UNIVERZITET U BEOGRADU
ELEKTROTEHNIČKI FAKULTET

Kepstar i osnovna frekvencija govornog signala

SEMINARSKI RAD

PREDMET: OBRADA GOVORNOG SIGNALA

Profesor: Autor:

dr Slobodan Jovičić **Nikola Jovanović 3125/2009**

Beograd, april 2010. godine

SADRŽAJ

Spisak slika

1. Uvod

Ljudska vrsta je tokom svog evolutivnog razvoja usavršila jedan poseban oblik sporazumevanja. Taj oblik sporazumevanja je komuniciranje putem govora. Za potrebe prenosa govora na daljinu i razvijanja sistema za prepoznavanje govora i govornika radi lakšeg i neposrednijeg komuniciranja sa računarom potrebno je detektovati razne parametre koji karakterišu govor svake osobe. Jedan takav parametar je osnovna (fundamentalna) frekvencija signala govora.

Rad je, zbog lakšeg razumevanja, organizovan tako da posle generalnog upoznavanja sa pojmovima spektra, kepstra i osnovne frekvencije govora, sledi kratak opis načina estimacije osnovne frekvencije govora i obrade dobijenih rezultata. U trećem poglavlju je definisan cilj ovog rada, a to je određivanje osnovne frekvencije govora pomoću kepstra.

Takođe u trećem poglavlju su opisani i metod i postupak analize odabranog govornog signala, dok su u četvrtom anaizirani dobijeni rezultati. Na kraju rada, u zaključnom poglavlju, sumirani su dobijeni rezultati.

S obzirom da je za pisanje ovog rada korišćena literatura na engleskom jeziku, kao i činjenica da se u ovoj oblasti koriste izrazi i skraćenice na engleskom jeziku, izostavljena je oznaka „engl.“ ispred engleskih termina.

2. Kepstar i osnovna frekvencija govora

2.1. Signal govora

Signal je namerno izazvan fizički proces koji u sebi nosi poruku. Govor je aperiodični signal koji spada u nestacionarne slučajne signale čiji se vremenski oblik i frekvencijski sadržaj kontinualno menjaju s vremenom. Upravo signali koji su delimično ili potpuno slučajni služe za prenos informacija, jer predvidiv signal ne nosi nikakvu informaciju.

U fizičkom, tačnije akustičkom, domenu govor se javlja u vidu zvuka. Posmatranje zvučnog pritiska kao signala znači njegovo pretvaranje u električni oblik. Da bi mogao da bude obrađen na računaru, govorni signal $x(t)$ mora da bude digitalizovan, odnosno diskretizovan i kvantizovan, nakon čega se dobija niz numeričkih podataka $x\ n$ konačne dužine.

Govorni signal je akustički (zvučni) signal koji se sastoji od niza glasova koji su povezani na lingvističkom nivou tako da imaju određeni smisao i značenje, odnosno sadrže informaciju. U širem smislu, glas je bilo kakav akustički signal (zvuk) koji se generiše govornim mehanizmom. Prema tome, glas obuhvata pevanje, normalan govor, šapat, onomatopeju, neartikulisane zvuke, itd. Fonem je osnovna jedinica u govornoj komunikaciji. Fonem je onaj glas čijom zamenom u reči ta reč menja lingvističko značenje. U našem fonetskom sistemu imamo 30 fonema [JOV99].

Dva su osnovna izvora zvuka kod generisanja glasova: oscilacije glasnica i frikcija vazdušne struje, a moguće su i kombinacije ovih izvora zvuka. Ako vazdušna struja nesmetano prolazi kroz vokalni trakt generisani akustički signal je kvaziperiodičan pa ima približno harmonijski spektar. Takve glasove nazivamo zvučni fonemi. Suženja vokalnog trakta stvaraju frikciju pa tako izgovoreni fonemi imaju šumni karakter.

Vokalni trakt se proteže od larinksa do usana i sastoji se iz tri osnovna dela: farinksa (ždrela), usne šupljine i nosne šupljine [JOV99]. Artikulacija je oblikovanje usne šupljine u svrhu proizvodnje glasova. U artikulatore spadaju: usne, jezik, zubi, donja vilica, tvrdo nepce, meko nepce i alveolarni rub iza gornjih zuba.

Najčešće proučavani parametri govora su: osnovna frekvencija govora i formantne frekvencije vokala. Ovi parametri su u direktnoj vezi sa fizičkim osobinama govornih organa pa tako osnovna frekvencija govora zavisi od dimenzija, mase i zategnutosti glasnica, a formantne frekvencije vokala od oblika vokalnog trakta.

2.2. Spektar govora

Predstava signala u frekvencijskom domenu naziva se spektar. Spektar se dobija kada se na signal primeni Furijeova transformacija (*Fourier Transformation, FT*) i predstavlja kompleksnu veličinu.

Spektrogram je uobičajen način predstavljanja spektralnog sadržaja nestacionarnog signala kakav je govor. Za dobijanje spektrograma, tačnije za dobijanje vremenski zavisne procene (estimacije) spektra nestacionarnih signala koristi se vremenski zavisna ili kratkovremena Furijeova transformacija (*Short-Time Fourier Transform, STFT*).

Ideja kod STFT je da se signal govora u vremenskom domenu izdeli na kratke intervale u okviru kojih se može smatrati da je stacionaran. Zatim se Furijeovom transformacijom odredi procena spektralne gustine snage govornog signala iz svakog intervala koja se naziva *kratkovremeni spektar*. Kratkovremeni spektri se izračunavaju za segmente govornog signala dužine 10÷20 ms.

Vremenski zavisna Furijeova transformacija (STFT) diskretnog signala $x\ n$ definiše se izrazom:

$$X\ n,\lambda\ = \sum_{m=-\infty}^{\infty} x\ m+n\ \ w\ m\ e^{-jm\lambda}\ , \tag{2.1}$$

gde je λ frekvencijska promenljiva, a $w\ m$ je prozorska funkcija ili prozor.

Sekvenca elemenata iz niza $x\ n$ koja je izdvojena prozorom $w\ m$ je označena sa $x_{SEG}\ n$. Za svako $x_{SEG}\ n$ vrši se procena spektralne gustine snage pomoću diskretne Furijeove transformacije (*DFT*).

Diskretna Furijeova transformacija je niz konačne dužine *N* koji se dobija uniformnim odabiranjem jedne periode Furijeove transformacije sekvence $x_{SEG}\ n$. Dobijeni odbirci se nazivaju DFT koeficijenti.

Izraz za izračunavanje koeficijenata diskretne Furijeove transformacije jednog izdvojenog segmenta $x_{SEG}\ n$ iz niza $x\ n$ glasi:

$$X_{SEG}\ \ k\ = \sum_{n=0}^{N-1} x_{SEG}\ \ n\ e^{-j\frac{2\pi k}{N}n}\ ,\ 0 \le k \le N-1\ , \tag{2.2}$$

gde je $x_{SEG}\ \ n\ = x\ n+m\ \cdot w\ m$ pri čemu važi da je $w\ m\ \neq 0$ za $m \in\ 0, N$ i $w\ m\ = 0$ za $m \notin\ 0, N$.

Kao posledica periodičnosti FT i DFT će biti periodična sa periodom *N*. Za realne signale DFT koeficijenti zadovoljavaju uslov konjugovane simetrije pa je dovoljno posmatrati spektar realnog signala samo u opsegu frekvencija $0 \le f < f_S/2$, odnosno dovoljno je izračunati samo $N/2$ DFT koeficijenata. Za efikasno izračunavanje DFT koristi se grupa algoritama koja se zove brza Furijeova transformacija (*Fast Fourier Transform, FFT*).

Razmak između članova DFT niza naziva se frekvencijska rezolucija:

$$\Delta f = \frac{f_S}{N}\ , \tag{2.3}$$

gde je f_S frekvencija odabiranja, a N dužina DFT niza. Prema tome, DFT koeficijenti postoje na frekvencijama $f_K = k \cdot \frac{f_S}{N} = k \cdot \Delta f$, gde je $0 \le k \le N-1$.

Vremenska rezolucija Δt je recipročna vrednost Δf pa važi $\Delta t = 1/\Delta f$. Izbor dužine prozorske funkcije $w(m)$ je kompromis između rezolucije u vremenskom i rezolucije u frekvencijskom domenu jer smanjenje dužine prozora znači povećanje vremenske rezolucije i obratno.

Oblik prozora treba izabrati tako da se smanji curenje spektra. Curenje spektra se javlja u slučaju da frekvencija spektralne komponente nije celobrojni umnožak rezolucije. Tada se spektralna komponenta nalazi u intervalu između dve frekvencije iz skupa f_K pa joj se amplituda ne može tačno odrediti. Kako bi se dobila bolja procena spektra prozori mogu da se preklapaju. Poželjna je mala dužina prozora da bi se smanjio uticaj promena parametara govora unutar intervala koji se analizira. Što je prozor duži, veće su promene parametara, pa je veće i odstupanje od početno pretpostavljenog stacionarnog modela.

