UNIVERZITET U NIŠU
ELEKTRONSKI FAKULTET

Seminarski rad

Predmet: Analiza multimedijalnih podataka

Klasifikacija muzičkih instrumenata

Student:

Dimitrije Mitić 822

Mentor:

dr Milena Stanković

Sadržaj

1.Uvod

Pod pojmom pribavljanja informacija iz muzičkih sadržaja (*Music Information Retrieval – MIR*), smatra se relativno nova, interdisciplinarna nauka koja se bavi pribavljanjem relevantnih informacija iz muzičkih komada, kako bi se, uz pomoć ovih podataka, omogućilo efikasno pronalaženje i pribavljanje određenog tipa muzičkog sadržaja iz veoma obimnog muzičkog dataset-a kakvi danas postoje na internetu [1]. Dobar deo glavnih problema kojima se MIR bavi se svode na određenu vrstu problema klasifikacije, kao što su: klasifikacija muzičkog žanra, klasifikacija raspoloženja (da li neki muzički komad spada u ozbiljne, veselije, usporenije itd.), klasifikacija muzike na osnovu autora ili mesta nastanka, itd. Glavni razlozi za nagli razvoj MIR-a i njegove raširene primene su: razvoj tehnika za kompresiju audio signala, rastuća procesorska moć računara koja je omogućila brzo ekstraktovanje korisnih karakteristika iz audio signala, pojava muzičkih striming servisa kao što su: *Spotify* ili *Deazer* koji su omogućili da su korisniku skoro uvek i svuda dostupne baze podataka sa velikim količinama raznovrsnog audio materijala.

Klasifikacija muzičkog sadžaja na osnovu instrumenata koji se javljaju u istom, je jedan od problema kojim se bavi MIR. Kod ovakve vrste klasifikacije potrebno je pozabaviti se sledećim glavnim pitanjima: koje karakterisitike audio signala najbolje opisuju kvalitet tona kada je reč o različitim instrumentima, a drugo pitanje se odnosi na to koju vrstu klasifikatora treba upotrebiti kako bi se napravila efikasna predikcija, odnosno preslikavanje iz domena karakteristika na domen odgovarajućih oznaka (u ovom slučaju vrste instrumenata), tako da greška kod predviđana bude minimalna. U velikom delu radova na ovu temu, korišćene su karakteristike koje se odnose na boju zvuka (*timbre features*). Boja zvuka (*timbre*) predstavlja jednu od ključnih stvari kojom se opisuju muzički komadi, iz razloga što različiti izvori zvuka proizvode zvuke koji se međusobno razlikuju upravo po boji. Na osnovu ove grupe karakteristika, moguće je razlikovati zvuk koji proizvodi npr. klarinet od onoga koji proizvodi klavir (čak i slučaju da je odsvirani tonovi imaju istu visinu i glasnost) ili od ljudskog glasa. Neke karakteristike iz ove grupe se pribavljaju iz zvuka koji je predstavljen u vremenskom domenu, dok je za neke potrebno prebacivanje u frekvencijski domen uz pomoć Furijeove analize. Neke najpoznatije karakteristike iz *timbre* grupe, koje se najčešće upotrebljavaju kod raznih muzičkih, ali i drugih zvučnih klasifikacija, kao i prepoynavanje govora su: *Spectral Centroid, Spectral Bandwidth, Zero Crossing rate, Mel Frequency Spectral Coeficients* (*MFFCs*) i dr [1]. Neke od karakteristika koje se odnose na boju zvuka, je potrebno analizirati kako evoluiraju kroz vreme, u ovu vrstu karakteristika integrisanih u vremenu: spadaju spektrogrami, Mel-spektrogrami, kao i MFCC-i. Kada je reč o mehanizmima koji se koriste za samu klasifikaciju, autori se veoma često oslanjaju na *deep leaning* tehnologije koje vrše klasifikaciju na osnovu ekstrakovanih prethodno pominjanih karakteristika zvuka pretvorenih u tzv. vektor karakteristika (*feature vector*). Sve je više radova koji predlažu klasifikaciju uz pomoć konvolucionih neuronskih mreža (CNN) koje se koriste kod klasifikacije slika. Kod ovog pristupa se neke vrste spektrograma koriste umesto slika kao ulaz za CNN.

U ovome radu eksperimentisano je sa automatskim prepoznavanjem predominatnog instrumenata uz pomoć običnih i Mel spektrograma koji se pribavljaju iz muzilkih komada

(podeljenih na nepreklapajuće segmente u trajanju od 1s), kao i uz pomoć MFCC komponente. Kao klasifikator je korišćena duboka neuraonska mreža, konkretno CNN sa više slojeva i filtera. Sam rad je podeljen na dve celine. U prvoj je data teorijska pozadina koja se odnosi na karakteristike koriščene u klasifikaciji. U ovome delu će više reči biti o Furijeovim transforamcijama i frekvencijskom domenu u kojima se mogu predstavaljati audio signali, takođe i o *Short-time* Furijeovoj transformaciji kao načinu za kreiranje spektrograma. Drugi deo ovog rada se bavi analizom samog procesa klasifikacije, preprocesiranja audio signala, arhitekturom klasifikatora, kao i analizom samih rezultata.

2. Karakterstike zvuka korišćene u klasifikaciji

Proces odabira adekvatnih karakteristika zvuka je ključan prilikom kreiranja sistema za klasifikaciju, iz razloga što od tog izbora zavisi da li će iz zvuka biti izvučene one informacije potrebne za razlikovanje zvuka proizvedenog od različitih izvora. Svaki muzički zvuk se može opisati sa četiri perceptualne karakteristike: visina (*pitch*), glasnost, trajanje i boja (*timbre*). Prve tri karakteristike su usko vezane i jasno se mogu meriti kroz fizičke veličine kao što su redom: frekvencija, amplituda i vreme. Uz pomoć boje zvuka ljudski mozak može da razlikuje dva zvuka sa istom visinom i glasnošću. Boja zvuka je znatno kompleksnija karakteristika i ne može se opisati kroz samo jednu fizičku karakteristiku signala. Proučavajući način na koji ljudski mozak percipira boju zvuka, došlo se je do zaključka da ista najviše zavisi od karakteristika koje se tiču frekvencijskog spektra datog signala (distribucija energije po frekvencijama, promena te energije kroz vreme, zastuplejnost energije u harmonicima na višim frekfencijama i dr.) [2]. U ovom radu je isprobano korišćenje nekoliko tipova spektrograma kao glanvih karakteristika koje pokazuju na koji način frekvencijski spektar audio signala varira kroz vreme. U nastvaku ovog poglavlja biće objašnjen način na koji se početni audio signal prebacuje iz vremenskog u frekvencijski domen uz pomoć Furijeove transformacije, kao i niz drugih transformacija koje su potrbne da bi se uhvatila promena energije u harmonicima kroz vreme.

2.1 Furijeova transformacija

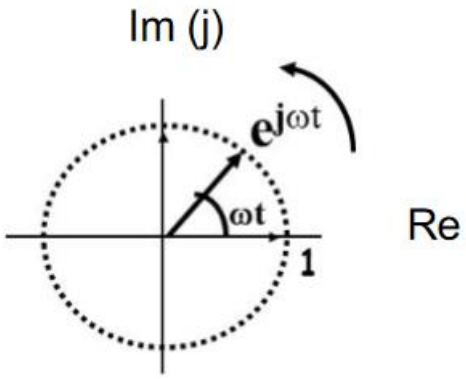
Pod Furijeovom transformacijom se podrazumeva matematička metoda kojom se konkretan signal, predstavljen svojom funkcijom (u našem slučaju je to audio signal), može predstaviti kao potencijalno beskonačna suma bazičnih sinusoidnih funkcija [3]. Ovo znači da se jedan složeni signal može predstaviti kroz niz sinusoidnih, prostih, signala koji imaju različite frekvencije, amplitude i faze. Korišćenjem Furijeove transformacije moguće je prebaciti signal iz vremenskog domena u frekvencijski, samim tim je moguće analizirati spektar frekvencija koje sačinjavaju taj originalni signal. Kroz izraz 1 dat je matematički oblik Furijeove transformacije, dok je izrazom 2 dat prikaz inverzne transforamcije kojom se uz pomoć frekvencijskog opsega dolazi do originalnog signala u vremenskom domenu [4]:

$$(1) \qquad \mathcal{F}\{x(t)\} = F(j\omega) = \int_{-\infty}^{+\infty} f(t)e^{-j\omega t}dt$$

$$(2) \qquad \mathcal{F}^{-1}\{F(j\omega)\} = f(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} F(j\omega)\, e^{j\omega t}d\omega$$

Pri čemu je *x(t)* složeni signal predstavljen u vremenskom domenu, $F(j\omega)$ je Furijeova transformacija tog signala, ω je ugaona frekvencija izražena u radijanima po sekundi, *e* je Ojlerov broj, dok je *j* oznaka za imaginarni broj. Ugaona frekvencija se još može izraziti preko izraza: $\omega = 2\pi f$, pri čemu je *f* frekvencija sinusoidnog signala izražena u Hercima (Hz) i recipročna je periodu oscilovanja T ($f = 1 / T$). Sam eksponencijalni izraz $e^{-j\omega t}$ se koristi za predstavljanje prostog sinusiodnog signala koji osciluje sa datom frekvencijom ω, tj njime se

predstavlja frekvencijska komponenta originalnog signala. Ovaj izraz se interpretira kao vektor, intenziteta 1 koji rotira u smeru kazaljke na satu u kompleksnoj ravni, brzinom ω radijana u sekundi, pri čemu su $sin(\omega t)$ i $cos(\omega t)$ projekcije ovog vektora na imaginarnoj i realnoj osi respektivno. Ovo je priakzano na slici 1. Iz izraza 2 se može videti da se signal rastavlja na sumu svojih frekvencijskih komponenti, pri čemu sama vrednost $F(j\omega)$ služi kao težinski faktor date frekvencijska komponente.



