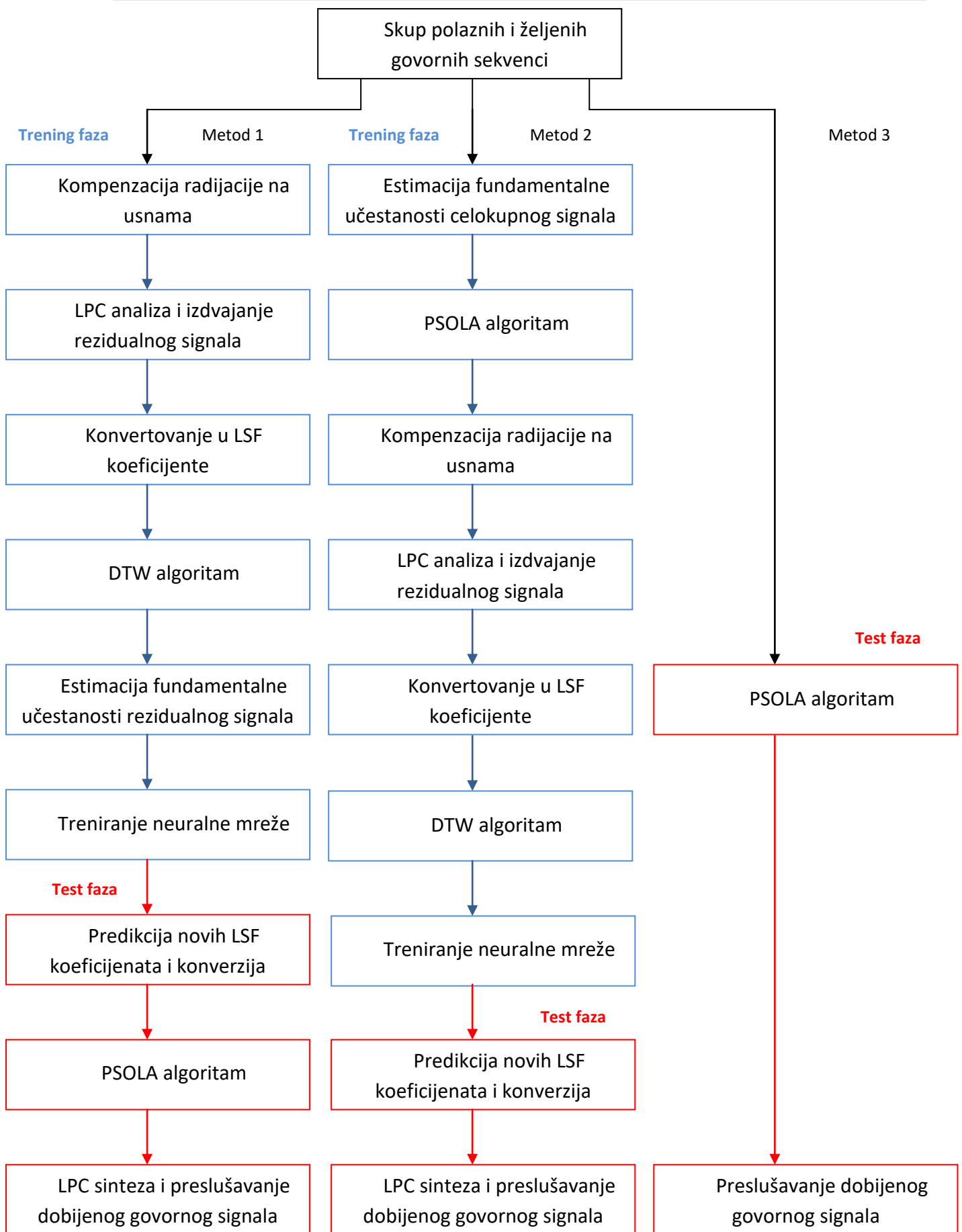
Krajnji segment koji predstavlja preklapanje i dodavanja baziran je na *Overlap Add* metodi koja predstavlja efikasno izračunavanje diskretne konvolucije veoma dugačkog signala i niskofrekventnog *FIR* filtra. Ideja je da se problem podeli na više manjih problema, odnosno da se ulazni signal podeli na male delove, što se već podrazumeva prethodnom obradom. Sistem je sledeći, signal se podeli na kratke delove, provlači se kroz odabrani filter i tako svaki obrađeni frejm slaže u novi dobijeni signal, čime se dobija krajnji signal. Postoje i druge varijacije *PSOLA* algoritma kao što su *FD-PSOLA* ( *Frequency Domain PSOLA* ) i *LP-PSOLA* ( *Linear Predictive PSOLA* ) i oni predstavljaju teoretski pogodnije pristupe modifikacije fundamentalne učestanosti jer omogućavaju nezavisnu kontrolu nad spektralnom anvelopom sintetizovanog signala. *FD-PSOLA* koristi se samo za skaliranje fundamentalne učestanosti i *LP-PSOLA* se koristi u rezidualnim vokoderima. Visina tona

može se odrediti samo za zvučne delove signala, primenjivanjem na bezvučnim delovima signala može se dobiti samo tonski šum.

Dati algoritam može se realizovati na više načina. Može se implementirati kod samostalno na bazi datih teorijskih osnova, međutim, kako je ova tema jako popularna i istražena u poslednje vreme, došlo je do razvoja ovog algoritma u širokom spektru, primenjuje se u raznim oblastima, pa je stoga došlo i do razvoja jednog od jako korisnih programa koji se koristi u analizi i sintezi govornog signala, a taj program nosi naziv *Praat*, čiji su izumitelji profesori kolumbijskog univerziteta. Jedan od razloga što se ljudi odlučuju da koriste ovaj program u obradi govornog signala nije samo to što on pokriva dosta mogućnosti i opcija za obradu, već i zato što je besplatan. Praat je dizajniran tako da se dobro uklapa uz programske jezike kao što su *MATLAB*, *SPSS*, *Excel*, itd. Predstavlja jedan od verovatno najopsežnijih paketa alata za fonetska istraživanja širom sveta i podmlađuje se neverovatnom brzinom. Često isti zahtevi mogu biti odrađeni na više načina, kao što je npr. ekstrakcija fundamentalne učestanosti koja se može uraditi na četiri načina: autokorelacionom metodom, kroskorelacionom metodom, *SPINET* metodom i subharmoničnim sumiranjem. Pomoć je dostupna za svaki od algoritama, objašnjavanje značenja mnogih vrednosti parametara koji se mogu specificirati za svaki algoritam i pružanje referenci na literaturu su dostupni u velikoj meri. Više o samoj primeni ovog algoritma i postupku transformacije govora biće reči u nastavku u poglavljima Metodologija rada i Implementacija i rezultati.

### 3 METODOLOGIJA RADA

Kao što je već poznato, cilj ovog rada je što bolja implementacija algoritma transformacije govora baziranog na promeni fundamentalne učestanosti i LPC analize i sinteze signala. Na raspolaganju je set podataka koji se sastoji od 115 rečenica izgovorenih od strane muške i ženske osobe. Ideja je iskoristiti jedan deo seta za treniranje neuralne mreže, dok se ostatak može iskoristiti kao test skup. 100 nasumično izabranih parova datih rečenica uzeto je kao trening skup, dok se preostalih 15 rečenica koristi kao test skup podataka. Jasno se može uočiti da su date govorne sekvence jako dobrog kvaliteta, sa malom količinom smetnji, što se i moglo očekivati pošto su snimane u studijskim uslovima. Pristupanjem rešavanju ovog problema treba uzeti u obzir sve do sada teorijski navedene informacije i potrebne algoritme i to ukombinovati u jednu kompaktnu celinu koja će uspešno rešavati zadati problem. Implementacija ovog algoritma za transformaciju govora rađena je u programskom jeziku MATLAB, verziji R2017a. Takođe, potrebno je u datom programskom jeziku imati instaliran *Deep Learning Toolbox* paket kako bi bilo omogućeno treniranje neuralne mreže. Pored ovih dodatnih paketa, potrebno je instalirati i *Praat.exe* softverski paket za brzu i preciznu analizu i sintezu govornog signala u kom će se efikasno izvršavati deo koji se odnosi na *PSOLA* algoritam. Samo rešavanje sastoji se iz 3 metode koje će biti upoređivane. Prva metoda odnosi se samo na menjanje fundamentalne učestanosti govornog signala, pri čemu ostala svojstva ostaju ista. Za ovu metodu nije potrebno treniranje neuralne mreže kao za druge dve metode. Što se tiče druge dve metode, one se odnose na promenu fundamentalne učestanosti uz promenu i ostalih svojstava govornog signala, a to je omogućeno uz pomoć neuralne mreže. Neuralna mreža prima kao obeležja LPC koeficijente svakog para polaznog i željenog signala trening skupa koji su prethodno konvertovani u LSF koeficijente radi stabilnosti. Ova neuralna mreža trenira se za svaki od ova dva metoda za broj LPC koeficijenata jednak 16, 20 i 24 gde se u zavisnosti od povećavanja broja koeficijenata povećava i broj neurona u skrivenom sloju respektivno 27, 34 i 50. Neuralna mreža ima jedan skriven sloj sa datim brojem skrivenih neurona. Nakon treniranja neuralne mreže, pristupa se testiranju pri čemu se kao ulaz prosleđuje polazni test govorni signal, a potrebno je dobiti željeni signal. Na sledećoj strani nalazi se blok dijagram implementacije ovog algoritma o čijim će detaljima biti reči u nastavku.