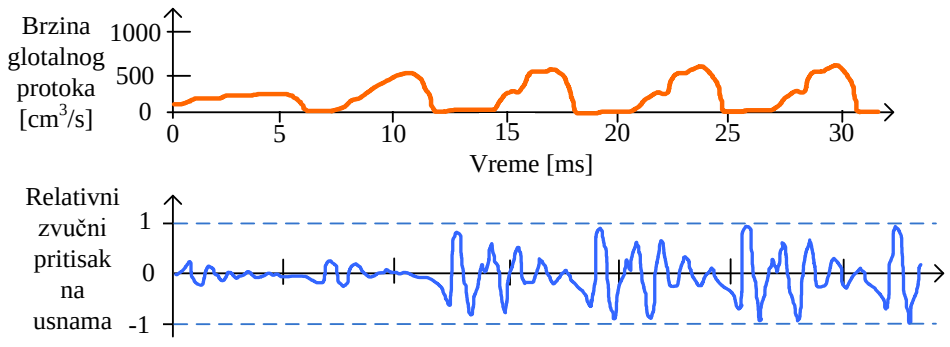
Jedna od prozorskih funkcija koja se najčešće koristi je Hamingov (*Hamming*) prozor. Koeficijenti Hamingovog prozora se izračunavaju na osnovu sledeće formule:

$$w(m) = \begin{cases} 0{,}54 - 0{,}46 \cdot \cos\left(\frac{2\pi m}{N-1}\right), & 0 \le m \le N-1, \\ 0, & inače. \end{cases} \tag{2.4}$$

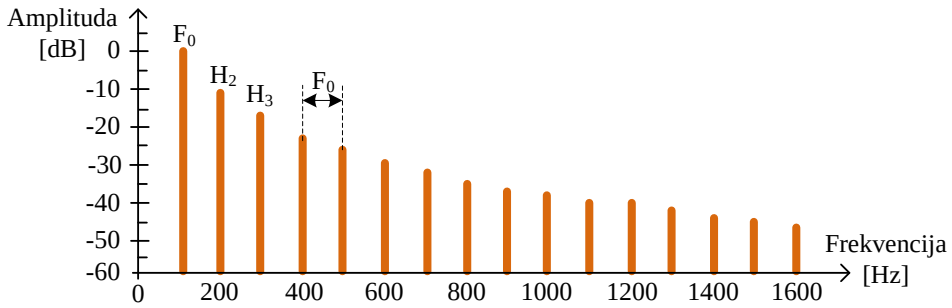
2.3. *Osnovna frekvencija govora*

Frekvencija oscilovanja glasnica, je tzv. osnovna (fundamentalna) frekvencija glasa i obeležava se sa F_0. Zvuk nastao vibracijom glasnica se oblikuje u vokalnom traktu. Vibracije glasnica periodično prekidaju vazdušnu struju i od nje stvaraju tzv. glotalni protok.

Glotalni protok ima impulsni karakter sa približno trougaonom formom glotalnog impulsa (slika 2.1). Odavde sledi da je spektar povorke glotalnih impulsa bogat harmonicima (multiplima osnovne frekvencije), kao što se vidi sa slike 2.2.



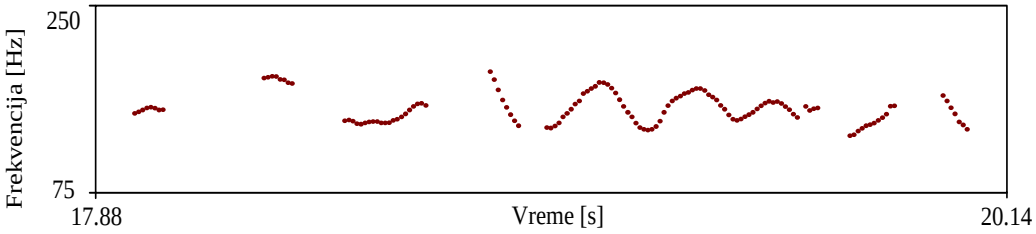
Slika 2.1. Protok vazduha kroz glotis i zvučni pritisak na usnama kod vokala /a/ [RAB78].



Slika 2.2. Skica harmonijskog spektra povorke glotalnih impulsa.

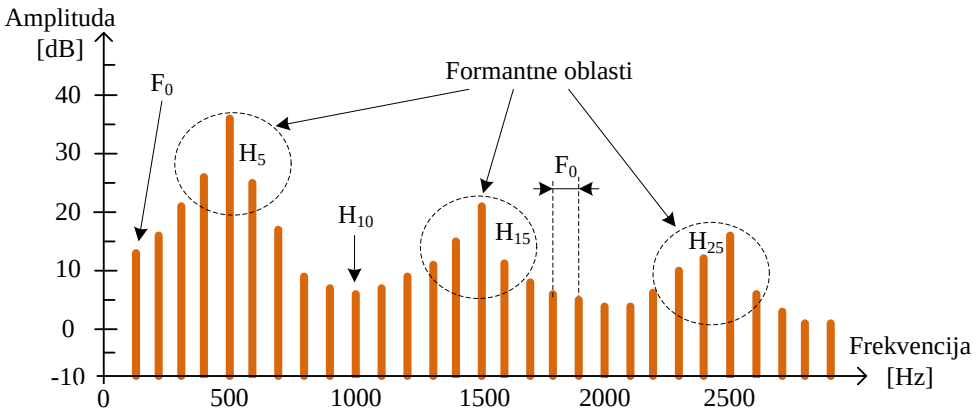
Zategnutost glasnica određuje veličinu osnovne frekvencije tako što veća zategnutost rezultuje višim frekvencijama vibriranja. Osim toga, primećeno je da veće i masivnije glasnice vibriraju na nižim frekvencijama. Osnovna frekvencija (F_0) je, prema [JOV99], kod muškaraca obično u opsegu 80÷180 Hz, kod žena 180÷230 Hz, dok je kod dece 230÷300 Hz.

F_0 određuje visinu glasa (tzv. *pitch*) tako što se visoke frekvencije vibriranja glasnica percipiraju kao veća visina (tzv. visok glas), a niske frekvencije vibriranja kao manja visina (tzv. dubok glas). Osim od glasnica, F_0 zavisi i od prozodijskih obeležja govora, kao što su akcenti ili emocije. Prema tome, osnovna frekvencija glasa je nestacionarna i menja se tokom govora (slika 2.3).



Slika 2.3. Primer vremenske zavisnosti F_0, tzv. F_0 kontura.

Na izlazu govornog mehanizma (na otvoru usta) vazdušna struja iz pluća ima istaknute delove spektra na rezonantnim frekvencijama vokalnog trakta i spuštenim delovima spektra između njih. Oblasti pojačanja snage govornog signala u spektru glasa uzrokovane akustičkom rezonancom nazivaju se formanti (slika 2.4) i definišu se svojom centralnom (formantnom) frekvencijom. Treba naglasiti da harmonik, tj. umnožak osnovne frekvencije, ne mora da se podudara sa centralnom frekvencijom formanta.



Slika 2.4. Formantne oblasti u spektru vokala.

2.4. Kepstar

1963. godine, naučnici *B. P. Bogert, M. J. R. Healy* i *J. W. Tukey* publikovali su rad o detekciji eha u signalu u kome su uveli nove pojmove za opisivanje postupka koji su primenili. Iskoristili su anagrame ustaljenih termina, pa su tako dobili *kepstar* (*cepstrum*) umesto spektar. Slično, termin *kvefrencija* (*quefrency*) je uveden za naziv nezavisne promenljive kepstra koja ima dimenziju vremena, dok termin *liftering* označava filtriranje u kvefrencijskom domenu. Oni su definisali kepstar signala kao spektar snage logaritma spektra snage signala [BEN08]. Kasnije je prihvaćena definicija koja kaže da je kepstar inverzna FT logaritma modula tj. magnitude FT signala.

Kompleksni kepstar $\hat{x}\ n$ se dobija inverznom z transformacijom izraza $\log\ X\ z$, gde je $X\ z$ z transformacija signala $x\ n$ dobijena iz:

$$X\ z = \mathbb{Z}\ x\ n\ = \sum_{n=-\infty}^{\infty} x\ n\ z^{-n}\ , \tag{2.5}$$

gde je z kompleksna promenljiva. Z transformacija je generalizacija Furijeove transformacije diskretnih signala [MIL99]. Smenom $z = e^{j\omega}$ u (2.5) izvodi se izraz za Furijeovu transformaciju.

Kompleksni kepstar nije kompleksan, naime ako je $x\ n$ realno, i $\hat{x}\ n$ je realno. Odrednica „kompleksni“ se zapravo odnosi na kompleksni logaritam koji se koristi za izračunavanje kompleksnog kepstra [BEN08]. Kompleksni logaritam je definisan kao:

$$\hat{X}\ e^{j\omega}\ = \log\ X\ e^{j\omega}\ = \log\left|X\ e^{j\omega}\right| + j \cdot \arg\ X\ e^{j\omega}\ . \tag{2.6}$$

Realni kepstar ili samo kepstar $c\ n$ predstavlja parni deo kompleksnog kepstra $\hat{x}\ n$:

$$c\ n\ = Ev\ \hat{x}\ n\ = \frac{\hat{x}\ n\ + \hat{x}\ -n}{2}\ , \tag{2.7}$$

pri čemu važi:

$$c\ n\ = \mathbb{Z}^{-1}\ \log\left|X\ z\right|\ = \mathbb{F}^{-1}\ \log\left|X\ e^{j\omega}\right|\ = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X\ e^{j\omega}\right|\ e^{j\omega n} d\omega\ . \tag{2.8}$$

Kompleksni kepstar čuva informacije i o magnitudi i o fazi prvobitnog spektra omogućavajući rekonstrukciju signala, dok kepstar sadrži informacije samo o magnitudi (modulu) $\left|X\ e^{j\omega}\right|$ prvobitnog spektra signala.