Slika 1 Trigonometrijski oblik izraza $e^{j\omega t}$

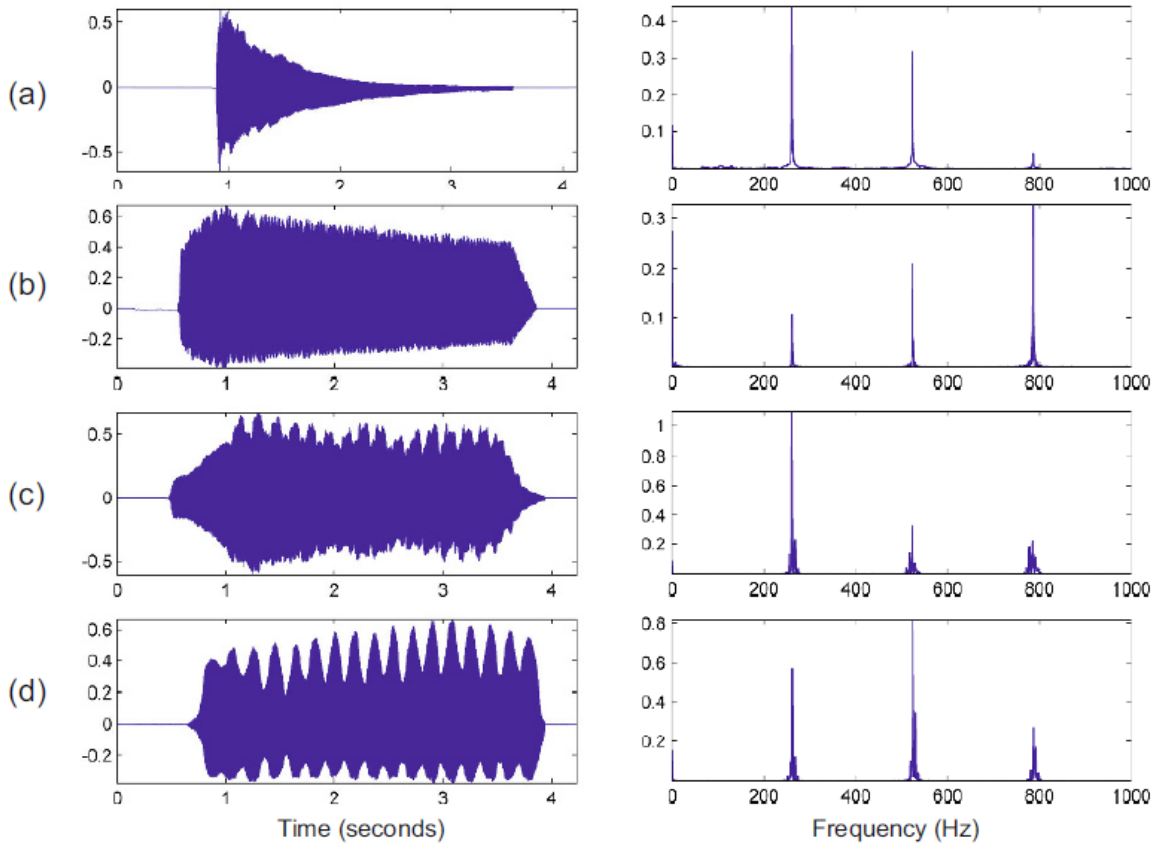
Vrednost funkcije $F(j\omega)$ je kompleksan broj i u koliko taj broj posmatramo u njegovoj polarnoj formi datoj kroz sledeći izraz [5]:

$$(3) \qquad F(j\omega) = |F(j\omega)|e^{j\varphi_\omega}$$

gde moduo ovog broja, $|F(j\omega)|$, predstavlja koliko je frekvencija ω zastupljena u originalnom signalu i služi za kreiranje magnitudnog spektra datog signala, dok argument (ugao) φ_ω $(arg(F(j\omega)))$ označava pomeraj tj fazu koju ima data sinusoidna komponenta i koristi se kod kreiranja faznog spektra. Magnitudni i fazni spektar zajedno čine spektar signala. Kod inverzne Furijeove transformacije suma prostih sinusoida za svaku frekvenciju ω, pri čemu se svaka sinusoida množi sa odgovarajućom vrednošću $|F(j\omega)|$ i dobija pomeraj $arg(F(j\omega))$, daje u konačnici početni signal. U spektralnoj analizi audio signala često se koristi kvadrat magnitude ($|F(j\omega)|^2$) kojim se izražava koliko je koja frekvencija energetski zastupljena u originalnom signalu. Ova vrednsot se npr koristi kod kreiranja spektrograma, da bi se posmatralo kako energija određenih frekvencija varira kroz vreme.

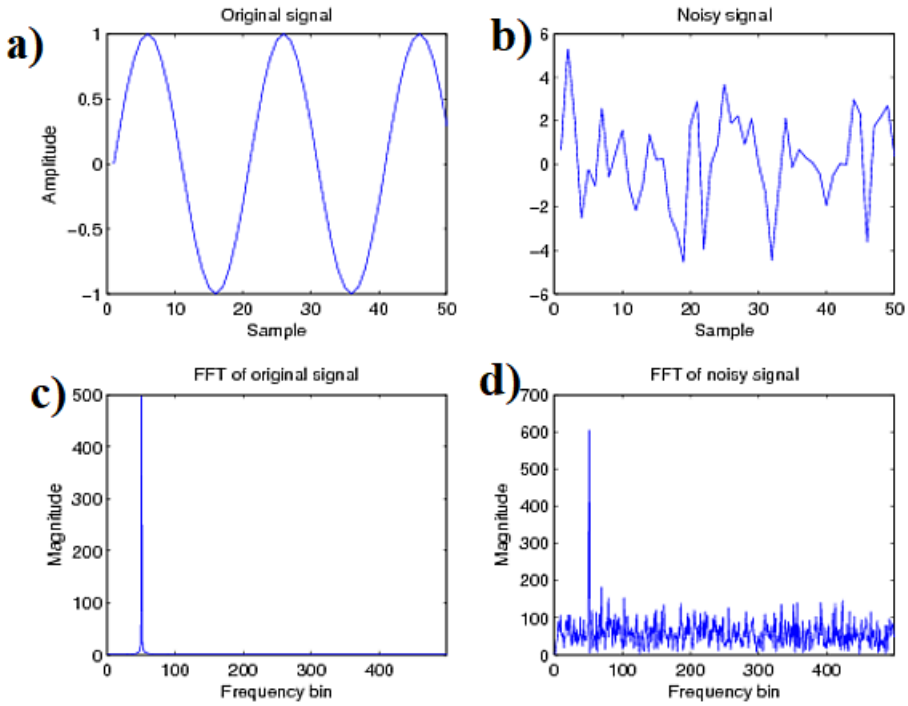
Na slici 2 je dat prikaz originalnog signala i odgovarajućeg magnitudnog spektra za ton C4 izveden od strane: a) klavira, b) trube, c) violine, d) flaute. Posmatrajući samo vremenske domene datih signala, ne mogu se izvući nikakva saznananja o sličnosti ili razlikama datih signala. Međutim, u koliko posmatramo frekvencijski domen, može se primetiti da C4 ton sadrži izrazitu frekvencijsku komponentu sa 262 Hz, što se moze videti posmatrajući vrhove magnitudskog spektra. Takođe se može videti da su naredne frekvencije koje su zastupljene celobrojni umnožci ove fundamentalne frekvencije od 262 Hz (524 Hz, 768 Hz,..). Može se takođe zapaziti jedna informacija koja se tiče same boje zvuka, a to je da kod različitih

instrumenata, za odsvirani isti ton, postoje iste frekvencijske komponente, ali se distribucija energije medju njima razlikuje od instrumenta do instrumenta što se može videti iz visina magnituda za istu frekvenciju na magnitudnim spektrima instrumenata.



Slika 2 Prikaz signala u vremenskom i frekvencijskom domenu za odsvirani C4 ton na: a) klaviru, b) trubi c) violini d) flauti [5]

Značaj Furijeove transformacije može se još pokazati i kroz naredni primer. U koliko imamo neki početni signal, u ovom primeru to je običan sinusoidni signal (slika 3a) sa frekfencijom 50 Hz, i u koliko ovakvom signalu dodamo određeni šum, forma signala se menja što se vidi na slici 3b. Posmatrajući signal sa šumom u vremenskom domenu, ne možemo izvući nikakvo saznanje o originalnom signalu, niti o prirodi samog šuma tj šta su u signalu korisne informacije, a šta ne. Međutim ukoliko uporedimo magnitudne spektre originalnog i signala sa šumom (slike 3c i 3d), jasno se vidi da je u signalu najviše zastupljena jedna frekfencija (50 Hz), dok je šum u stvari konstantna niska energija distribuirana kroz sve frekvencije. Sada je moguće, upotrebom filtera koji bi isključio ove slabo zastupljene frekfencije iz spektra i upotrebom inverzne Furijeove transformacije, ponovo doći do početnog signala.