Kao što je rečeno, potrebno je date algoritme opisane u prethodnom poglavlju ukombinovati u jednu kompaktnu celinu za efikasno rešavanje problema. Predlog konstrukcije algoritma dat je odgovarajućim blok dijagramom. Najpre, na samom početku na raspolaganju su govorne sekvence jednog i drugog pola koje će se koristiti u daljoj obradi. U ovom radu razmatrana su tri metoda obrade, a prvo će biti objašnjen metod broj 3.

Metod 3 bazira se samo na promeni fundamentalne učestanosti date govorne sekvence i ne zahteva postojanje neuralne mreže. Polazna test sekvenca dovodi se na segment obrade koji predstavlja *PSOLA* algoritam. Ovaj algoritam vrši mapiranje fundamentalne učestanosti sa polazne na željenu učestanost. Najpre se signal deli na male vremenske trenutke, nakon čega se vrši sužavanje ili širenje signala u zavisnosti od toga da li želimo višu ili nižu krajnju fundamentalnu učestanost i na kraju se dolazi do *Overlap Add* algoritma koji predstavlja diskretnu konvoluciju signala sa određenim filtrom. Ostvarenje ovog algoritma omogućeno je uz pomoć Praat softverskog paketa za fonetsku analizu i sintezu. Naime, implementirana je skripta u okviru softverskog paketa, koja na osnovu ulaznog signala koji se obrađuje, njegove dužine i skala faktora polazne i željene fundamentalne učestanosti uz par koraka koji predstavljaju prethodno navedene korake brzo i precizno izvršava željenu radnju. Skala faktor je veći od 1 ako je u pitanju prelazak sa niže na višu fundamentalnu učestanost, a manji od 1 u suprotnom slučaju. Više o ovom algoritmu biće priče u narednom poglavlju Implementacija i rezultati. Što se tiče skala faktora i fundamentalnih učestanosti, prilikom treniranja neuralne mreže za ostala dva metoda, pored LPC koeficijenata potrebnih za njeno treniranje, čuvaju se i vektori koji predstavljaju procenjene fundamentalne učestanosti polaznih i željenih govornih sekvenci. Ove procene bazirane su na autokorelacionoj metodi procene fundamentalne učestanosti opisane u prethodnom poglavlju. Ovako dobijeni vektori koriste se kao ulazi u *Log Linear Transform* funkciju koja date vektore prebacuje u logaritamsku raspodelu, nalazi njihove srednje vrednosti i standardne devijacije, pa na osnovu datih informacija i date polazne fundamentalne učestanosti test signala prediktuje vrednost željene fundamentalne učestanosti koja se koristi u skala faktoru. Nakon izvršavanja algoritma dobija se signal sa promenjenom fundamentalnom učestanosti, ali bez promene ostalih svojstava, zbog čega se odaje utisak samo povišenog, odnosno sniženog tona koji ne deluje prirodno, bez promene boje i ostalih osobina samog glasa, zbog čega i ovaj metod predstavlja jednu od najlošijih implementacija.

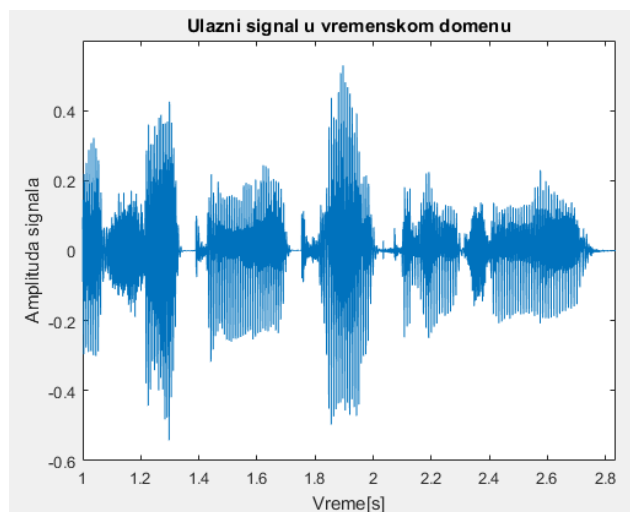
Kada je u pitanju metod 1, eksperimentalnim putem i upoređivanjem rezultata, došlo se do zaključka da se bolji rezultati dobijaju ako se izvrši kompenzacija radijacije na usnama. To je jedna pojava o kojoj je bilo reči u delu o modeliranju govornog signala, a koja u mnogome utiče na sam glas i njegovo stvaranje. Modeliranje kompenzacije radijacije na usnama uspostavljeno je provlačenjem početnog signala kroz filter sa parametrima za B jednakim 1 i -0.9375 i A jednakim jedinici. Nakon kompenzacije radijacije na usnama, vrši se LPC analiza signala kojima se dobijaju koeficijenti linearnog prediktora i rezidualni deo signala. Kako bi se očuvala stabilnost, dati koeficijenti konvertuju se u LSF koeficijente pre treniranja neuralne

mreže. Takođe, pre nego što se dati koeficijenti proslede neuralnoj mreži, provlače se kroz DTW algoritam koji je opisan ranije. Vršiti se dinamičko vremensko savijanje u kome dolazi do poravnavanja date dve sekvence koje se koriste. Prilikom izgovaranja rečenica, prosto je nemoguće da se za dato vreme sve reči izgovore potpuno isto, neke se izgovore duže, neke kraće, stoga ovaj algoritam služi da popravi ovakve nuspojave nastale snimanjem sekvenci kako bi procena bila što preciznija. Ovako istrenirana mreža koristi se za predikciju novih LSF koeficijenata test sekvence. Nakon predikcije koeficijenti se ponovno konvertuju u LPC koeficijente koji se nadalje koriste. Onda se koristi već pomenuta funkcija za predikciju nove fundamentalne učestanosti koja se zajedno sa ovako dobijenim signalom prosleđuje *PSOLA* algoritmu koji vrši mapiranje datog signala na željenu fundamentalnu učestanost. Na samom kraju vrši se LPC sinteza dobijenog signala koja predstavlja inverznu akciju od LPC analize, nakon čega se kranji signal može preslušati.

Metod 2 ima dosta sličnosti sa metodom 1. Najpre, polazi se od istih signala, pri čemu se odmah vrši procena fundamentalne učestanosti sekvenci na celokupnom signalu, za razliku od metode 1 u kojoj se procena vršila na rezidualnom delu signala. Nakon toga, signal se prosleđuje u *PSOLA* algoritam koji vrši mapiranje fundamentalne učestanosti na željenu. Na ovako dobijenom signalu vrši se kompenzacija radijacije na usnama, LPC analiza, konvertovanje koeficijenata u LSF i DTW algoritam. Sada se neuralna mreža trenira LSF koeficijentima signala već promenjene fundamentalne učestnosti i koeficijentima željenog signala. Nakon treniranja mreže, dolazi se do faze testiranja u kojoj se postupak ponavlja sa test sekvencom. Radi se procena fundamentalne učestanosti test sekvence, procena željene učestanosti i nakon toga mapiranje test sekvence na željenu učestanost. Onda se vrši kompenzacija radijacije na usnama, procenjuju se LPC koeficijenti i prosleđuju se neuralnoj mreži koja prediktuje nove LSF koeficijente koji se potom konvertuju ponovo u LPC koeficijente uz pomoć kojih se LPC sintezom dolazi do konačnog signala koji se može preslušati. Ono što se može zaključiti je to da su rezultati dobijeni ovim dvema metodama doveli do znatno boljih procena nego kada je to u pitanju sa metodom 3, baš zato što se uzimaju u obzir i ostale osobine govornog signala. Nešto više o samoj implementaciji ovih algoritama i detaljnom opisu biće priče u sledećem poglavlju koje nosi naziv Implementacija i rezultati.