2.5. Kepstar govora

Signal govora $s\ n$ je rezultat konvoluiranja pobude (ekscitacije) $e\ n$ i impulsnog odziva $v\ n$ vokalnog trakta:

$$s\ n\ = e\ n\ * v\ n\ = \sum_{k=-\infty}^{\infty} e\ k\ v\ n-k\ \ , \tag{2.9}$$

i bilo bi korisno razdvojiti („dekonvoluirati“) ove dve komponente govornog signala. Iako je to u opštem slučaju neizvodljivo, dekonvolucija se može primeniti na govorni signal jer pobuda i odziv imaju veoma različite spektre.

Konvolucija dva niza u vremenskom domenu postaje proizvod njihovih z transformacija u transformacionom domenu [MIL99]. Prvi korak u kepstralnoj dekonvoluciji pretvara proizvod dva spektra u sumu dva signala. Tražena transformacija je logaritamska (izrazi 2.10 i 2.11). Ako su dobijeni signali-sabirci dovoljno spektralno različiti mogu se razdvojiti linearnim filtriranjem.

Računski koraci za određivanje kompleksnog kepstra $\hat{s}\ n$ iz govornog signala $s\ n = e\ n * v\ n$ su:

$$S\ z = \mathbb{Z}\ s\ n = \mathbb{Z}\ e\ n * v\ n = \mathbb{Z}\ e\ n \cdot \mathbb{Z}\ v\ n \quad , \tag{2.10}$$

$$S\ z = E\ z \cdot V\ z \quad , \tag{2.11}$$

$$\log\ S\ z = \log\ E\ z \cdot V\ z = \log\ E\ z + \log\ V\ z \quad , \tag{2.12}$$

$$\hat{s}\ n = \mathbb{Z}^{-1}\ \log\ S\ z = \mathbb{Z}^{-1}\ \log\ E\ z + \mathbb{Z}^{-1}\ \log\ V\ z \quad , \tag{2.13}$$

$$\hat{s}\ n = \hat{e}\ n + \hat{v}\ n \quad , \tag{2.14}$$

gde je $E\ z$ z-transformacija pobude, a $V\ z$ je funkcija prenosa vokalnog trakta.

Pošto se formantna struktura $V\ e^{j\omega}$ menja sporo po frekvencijama u odnosu na harmonike i šum u $E\ e^{j\omega}$, doprinosi (kontribucije) $E\ e^{j\omega}$ i $V\ e^{j\omega}$ se mogu linearno razdvojiti posle inverzne FT.

Pristup izračunavanja kepstra se sastoji u tome da FT zameni z transformaciju. Pošto je u pitanju govor, potrebno je segmentirati signal $s\ n$ i izračunati STFT pomoću DFT. Tipično se kepstar računa jednom svakih 10÷20 ms pošto se parametri pobude u normalnom govoru ne menjaju naglo [PET02]. S obzirom na to da se analiziraju segmenti ulaznog signala izdvojeni prozorom kao u 2.2 pri čemu važi da je $x_{SEG}\ n = x\ n+m \cdot w\ m$, na kraju se dobiju sledeći izrazi (indeks SEG je izostavljen):

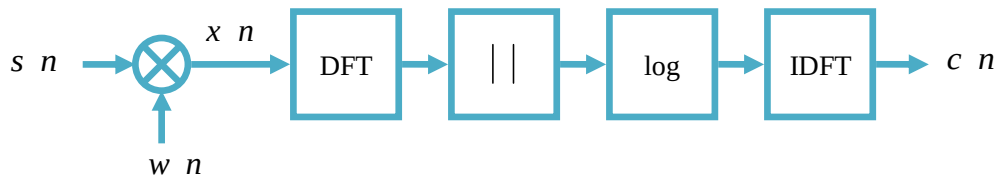
$$X\ k = \sum_{n=0}^{N-1} x\ n\ e^{-j\frac{2\pi k}{N}n} \quad , \quad 0 \le k \le N-1 \ , \tag{2.15}$$

$$\hat{X}\ k = \log\ X\ k \quad , \quad 0 \le k \le N-1 \ , \tag{2.16}$$

$$\hat{x}\ n = \frac{1}{N}\sum_{k=0}^{N-1} \hat{X}\ k\ e^{j\frac{2\pi k}{N}n} \quad , \quad 0 \le n \le N-1 \ , \tag{2.17}$$

$$c\ n = \frac{1}{N}\sum_{k=0}^{N-1} \log\left|X\ k\right| e^{j\frac{2\pi k}{N}n} \quad , \quad 0 \le n \le N-1 \ . \tag{2.18}$$

Čitav postupak izračunavanja kepstra je prikazan na slici 2.5.



Slika 2.5. Postupak izračunavanja kepstra.

2.6. Primena kepstra

Kepstralna analiza se primenjuje kod kepstralnog vokodera, detekciju F_0 i formanata. Za potrebe detekcije fundamentalne frekvencije, kojom se bavi ovaj rad, kompleksni kepstar je suvišan i dovoljno je poznavati realni kepstar $c\ n$.

Glavna smetnja široj primeni kepstralnog vokodera je dvostruko izračunavanje DFT-a i u predajniku i u prijemniku, uz estimaciju F_0. Tu su, prema [OSH00], i: pojačanje niskih nivoa tj. isticanje šumnih delova spektra zbog operacije logaritmovanja, i potreba za upotrebom adaptivnog prozora za izdvajanje pobudne komponente kepstra kako se ne bi izgubilo previše informacija iz govora sa visokom F_0.

U kepstru zvučnog govornog segmenta na poziciji koja odgovara osnovnom periodu titranja glasnica postoji pik, dok u kepstru bezvučnog govornog segmenta takav pik ne postoji. Upravo ovo svojstvo kepstra može se koristiti kao osnova za određivanje zvučnosti govornog segmenta kao i za određivanje osnovnog perioda ($1/F_0$) zvučnih segmenata govornog signala.

2.7. Estimacija F_0

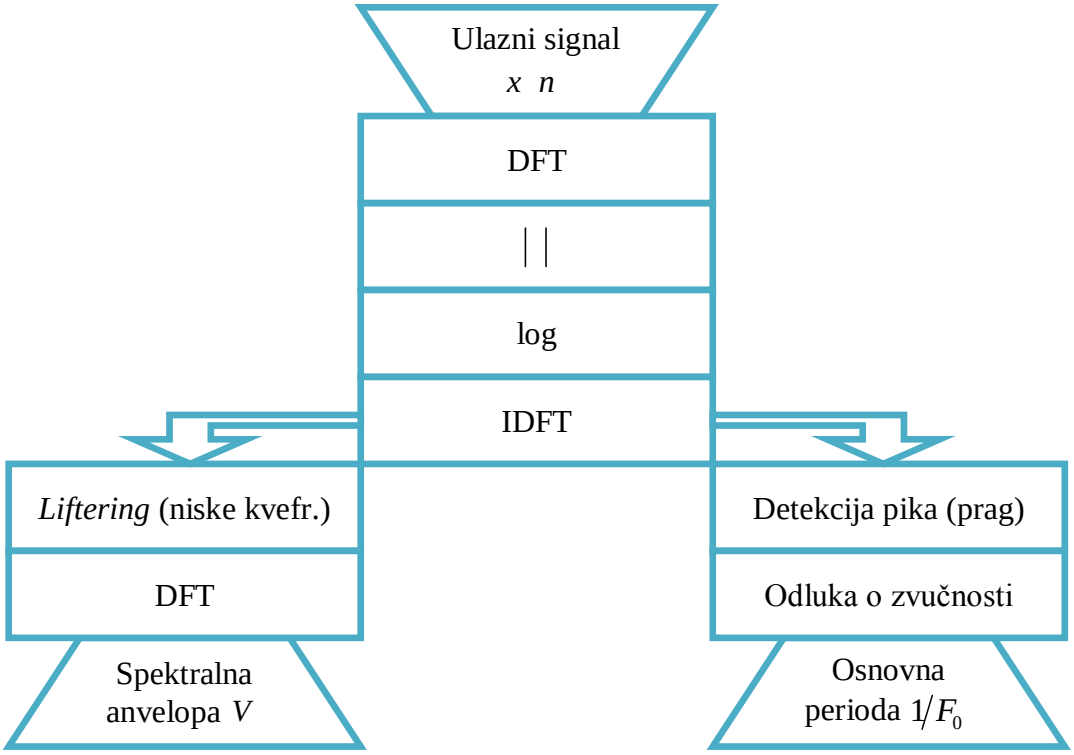
Najtačnije i najnepraktičnije metode koriste ultrazvuk, fotografisanje glasnica i merenje impedanse larinksa kontaktnim mikrofonima. Ipak, najveći broj detektora F_0 koristi algoritme koji na ulaz dobijaju samo signal govora. Oni obično vrše odluku o zvučnosti i, prema [OSH00], razlikuju četiri klase glasova: zvučne, bezvučne, kombinovane (npr. glas /z/) i *nonspeech* (tišina, pauza, predah, pozadinski šum). Procene zvučnosti mogu imati tačnost do 95% ako je SNR (odnos signal-šum) iznad 10 dB, ali podbacuju za SNR ispod 0 dB [OSH00].