Slika 3 Signal sa i bez šuma u vremenskom i frekvencijskom domenu

Jednačine za Furijeovu transformaciju 1 i 2 su primenljive na analogni, kontinualni signal, pri čemu su *x(t)* i $F(j\omega)$ kontinualne funkcije od vremena (*t*) i od frekvencije (ω) respektivno. Kada se radi o digitalnoj obradi signala, računari ne mogu da rade sa kontinualnim vrednostima tj. sa analognim signalima. Mora se pronaći način da se ove vrednosti nekako diskretizuju. Prvi korak u digitalnoj obradi signala jeste konverzija iz analognog u digitalni oblik. Ovo se obavlja uz pomoć ekvidistantnog semplovanja, tj pribavljanjem i kodiranjem vrednosti nekog signala u jednakim vremenskim razmacima. Ovo se može takođe porikazati u vidu sledećeg matematičkog izraza [5]:

$$(4) \qquad x[n] = x(nT_s)$$

Gde je *x(t)* originalni, analogni signal, $x[n]$ predstavlja semplovanu vrednost originalnog signala u trenutku *nTs*, gde je *Ts* perioda semplovanja, a *n* vremenski indeks. Na osnovu periode *Ts* definiše se i frekvencija semplovanja *Fs = 1 / Ts*, koja označava broj pribavljanjenih semplova u sekundi. Prema teoremi semplovanja, originalni signal *x(t)* se može prilično uspešno rekonstruisati iz njegove diskretne verzije $x[n]$, u koliko *x(t)* ne sadrži u sebi frekvencije više od *Fs* / 2, u protivnom može doći do anomalije poznatije kao *aliasing*. Sama vrednost *Fs* / 2 je poznatija kao Nikvistova frekvencija. Kroz formule 5 i 6 dat je način za računanje diskretne Furijeove transformacije i inverzne Fureijeove trasnsformacije respektivno [5]:

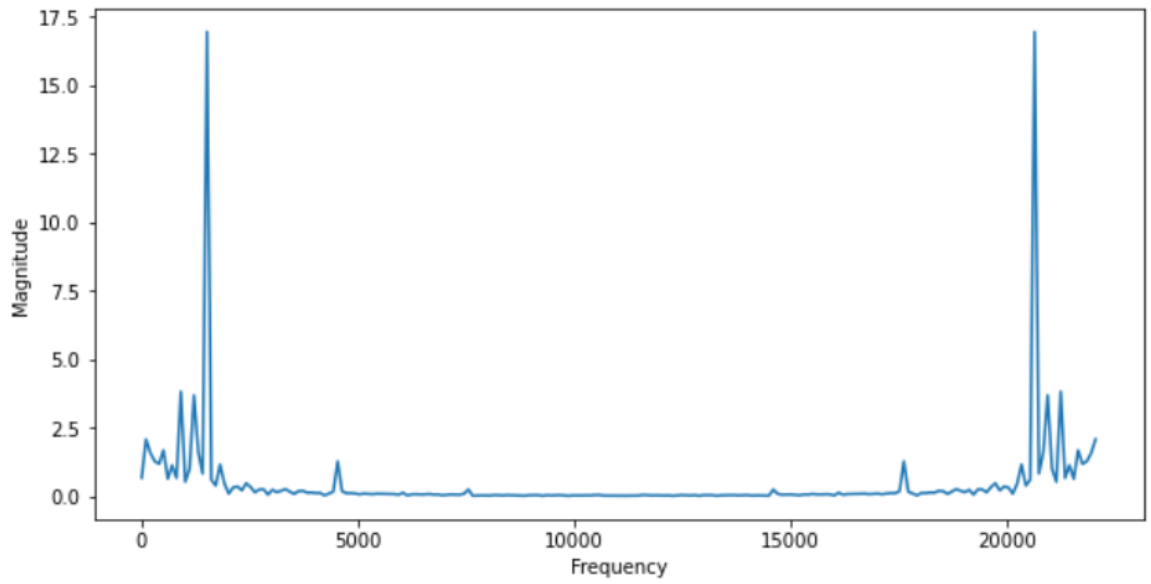
$$(5) \qquad X[k] = \sum_{n=0}^{n=N-1} x[n]\, e^{\frac{-j2\pi}{N}kn}$$

$$(6) \qquad x[n] = \sum_{k=0}^{k=N-1} X[k] e^{\frac{j2\pi}{N} kn}$$

Gde je *x[n]* diskretni signal, *N* je broj semplova originalnog signala, kao i broj Furiejovih koeficijenata koji se računaju (*n* i *k* uzimaju vrednosti iz opsega od 0 do *N*-1). Broj *N* se uvodi da bi se ograničilo sumiranje (smatra se da svi semplovi nakon N, $x[N]$, $x[N+1]$, ... imaju vrednost 0), kao i da bi se diskretizovala sama vrednost ω. Formula koja se koristi da bi se odgovarajući koeficijent *k* za koji je izračunata vrednost Furijeove transformacije $X[k]$, pretvorio u odgovarajuću realnu frekvenciju je sledeća [5]:

$$(7) \qquad F_{coef}[k] = \frac{k\, F_S}{N}$$

Za realne vrednosti *x[n]*, diskretna Furijeova transformacija iskazuje svojstvo simetrije, tako da koeficijenti iznad *N* / 2, postaju redundatni i mogu se zadržati samo koeficijenti $X[k]$ gde $k \in [0, \lfloor N/2 \rfloor]$, odnosno samo koeficijenti koji odgovaraju vrednostima frekvencije do Nikvistove granične frekvencije. Na slici 4 je dat primer magnitudnog spektra tona odsviranog od strane klairenta, gde je frekvencija semplovanja 22050 Hz, gde se može videti da je funkcija simetrična u odnosu na Nikvistovu frekvenciju koja je u ovom slučaju 11025 Hz.



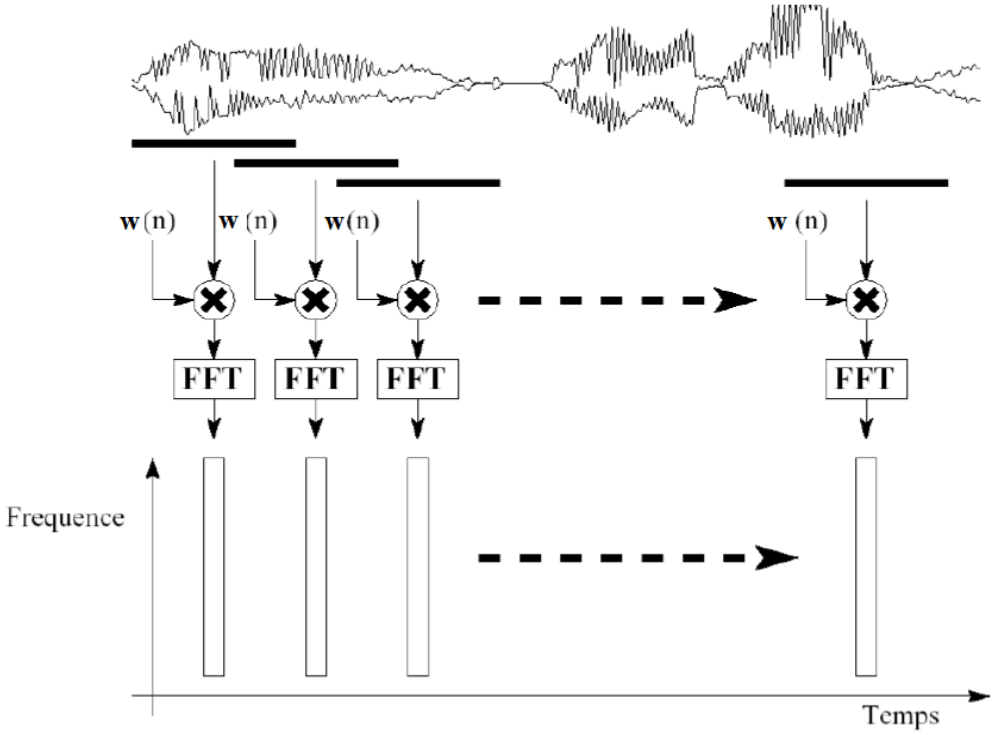
Slika 4 Magnitudni spektar tona klarineta semplovanog sa frekvencijom od 22050 Hz

Gaus i Furije su početkom 19. veka pronašli efikasan algoritam za brzo računanje diskretne Furijeove transforamcije, poznatiji kao brza furijeova transformacija (Fast Fourier Transformation – FFT), koju su kasnije James Cooley i John Tukey optimizovali. Ovaj algoritam, koji se i danas masovno koristi kako u telekomunikaciji, tako i u velikom delu multimedijalnih sistema, oslanja se na rekurziji kojom se složenost tj broj operacija umanjuje sa N^2 na $N \log_2 N$. Pri čemu je N broj semplova u audio fajlu.

Furijeova transformacija transformiše signal iz vremenskog u frekvencijski domen, pri čemu se prilikom ove vrste transformacije gube informacije koje se tiču vremena, odnosno o tome u kojem vremenskom trenutku je koja frekvencijska komponenta aktivna. U sledećem poglavlju će više reči biti o načinu na koji se mogu pribaviti informacije o variranju energije određenih frekvencija kroz vreme, odnsono kako se frekvencijski spektar menja u funkciji vremena.