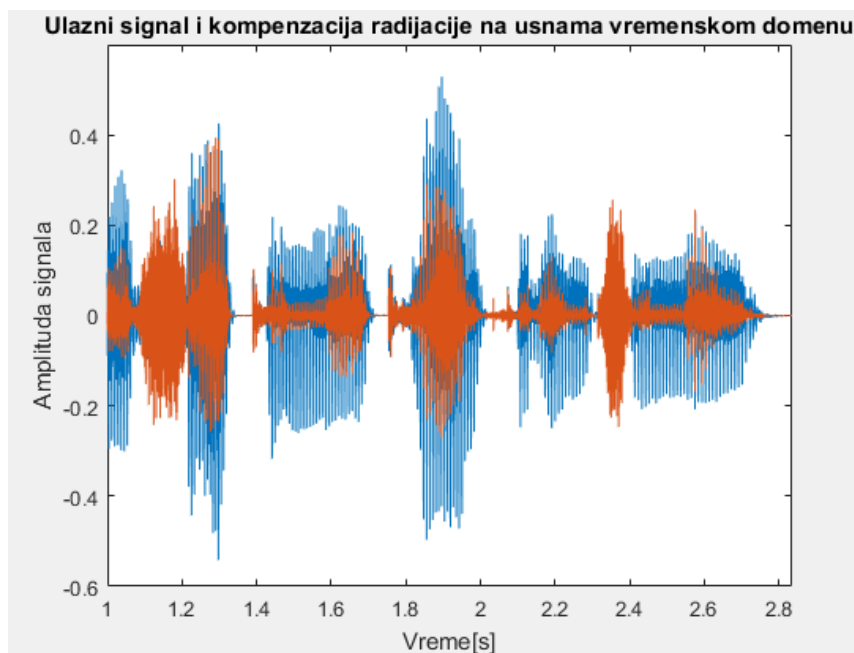
### 3 IMPLEMENTACIJA I REZULTATI

U ovom delu biće više reči o samoj implementaciji algoritma, njegovim detaljima i dobijenim rezultatima posmatranim na jednom egzaktnom primeru. Na samom početku moguće je izabrati željeni metod, broj LPC koeficijenata, tip prozora koji se koristi, broj odbiraka za treniranje, prisutnost ili odsutnost kompenzacije radijacije na usnama kao i da li želite trenutno da trenirate neuralnu mrežu ili ne. Ova opcija se daje zato što već postoje istrenirani modeli za svaku od kombinacija izabranih parametara iz razloga što treniranje mreže traje dugo, nešto više od sat vremena po modelu, stoga da bi se korisniku smanjilo utrošeno vreme posmatrajući ovu implementaciju, na raspolaganju su istrenirani modeli, a ovde će biti opisan celokupni algoritam, zajedno sa pojedinostima koji su bitni za treniranje date mreže. Potom je na korisniku opcija da na osnovu izabranog metoda odredi koji će se folder odnositi na polazne sekvence, a koji na željene sekvence za treniranje. U zavisnosti od broja LPC koeficijenata posmatranih u datoj verziji algoritma, bilo to 16, 20 ili 24 koeficijenta, koristi se odgovarajući broj neurona u skrivenom sloju respektivno 27, 34 i 50. Neuralna mreža modelovana je kao *feedforward* mreža uz pomoć ugrađene funkcije *newff()* i sastoji se iz jednog ulaznog i jednog izlaznog sloja, kao i jednog skrivenog sloja, sa podešenim parametrima *trainFcn='trainscg'*, *maxfail=100000*, *epochs='100000'* i *time=420* nakon čega se pristupa treniranju. Na početku, pre samom procesa treniranja, pri svakoj iteraciji algoritma za pripremu obeležja za treniranje neuralne mreže vrši se učitavanje odgovarajućih parova govornih sekvenci uz pomoć ugrađene funkcije *audioread()*. Na Slici 10 prikazan je izgled jednog od ulaznih signala u vremenskom domenu kada je u pitanju rečenica "Do you know that you are shaking my confidence in you?" izgovorena od strane muške osobe.



Slika 10 – Prikaz ulaznog govornog signala u vremenskom domenu

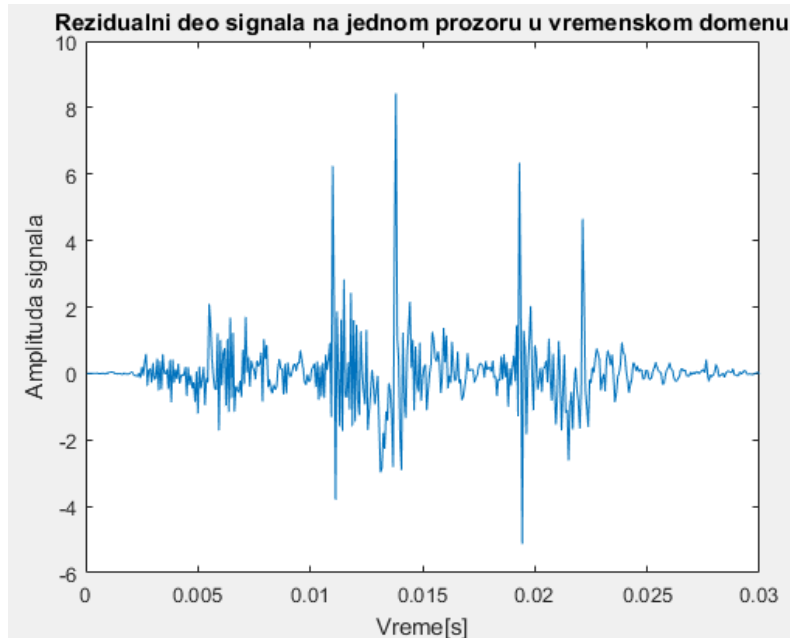
U ovom delu biće detaljno opisan postupak rešavanja prve metode, dok je postupak za rešavanje druge metode analogan rešavanju metode 1 sa izmenjenim redosledom dešavanja, kao i par promena na kojima će biti stavljen akcenat. Što se tiče metode 3, ona predstavlja samo jedan od delova koji su obuhvaćeni ovim dvema metodama, stoga je samo analiziranje njega suvišno. Krajnji rezultati sve tri metode biće upoređeni nakon opisa datih algoritama. Na osnovu odabira parametra na samom početku koji se tiče kompenzacije radijacije na usnama, pristupa se ovom filtriranju provlačenjem početnog signala kroz filter sa parametrima za B jednakim 1 i -0.9375 i A jednakim jedinici pomoću ugrađene funkcije *filter()*. Nakon date kompenzacije dobija se signal dat na Slici 11.



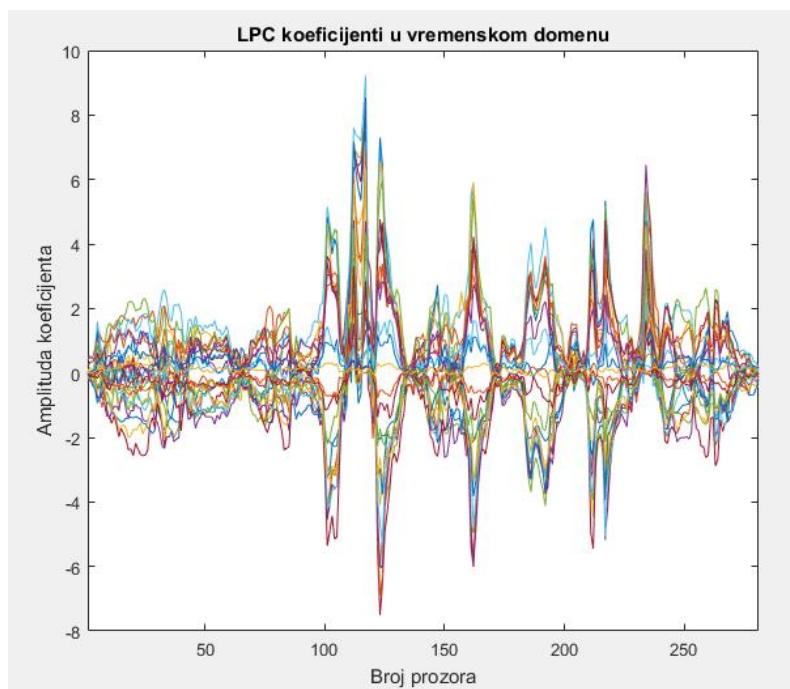
Slika 11 – Prikaz ulaznog signala i signala sa kompenzacijom radijacije na usnama

Na datoj slici vidi se da se filtrirani signal razlikuje od polaznog signala, tj. da su potisnute neke komponente polaznog govornog signala. Nakon kompenzacije radijacije na usnama pristupa se LPC analizi. Ova analiza vrši se u okviru implementirane funkcije *LPC\_analiza()* kojoj se prosleđuju dati signali, kao i njihove učestanosti odabiranja, broj LPC koeficijenata, tip prozora, dužina prozora i koeficijent preklapanja prozora. Treba napomenuti da su signali snimani u studijskim uslovima sa učestanošću odabiranja od 16 kHz. Prozori su dužine 30 ms, sa 10 ms preklapanja, a funkcijom *windowChoice()* vrši se izbor tipa prozora za koji se najčešće uzima *Hanning window*. U samoj funkciji vrši se prozorovanje signala, pri čemu svaka kolona posmatranog vektora predstavlja jedan prozor govornog signala. Na svakoj koloni, odnosno prozoru govornog signala, radi se LPC analiza korišćenjem ugrađene funkcije *lpc()* čime se dobijaju *alpha* koeficijenti i *gain* pojačanja kao izlazi funkcije za svaki prozor. Filtriranjem datog prozora korišćenjem funkcije *filter()* sa parametrima *alpha* i *gain* kao

koeficijentima za B i A filtra, respektivno, dolazi se do rezidualnog dela signala datog prozora. Na Slici 12 može se videti izgled rezidualnog dela signala na primeru 100. po redu prozora u vremenskom domenu dok se na Slici 13 može videti kako se LPC koeficijenti menjaju od prozora do prozora.