U opštem slučaju, F_0 se određuje ili na osnovu periodičnosti u vremenskom domenu ili na osnovu pravilno razmaknutih harmonika u frekvencijskom domenu. Metodi za estimaciju F_0 zasnovani na frekvencijskom domenu su tačniji od onih u vremenskom domenu, ali imaju oko 10 puta složeniji račun.

Određivanje F_0 je važno kod mnogih aplikacija. Mnogi koderi govora sa niskim protokom (*low-rate voice coders*) zahtevaju tačnu estimaciju F_0 radi dobre rekonstrukcije govora, a neki koderi govora sa srednjim protokom (*medium-rate*) koriste F_0 da redukuju bitski protok uz očuvanje visoko kvalitetnog govora. F_0 konture su korisne za prepoznavanje i sintezu govora, a mogu pomoći gluvim osobama dok uče da govore.

Postupak određivanja osnovne frekvencije zasnovan na kepstru relativno je jednostavan i prikazan je na slici 2.6. Nakon izračunavanja kepstra u okolini očekivanog osnovnog perioda u kepstru se traži pik (maksimum, šiljak, špic). Ako je veličina pika iznad nekog prethodno postavljenog praga, ulazni govorni segment je verovatno zvučni, pri čemu pozicija pika predstavlja dobru estimaciju osnovnog perioda.

Dužina i relativna pozicija prozora značajno utiču na veličinu pika u kepstru. Obično se uzima dužina prozora tako da se, ako je moguće, u okviru segmenta nađu barem dve periode. Tako se za obradu govora muškog govornika dubljeg glasa tipično uzima prozor trajanja oko 40 ms, dok se za više glasove (viša F_0) mogu koristiti proporcionalno uži prozori [PET02].



Slika 2.6. Algoritam za estimaciju F_0 i $V\ e^{j\omega}$.

Potpuna tačnost određivanja F_0 se nikad ne može ostvariti zbog više uzroka: nestacionarne prirode govora, nepravilnog vibriranja glasnica, širokog opsega mogućih vrednosti F_0, uticaja vokalnog trakta i degradacija govora usled zašumljenosti.

2.8. Peglanje (smoothing)

Govor se često transformiše u skup parametarskih signala koji su tesno povezani sa pokretima artikulatora vokalnog trakta [OSH00]. Primer jednog takvog signala je F_0, parametar koji prati promenu jednog artikulatora – vibriranje glasnica. Postoje i parametri koji zavise od više artikulatora, npr. pozicije i amplitude formanata su određeni celokupnom konfiguracijom vokalnog trakta.

Vokalni trakt se sporo menja u poređenju sa frekvencijom odabiranja, jer tipični fonetski događaji traju duže od 50 ms, dok se govor može odabirati svakih 0,1 ms. Brze spektralne promene su ograničene na promene načina artikulacije na granicama fonema i na početak i kraj reči tj. na prelaze između govora i tišine. Prema tome, parametri govora obično sporo variraju i dopuštaju decimaciju [OSH00].

Kako se frejmovi analize (segmenti signala) tipično osvežavaju na svakih 10 ms, samo mali broj frejmova sadrži nagle izmene. Prema tome, neefikasno je prenositi parametre brzinom (protokom) kao što je 600 odbiraka/s. Stoga se vrši decimacija parametarskih signala što znači da se na dalju obradu prosleđuju samo vrednosti dobijene pododabiranjem (*downsampling*, *subsampling*). Praktični algoritmi za kodiranje i prepoznavanje koriste parametre sa 25÷200 odb/s u zavisnosti od primene [OSH00]. Ovo žrtvuje tačnost tokom brzih spektralnih promena, uz smanjenje memorijskih resursa i perceptualno jedva primetnu degradaciju rekonstruisanog signala.

Pod pretpostavkom da se traži kontura sporopromenljive komponente parametarskog signala za čuvanje, prenos ili dalju analizu, peglanje (*smoothing*) je neophodno kako pododabiranje (i decimacija u celini) ne bi dalo lažne rezultate zbog superponiranih brzih fluktuacija. Ako se parametar prosto decimira fiksnim faktorom, odnosno pododabira fiksnom frekvencijom odabiranja, onda je linearno filtriranje filtrom propusnikom niskih učestanosti najbolja opcija. Međutim, drugi načini predstavljanja parametarskih signala u redukovanom formatu su često uspešniji sa nelinearnim peglanjem.

Linearno filtriranje je posebno neodgovarajuće za F_0 konture kod kojih se smatra da F_0 tradicionalno uzima nulte vrednosti za bezvučne delove govora. Prelazi zvučno-bezvučno su nagli i linearno peglanje daje loše rezultate.

Osim u procesu decimacije, peglanje se koristi i za korekciju detektovanih vrednosti F_0. Estimatori formanata i F_0 su ozloglašeni zbog izolovanih pogrešnih procena ili „autsajdera“ (*outliers*) koji odstupaju od ostatka parametarske konture [OSH00]. Takve greške treba ispraviti postprocesiranjem, ali neke greške mogu opstati do izlaza. Linearni filtri jednako ponderišu sve odbirke signala i propagiraju uticaj greške na susedne delove ispeglane izlazne parametarske konture.

Jedna alternativa linearnom filtriranju je medijan (*median*) filtriranje koje čuva oštre diskontinuitete i eliminiše fine nepravilnosti i „autsajdere“. Peglanje se najčešće obavlja nad konačnim segmentom parametarskog signala, ali linearno peglanje kombinuje sve odbirke u segmentu da bi dalo ispeglani izlazni odbirak, dok medijan peglanje bira jedan odbirak obuhvaćen segmentom.

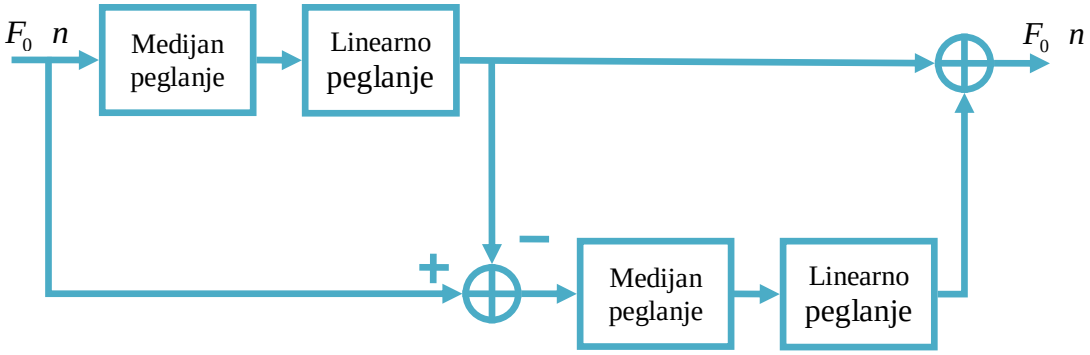
Kod medijan filtriranja (peglanja) u okviru svakog segmenta sa M odbiraka se odbirci ređaju po amplitudi bez obzira na vremenski redosled, a zatim se određuje izlazni odbirak odnosno medijan kao $M+1\,/2$-i od M poređanih odbiraka, gde je M neparan broj. Iznenadni diskontinuiteti (skokovi) konture su očuvani jer nema usrednjavanja. $M-1\,/2$ odbiraka pre i posle glavne konture ne utiče na izlaz pa je izlazna kontura kraća za $M-1$ vrednosti, osim ako se ulazna kontura ne dopuni (tzv. *padding*). Kontura se dopunjava na primer tako što se ponove vrednosti na njenim krajevima $M-1\,/2$ puta.

Medijan filtriranje dobro eliminiše autsajdere i dobro vrši peglanje, ali u slučaju zašumljenih signala ne obezbeđuje naročito gladak izlaz. Zato se često kombinuje sa linearnim filtrima da bi se dobio kompromisni izlaz – ispeglan i sa oštrim prelazima (tranzicijama). Kombinacija linearnog i medijan filtriranja se naziva nelinearno peglanje.

Pošto medijan filtriranje već obezbeđuje dobro peglanje, linearni filtar je niskog reda. Obično je linearni filtar simetrični *FIR* (*Finite Impulse Response*) filtar kako bi kašnjenja bila tačno kompenzovana [RAB78]. Prema [RAB78], opšte je prihvaćen (pre 30-ak godina) Haningov (*Hanning*) filtar sa impulsnim odzivom:

$$h\ n\ =\begin{cases}1/4 & , \quad n=0\\ 1/2 & , \quad n=1\\ 1/4 & , \quad n=2\end{cases} \tag{2.19}$$

koji je iskorišćen i u ovom radu. Postupak nelinearnog peglanja preuzet iz [RAB78] prikazan je na slici 2.7.



Slika 2.7. Postupak nelinearnog peglanja.

3. Sadržaj istraživanja

3.1. Cilj istraživanja

Cilj ovog rada je određivanje tačnije procena vrednosti osnovne frekvencije govora primenom kepstralne analize. Takođe je pokazano izdvajanje spektralne anvelope iz kepstra.

3.2. Metod istraživanja

Promena osnovnog perioda ($1/F_0$) titranja glasnica kroz vreme može se proceniti računanjem vremenski kratkotrajnog kepstra koji se zasniva na STFT. Korišćen je programski paket *MATLAB*, a komparacija rezultata je napravljena sa rezultatom koji daje softver *Praat* za isti snimak.