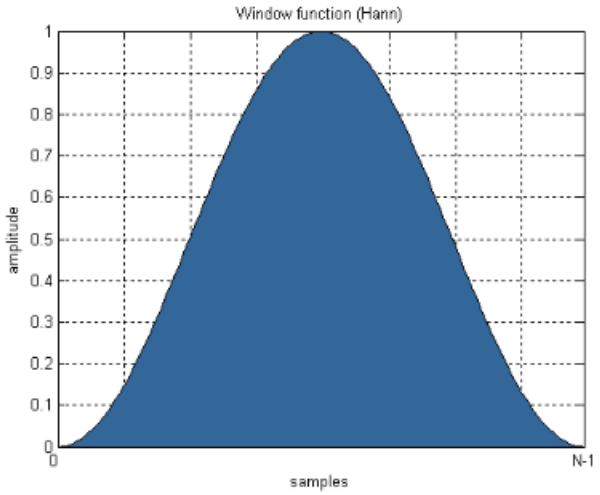
2.2 Kratkotrajna Furijeova Transformacija (*Short Time Fourier Transform - STFT)*

Kao što je već napomenuto, mana Furijeove transformacije je ta što pokazuje samo koje su frekvencijske komponente zastupljene u signalu, ali ne i kada se te komponente javljaju. Što praktično čini Furijeovu transformaciju neadekvatnom kada su u pitanju nestacionirani signali (signal čiji se frekvencijski spektar menja s vremenom) a takvi signali su npr muzika ili ljudski govor. Osnovna ideja kratkotrajne Furijeove transformacije (STFT-a) je da se signal izdeli na kratke segmente (frejmove), pri čemu bi se smatralo da je signal stacionaran u okviru jednog segmenta [6]. Nakon segmentiranja, Furijeova transformacija se primenjuje nad svakim segmentom, što znači da je STFT funkcija zavisna od vremena (vremesnkog segmenta) i od frekvencije. Samo segmentiranje signala se izvodi uz pomoć funkcije prozora koja ima vrednosti različite od nule u samo kratkom vremenskom opsegu, koji je jednak dužini segmenta. Ovakva funkcija se pomera kroz vreme i množi sa originalnim signalom, pri čemu se Furijeova transformacija odvija nad tim rezultatom množenja. Na slici 5 je dat proces dobijanje kratkotrajne Furijeove transformacije od početnog signala u vremenskom domenu. Na slici je sa *w* označena prozorska funkcija koja se množi sa konkretnim segmentom.



Slika 5 Kratkotrajna Furijeova transformacija

Kod STFT-a je važan i odabir prozorske funkcija, koja može biti pravougaona ili u obliku zvona [7]. Od odabira prozorske funkcije zavisi koliko će sam spektar koji se dobija kao rezultat FFT-a, biti osetljiv na anomaliju poznatu kao curenje energije spektra, koja se javlja kao posledica diskontinuiteta na krajevima signala, a koja se manifestuje kroz prisustvo nekih visokih frekvencija u spektru koje nisu zastupljene u originalnom signalu [6]. Prozorske funkcije koje imaju oblik zvona kao što su *Hann* i *Hamming* znatno umanjuju posledice ove anomalije time što izravnjavaju originalni signal (u našem slučaju segment) na njegovim krajevima i često se koriste u spektralnoj analizi muzičkog signala. Na slici 6 je dat izgled *Hann* funkcije koja je korišćena i u eksperimentalnom delu ovog rada.



Slika 6 Hann prozorska funkcija

Prilikom primene prozorske funkcije dolazi do gubitka informacija na krajevima signala. Ovaj problem se otklanja tako što se segmenti na koji se početni signal deli, preklapaju kao što je i prikazano na slici 5. Ovo dovodi do uvođenja još jednog bitnog parametra u računanju STFT-a, a to je pomeraj segmenta (*hop size*) koji predstavlja broj semplova za koji se prozor pomera u vremenu. Formula za izračunavanje diskretnog STFT-a X je sledeća [5]:

$$(8) \qquad X[m,k] = \sum_{n=0}^{n=N-1} x[n+mH]w[n]e^{\frac{-j2\pi}{N}kn}$$

Gde je $X[m,k]$ *k*-ti Furijeov koeficijent za *m*-ti vremenski segment. *N* je širina jednog segmenta u semplovima, *H* je *hop size* parametar. Pri čemu se *k* nalazi u intervalu [0, *K*], gde je *K* vrednost frekvencijskog indeksa za Nikvistovu graničnu frekvenciju, odnosno *K = N / 2*. Ova formula praktično pokazuje da se STFT izračunava tako što se za svaki fiksni vremenski segment *m* dužine N semplova, izračunava spektralni vektor dužine *K* +1 uz pomoć FFT-a.

U koliko bismo želeli da vidimo koje vreme tj koji sempl odgovara konkretnom vremenskom segmentu *m*, koristili bi formulu [5]:

$$(9) \qquad T_{coef}[m] = \frac{m\,H}{F_s}$$

Dok bi se rezolucija koeficijenta frekvencije na određenu konkretnu frekvenciju izračunavala uz pomoć formule 7, s tim što bi *N* označavalo broj semplova u jednom segmentu.

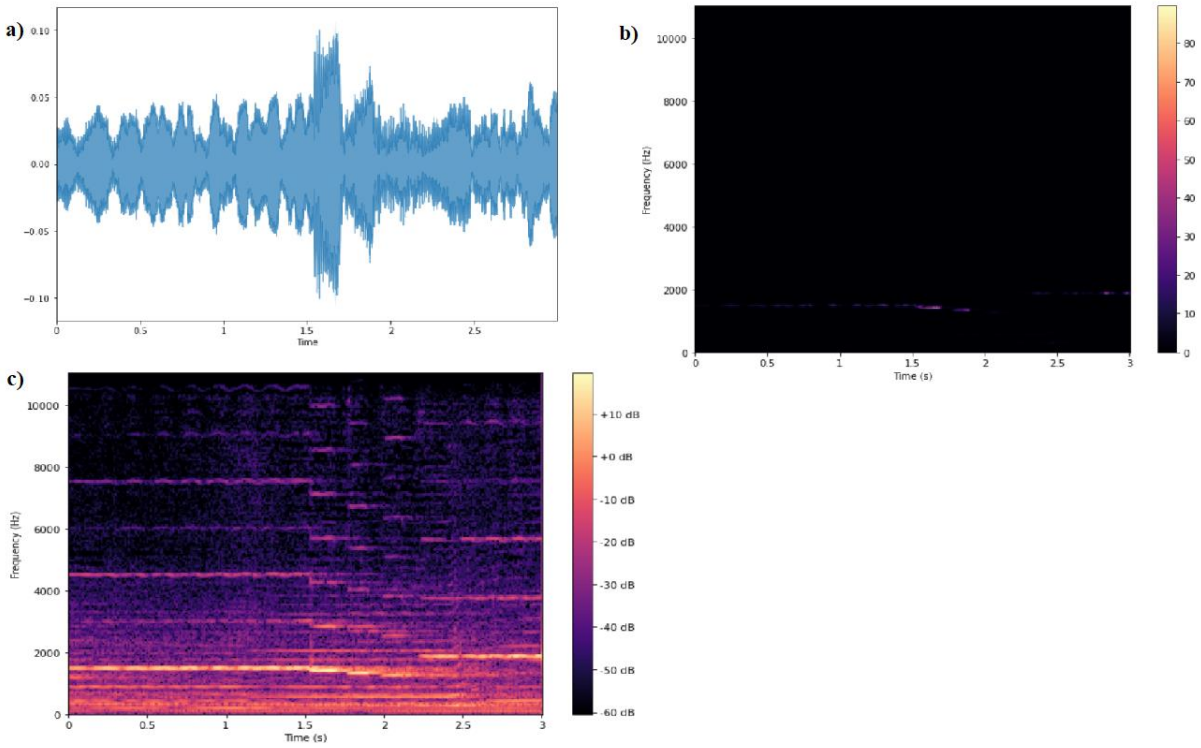
Iz formule 8 se vidi da su glavni faktori koji bitno utiču na razultat i koje je važno uzeti u obzir pre izračunavanja STFT-a: širina segmenta, pomeraj (*hop length*) i oblik prozorske funkcije. Širi segmenti omogućuju bolju frekvencijsku rezoluciju tj međusobno bliže frekvencije se mogu bolje razlikovati, iz razloga što dobijamo više frekvencijskih koeficijenata, dok je ovakva vrsta segmenata nepogodna sa aspekta vremenske rezolucije (veći broj vremeskih semplova se predstavlja kroz jedan vremenski koeficijent). Uži segmenti su bolji sa aspekta vremenske rezolucije, ali nisu efikasni sa frekvencijskog aspekta. Iz ovih razloga je potrebno naći neki kompromis kada je reč o širini segmenta. Vrednosti koje se najčešće koriste u spektralnoj analizi muzičkih sadržaja, za širinu vremenskog segmenta su: 2048, 1024 ili 512 [8]. Kada je reč o vrednosti pomeraja (*hop size*) u koliko imamo velikog preklapanja između susednih segmenata to dovodi do pojave redundantnih frekvencijskih vektora, dok je u slučaju manjeg preklapanja, moguće doći do gubitka informacija. Najčešće dužina pomeraja se uzima u odnosu na širinu segmenta i ta vrednost je obično *N / 2* ili *N / 4*.

Rezultat STFT-a su, kao i u slučaju FFT-a, kompleksne vrednosti. Kvadrat magnitude STFT-a je moguće vizualizovati kao spektrogram $Y[m,k]$, što se može predsstaviti kroz sledeći izraz [8]:

$$Y[m,k] = |X[m,k]|^2 \quad (10)$$

Spektrogram je 2D reprezentacija (slika) varijacije energija (ili magnitude) frekvencijskih komponenti signala kroz vreme [8]. Kod spektrograma horizontalna osa predstavlja vreme, vertikalna frekenciju, dok sam intenzitet boje u određenoj tački predstavlja kvadrat magnitude (energiju), za određenu frekvenciju u određenom trenutku.

