Izveštaj
Analiza podataka

Jelena Božičković, BI 23/2015, jelena.bozickovic@live.com

I. Baza podataka

Parkinsonova bolest (PB) predstavlja neurodegenerativno oboljenje centralnog nervnog sistema, koje uzrokuje delimičan ili potpun gubitak motornih funkcija, govora itd. S obzirom da bolest zahvata najčešće stariju populaciju, razvoj mobilnih aplikacija za dijagnostiku i monitoring bi znatno olakšao suočavanje sa bolešću svim obolelim osobama. Takve aplikacije bi na osnovu audio signala vršila prepoznavanje govora i utvrđivala da li osoba boluje od Parkinsonove bolesti ili ne. Pored teledijagnoze, omogućen bi bio i telemonitoring, odnosno praćenje napretka i stadijuma bolesti. Baza podataka na kojoj je rađena analiza podataka predstavlja javno dostupnu bazu sa XXX sajta, koju je omogućio departman za neurologiju sa Cerrahpasa medicinskog fakulteta u Istanbulu. Baza sadrži 40 ispitanika, od kojih 20 boluje od Parkinsonove bolesti, a ostalih 20 su zdravi ispitanici. Od svakog ispitanika prikupljeno je 26 audio snimaka, u kojima su posebno naglašeni samoglasnici, brojevi, reči i kratke rečenice. Ukupno u bazi postoji 1040 uzoraka, sa 26 obeležja koja su dobijena vremenskom i frekvencijskom analizom audio snimaka.

Sva obeležja u korišćenoj bazi su numerička i mogu se podeliti u 6 grupa. Frekvencijski, pulsni, amplitudski, glasovni, harmonijski i parametri piča. Frekvencijski parametri obuhvataju nekoliko načina opisivanja podrhtavanja glasa, *eng. Jitter*, koji opisuje frekvencijsku nestabilnost, pulsni parametri podrazumevaju broj pulseva, broj perioda, srednju vrednost perioda i njihovu standardnu devijaciju. Amplitudski parametri se odnose na nestabilnost amplitude, *eng. Shimmer* . Glasovni su oni koji opisuju bezvučne odbirke, broj pauza u govoru, kao i dužinu same pauze. Harmonijski opisuju autokorelaciju odbiraka, odnos govor-šum i obrnuto. Dok parametri piča opisuju jačinu zvuka i to po medianu, srednjoj vrednosti, standardnoj devijaciji, maksimumu i minimumu.

Tokom normalnog govora, ne bi trebalo da postoji nestabilnost ni u frekvencijskom ni u amplitudskom smislu, naročito kod izgovora samoglasnika.

II. Analiza podataka

Podaci su analizirani u programskom okruženju Matlab. Korišćenjem određenih statističkih alata, analizirana su sva obeležja i utvrđen je kako dinamički tako i interkvartilni opseg, provereno je da li postoje outlier-i, i određena je najveća korelacija među obeležjima.

A. Dinamički i interkvartilni opseg

Dinamički opseg predstavlja opseg u kom se nalaze svi uzorci, dok interkvartilni opseg predstavlja opseg između 25. i 75. percentila.

Tabela 1: Dinamički/interkvartilni opseg obeležja kod obolelih od Parkinsonove bolesti

	Obeležje 1	Obeležje 2	Obeležje 3	Obeležje 4	Obeležje 5
Frekvencijski parametri	**10.244/** 1.945	0.0007027/ 0.000537	6.063/ 1.028	**7.114/** 1.131	**18.190/** 3.084
Pulsni parametri	38.966/ 5.206	2.460/ 0.408	**16.076/** 2.824	**30.725/** 4.066	/
Amplitudski parametri	44.247/ 7.292	**48.226/ 8.472**	0.409/ 0.095	0.761/ 0.160	24.907/ 4.335
Glasovni parametri	364.935/ 74.599	**293.344/ 14.778**	372.305/ 57.025	/	/
Harmonijski parametri	88.158/ 33.328	11/ 1	60.298/ 18.738	/	/
Pič parametri	512.433/ 101.889	**1490/** 69	**1489/** 68	0.009/ 0.002	0.006/ 0.005

Tabela 2: Dinamički/interkvartilni opseg obeležja kod zdravih ispitanika

	Obeležje 1	Obeležje 2	Obeležje 3	Obeležje 4	Obeležje 5
Frekvencijski parametri	**14.186/** 1.672	0.0007011/ 0.0001	7.932/ 0.874	**13.461/** 0.890	**23.796/** 2.623
Pulsni parametri	38.689/ 6.799	2.618/ 0.526	**25.324/** 3.631	**72.152/** 4.872	/
Amplitudski parametri	40.298/ 6.246	**75.971/ 10.891**	0.458/ 0.107	0.867/ 0.182	27.723/ 5.148
Glasovni parametri	377.009/ 75.420	**161.770/ 28.273**	381.581/ 59.380	/	/
Harmonijski parametri	85/ 36.158	12/ 2	69.117/ 25.483	/	/
Pič parametri	426.400/ 175.762	**800/** 75.5	**749/** 71.5	0.008/ 0.003	0.004/ 0.001