3.3. Postupak istraživanja

Snimak sa muškim glasom reči „halo“ napravljen je sa frekvencijom odmeravanja od 11025 Hz u *wav* formatu. Dužina snimka *HALO.wav* je 0,535 s odnosno 0,535 × 11025 = 5898 odbiraka. Na svaki segment od 512 odbiraka ulaznog govornog signala primenjen je Hamingov prozor i FFT reda 512 tačaka. Prozori analize pomeraju za korak od 128 odmeraka i ukupno ih je 43.

Na vremenskoj (kvefrencijskoj) skali kepstra određen je položaj maksimuma (pika) kepstralne funkcije (ukoliko postoji). Ovaj maksimum definiše periodu T_0 osnovne frekvencije F_0 u analiziranom zvučnom govornom segmentu. Usvojen je opseg 80 Hz $\leq F_0 \leq$ 300 Hz što iznosi 3,33 ms $\leq T_0 \leq$ 12,5 ms. U ovom opsegu kvefrencija se traži kepstralni maksimum na bazi sopstvenog kriterijuma – praga iznad koga se pojavljuje kepstralni maksimum.

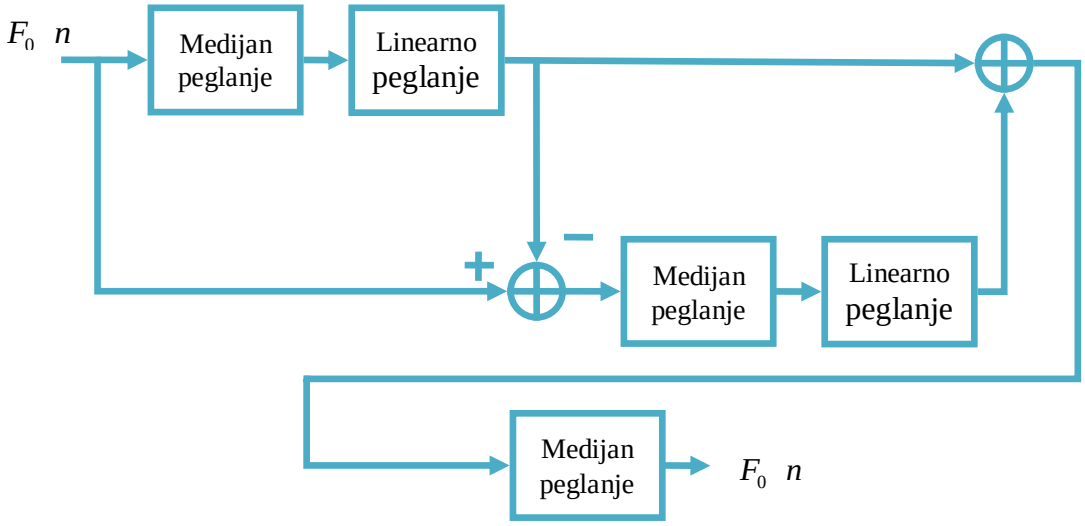
Dobijena F_0 kontura upoređena sa *Praat*-ovim rezultatom. U *Praat*-u je podešena dužina i pomak prozora tako da se dobiju 43 tačke u F_0 konturi, definisane su granice F_0 (80÷300 Hz) i koristi se podrazumevani *autocorrelation method*.

F_0 kontura dobijena u *MATLAB*-u postavljanjem praga detekcije je ispeglana na tri načina:

- Propuštanjem kroz medijan filtar dužine $M = 5$,

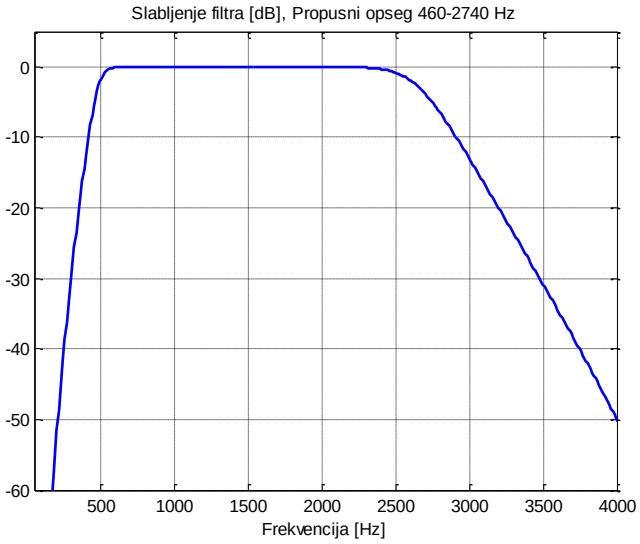
- Propuštanjem kroz nelinearni filtar, i
- Propuštanjem i kroz nelinearni i kroz medijan filtar dužine $M = 5$ (slika 3.1).

Kod nelinearnog filtriranja se takođe koristi medijan filtar dužine $M = 5$, a za linearno peglanje se koristi funkcija iz izraza (2.19). Ulazni signal medijan filtra je uvek proširivan tako da se očuva ista dužina signala i na izlazu. Da bi se skratilo vreme izračunavanja uvek su filtrirane nenulte vrednosti F_0 kontura.



Slika 3.1. Nelinearno i medijan peglanje.

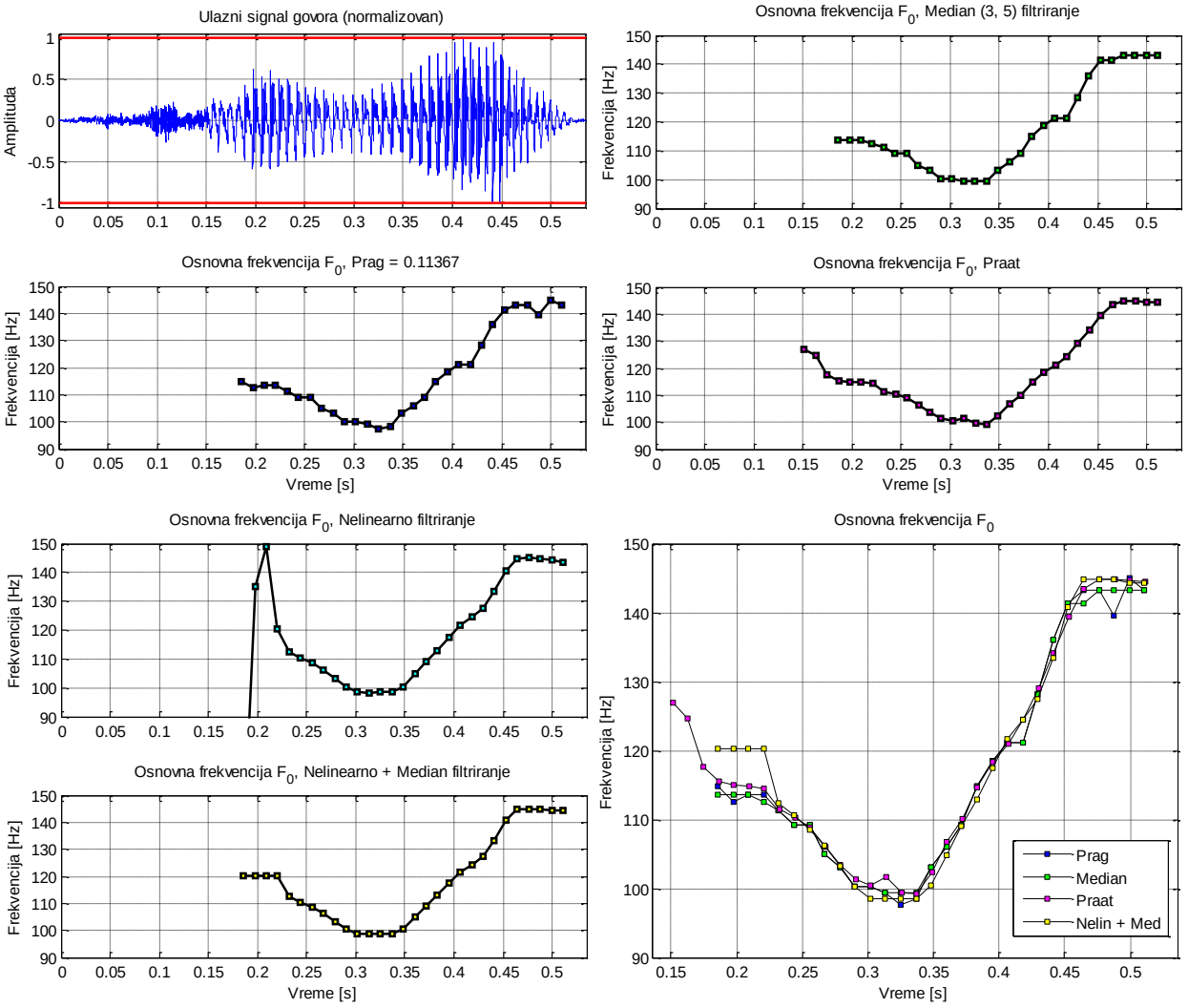
Originalni test snimak *HALO.wav* je modifikovan na dva načina. U jednom slučaju tako što mu je dodat beli šum tako da SNR bude 15 dB, a u drugom slučaju tako što je govor propušten kroz *IIR* (*Infinite Impulse Response*) filtar koji propušta opseg učestanosti između 460 i 2740 Hz bez slabljenja kao što je prikazano na slici 3.2. Prenosna karakteristika ovog filtra je odabrana tako da približno simulira tipični telefonski filtar iz [RAB78].