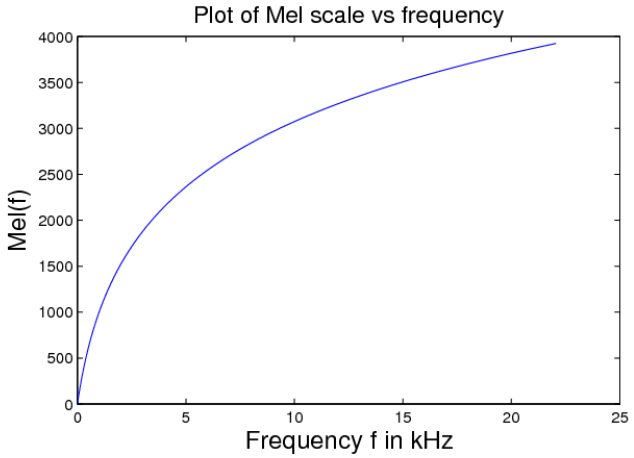
Na slici 7 je dat prikaz spektrograma za audio snimak muzike violončela u trajanju od 3s, gde je na slici 7a dat je prikaz ovog signala u vremenskom domenu, na slici 7b je dat prikaz energetskog spektrograma datog signala, dok je na slici 7c dat prikaz spktrograma pri čemu je vrednost energije logaritamski skalirana i izražena u decibelima. Širina segmenta koja je korišćena je 512, dok je pomeraj 256 semplova, a korišćena prozorska funkcija je *Ham.* Sa slike 7b se može videti da u koliko su vrednosti energije (kvadrati magnitude) izražene linearno, jedino što se može zaključiti je da postoje neke frekvencijske komponente ispod 2KHz. Ovakvo predstavljanje energije je neefikasno, jer zbog velike razlike u vrednostima, ne možemo da vidimo komponente zvuka koje su relevantne sa aspekta ljudkse percepcije zvuka. Mnogo je jednostavnije da energija bude logaritamski skalirana i izražena u decibelima (slika 7c), što je mnogo bliže načinu na koji ljudsko uho percipira zvuk. Na slici 7c se može jasno videti da je najdominantnija frekvencija (izražena najsvetlijom bojom) oko 1.7KHz, kao i da je ova frekvencija zastupljena do kraja audio snimka. Pored ove komponente, nesto energetski slabije i kraćeg trajanja, zastupljene su komponente koje predstavljaju harmonike osnovne frekvencije.



Slika 7 Audio signal muzike violončela,predstavljen: a) u vremenskom domenu b) preko spektrograma c) preko spektrograma gde je energija izražena logaritamski

2.3 Mel spektrogrami i MFCC

Ljudska percepcija frekvencije zvuka nije linearne, već logaritamske prirode. Što znači da ljudsko uho bolje pravi razliku između niskih frekvencija nego između visokih. Stevens, Volkman i Newman su 1937 na osnovu eksperimenata napravili perceptualnu skalu frekvencija na kojoj jednake distance između frekvencija takođe zvuče slušaocima jedanko distancirano [9]. Na slici 8 je dat odnos između Melove skale i linearne skale u Herzima.



Slika 8 Odnos između Melove i Herz-ove skale [10]

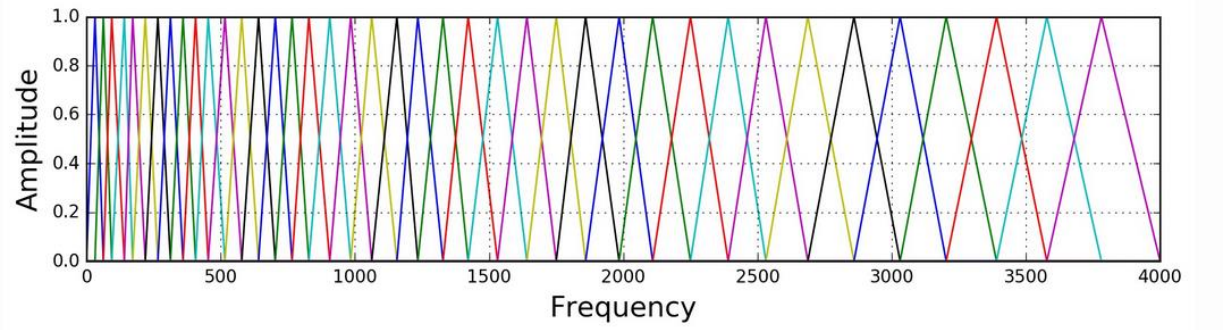
Sa slike 8 se vidi da frekvencije do 1KHz-a ljudsko uho može da prilično tačno razlikuje, dok za visoke frekvencije npr između 10 KHz i 11Khz, ljudsko uho ne pravi razliku. Konverzija vrednosti frekvencije iz Herz-ove (f) u Melovu skalu (m), obavlja se uz pomoć sledeće formule [11]:

$$(11) \qquad m = 2595 \log_{10}(1 + \frac{f}{700})$$

Obrnuta konverzija se obavlja preko formule 12:

$$(12) \qquad f = 700(10^{\frac{m}{2595}} - 1)$$

Utvrđeno je još i da se ljudsko uho ponaša kao niz filtera, iz razloga što detektuje samo frekvencijske komponente iz određenih opsega. S tim što su ovi filteri, u skladu sa Melovom skalom, nelinearno raspoređeni duž frekvencijske ose, tj mnogo ih više ima za niže, nego za više frekvencije. Sistemi za analizu i prepoznavanje govora ali i oni za analizu i klasifikaciju muzičkog sadržaja, često se oslanjaju i teže da imitiraju način na koji ljudsko uho obrađuje zvuk upravo primenom prethodno pomenutih frekvencijskih filtera. Kao i običan spektrogram i Mel spektrogram je vremensko – frekvencijski prikaz audio signala, s tim što se kod istog spektrograma vrednosti frekvencije skaliraju po Mel skali. Kod Mel spektrograma se, kao i kod običnih spektrograma, takođe vrši segmentiranje i računanje FFT-a za preklapajuće segmente. S tim što se kod Mel spektrograma, na rezultat FFT-a primenjuje niz Mel filtera, pri čemu su ovi filteri međusobno jednako udaljeni na Melovoj skali. Na slici 9 je dat prikaz Melovih trougaonih filtera. Sa slike se vidi da je više užih filtera koncentrisano u nižim frekvencijama, dok je manje širih filtera skoncentrisano u višim frekvencijama, što praktično imitira način funkcionisanja ljudskog sluha.



Slika 9 Melov filterbank koji se sastoji od 40 trougaonih filtera [12]

U koliko konkretan filter označimo sa H_m gde je m indeks tog filtera iz skupa (Melov koeficijent), ovaj filterbank možemo predstaviti matematički kao [12]:

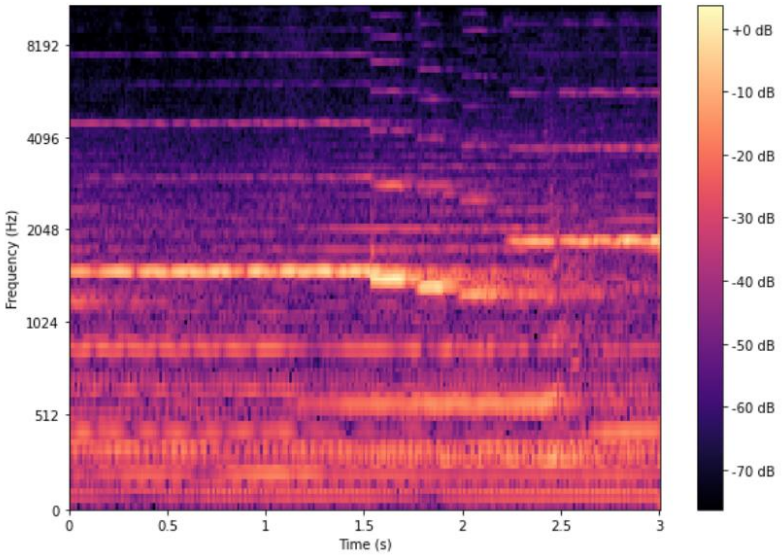
$$(13) \qquad H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{2(k-f(m-1))}{f(m)-f(m-1)}, & f(m-1) \le k \le f(m) \\ \frac{2(f(m+1)-k)}{f(m+1)-f(m)}, & f(m) < k \le f(m+1) \\ 0, & k > f(m+1) \end{cases}$$

Gde je *f* skup diskretnih frekvencija gde svaka vrednost predstavlja vrh jednog filtera, pri čemu je $m \in [1,\ M+1]$, gde je *M* ukupan broj filtera.

Melov spektar za odgovarjaući magnitudni spektar *X*[k] se dobija tako što se isti magnitudni spektar množi sa svakim od melovih filtera H_m [12]:

$$(14) \qquad s[m] = \sum_{k=0}^{N-1} |X[k]|^2 H_m[k] \qquad 0 \le m \le M-1$$

Kao rezultat ove operacije na kraju se dobija po jedna izračunata vrednost za svaki od filtera. Kao što se može primetiti, Melov spektrogram zahteva još jedan dodatni parametar u odnosu na standardni spektrogram, a to je sam broj filtera. Brojem filtera se frekvencijski opseg deli na međusobno ekvidistantne intervale po Melovoj skali, s tim što sada uz pomoć spektrograma možemo da posmatramo promenu energiju u ovim frekvencijskim intervalima, kroz vreme. Na slici 10 je dat prikaz Melovog spektroragrama za već spominjani audio snimak muzike violončela sa slike 7, svi korišćeni parametri su isti, s tim da je broj korišćenih Melovih filtera 128.