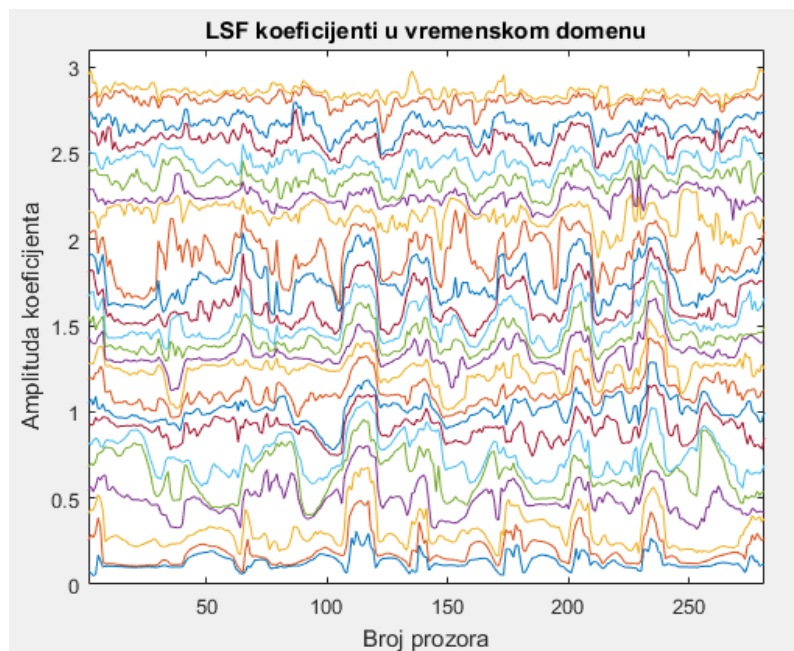


Slika 12 – Prikaz rezidualnog dela signala na 100. prozoru u vremenskom domenu



Slika 13 – Prikaz promene LPC koeficijenata u vremenskom domenu u zavisnosti od prozora

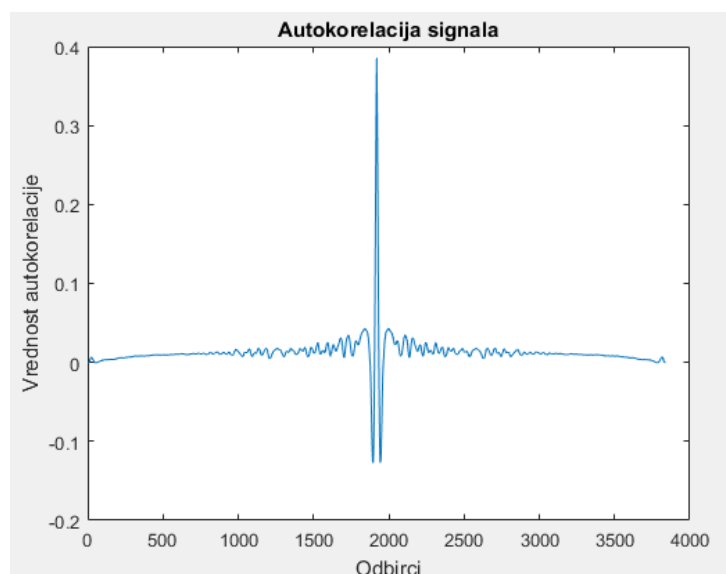
Nakon LPC analize signala dobijeni koeficijenti konvertuju se u LSF koeficijente radi stabilnosti jer su mnogo robusniji u odnosu na LPC koeficijente. Data konverzija vrši se uz pomoć implementirane funkcije *lpc2lsf()* tako što se svaki prozor LPC koeficijenata provlači kroz ugrađenu funkciju *poly2lsf()*. Na Slici 14 može se videti izgled LSF koeficijenata izračunatih na svakom prozoru.



Slika 14 – Prikaz LSF koeficijenata izračunatih na svakom prozoru

Nakon konvertovanja, ovakvi signali polaznih i željenih sekvenci propuštaju se kroz DTW algoritam koji je detaljno opisan u prethodnom poglavlju. Ovaj algoritam služi da popravi neke delove signala, odnosno da izvrši poravnanje datih sekvenci u vremenskom domenu. To znači da je moguće, što je sasvim realan slučaj, da neke reči nisu izgovorene istovremeno, istom dužinom, stoga je potrebno izvršiti njihovo poravnanje. Ovaj algoritam bazira se na sličnosti datih signala računajući njihovu euklidsku distancu kao i minimume susednih odbiraka i tako dolazeći do krajnjeg rezultata. Za ovaj deo obrade implementirana je funkcija *dtws()* kojoj se prosleđuju LSF koeficijenti polazne i željene sekvence. Na samom početku računa se euklidsko rastojanje za svaki odbirak dato relacijom (18), nakon čega se na osnovu relacije (19) odgovarajućim elementima dodaju pronađeni minimumi. Na osnovu ovako izračunatih formula, dolazi se do pronalaženja odgovarajućih indeksa na osnovu kojih se vrši poravnavanje signala. Ovako poravnati signali spremni su za ulazak u neuralnu mrežu. Svaki od ovih signala pakuje se u vektor obeležja koji će služiti za treniranje neuralne mreže. Takođe, pored vektora koeficijenata, prave se i vektori procenjenih fundamentalnih učestanosti polaznih i željenih sekvenci. Procena fundamentalnih učestanosti bazira se na primeni autokorelacione metode. Rezidualni signal kao i frekvencija odabiranja prosleđuju se

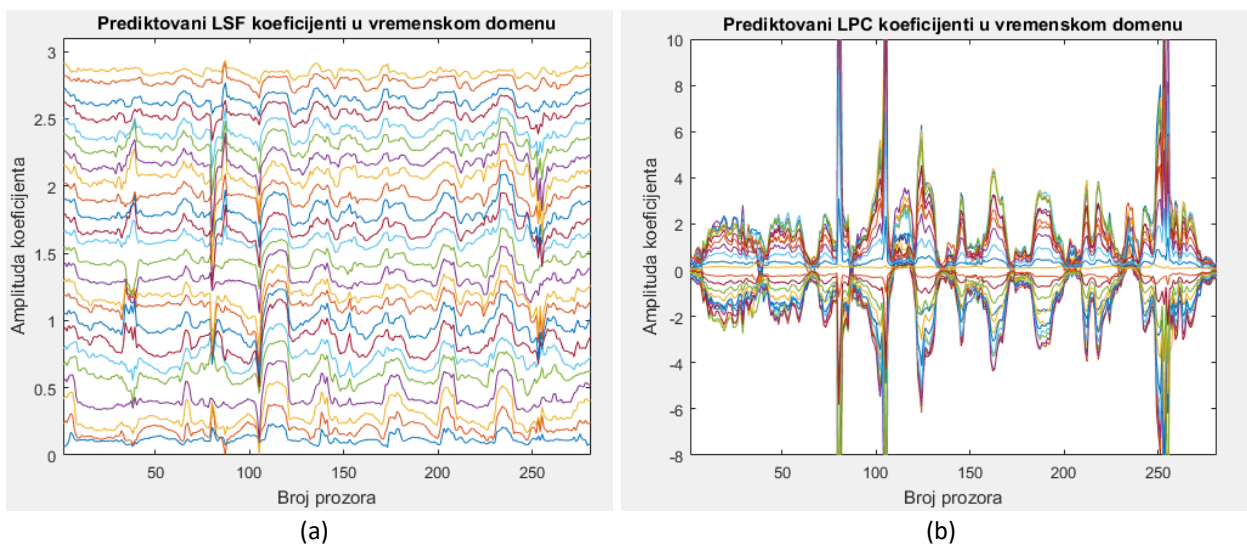
implementiranoj funkciji *Pitch\_estimation()* koja za ulogu ima da proceni date učestanosti. Procena se radi na prozorima dužine 120 odbiraka sa 10 odbiraka preklapanja. Na svakom prozoru vrši se računanje autokorelacione funkcije na osnovu implementirane funkcije *pcorr()* kojoj se prosleđuju prozor i frekvencija odabiranja. Data funkcija unutar glavne funkcije bazira se na računanju autokorelacije na sledeći način. Najpre, prozor se filtrira Batervortovim niskofrekventnim filtrom, sudeći po tome da znamo da se fundamentalna učestanost odrasle osobe može nalaziti u opsegu otprilike od 60 Hz do 320 Hz. Onda se nalazi maksimum signala u okviru prve trećine i treće trećine signala. Od tako dva dobijena maksimuma uzima se manji kao prag za klipovanje signala. Kao metoda klipovanja signala uzima se metoda centralnog klipovanja, gde se signalima koji su po apsolutnoj vrednosti manji od praga  $CL$  upisuje nula, dok se signalima koji su veći ili za negativne delove manji od praga, dati prag oduzima. Na ovaj način se omogućava to da se smanjuje broj pikova autokorelacione funkcije koji omogućavaju tačniju procenu fundamentalne učestanosti jer se ovim putem uklanjaju pikovi funkcije koji ne potiču od glasnica, već od nekih drugih delova. Na svakom prozoru računa se energija signala na osnovu koje se eksperimentalno dolazi do praga za ocenjivanje zvučnog i bezzvučnog dela signala. Na osnovu informacije o tome da li je u pitanju zvučni ili bezzvučni deo signala, kao i informacije da li je procenjena učestanost u opsegu između 60 Hz i 320 Hz, data procena se uzima kao validna ili se odbacuje. Na ovaj način dobijamo procenu za svaki prozor. Paralelno sa ovom procenom, kada se posmatra trenutno dobijena procena prozora, posmatraju se i prethodne dve procene, gde se kao konačna procena uzima medijana date tri procene kako bi konačna procena bila robusnija. Na Slici 15 vidi se jedna od procena autokorelacione funkcije prozora govornog signala.



Slika 15 – Procena autokorelacione funkcije jednog od prozora rezidualnog dela govornog signala



Procena fundamentalne učestanosti dobija se na osnovu pozicije pika procene autokorelacione funkcije. Kada se pozicija pika ukombinuje sa postavljenim pragovima za minimalnu i maksimalnu vrednost fundamentalne učestanosti i podeli sa frekvencijom odabiranja, dobija se data procena. Nakon ove funkcije, vektori obeležja su spremni. Vrš se treniranje neuralne mreže sa gore navedenim parametrima u zavisnosti od izabrane metode i broja LPC koeficijenata. Ovim je trening faza završena. Nakon trening faze nastupa test faza. Za testirajući signal ceo proces se ponavlja do dobijanja LSF koeficijenata. Nakon toga, dobijeni LSF koeficijenti prosleđuju se istreniranoj neuralnoj mreži koja prediktuje nove vrednosti LSF koeficijenata. Ova predikcija omogućena je prostim pozivanjem neuralne mreže sa dati parametrima *netLSF()*. Nakon predikcije, potrebno je procenjene koeficijente vratiti u LPC domen. To se obavlja uz pomoć implementirane funkcije *lsf2lpc()*. Ova funkcija na svakom prozoru vrši konverziju datih koeficijenata na osnovu ugrađene funkcije *lsf2poly()* i *polystab()* kojima se prosleđuju koeficijenti respektivno. Ovim je uspešno izvršena konverzija ponovo u LPC koeficijente. Na Slici 16 mogu se videti prediktovani LSF i LPC koeficijenti.