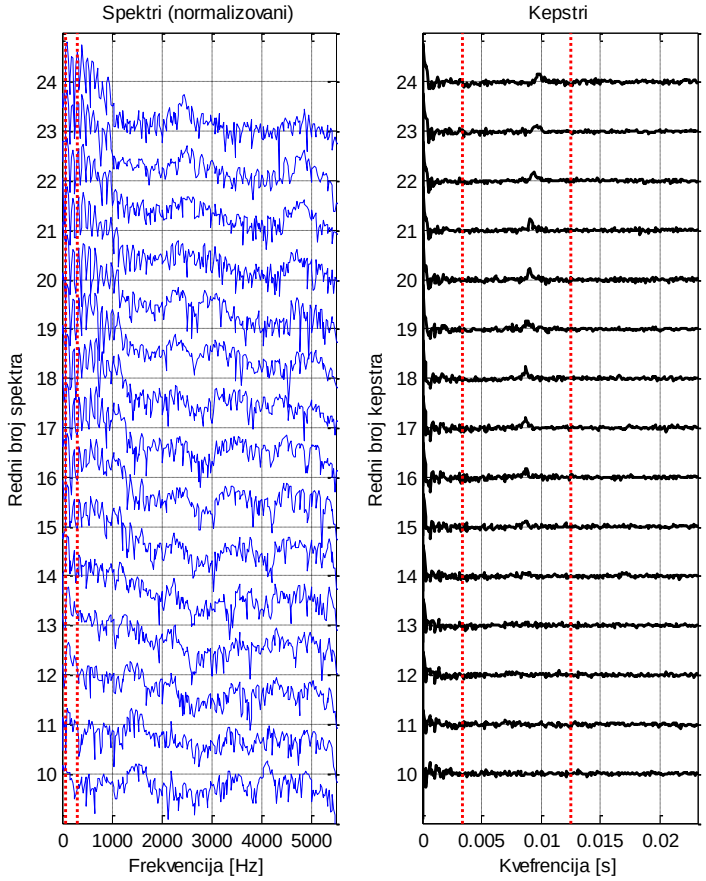
Slika 3.2. Funkcija prenosa digitalnog filtra propusnika opsega.

4. Analiza rezultata

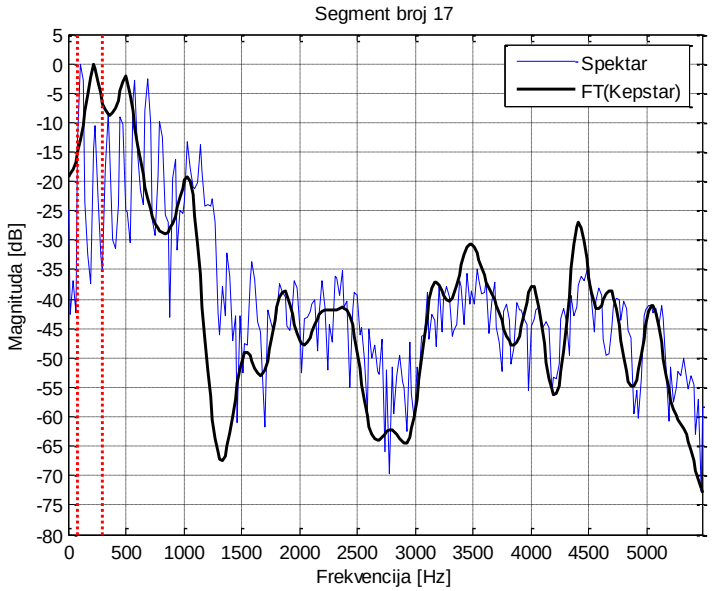
Na slici 4.1 prikazani su rezultati, odnosno dobijene F_0 konture za originalni test snimak *HALO.wav*. Prag detekcije je postavljen na 43% vrednosti najvećeg kepstralnog pika iz opsega 3,33÷12,5 ms, tj. vrednost praga je 0,11367. Odmah treba istaći da je korišćena druga vrednost praga za svaki od tri analizirana slučaja ulaznog govornog signala.



Slika 4.1. Ulazni signal *HALO.wav* i F_0 konture.

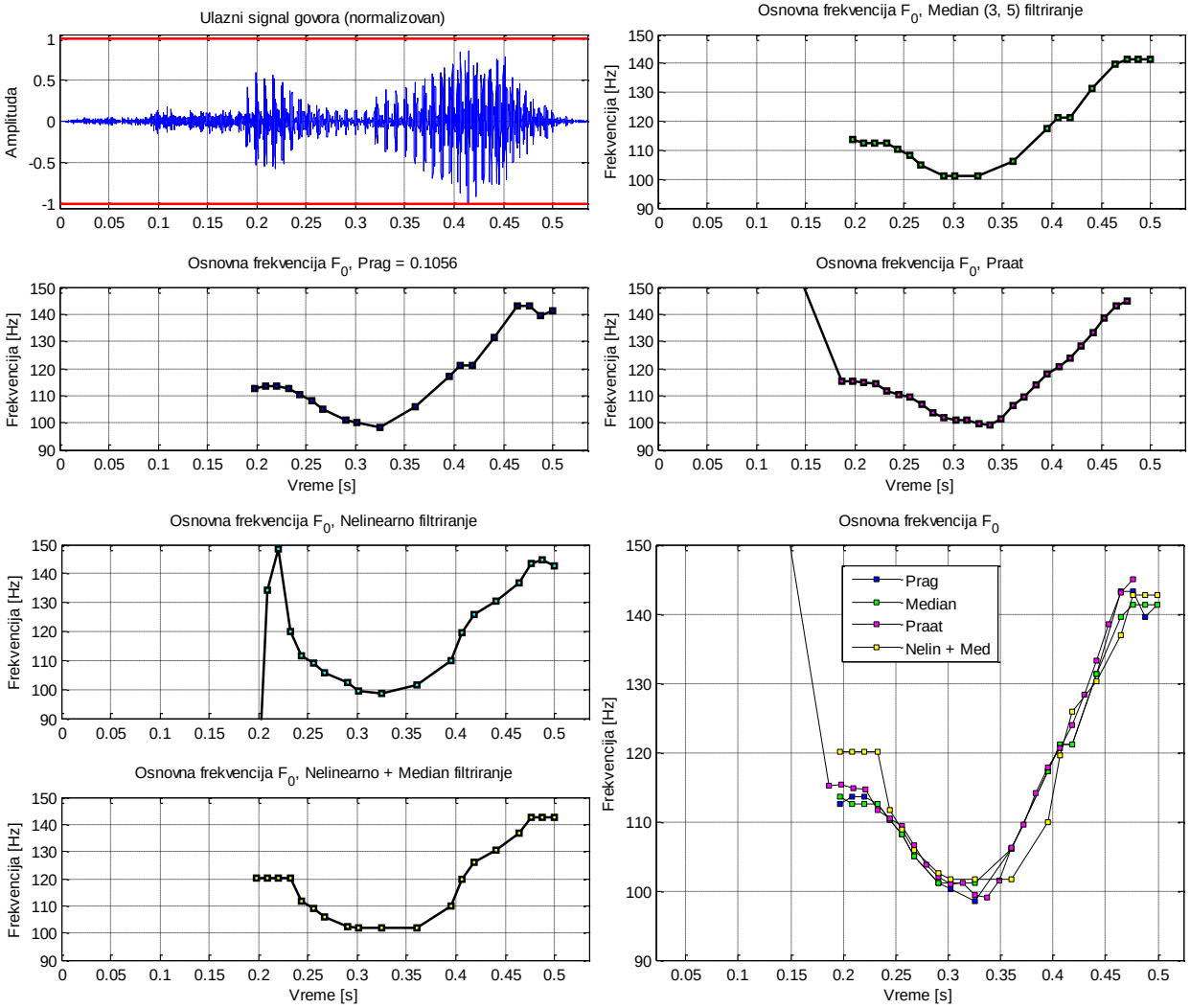


Slika 4.2. 15 kratkovremenih spektara i odgovarajućih kepstara signala *HALO.wav*.