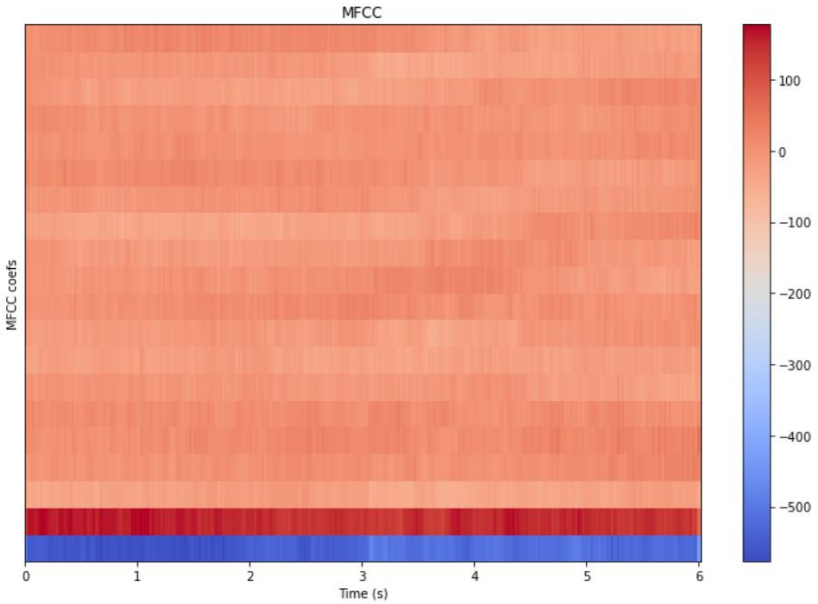


Slika 10 Mel spectrogram za audio isečak muzike violončela

Kod Mel spektrograma, usled preklapanja filtera, dolazi do pojave da imamo korelirajuće količinine energije u susednim Melovim koeficijentima. U koliko na vrednosti Melovog spektra za određeni segment, najpre primenimo logaritam, a zatim transformaciju pozantiju kao diskretna kosinusna transformacija (*Discrete cosine transform – DCT*) dobijamo set cepstralnih koeficijenata koji predstavljaju set međusobno nezavisnih vrednosti kojima se opisuje spektralna envelopa za taj konkretni segment [14]. Koeficijenti koji se dobijaju kao rezultati ove operacije se nazivaju Mel-frekvencijski cepstralni koeficijenti (*MFCC*), koji predstavlja najčešće korišćenu karakteristiku kada je reč o prepoznavanju glasa, ali sve više i u klasifikaciji muzičkog sadržaja.Sama transformacija Melovog magnitudnog spektra u MFCC se može opisati sledećom formulom [13]:

$$(15) \qquad c[n] = \sum_{m=0}^{M-1} \log_{10}(s[m]) \cos\left(\frac{\pi n(m-0.5)}{M}\right)$$

Gde su $c[n]$ Melovi cepstralni koeficijenti, pri čemu $n \in [0, C-1]$, gde je C ukupan broj Melovih cepstralnih koeficijenata koje računamo, $s[m]$ je Melov spektar, gde je m indeks konkretnog Melovog koeficijenta (filtera). Na slici 11 se može videti izračunati MFCC za već korišćeni isečak muzike violončela, pri čemu je računato 20 MFCC koeficijenata. Posmatrajući slike 8, 10 i 11, mogu se zapaziti sličnosti u formi između običnog i Melovog spektrograma za isti audio signal, dok MFCC ne pokazuje ovu sličnost iz razloga što predstavlja signal u jednom apstraktnijem domenu, preko spektralne envelope datog signala.



Slika 11 MFCC za audio isečak muzike violončela

3. Klasifikacija muzičkog sadržaja na osnovu predominantnog instrumenta

U procesu klasifikacije muzičkog sadržaja na osnovu instrumenta koji je predominantan u istom, eksperimentisano je sa karakteristikama koje se odnose na boju zvuka, opisanih u prethodnom poglavlju, a to su: običan energetski spektrogram, Melov spektrogram i MFCC. Dataset koji je korišćen u klasifikaciji je IRMAS biblioteka, koja se sastoji od muzičkih isečaka u trajanju od 3s proizvedenih od različitih instrumenata, pri čemu isečci mogu pripadati različitim muzičkim žanrovima. Pre pribavljanja odgovarajućih karaktereistika iz ovih isečaka, potrebno je iste preprocesirati. U ovu fazu spada proces downsampling-a audio fajlova sa inicijalnih 44.1 KHz na 22.05KHz, kao i konvertovanje stereo u mono audio signal uzimanjem prosecne vrednosti za oba kanala. Ovo preprocesiranje je potrebno da bi se standardizovao input za klasifikator i kako bi se poboljšale performanse klasifikacije. Sam audio input se segmentira na segmente od 1s za koje se potom računaju spektrogrami, iz razloga poboljšanja rezultata klasifikacije, kao i zbog obezbeđivanja većeg broja primeraka za klasifikaciju. Iz razloga što sam dataset nije izbalansiran kada je reč o broju primeraka određenih instrumenata, ubačen je i korak augmentacije audio sadržaja koji za cilj ima da primeni određeni niz transformacija (promena visine tona, vremensko ubrzavanje ili skraćivanje i dr.) nad originalnim muzičkim isečkom, pri čemu informacije o predominantnom instrumentu i dalje ostaju iste, a sam dataset se proširuje. Za klasifikaciju je korišćena konvoluciona neuronska mreža (*Convolutional neural network – CNN*) kojoj se na ulaz dovode odgovarajući spektrogrami, koje ovaj klasifikator analizira primenom svojih konvolucionih i *max pooling* filtera. Kao rezultat ovaj klasifikator daje niz realnih vrednosti pri čemu svaka predstavlja verovatnoću da se dati instrument javlja u konkretnom muzičkom isečku. Nakon procesa učenja vrši se testiranje ovog sistema i to za muzičke isečke koje sadrže samo jedan predominanti instrument, a zatim i na polifone muzičke isečke različite dužine trajanja. U nastavku ovog poglavlja će više reči biti o samom datasetu, audio augmentaciji, samom procesu klasifikacije kao i konfiguraciji klasifikatora. Takođe će biti data i analiza rezultata klsifikacije.

3.1 Dataset

IRMAS[1] (Instrument Recognition in Music Audio Signals) dataset je korišćen kao materijal na kome bi se procenile performanse predloženog sistema za klasifikaciju. Sam materijal je, na osnovu autora dataseta, podeljen na onaj za učenje i onaj za testiranje. Materijal za učenje se sastoji od .wav fajlova gde je svaki trajanja 3s i gde se u svakom fajlu javlja jedan od sledećih instrumenata: violončelo, klarinet, flauta, akustična gitara, električna gitara, orgulje, klavir, saksofon, truba ili violina. U okviru podele fajlova na osnovu prisutnih instrumenata, fajlovi se mogu klasifikovati i na osnovu muzičkog žanra kome pripadaju kao što su: klasika, pop-rok,

[1] Link za IRMAS Dataset: https://www.upf.edu/web/mtg/irmas

country, latino. Kroz tabelu na slici 12 dat je prikaz distribucije fajlova u datasetu na osnovu instrumenata koji je predmominantan u njima.

Instrument	Skraćenica	Broj primeraka - trening	Broj primeraka - testiranje
Violončelo	cel	388	111
Klarinet	cla	505	62
Flauta	flu	451	163
Akustična gitara	acg	637	535
Električna gitara	gel	760	942
Orgulje	org	682	361
Klavir	pia	721	995
Saksofon	sax	626	326
Truba	tru	577	167
Violina	vio	580	211

Slika 12 IRMAS dataset - instrumenti koji su korišćeni u eksperimentu

IRMAS dataset obezbeđuje 5927 muzičkih isečaka fiksne dužine od 3s u kojima se javlja samo jedan predominantan instrument. U ovom radu, 80% ovog skupa se koristi za treniranje sistema (konvolucione neuronske mreže), dok se ostalih 20% koristi za testiranje performasni sistema kada je reč samo o klasifikaciji monofonih audio sadržaja. Takođe, IRMAS obezbeđuje 1830 primeraka za testiranje, pri čemu fajlovi koji pripadaju ovom skupu nisu fiksne dužine (mogu biti između 5 i 20s). Ovi fajlovi, koje je autor dataseta označio kao test podatke, su polifoni (u njima je prisutno 1 ili više dominatntnih instrumenata) i u ovom radu se koriste za testiranje sistema kada je reč o klasifikaciji polifonih audio sadržaja.

Na slici 12 se može videti da dataset nije u potpunosti izbalansiran, odnosno da za neke instrumente imamo veći broj primeraka nego za druge. Ovaj problem je moguće razrešiti na više načina kao što su: undersempling – gde za svaki instrument uzimamo onoliki broj primeraka, koliko ima instrument sa najmanje primeraka (u našem slučaju bi to bilo 388), ili oversempling gde je potrebno koristiti kopije da bi imali isti broj primeraka za svaki instrument. Problem prvog pristupa je taj, što odbacujemo veliki broj primeraka na kojima bi sistem moga da uči, dok je problem oversemplinga taj, što je zbog kopija, sistem isuviše dobro naučen da prepoznaje određene primerke, dok na novim primercima, prilikom testiranja, sistem pokazuje slabije rezultate.

U ovom radu rešenje za ovaj problem neizbalansiranosti je potraženo u tzv augmentaciji muzičkih fajlova. Ovo je ostvareno tako što se nad konkretnim muzičkim fajlom prilikom faze učenja sistema, nasumično primenjuje jedna od transforamcija kao što su: dodavanje šumova, povećanje ili smanjenje visine tonova, vremensko razvlačenje ili sažimanje audio fajla, ili kombinacija prethodna dva. Korišćenjem ovakve augmentacije se sam deo dataseta za učenje dinamički proširuje.