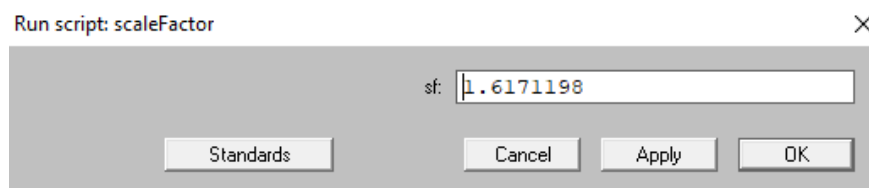


Slika 16 – Prikaz prediktovanih LSF i LPC koeficijenata na prozorima u vremenskom domenu

Jednom kada su dobijeni koeficijenti, može se pristupiti *PSOLA* algoritmu. Pre toga, potrebno je prediktovati željenu fundamentalnu učestanost govornog signala. Ovo je omogućeno uz pomoć implementirane funkcije *logLinearTransform()* kojoj se prosleđuju vektori procenjenih fundamentalnih učestanosti trening signala, kao i procenjena učestanost test signala. Dati vektori prebacuju se u logaritamsku raspodelu, nalaze se srednje vrednosti i standardne devijacije datih vektora, pa se na osnovu njih prediktuje očekivana fundamentalna učestanost. Nakon predikcije, procena se iz logaritamskog domena vraća u linearni domen. Ovim se dobija precizna procena tražene fundamentalne učestanosti čime je izbegnuto treniranje još jedne neuralne mreže koje bi dovelo do dodatne vremenske kompleksnosti ovog algoritma. Kada je procena napravljena, moguće je odrediti skala faktor fundamentalnih učestanosti koji se računa kao količnik tražene i polazne fundamentalne

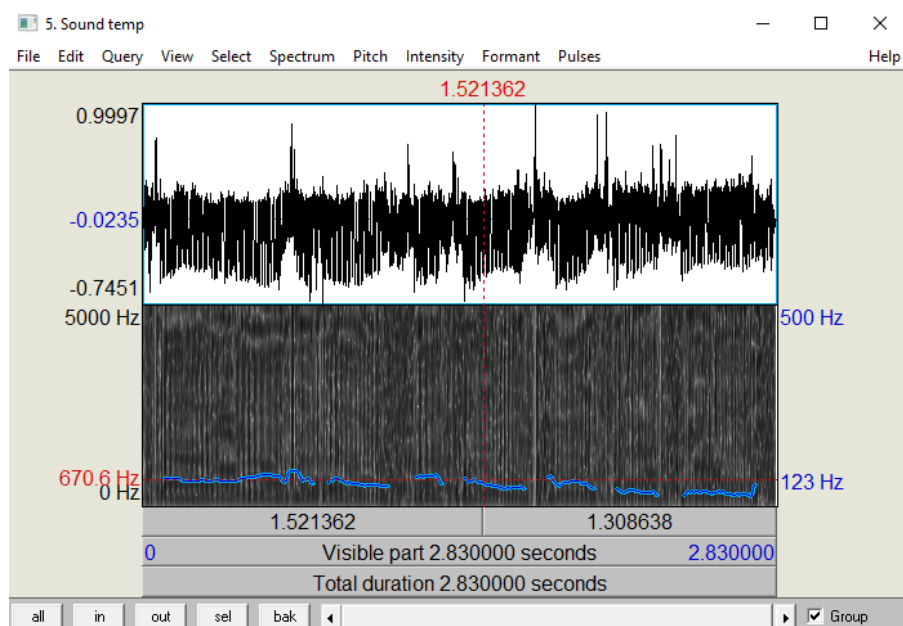


učestanosti. Skala faktor je veći od jedinice onda kada je u pitanju prelazak sa niže na višu fundamentalnu učestanost, a manji u suprotnom slučaju. Za promenu fundamentalne učestanosti polaznog signala koristi se softverski paket *Praat* u kome je implementiran *PSOLA* algoritam. U svakom pokretanju ovog algoritma pri treniranju neuralne mreže za metod 2 i pri testiranju pozivom funkcije `system( 'Praat.exe --run PSOLA.praat' + ' ' + sf )` izvršava se skripta koja obavlja ovaj algoritam. On radi po principu koji je objašnjen u poglavlju pre. Cilj je izdeliti signal na kratke vremenske delove, izvršiti dodavanje ili oduzimanje nekih delova signala u zavisnosti od toga da li želimo višu ili nižu krajnju učestanost, nakon čega se *Overlap Add* metodom radi diskretna konvolucija signala sa filtrom, čime se dobija konačni izgled signala. Radi pojašnjenja same strukture i procesa, u sledećem delu biće detaljno prikazani koraci rešavanja u ovom softverskom paketu. Najpre se vrši unos skala faktora prethodno izračunatog na Slici 17.



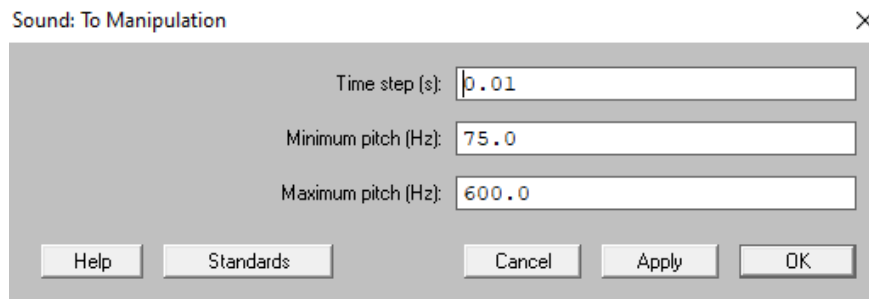
Slika 17 – Unos odgovarajućeg skala faktora

Opcijom *Open>>Read from file* u softverskom paketu izabrati signal koji se testira. Ovom opcijom otvara se niz funkcija sa desne strane kojima se može obrađivati dati signal. Primititi da je dijapazon funkcija znatno veliki. Najpre, signal se može prikazati u vremenskom domenu klikom na funkciju *View&Edit*. Pored ove opcije, postoji i mnogo drugih opcija kao što su prikaz procene fundamentalne učestanosti na zvučnim delovima signala, prikaz formanta, preslušavanje signala itd. Pogledati Sliku 18.



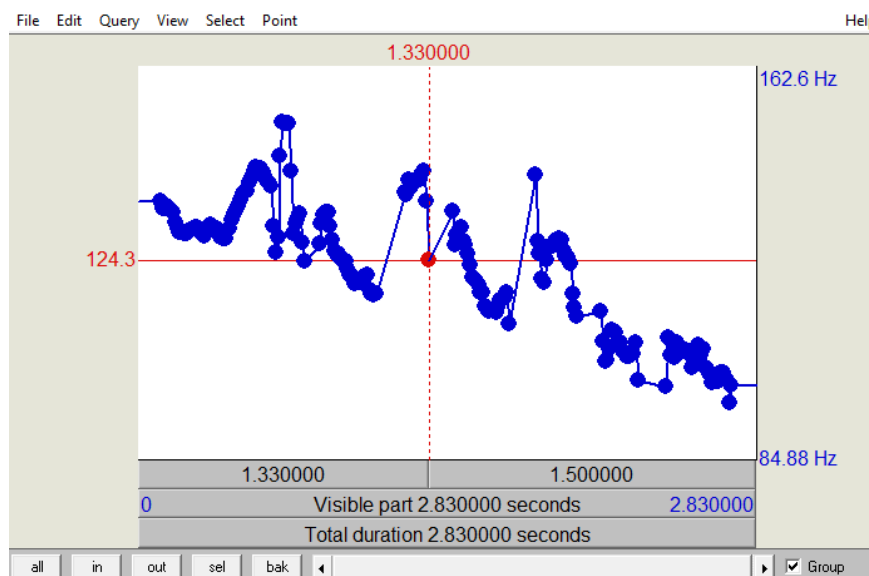
Slika 18 – Prikaz ulaznog signala za obradu

Klikom na funkciju *Manipulation*, otvara se opcioni prozor u kome treba izabrati *To Manipulate* čime se otvara prozor na Slici 19 u koji treba uneti odgovarajuće podatke kao što su minimalna i maksimalna fundamentalna učestanost i vremenski korak.



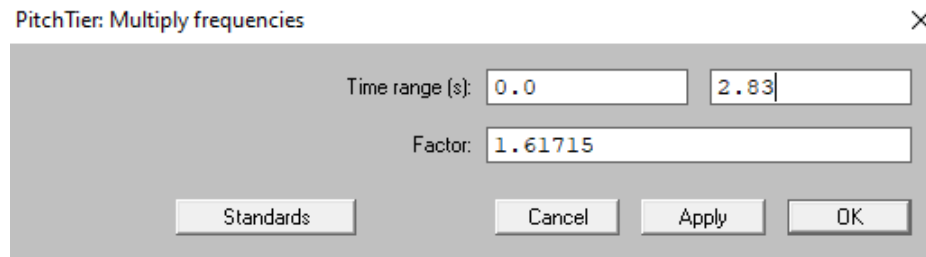
Slika 19 – Unos odgovarajućih parametara u funkciji *Manipulate*

Nakon ove akcije, u komandnom prozoru pojavljuje se temporarni signal za manipulaciju. Njegovim selektovanjem i klikom na funkciju *Extract Pitch Tier* dobijamo procenu fundamentalne učestanosti koja se može prikazati Slikom 20.