Ukoliko uporedimo dinamičke opsege frekvencijskih parametara kod obolelih i zdravih ispitanika, prikazanih u tabeli 1 i 2, možemo zaključiti da manje frekvencije kod obolelih ispitanika ukazuju na brže promene u signalu, što je i logično s obzirom da glas više drhti kod obolelih nego zdravih osoba. Kod pulsnih parametara najviše se razlikuju obeležja 3 i 4, odnosno, srednja vrednost i standardna devijacija perioda. Oba obeležja su mnogo manja kod osoba koje pate od PB. Amplitudska nestabilnost se ne

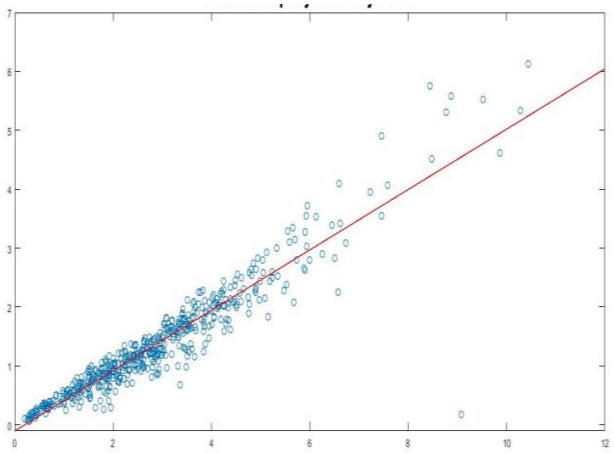
razlikuje mnogo u 2 klase, osim drugog obeležja koji predstavlja amplitudu u decibelima. Vidimo da zdravi ispitanici mogu mnogo jači intenzitet zvuka da proizvedu, skoro pa duplo jači od zvuka koji proizvode obolele osobe. Intuitivno možemo pretpostaviti da osobe obolele od PB prave mnogo više pauza u govoru. Obeležje 2 glasovnih parametara upravo to i potvrđuje. Pomoću obeležja 2 i 3 pič parametra jasno možemo razlikovati klase, kod obolelih osoba njihova vrednost je duplo veća.
Obeležje kojim najbolje možemo razdvojiti 2 klase ispitanika, sudeći po interkvartilnom opsegu, je broj pauza u govoru.
Proverom modusa po obeležjima utvrđeno je da se u više od 40% uzoraka pojavljuje vrednost nula za odnos signal-šum, što je po mom mišljenu logično jer u govoru ne bi trebao da postoji šum, osim kod bezvučnih suglasnika, koji nisu bili predmet ispitivanja u ovoj bazi.

B. Outlier-i

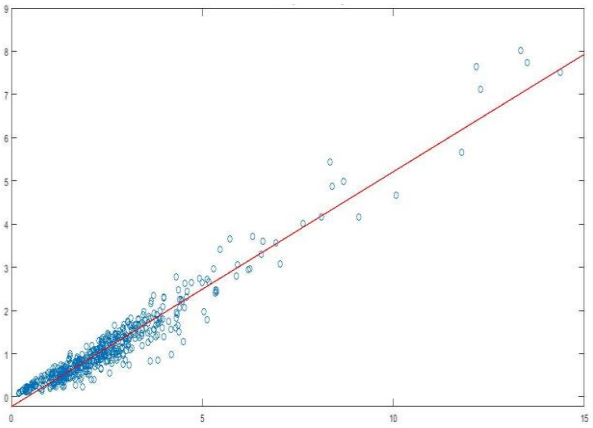
Većina obeležja sadrži vrednosti koje odstupaju od uobičajenih, međutim obeležje amplitudske nestabilnosti kod obolelih od PB ne prikazuje outlier-e, što se može tumačiti tako da ne postoje vrednosti tog obeležja koje odstupaju od opsega u kom se nalazi većina vrednosti. Sa druge strane, kod zdravih ispitanika, postoje vrednosti amplitudskog obeležja koje se tumače kao outlier-i. Razlog tome zdravi ispitanici mogu da povise ton iznad određene granice, čime se povećava amplituda govornog signala, dok oboleli od PB ne uspevaju da učine isto.

C. Korelacija

Najveću korelaciju među obeležjima imaju obeležja frekvencije i to obeležje 1, tj. varijacija osnovne frekvencije, i obeležje 3, tj. apsolutna razlika između frekvencije jedne periode i njena 2 najbliža suseda. Korelacije za klase prikazane su na slikama 1 i 2.