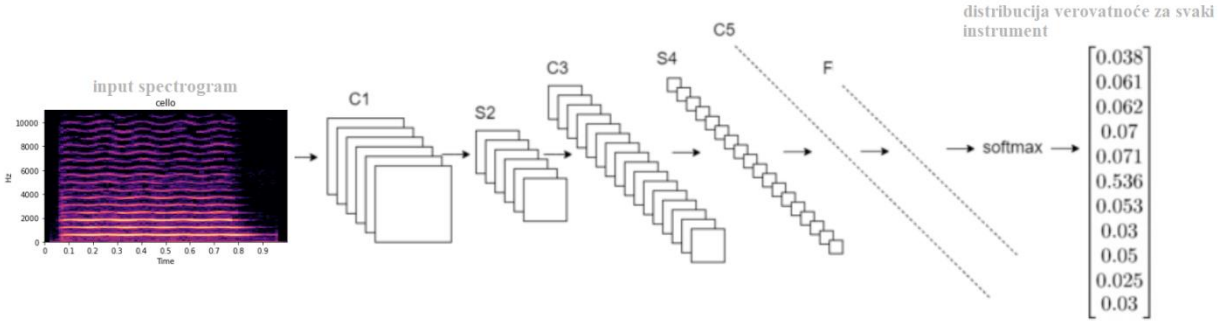
Sa aspekta performansi klasifikatora, u procesu učenja sistema, najbolji rezultati su ostvareni u koliko se muzički isečak podeli na segmente dužine 1s. Pri čemu se spektrogram računa za svaki od ovih isečaka posebno. Ovo takođe utiče na povećanje broja primeraka za učenje. Kod testiranja polifonih muzičkih primeraka, isti se takođe deli na segmente trajanja 1s, pri čemu

se, za razliku od faze učenja gde se segmenti ne preklapaju, vrši preklapanje segmenata za 0.5s. Na kraju testiranja se vrši agregacija rezultata dobijenih za svaki pojedinačni segment.

3.2 Arhitektura klasifikatora

Konvolucione neuronske mreže (CNN) su *deep learning* mreže dizajnirane po ugledu na ljudski vizuelni korteks, koje služe za klasifikaciju slika u kompujeterskoj viziji i za prepoznavanje objekata. Kod ovakve vrste neuronskih mreža, svaki nivo (*layer*) ima 3D strukturu, pri čemu dimenzija odgovaraju broju karakteristika. U input nivou, ove karaktersitike su RGB kanali slike, dok se u skrivenim nivoima, karakteristike odnose na mape koje služe za pronalaženje određenih oblika na slici. Konvoluciona mreža se može posmatrati kao niz naslaganih skrivenih nivoa (*hidden layers*) pri čemu se svaki sastoji od: konvolucionih nivoa, koje prati aktivaciona funckija (dodaje nelinearnost u neuronsku mrežu), nakon čega mogu ići *pooling* nivoi. Uloga konvulicionih nivoa je da uz pomoć različitih filtera detektuje određene oblike na slici, pri čemu početni konvolucioni nivoi uočavaju primitivnije oblike kao što su npr horizontalne ili vertikalne linije, dok kasniji nivoi uočavaju kompleksnije oblike. Uloga *pooling* nivoa je da kompresuje sliku koja je nastala kao rezultat primene konvolucije, kako bi se ekstraktovali samo dominantni elementi i smanjila komputaciona moć potrebna za dalju obradu [15].

U ovom radu, osnovna ideja je da se na ulaz jedne duboke (sa više nivoa) konvolucione mreže dovode spektrogrami konkretnih muzičkih segmenata, kako bi mreža detektovala one oblike na spektrogramu koji su zajednički za sve muzičke komade u kojima je predominantan konkretan muzički instrument. Na ovo rešenje se još može gledati kao na pokušaj da se upotrebi CNN kao standardna tehnika u klasifikaciji slika, na vizuelnu reprezantaciju zvuka tj na spektrogram. Na slici 13 je predstavljen predloženi sistem za klasifikaciju. Na slici su sa C označeni konvolucioni, dok su sa S označeni max pooling nivoi.



Slika 13 Predloženi system za klasifikaciju

Za preprocesiranje i ekstrakciju spektrograma korišćena je Python biblioteka Librosa, dok su za samu implementaciju klasifikatora koriščene bibliotke Tensorflow i Keras. Kada je reč o konfigurisanju konvolucione neuronske mreže, eksperimentisano je sa različitim brojem nivoa, kao i različitim brojem koriščenih filtera u nivoima, koji su korišćeni u raznim radovima na ovu temu ([16] [17]), pri čemu je najbolji rezultat dala konfiguracija data na slici 14. U ovoj konvolucionoj mreži imamo 4 ciklusa od dva vezana konvoluciona nivoa, koji prati jedan *max pool* nivo, pri čemu se u svakom ciklusu broj konvolucionih filtera povećava 2 puta. Takođe,

da bi se izbegao čest problem *overfiting*-a, koristi se dropout tehnika, kod koje imamo odbacivanje nekih čvorova u toku procesa učenja. Kao aktivaciona funkcija, nakon svakog konvolucionog nivoa, korišćena je *Leaky ReLU* funkcija, koja je pokazala bolje rezultate nego klasična *ReLU* funkcija

Dimenzije inputa	Opis
1 x 257 x 87	Spektrogram
32 x 257 x 87	3 x 3 konvolucija, 32 filtera
32 x 257 x 87	3 x 3 konvolucija, 32 filtera
32 x 257 x 87	3 x 3 Max pooling
32 x 86 x 29	Dropout (0,25)
64 x 86 x 29	3 x 3 konvolucija, 64filtera
64 x 86 x 29	3 x 3 konvolucija, 64filtera
64 x 86 x 29	3 x 3 Max pooling
64 x 29 x 10	Dropout (0,25)
128 x 29 x 10	3 x 3 konvolucija, 128 filtera
128 x 29 x 10	3 x 3 konvolucija, 128 filtera
128 x 29 x 10	3 x 3 Max pooling
128 x 10 x 4	Dropout (0,25)
256 x 10 x 4	3 x 3 konvolucija, 256 filtera
256 x 10 x 4	3 x 3 konvolucija, 256 filtera
256 x 10 x 4	3 x 3 Max pooling
256 x 4 x 2	Dropout (0,25)
256 x 4 x 2	Global Max Pooling
256	Izravnan i potpuno povezan nivo
1024	Potpuno povezan nivo
1024	Dropout (0,33)
1024	Softmax nivo – output: 10

Slika 14 Konfiguracija korišćene konvolucione neuronske mreže

Na slici 14 se vidi da su dimenzije spektrograma 257 x 87, pri čemu je visina spektrograma 257 posledica toga da je parametar širina segmenta, kod računanja STFT-a, 512 semplova (pola od ovoga je koeficijent Nikvistove frekvencije). Širina spektrograma je posledica toga da su muzički isečci u trajanju 1s, pri čemu je učestanost semplovanja 22050 Hz, a parametar *hop size* je pola segmenta, odnosno 256 semplova. U slučaju da se za klasifikaciju koristi Mel Spektrogram, dimenzije ulaza bi bile (1 x 128 x87), s obzirom da je broj korišćenih Melovih spektrograma 128. Ako se na ulaz dovodi MFCC, dimenzije ulaza bi bile (1 x 20 x 87). Važno je naglasiti, da manje dimenzije podrazumevaju brži proces učenja mreže. Kao aktivaciona funkcija poslednjeg nivoa, korišćena je *sigmoid* funkcija, koja pogoduje kod tzv. *multilabel* klasifikacije koju imamo kod polifonih muzičkih sadržaja, pri čemu klasifikator ne gubi kod performasni kod klasifikacije monofonih muzičkih sadržaja, iz ovog razloga ovaj zadnji nivo ima 10 realnih vrednosti u rasponu od 0 do 1.