Slika 20 – Prikaz procenjene fundamentalne učestanosti

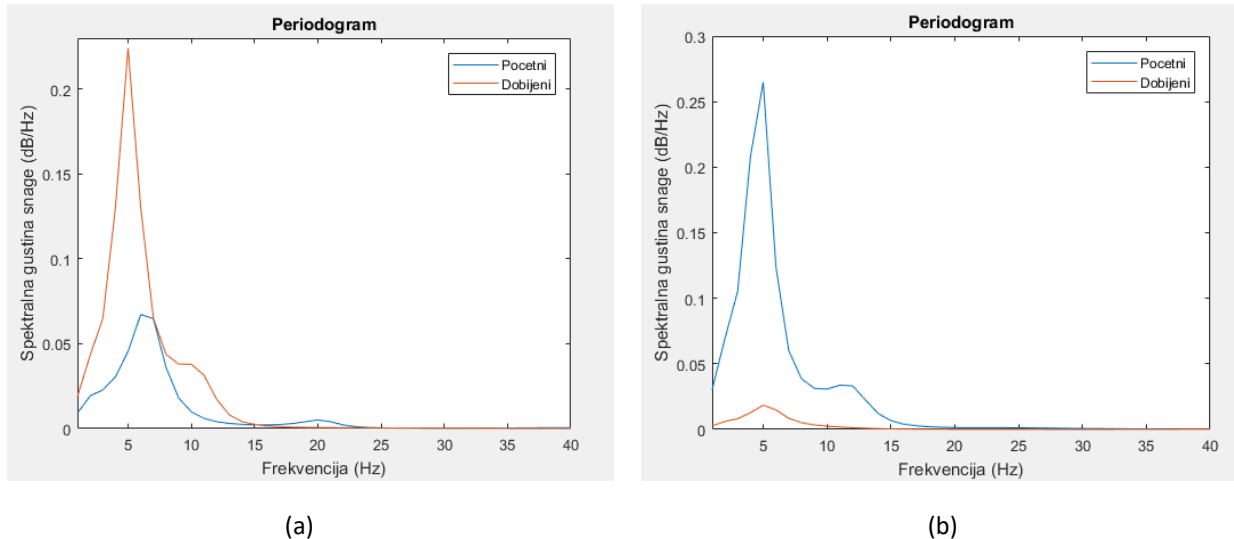
Selektovanjem ovog rezultata i klikom na funkciju *Modify* otvara se prozor sa ponuđenim izborima od kojih se bira opcija *Multiply Frequencies* gde se unose podaci o dužini trajanja signala, kao i skala faktoru koji je prethodno izračunat. Dati prozor za popunjavanje može se videti na Slici 21.



Slika 21 – Unos odgovarajućih parametara naredbe *Multiply Frequencies*

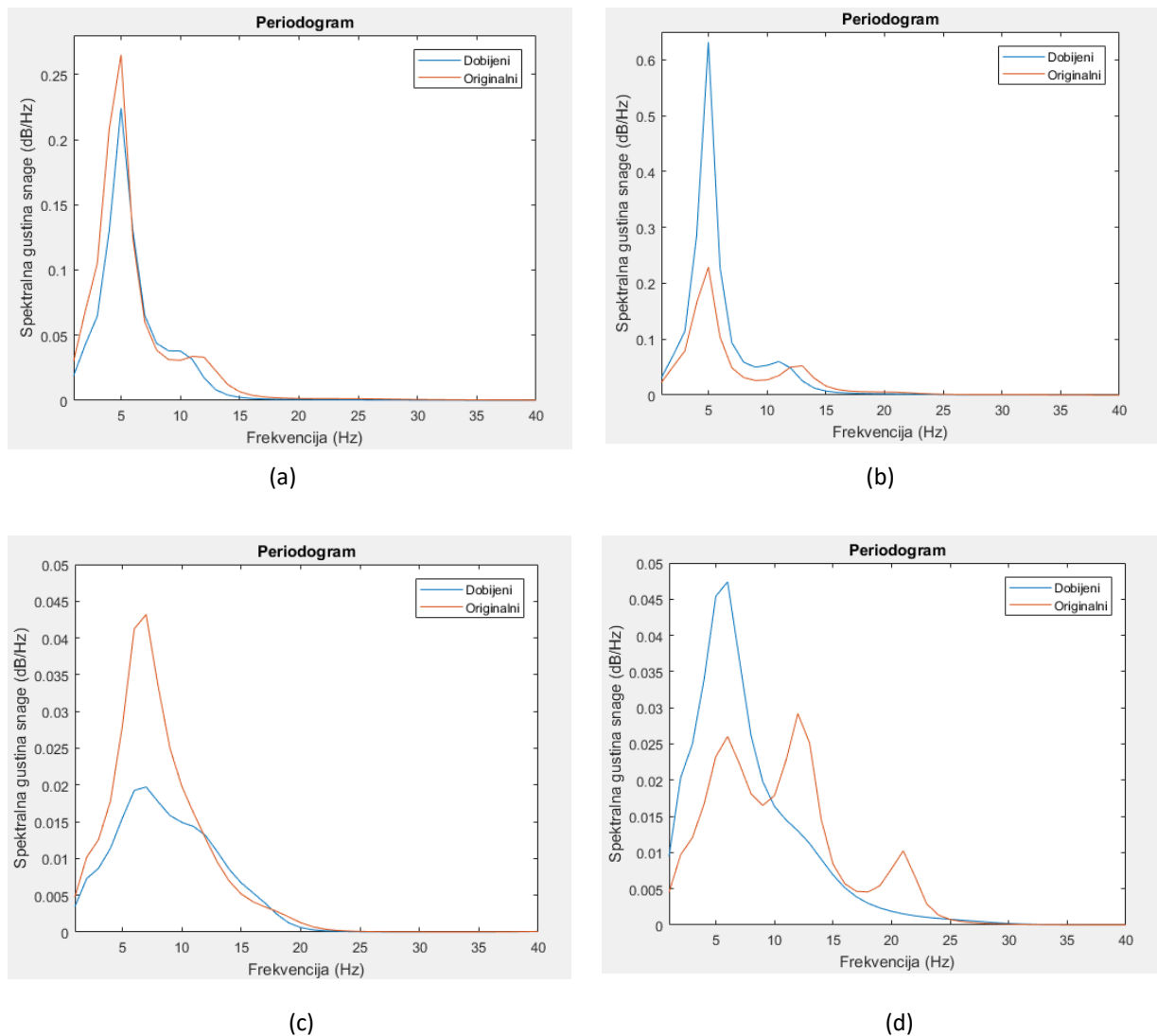
Selektovanjem temporarnog signala za modifikaciju i signala dobijenog prethodnom naredbom otvara se ponuđena opcija za izvršenje naredbe *Replace Pitch Tier* koja vrši zamenu fundamentalne učestanosti početnog signala i željene prediktovane fundamentalne učestanosti. Na samom kraju, selektovanjem tako promenjenog signala i klikom na funkciju *Get resynthesis (overlap add)* vrši se poslednji deo *PSOLA* algoritma koji se odnosi na *Overlap Add* metodu, odnosno diskretnu konvoluciju signala. Ovako obrađen signal predstavlja željeni rezidualni deo signala. Nakon ove obrade, ostaje samo još da se izvrši *LPC* sinteza dobijenog signala. Ova sinteza omogućena je pozivom implementirane funkcije *LPC\_sinteza( )* kojoj se prosleđuju prediktovani koeficijenti, dobijeni rezidualni deo signala, učestanost odabiranja, tip prozora, dužina prozora i koeficijent preklapanja prozora. Sama funkcija predstavlja inverzni oblik *LPC* analize kojom se uz pomoć ugrađene funkcije *filter( )* na svakom prozoru signala vrši sinteza čime se dobija konačni oblik signala. Ovako dobijeni oblik signala moguć je za preslušavanje i predstavlja rešenje metode 1. Ono što je prednost ovog softverskog paketa su preciznost i brzina izvršavanja, pri čemu se prikazivanje pojedinih delova njegove implementacije, kao što je ovde slučaj, mogu izostaviti, već se ceo algoritam izvršava pozivanjem skripte *PSOLA.praat* koja sadrži sve gore pomenute korake. Ono što je razlika između druge i prve metode jeste u drugačijem pristupu problemu, ali suština rešavanja je ista. Osvrtom na blok dijagram na strani 24 stiže se uvid u razlike ove dve metode. Razlika je u redosledu izvršavanja pojedinih akcija, kao i u tome što se u metodi 2 obrada vrši na celokupnom signalu. Nakon treniranja mreže sledi test faza. Na test signalu radi se procena fundamentalne i predikcija željene učestanosti, nakon čega radi se *PSOLA* algoritam na celom signalu, a potom kompenzacija radijacije na usnama. Sledi *LPC* analiza i konvertovanje u *LSF* koeficijente i predikcija novih *LSF* koeficijenata uz pomoć istrenirane neuralne mreže. Nakon toga, moguće je *LPC* sintezom doći do konačnog signala. Što se tiče treće metode, ona predstavlja i najjednostavniju metodu. Ona podrazumeva samo upotrebu *PSOLA* algoritma na celokupnom signalu, pri čemu se dobija novi signal drugačije fundamentalne učestanosti. Razlog za korišćenje ove metode je ideja da se ukaže na razliku u rezultatima kada je u pitanju samo promena fundamentalne učestanosti i kada uz to dolazi i promena osobina koje potiču od vokalnog trakta upotrebom *LPC* analize i sinteze signala. Kako je ovaj problem specifičan, teško je odrediti egzaktno koliko je ovaj algoritam uspešan, osim nekih subjektivnih procena. Ono što je moguće posmatrati kao jedan vid uspešnosti rešenja jeste poklapanje spektralnih gustina snage originalnih željenih signala i dobijenih željenih signala. Takođe, moguće je preslušavanjem utvrditi stepen uspešnosti i poklapanja

sa originalnim signalom, međutim, sve je to deo neke subjektivne procene. Kada je u pitanju spektralna gustina snage signala, korišćena je ugrađena funkcija *pylear()* kojoj se prosleđuju signal i broj LPC koeficijenata. Na Slici 22 mogu se videti neki od primera spektralne gustine snage polaznog i dobijenog signala kada je u pitanju promena sa muškog na ženski glas i promena sa ženskog na muški glas.