Sl. 1. Grafik rasipanja obeležja 1 i 3 kod klase obolelih od PB



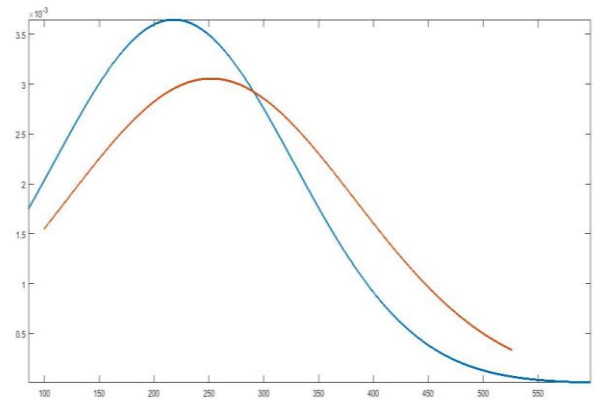
Sl. 2. Grafik rasipanja obeležja 1 i 3 kod klase zdravih ispitanika

Prema slikama 1 i 2 možemo zaključiti da je najveći broj uzoraka skoncentrisan na nižim vrednostima, da je zavisnost linearna, kao i da su obeležja pozitivno korelisana. Faktor korelacije prikazanih obeležja u obe klase je približno 0.86.

D. Prvi pič parametar – median procene jačine zvuka

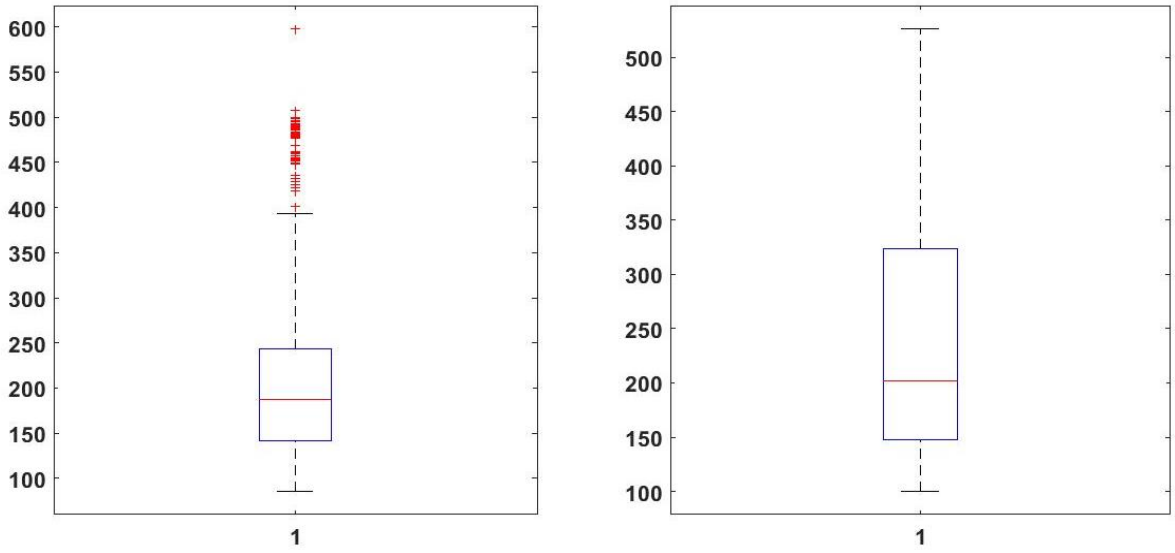
Na slici 4 prikazani su boxplot grafici na kojima su ilustrovani dinamički opseg, interkvartilni opseg, median i outlier-i. Poređenjem grafika možemo zaključiti da klasa zdravih ispitanika ima veći i dinamički i interkvartilni opseg, da ne postoje vrednosti koje se mnogo razlikuju od vrednosti u tim opsezima. Median ovog obeležja je približno isti za obe klase i iznosi približno 200. Klasa obolelih od PB ima mnogo uži i dinamički i interkvartilni opseg, kao i mnoštvo outliera. Grafici se poklapaju sa hipotezom da kad čovek, koji nije oboleo od PB, govori jačina njegovog govora može da se povećava ili smanjuje, ali uvek ostaje u određenim granicama koje sam čovek dozvoli. Međutim, kada osoba koja je obolela od PB naglo povisi ili smanji jačinu govora, dolazi do podrhtavanja glasa, pri čemu određeni vokali dobijaju veću jačinu. Ti vokali su na ovom grafiku predstavljeni kao outlier-i, jer odstupaju od osnovne jačine govora.

E. Raspodele prvog i drugog pič parametra



Sl. 3. Normalne raspodele parametara mediana i srednje vrednosti jačine zvuka

Plava raspodela predstavlja raspodelu klase 1, obolelih



Sl. 4. Box plot grafici kod klase obolelih od Parkinsonove bolesti, i klase zdravih ispitanika, respektivno.

ispitanika, a crvena klasu 2. Srednje vrednosti obe klase su približno jednake, a varijansa crvene je veća od varijanse plave. Plava brže opada, ali ima i širi opseg vrednosti.

III. ZAKLJUČAK

Među datim obeležjima postoje ona koja se veoma razlikuju između klasa. Korišćenjem tih obeležja bilo bi moguće napraviti klasifikator koji bi klasifikovao one audio signale koji potiču od osoba obolelih od Parkinsonove bolesti, i signale zdravih ispitanika.