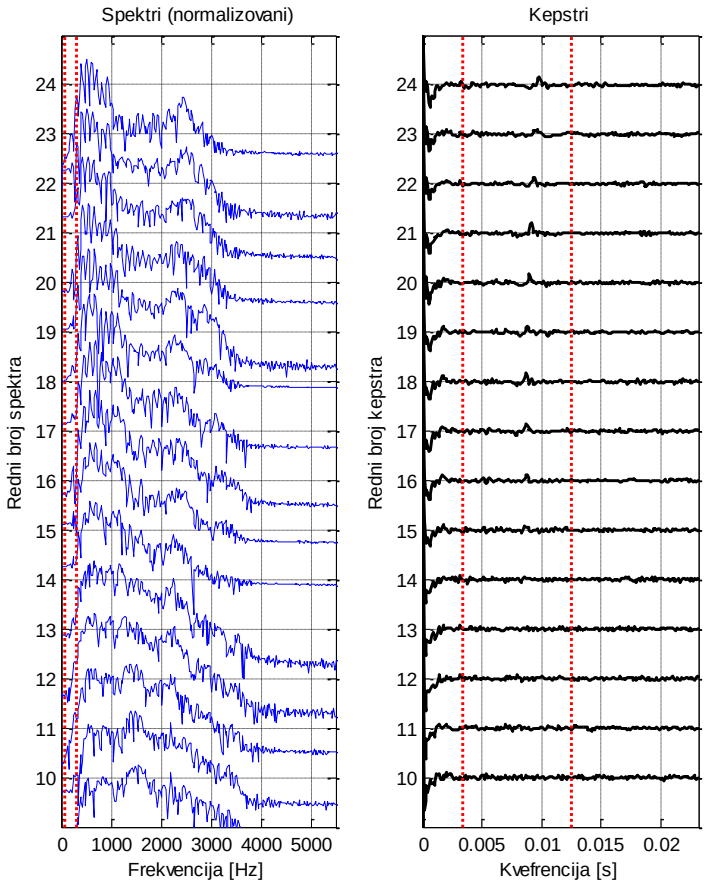


Slika 4.3. Spektar 17-og segmenta signala *HALO.wav* i odgovarajući skalirani spektar niskokvefrencijskog kepstra.

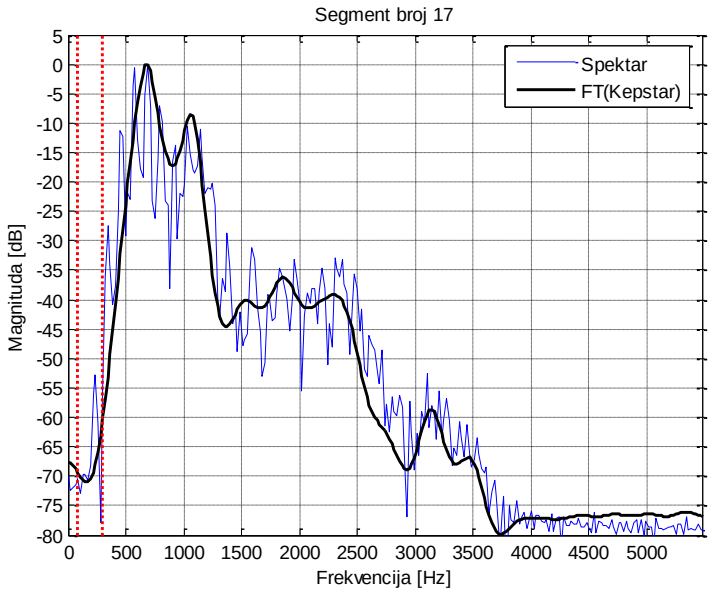
Na slici 4.4 prikazani su rezultati, odnosno dobijene F_0 konture za modifikovani test snimak *HALO_Propusni_opseg_460-2740_Hz.wav*. Prag detekcije je postavljen na 52% vrednosti najvećeg kepstralnog pika iz opsega 3,33÷12,5 ms, tj. vrednost praga je 0,1056.



Slika 4.4. Ulazni signal *HALO_Propusni_opseg_460-2740_Hz.wav* i F_0 konture.

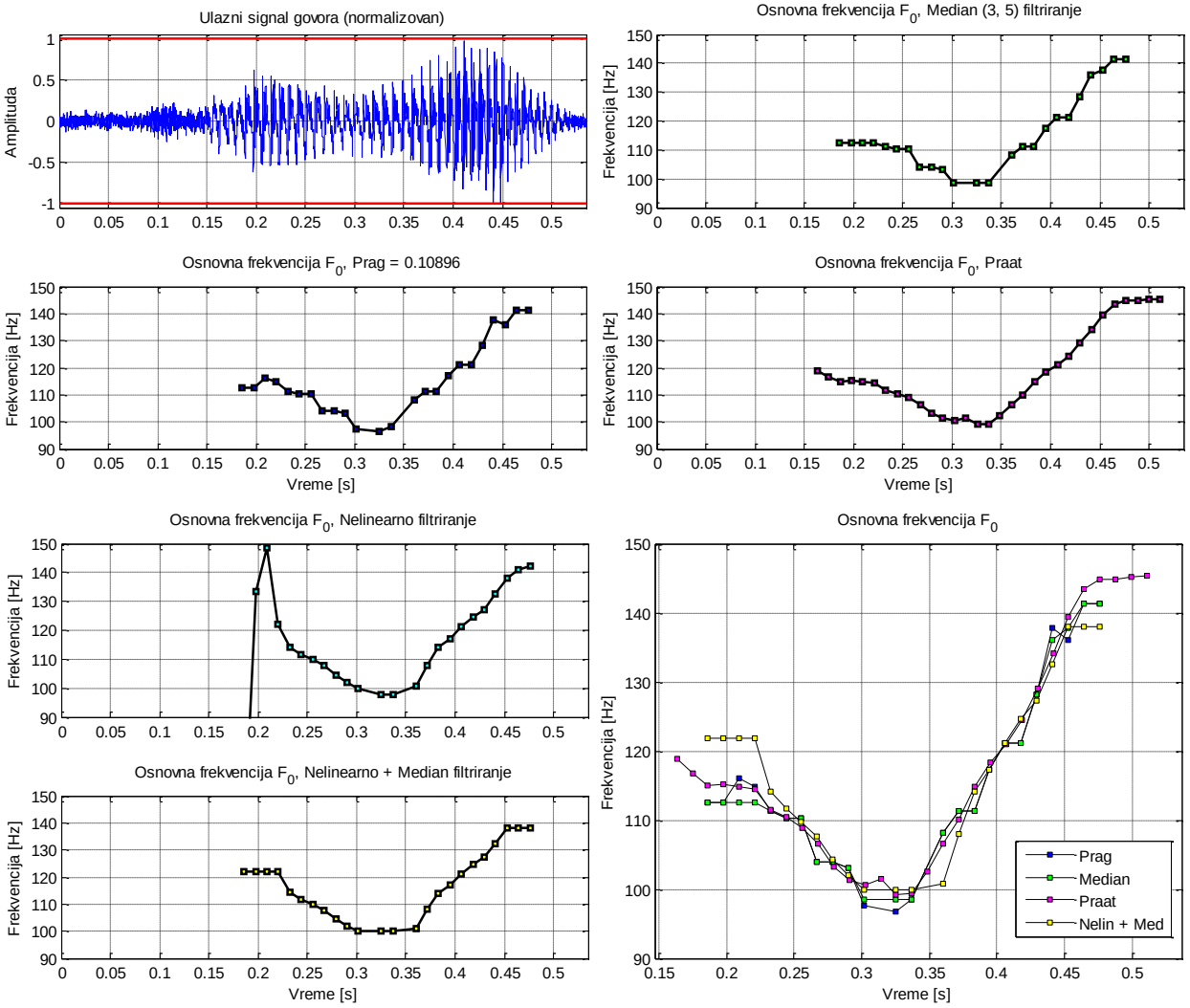


Slika 4.5. 15 kratkovremenih spektara i odgovarajućih kepstara signala *HALO_Propusni_opseg_460-2740_Hz.wav*.

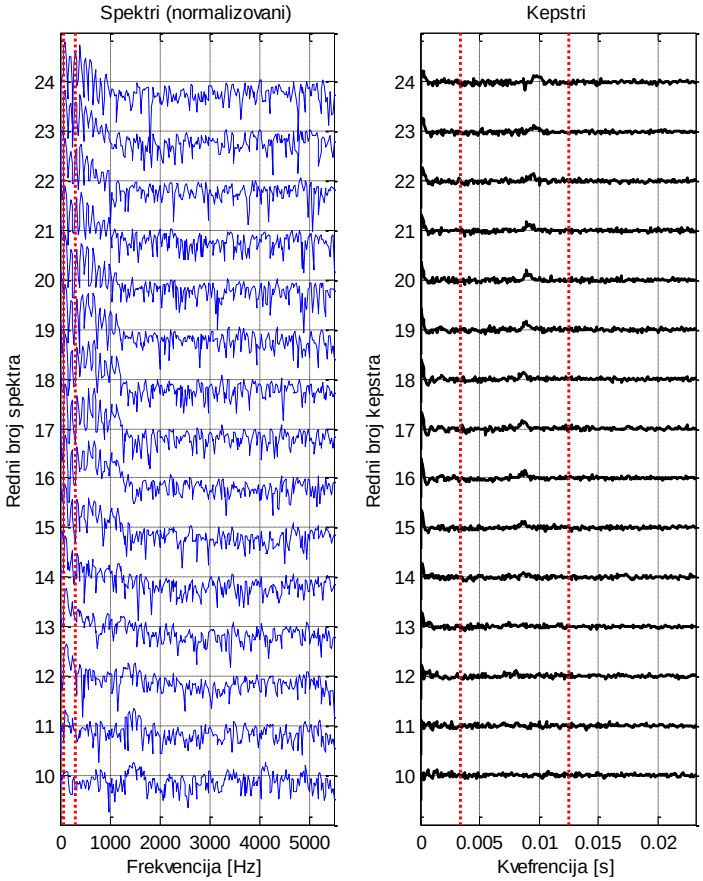


Slika 4.6. Spektar 17-og segmenta signala *HALO_Propusni_opseg_460-2740_Hz.wav* i odgovarajući skalirani spektar niskokvefrencijskog kepstra.