Slika 22 – (a) Prikaz spektralnih gustina snage početnog i dobijenog govornog signala kada je u pitanju prelaz sa muškog na ženski glas; (b) Prikaz spektralnih gustina snage početnog i dobijenog govornog signala kada je u pitanju prelaz sa ženskog na muški glas

Sa datih slika može se uočiti da se prilikom transformacije govora sa muškog na ženski glas dobija spektralna gustina snage veće amplitude kao na Slici 22 (a), kao što se i moglo očekivati. U suprotnom slučaju, na Slici 22 (b) početni signal predstavlja ženski glas, dok dobijeni signal predstavlja muški glas, pa je stoga dobijena manja spektralna gustina snage. Ovaj prikaz ima ulogu da pokaže kako se menjaju osobine spektra prilikom korišćenja prve i druge metode, dok se ovakve promene ne vide u metodi 3 koja se bazira samo na promeni fundamentalne učestanosti. Kako bi se uporedili dobijeni rezultati sa originalnim sekvencama, može se pogledati koeficijent preklapanja spektralnih gustina snage oba signala. Na Slici 23 na sledećoj stranici nalaze se poređenja nekih od rezultata datih u oba slučajeva.



Slika 23 – (a) Prikaz preklapanja spektralnih gustina snage za primer željenog ženskog glasa sa velikim procentom poklapanja; (b) Prikaz preklapanja spektralnih gustina snage za primer željenog ženskog glasa sa lošijim procentom poklapanja; (c) Prikaz preklapanja spektralnih gustina snage za primer željenog muškog glasa prosečnim poklapanjem; (d) Prikaz preklapanja spektralnih gustina snage za primer željenog muškog glasa sa lošijim poklapanjem

Na datim slikama može se uočiti da je procenat poklapanja spektralnih gustina snage veći kada je u pitanju transformacija govora sa muškog na ženski glas, mada i rezultati u suprotnom slučaju daju zadovoljavajuće performanse. Ono što se, takođe, može primetiti je da iako postoje dobri rezultati u oba slučaja, postoje i neki test primeri na kojima ovaj algoritam ne radi baš sa velikom preciznošću. Razlog tome je što nisu sve govorne sekvence podjednako pogodne za ovaj algoritam. Poznato je da je LPC analiza i sinteza uspešna na onim delovima signala koji imaju izražene rezonantne i slabo izražene antirezonantne učestanosti, a da je procena lošija ako je signal bogat jako izraženim antirezonantnim učestanostima koji se kriju u nekim fonemama kao što su 'm' i 'n', stoga signali koji se sastoje

u većoj meri od ovakvih svojstava daju lošije performanse. Takođe, usled analize i sinteze signala u bilo kom koraku algoritma može doći do gubitka određene količine informacija, nastanka određene distorzije koja dodatno dovodi do pogoršanja performansi. Ono što je potrebno naglasiti je da se dobijaju bolji rezultati kada je u pitanju algoritam koji u sebi sadrži veći broj LPC koeficijenata, jer je tada i procena tačnija. Takođe, kada je prisutna kompenzacija radijacije na usnama, dobijaju se bolji rezultati po subjektivnoj proceni nakon preslušavanja. Dobijeni signali su realniji i više liče na željeni signal, samim tim što su ovom kompenzacijom propuštene samo niske učestanosti koje odgovaraju opsegu od interesa. Prilikom preslušavanja rezultata bez ove kompenzacije, dobijaju se znatno viši glasovi koji ne odgovaraju u tolikoj meri željenim glasovima. Kako bi procena bila upotpunjena, rezultati sva tri algoritma predstavljeni su slušaocima na kojima je bio zadatak da na osnovu svoje subjektivne procene odrede uspešnost ove implementacije. Slušaocima su na raspolaganju bili rezultati metode 1, metode 2 i metode 3, nakon čijih preslušavanja su pristupili popunjavanju *Google Forms* upitnika u kome je bio zadatak oceniti svaki od ovih rezultata za transformaciju sa muškog na ženski govor i obrnuto u tri kategorije: Prepoznavanje pola, koja se odnosi na subjektivnu procenu da li je u pitanju muški ili ženski glas, Razumljivost teksta, koja se odnosi na raspoznavanje izgovorene rečenice, i Kvalitet signala, koja se odnosi na kvalitet dobijenog signala u vidu količine prisutnog šuma i poklapanja dobijenog govornog signala sa željenim govornim signalom. Upitniku je pristupalo 23 ispitanika koji su za zadatak imali da ocene date performanse ocenom od 1 do 5 po sopstvenom nahođenju, nakon čega je izvedena prosečna ocena za svaku od metoda. Rezultati upitnika nalaze se u Tabeli 1.

	MUŠKI GLAS → ŽENSKI GLAS			ŽENSKI GLAS → MUŠKI GLAS		
	Metoda 1	Metoda 2	Metoda 3	Metoda 1	Metoda 2	Metoda 3
<b>Prepoznavanje pola</b>	4.7/5	4.78/5	3.48/5	4.74/5	4.96/5	2.83/5
<b>Razumljivost teksta</b>	4.43/5	4.61/5	4.87/5	4.57/5	4.61/5	4.61/5
<b>Kvalitet signala</b>	3.7/5	3.74/5	4.48/5	4.3/5	3.91/5	3.83/5

Tabela 1 – Prikaz dobijenih rezultata upitnika kao ocena performansi dobijenih rezultata

Na osnovu tabele koja predstavlja prosečne ocene rezultata 23 ispitanika može se izvesti zaključak da metod 3 ima dobre performanse kada je u pitanju razumljivost teksta i kvalitet signala, sudeći po tome da se na njemu radi samo promena fundamentalne učestanosti bez dodatne LPC analize i sinteze signala, međutim, kada je u pitanju prepoznavanje pola, odnosno to koliko dobijeni signal liči na ženski ili muški glas, ima loše performanse. Razlog ovome je to što se vrši samo promena fundamentalne učestanosti bez promene drugih osobina govornog signala, čime se samo stiče utisak da je isti govornik pričao višim, odnosno nižim glasom, sa istim osobinama koje potiču od vokalnog trakta, tako da se ne stiče dovoljan utisak o dobijanju glasa suprotnog pola. Kada su u pitanju druge dve metode, ova kategorija daje bolje rezultate, jer ovim putem dolazi i do menjanja ostalih osobina govornog signala, koje dovode do potpunijeg utiska o glasu suprotnog pola. Kada je u pitanju razumljivost glasa, kao što je već pomenuto, metoda 3 ima najbolju ocenu, dok druge dve metode imaju nešto nižu ocenu samim tim jer u njihovoj implementaciji dolazi do dodatne obrade govornog signala koja dovodi do pojave nekih nuspojava i nejasnoća u govoru koje običnom slušaocu mogu naneti osećaj lošeg raspoznavanja izgovorene rečenice. Kada je u pitanju kvalitet signala, najbolje performanse ima metod 3 za prelaz sa muškog na ženski glas, zbog već pomenutih performansi, dok to već za obrnuti primer nije slučaj. Za obrnuti primer dobija se glas koji se može opisno okarakterisati kao "ženski robot" koji u mnogome ne liči na željeni glas. Ostale dve metode imaju dobre ocene kada je u pitanju kvalitet signala, ali se i ne može reći da su one sjajne. Dobijeni govorni signal liči na željeni signal, ali ne u tolikoj meri da bi dobio veoma visoku ocenu, a i može se osetiti prisustvo distorzija koje nastaju prilikom same odbrane. Ono što se može zaključiti na osnovu posmatranja cele tabele je da generalno najbolje performanse dobija metoda 2 kada je u pitanju prelaz sa ženskog na muški glas, gde se prepoznavanje pola gotovo u 100% procena pokazalo željenim ishodom, dok je razumljivost teksta veoma visoka i kvalitet prilično zadovoljavajući.

## 4 DISKUSIJA

Kao što se moglo zaključiti iz prethodnih rezultata, ovaj algoritam daje dosta zadovoljavajuće rezultate, pogotovu kada su u pitanju metoda 1 i metoda 2. Ono što je već poznato je to da je na raspolaganju bio set podataka od 115 rečenica izgovorenih na engleskom jeziku od stranem muške i ženske osobe. Jedan deo tih rečenica iskorišćen je da treniranje neuralne mreže, dok je ostatak korišćen za testiranje. Ideja je bila izvršiti transformaciju govora promenom fundamentalne učestanosti i LPC analizom i sintezom signala, uz propratne dodatne obrade koje je zahtevao ovaj algoritam. Očekivani rezultat bio je precizno transformisan polazni govor tako da on u što većoj meri liči na željeni govor suprotnog pola. Na osnovu toga dolazi se do zaključka da se implementacija ovog algoritma može okarakterisati kao uspešna, sudeći po rezultatima prikazanim u prethodnom poglavlju, procentu poklapanja spektralnih gustina snage i rezultata upitnika. Ono što je možda jedan od problema ovog algoritma je to što je potrebno dugo treniranje neuralne mreže kada su u pitanju određeni modeli, za svaki model potrebno je više od sat vremena treniranja. Takođe, ovaj algoritam uspešno vrši transformaciju govora na željeni govor uz dosta veliki set podataka koji je korišćen za treniranje i testiranje mreže, snimljen u jako dobrim studijskim uslovima, sa kvalitetnim mikrofonom, što je često neizvodljivo, osim ako nismo dobro opremljeni. Pitanje je koliko bi dobro ovaj algoritam radio kada ne bi bilo dovoljno podataka za ekstrakciju obeležja i treniranje neuralne mreže. Još jedan od problema koji se može javiti je dostupnost seta podataka različitih osoba koje izgovaraju date rečenice, čime neuralna mreža ne bi uspeła lepo da nauči da prepozna željeni govor, već bi se takoreći zbunila u treniranju. Takođe, iako je implementiran DTW algoritam obrade signala, potrebno je u što većoj meri izgovoriti rečenice istom brzinom i u isto vreme, što predstavlja dodatni problem, jer već u ovakvoj implementaciji postoje neki rezultati koji i nisu baš dobri, a sama pojava nekog od ovakvih problema bi dodatno pogoršala rezultate. Ono što bi se dodatno moglo reći o ovom algoritmu je to što on ne radi u realnom vremenu, već je potrebno da ima dostupan set podataka koje će obrađivati. Danas se dosta traže programi koji brzo i efikasno, u realnom vremenu, vrše preciznu procenu i transformaciju signala, što sa ovim algoritmom nije slučaj. Jedan od takvih primera je svetski poznata aplikacija "Talking Tom" koja je prethodnih godina doživela jako veliku popularnost i proslavila slovenački par inženjera koji su izumeli ovu aplikaciju.

Kada je u pitanju poređenje algoritma sa literaturom, [1] *Mark Tse* sa kolumbijskog univerziteta u svom pomenutom radu bavi se rešavanjem ovog problema na sličan način. Takođe se vrši treniranje neuralne mreže datim parovima LSF koeficijenata dobijenih nakon konverzije LPC koeficijenata pri čemu se mapiranje fundamentalne učestanosti vrši na



eksitacionom delu signala *PSOLA* algoritmom, što predstavlja metodu 1 u ovom radu, pored koje postoje još dve metode. Nakon mapiranja fundamentalne učestanosti LPC sintezom na isti način dolazi se do krajnjeg signala. Ono što je razlika u odnosu na ovaj rad je to što se klasifikovanje na zvučni i bezvučni deo govornog signala vrši na osnovu energije eksitacionog signala i koeficijenta refleksije, poređenjem sa odgovarajućim pragom, što u potpunosti nije slučaj u ovom radu. Takođe, procena fundamentalne učestanosti bazira se na kepstralnoj dekonvoluciji eksitacionog signala i na proceni autokorelacionom metodom, dok se ovaj rad oslanja samo na autokorelacionu procenu. Uvidom u tabelu procene rezultata ovog algoritma na osnovu ocena 20 ispitanika dolazi se do zaključka da su rezultati dobri, ako ne u neki segmentima i bolji od rezultata u ovom radu. Ono što je sličnost sa ovim radom, takođe, je to što je poklapanje rezultata za prelazak sa ženskog na muški glas manje nego preklapanje u obrnutom slučaju. [2] *Permanallur Ranganathan* u svom radu bazira se na transformaciji glasa na osnovu LPC analize govornog signala, pri čemu se obrada vrši samo na jednoj izgovorenoj fonemi, a ne na celoj rečenici i to samo kada su u pitanju vokali. LPC analiza signala radi se na celokupnom signalu, koji nije u velikoj meri stacionaran, pa su dobijeni rezultati neprecizniji. Ovim pristupom nije moguće vršiti obradu na dužim govornim signalima koji sadrže više od jedne foneme, a kamoli celih rečenica. Za ovakvu obradu potrebno je treniranje neuralne mreže koje u ovom slučaju uopšte nema, pa je nemoguće uraditi predikciju željenih LPC koeficijenata, već se koeficijenti početnog signala multipliciraju nekim brojem kako bi došlo do njegove promene, pri čemu se dobija drugačiji signal, ali mogućnost dobijanja željenog signala je isključena, osim ako se ne vodimo samo pukim nagađanjem i eksperimentisanjem, što oduzima dosta vremena i nije efikasno. [3] *Tomoki Toda* u svom radu izložio je jednu od novijih metoda spektralne konverzije govora. Za spektralnu konverziju između govornika koristi se GMM model (*Gaussian Mixture Model*) združene gustine verovatnoće polaznog i željenog signala. Ova konvencionalna metoda pretvara spektralne parametre frejm po frejm na osnovu minimalne srednje kvadratne greške. Iako je metod efektivan, dolazi do pogoršanja kvaliteta govora usled problema kao što je to da odgovarajući spektralni pokreti nisu uvek uzrokovani procesom konverzije na frejmovima ili da su pretvoreni spektri preterano zaglađeni statističkim modeliranjem. Kako bi se ti problem rešili, predložen je metod baziran na konverziji zasnovanoj na proceni maksimalne verodostojnosti trajektorije parametara gde je efekat prekomernog smirivanja ublažen uzimajući u obzir karakteristiku globalne varijanse konvertovanih spektara . Eksperimentalni rezultati pokazuju da se performanse konvertovanog signala mogu drastično poboljšati predloženom metodom sa strane kvaliteta govora i tačnost konverzije čime se mogu dobiti i bolji rezultati nego što je slučaj sa algoritmom u ovom radu.

## 5 ZAKLJUČAK

Na osnovu svega izloženog do sada, jasno se može potvrditi da je dati algoritam odgovorio na sve potrebne segmente koji su od ključnog značaja za uspešno rešavanje ovog problema. Kako je ideja transformacije govora sa glasa jednog pola na glas suprotnog pola promenom fundamentalne učestanosti i primenom *LPC* analize i sinteze odlukom ispitanika dobro ocenjena, ta pretpostavka se i potvrđuje. Činjenica je da su metoda 1 i metoda 2 davale bolje rezultate nego metoda 3, što je i bila poenta same primene metode 3, da se ukaže na njena ograničenja i istaknu dobre osobine prve dve metode. Ono što je ograničenje ovog algoritma je to što zahteva unapred prikupljen set podataka kako bi uspešno izvršio treniranje neuralne mreže. Jedno od mogućih unapređenja ovog algoritma je njegova modifikacija i primena u realnom vremenu, čime bi se ovaj algoritam mogao koristiti u razne svrhe, samim tim jer je potražnja za ovakvim implementacijama u realnom vremenu danas jako velika. Takođe, algoritam bi se mogao dalje istražiti po pitanju transformacije govornog signala na neke druge željene govorne signale ili na više različitih govornih signala nekom od dodatnih metoda klasifikacije.

Svedoci smo sve većeg razvitka tehnologije poslednjih godina, a pogotovu kada su u pitanju razni programi i uređaji koji se baziraju na nekoj promeni signala, njegovoj obradi, analizi i sintezi, stoga bi se primena ovakvog algoritma mogla ogledati u raznim aplikacijama za konverziju govora, prepoznavanje, identifikaciju, aplikacijama za decu i odrasle, zaštitu, generisanje glasova likova u crtanom filmu i u mnogim drugim oblastima.

## 6 LITERATURA

- [1] Mark Tse, "Voice Transformation", EE6820 Speech and Audio Processing Project Report, 2003. godina
- [2] Permanallur Ranganathan Gurumoorthy, "LPC based Voice Morphing", 2008. godina
- [3] Tomoki Toda , Alan W. Black , Keiichi Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory"
- [4] L.R.Rabiner, R.W.Schafer, "Digital Processing of Speech Signal", 1978. godina
- [5] Yurika Permanasari, Erwin H. Harahap, Erwin Prayoga Ali, "Speech Recognition using Dynamic Time Warping", 2019. godina
- [6] Sami Lemmetty, "Review of Speech Synthesis Technology", 1999. godina
- [7] Pascal van Lieshout, "Praat short tutorial", 2003. godina
- [8] prof. dr Željko Đurović, Autorizovane beleške sa predavanja na predmetu Obrada i prepoznavanje govora 13E054OPG, Univerzitet u Beogradu - Elektrotehnički fakultet, 2020. godina

## PRILOG A

Propratni delovi diplomskog rada koji su korišćeni u izradi dati su u prilogu sledećim redosledom:

1. programski kod u MATLAB-u R2017a,
2. slike korišćene u izradi diplomskog rada,
3. prezentacija diplomskog rada,
4. istrenirani modeli neuralne mreže,
5. dostupni set podataka,
6. *Google Forms* upitnik,
7. CD sa prethodno navedenim priložima
8. i drugo.