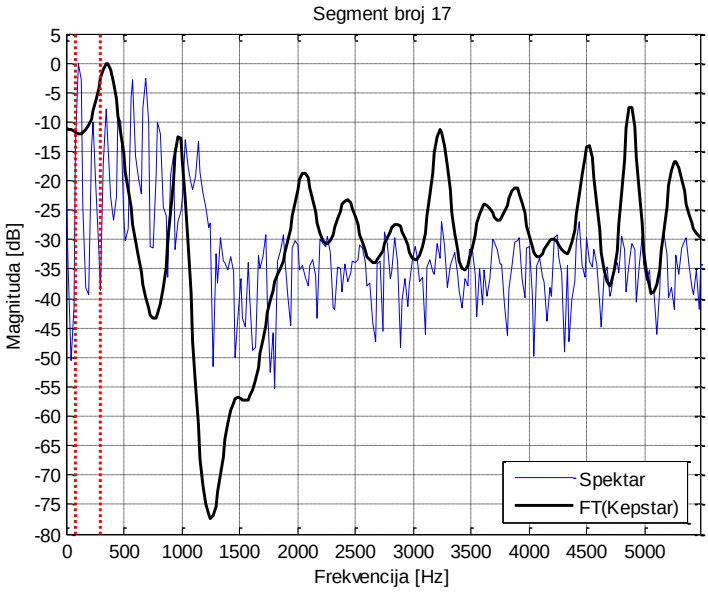
Na slici 4.7 prikazani su rezultati, odnosno dobijene F_0 konture za modifikovani test snimak *HALO_SNR_=_15_dB.wav*. Prag detekcije je postavljen na 48% vrednosti najvećeg kepstralnog pika iz opsega 3,33÷12,5 ms, tj. vrednost praga je 0,10896.



Slika 4.7. Ulazni signal *HALO_SNR_=_15_dB.wav* i F_0 konture.



Slika 4.8. 15 kratkovremenih spektara i odgovarajućih kepstara signala *HALO_SNR_=_15_dB.wav*.



Slika 4.9. Spektar 17-og segmenta signala *HALO_SNR_=_15_dB.wav* i odgovarajući skalirani spektar niskokvefrencijskog kepstra.

Upoređujući F_0 konture dobijene programom *Praat* i kodom napisanim u *MATLAB*-u uočava se da u oba slučaja postoje problemi sa detekcijom F_0 na prelazu sa glasa /h/ na /a/, odnosno na prelazu sa bezvučnog (šumnog) fonema na zvučni. Na tom istom prelazu se pri nelinearnom filtriranju F_0 kontura dobijenih kodom napisanim u *MATLAB*-u javljaju veliki „autsajderi“ koji moraju da se ispeglaju dodatnim medijan filtrom koji ipak ne odrađuje baš dobar posao kao što se može videti na slikama 4.1, 4.4 i 4.7.

Takođe, sa navedenih slika se uočava da su vrednosti koje vraća *Praat* bolje izdržale dodavanje šuma i filtriranje, mada se kod filtriranja na 0,02367 ms javila jedna lažna vrednost F_0 od čak 266,628 Hz koja daleko odskače od ostatka konture.

Kada se uporede F_0 konture posle medijan peglanja i kombinacije nelinearnog i medijan peglanja sa konturom iz *Praat*-a na slici 4.1 primećuje se da se medijan peglanje bolje nosi sa prelazom /h/-/a/ ali da medijan peglanje kome prethodi nelinearno peglanje ipak ima glađi izlaz. Međutim, kada se ispituje filtrirani signal govora medijan peglanje daje ponajbolje rezultate. Slično se može reći i za zašumljeni signal, mada kontura posle medijan peglanja nije dovoljno glatka.

Na slikama 4.2, 4.5 i 4.6 su prikazani spektri i kepstri ovog problematičnog /h/-/a/ prelaza i jasno je zašto dolazi do grešaka u detekciji: pikovi u posmatranom opsegu kvefrencija, obeleženom tačkastim crvenim linijama, tada nisu uski već su ponegde prošireni i zaobljeni što je otežalo određivanje njihovih tačnih položaja na kvefrencijskoj osi.

Potrebno je napomenuti da su na slikama 4.3, 4.6 i 4.9 spektri niskokvefrencijskog kepstra dobijeni tako što je opseg kvefrencija od 0 do 3,33 ms dužine 36 odbiraka dopunjen nulama i konvoluiran sa Hamingovim prozorom dužine 512 da bi se zatim izvršila brza Furijeova transformacija. Tako dobijene vrednosti estimirane raspodele snage signala u frekvencijskom domenu su logaritmovane ($\log_{10}$), pomnožene sa 20 i podešene tako da je najveći nivo na 0 dB.

Da bi se mogli uporediti nivoi običnog spektra i spektra niskokvefrencijskog kepstra, ovaj „spektar kepstra“ je morao da bude skaliran, tačnije pomnožen sa $20 \cdot 1/\log 10$ kao i odgovarajućim faktorom skaliranja i to: u slučaju snimka *HALO.wav* faktor je 43, za *HALO_ Propusni_opseg_460-2740_Hz.wav* faktor je 28, a za *HALO_SNR_=_15_dB.wav* faktor je 90. Tek nakon skaliranja i podešavanja „spektra kepstra“ tako da mu najveća vrednost bude 0 dB utvrđeno je da je spektar niskokvefrencijskog kepstra zapravo gruba aproksimacija spektralne anvelope običnog spektra.

Problem je međutim u tome što nije utvrđeno na osnovu čega se može izabrati vrednost faktora skaliranja, pa su navedene vrednosti odabrane metodom pokušaja i pogrešaka, dakle subjektivno. Osim toga, uočeno je da, baš kao što je navedeno u [OSH00] i u odeljku 2.6 ovog rada, „spektar kepstra“ ima prenaglašene vrednosti na visokim frekvencijama što je posebno uočljivo na slici 4.9 gde je ulazni govorni signal zašumljen, a „spektar kepstra“ ima čitav niz velikih pikova.

5. *Zaključak*

U većini slučajeva se, uz adekvatnu dužinu prozora, osnovna frekvencija može relativno pouzdano odrediti na osnovu pozicije i amplitude pikova kepstra. Dakle, prisutnost istaknutog pika u kepstru je vrlo jaka indikacija da je odgovarajući govorni segment zvučan. Međutim, odsutnost pika ili postojanje pika male vrednosti ne mora nužno da znači da je taj govorni segment bezvučan.

Na žalost, kao što je to često slučaj u obradi govora, postoje brojni specijalni slučajevi i kompromisi o kojima treba voditi računa kod dizajna algoritama za određivanje osnovne frekvencije. Dodatna logika potrebna za obradu specijalnih slučajeva zahteva značajnu količinu dodatnog koda u programskim implementacijama. Algoritam zasnovan na kepstru nije izuzetak.

Problem je granica bezvučnog i zvučnog glasa, kao i degradacija govornog signala. Zbog toga je neophodno koristiti adaptivni prag za detekciju F_0 i različite vrste postprocesiranja od kojih se medijan peglanje pokazalo kao pristojno, kompromisno rešenje.

Na kraju, prikazani algoritmi su testirani na snimku samo jedne kratke reči pa neko buduće ispitivanje treba proširiti većim fondom reči koje će izgovoriti više osoba, i muških i ženskih. Takođe, nekad se i dužina prozora menja i prilagođava signalu tokom obrade na osnovu prethodnih (ili usrednjenih) estimacija osnovne frekvencije.

Osim toga, umesto praćenja trenutnih promena F_0 mogu se posmatrati njene statističke osobine, npr. srednja vrednost i standardna devijacija, koje opisuju F_0 tokom dužeg vremenskog intervala. Osnovna frekvencija je jedan od robusnijih akustičkih parametara jer ostaje gotovo neizmenjena prilikom prenosa govornog signala pod lošim uslovima kao što je npr. telefonski razgovor.

Literatura

[BEN08] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang (Eds.), *Springer Handbook of Speech Processing*, Springer-Verlag Berlin Heidelberg, 2008.

[BOE09] Paul Boersma, David Weenink, *Manual for the Praat: doing phonetics by computer*, program version 5.1.29, www.praat.org, 2010.

[GIL05] Amos Gilat, *Uvod u MATLAB 7*, prevod drugog izdanja, Mikro knjiga, 2005.

[JOV99] Slobodan T. Jovičić, *Govorna komunikacija: fiziologija, psihoakustika i percepcija*, Nauka, Beograd, 1999.

[MAT04] MATLAB – The Language of Technical Computing, The Help documentation, program version 7.0.0.19920 (Release 14), The MathWorks, Inc. 1984 – 2004.

[MIL99] Ljiljana Milić, Zoran Dobrosavljević, *Uvod u digitalnu obradu signala*, Elektrotehnički fakultet, Beograd, 1999.

[OSH00] Douglas O'Shaughnessy, *Speech Communications: Human and Machine*, Second Edition, IEEE Press, 2000.

[PET02] Davor Petrinović, *Digitalna obrada govora* – skripta, Fakultet elektrotehnike i računarstva, Zagreb, 2002.

[POP03] Miodrag Popović, *Digitalna obrada signala*, III izdanje, Akademska misao, Beograd, 2003.

[RAB78] Lawrence R. Rabiner, Ronald W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.