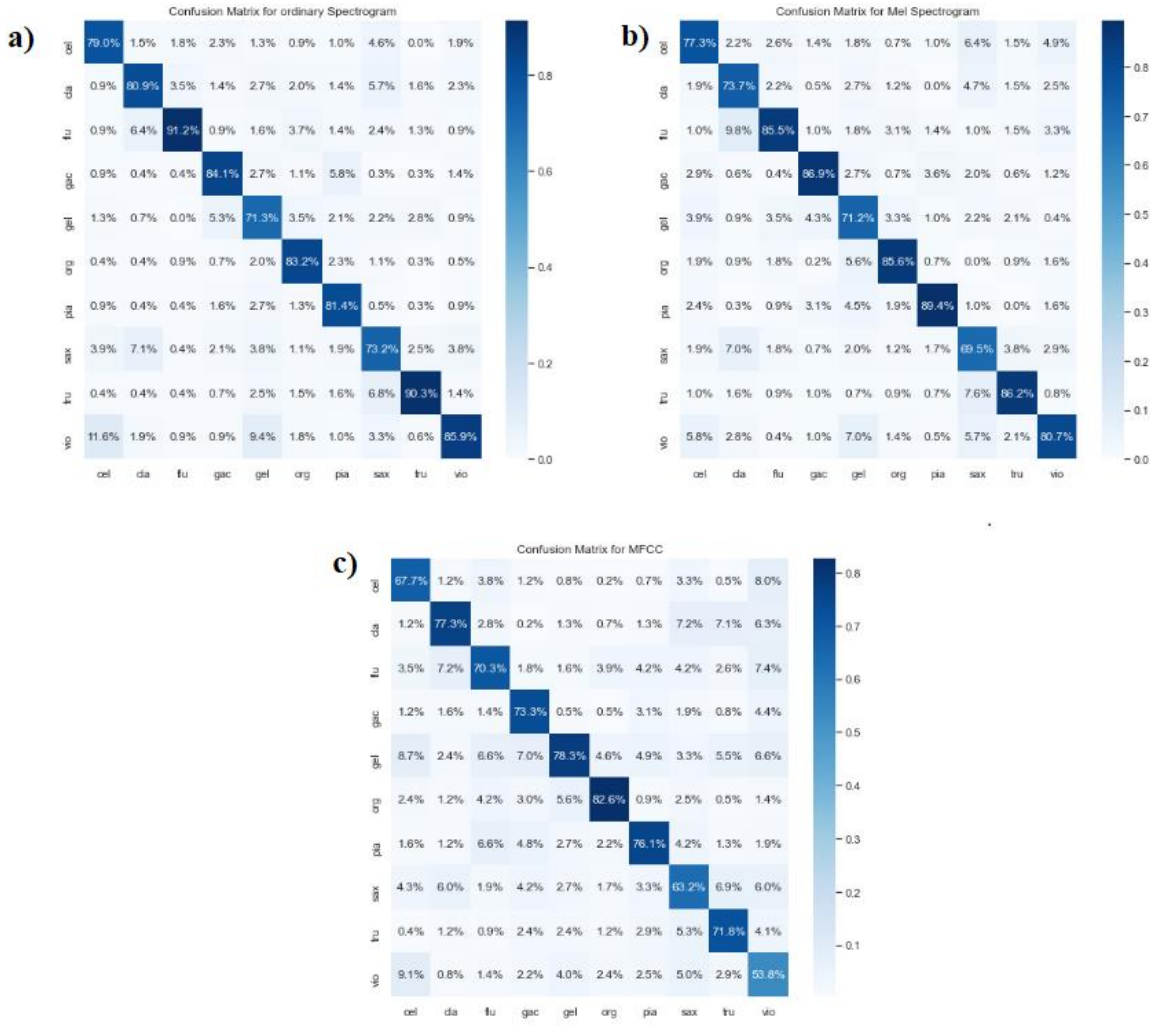
3.3 Analiza rezultata eksperimenta

U koliko se ograničimo na analizu klasifikacije monofonog muzičkog sadržaja uz pomoć klasifikatora prikazanog na slici 13, najbolji rezultat je postignut u koliko se za klasifikaciju koriste regularni spektrogrami. Korišćenjem spektrograma, dobijenog uz pomoć STFT-a, sistem je uspeo da postigne ukupnu tačnost od 86%. Na slici 15a je dat prikaz matrice zabune za klasifikaciju na osnovu običnih energetskih spektrograma. Kod ovih matrica u redovima su označeni primerci koji pripadaju konkretnoj klasi instrumenata, dok kolone označavaju kako je

sitem klasifikovao primerke koji pripadaju konretnoj klasi (označene redom tabele). Eksperimentisano je sa više vrednosti STFT parametara, pri čemu je sledeća kombinacija dala najbolji rezultat:

- Dužina segmenta (u semplovima): 512
- Dužina preksoka – *hop length* (u semplovima): 256
- Tip prozorske funkcije: *Hann*

Ono što se može videti na slici 15 (na sve 3 matrice zabune), je da se najveća greška u klasifikaciji javlja između instrumenata koji pripadaju istoj porodici, npr. dešava se da je violončelo pogrešno procenjeno kao violina, klarinet kao flauta ili saksofon, truba kao saksofon, gitara kao električna gitara itd. Iz ovoga se zaključuje da instrumenti iz iste porodice instrumenta (žičani, duvački, instrumenti sa tipkama) imaju određene zajedničke karakteristike koje se odnose na boju zvuka, zbog kojih klasifikator, kao i ljudsko uho, u nekim slučajevima ne može da napravi jasnu razliku između njih.



Slika 15 Matrice zabune za klasifikatore koji koriste: a) Energetski spektrogram 2) Mel Spektrogram 3)MFCC-a

Kroz sliku 15b prikazane su perfomanse klasifikatora koji koristi Mel Spektrogram kao ulaz, pri čijem računanju su korišćeni isti parametri kao kod običnog energetskog spektrograma, s tim da je odabrani broj Melovih filetra 128. Tačnost ovakvog klasifikatora je 85%, što ukazuje na to da nema nekih relevantnih gubitaka u odnosu na običan energetski spektrogram, odnosno da se korišćenjem Melovih filtera, osnovni spektrogram može kompresovati pri čemu se ne gube bitne informacije (locirane u nižim frekvencijama) koje se odnose na boju zvuka, uz pomoć kojih se vrši klasifikacija muzičkog sadržaja na osnovu predominantnih instrumenata.

Kada je reč o klasifikaciji korišćenjem MFCC-a, prikazanoj kroz matricu zabune na slici 15c, sistem ovde pokazuje nešto slabije rezultate, što se vidi iz podatka da je ukupna tačnost, klasifikatora sa ovim ulazom 72%. Broj MFCC koeficijenata koji je korišćen u ovom radu je 21, koji se inače često koristi kod sistema za raspoznavanje govora koji koriste MFCC. Može se zaključiti da, iako MFCC pokazuje dobre rezultate i nezaobilazan je faktor u prepoznavanju govora, ova metoda sadrži manje informacija o boji zvuka (verovatno izgubljenih u transforamcijama dekorelacije) u odnosu na Melov i običan energetski spektrogram iz kojih se sam MFCC dobija.

Kod polifone klasifikacije, s obzirom da se IRMAS dataset sastoji od materijala različitog trajanja, svaki primerak je podeljen na preklapajuće segmente u trajanju od 1s (0.5s je preklapanje) na osnovu kojih se kreiraju spektrogrami koji se upotrebljavaju u klasifikaciji. Problem koji se dalje nameće je, na koji način izvršiti agregaciju rezultata dobijenih za ove segmente. Eksperimentisano je sa sledećim vrstama agregacija predloženih u radu [17]: uzimanjem prosečne vrednosti za sve segmente, za svaki od klasa (instrumenata), druga metoda je sumiranje svih dobijenih rezultata po klasama, za sve segmente, nakon cega se svaka od ovih dobijenih 10 vrednosti deli sa najvecom dobijenom sumom. Takođe, pošto je aktivaciona funkcija zadnjeg nivoa CNN-a *sigmoid*, rezultat svake klasifikacije su 10 brojeva, svaki između 0 i 1. Ovo znači da je potrebno ustanoviti konkretan prag tj granicnu vrednost na osnovu koje bi odredili da li je instrument zastupljen u nekom muzičkom komadu ili nije. U slučaju monofone klasifikacije, potrebe za ovim ne bi bilo, jer bi se uzimao indeks (instrument) koji ima najveću vrednost. U ovom radu je eksperimentisano sa obe agregacione metode i sa više različitih vrednosti praga, pri čemu se pokazalo da druga alternativa daje malo bolje rezultate u odnosu na alternativu 1, pri čemu je korišćen prag koji iznosi 0,45 .

Još jedan problem koji je morao biti rešen kod klasifikacije polifonih instrumenata jeste koju metriku koristiti za ocenu perfomansi klasifikatora. Kod tzv. *multilabel* klasifikacije, kakva je klasifikacija polifonih muzičkih komada, predikcije se ne mogu striktno podeliti na samo tačne i netačne, već mogu biti i delimično tačne (npr u koliko je sistem pogodio 2 od 3 instrumenata koji se nalaze u nekom komadu, to je bolje nego da nijedan nije pogodjen). Ideja je da se iskoriste postojeće metrike za najprostiju, binarnu klasifikaciju, za svaku klasu (u ovom slučaju instrument) nakon čega bi se računao prosek nad svim klasama. U ove elementarne metrike spadaju:

- Preciznost – sposobnost klasifikatora da negativne primerke ne klasifikuje kao pozitivne. Odnosno, preciznost govori koliko je selektovanih elemenata (procenjenih da pripadaju određenoj klasi) relevanto (stvarno pripadaju toj klasi).

- Opoziv (*Recall*) – sposobnost klasifikatora da pronađe sve pozitivne primerke, tj. koliko je relevantnih elemenata selektovano.
- *F1 Score* – predstavlja ponderisanu harmonijsku sredinu za parametre preciznost i opoziv. Vrednost F1 na najefikasniji način, kroz jednu vrednost, kombinuje preciznost i opoziv.

Prethodno pominjane metrike se mogu računati na:

- Makro nivou – najpre se sračunavaju metrike za svaku od klasa, nakon čega se uzima aritmetička sredina nad svim klasama. Ovakav način računanja nije pogodan kada imamo neizbalansirane klase (velike razlike u broju primeraka po klasama), iz razloga što jednako tretira sve klase.
- Mikro nivou – najpre se vrši agregacija činioca na nivou svih klasa, nakon čega se računa ukupna metrik, na ovaj način u računanju metrike svi primerci imaju jednakog udela, za rayliku od makro nivoa gde sve klase imaju jednakoog udela. Ovaj pristup je povoljniji za nebalansirane klase kao i za *multilabel* klasifikaciju, gde se najčešće i koristi.

U ovom radu se kao glavna metrika za merenje performani sistema kada je reč o polifonom sadržaju koristi mikro *F1 Score*. U koliko sa y i $\hat{y}$ označimo skup predviđenih i stvarnih oznaka respektivno. Mikro preciznost, *Recall* i *F1 Score* se definišu kroz sledeće matematičke izraze respektivno:

$$P(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|y|} \quad (16)$$

$$R(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|\hat{y}|} \quad (17)$$

$$F_1(y,\hat{y}) = \frac{2P(y,\hat{y})R(y,\hat{y})}{P(y,\hat{y})+R(y,\hat{y})} \quad (18)$$

Na slici 16 su priakzani rezultati performasni sitema nakon testiranja nad IRMAS polifonim muzičkim primercima. Može se videti da sistem pokazuje bolje perfomanse kod klasifikacije monfonih u odnosu na polifone muzičke sadržaje. Jedan od razloga je i to što je klasifikator treniran nad monofonim sadržajem, što mu omugućuje da uspešno prepozna najdominantniji instrument u polifonom sadržaju, ali manje uspešno prepoznaje prateće instrumente. Još jedan od razloga manje uspešne klasifikacije, u odnosu na monofoni sadržaj, je sam IRMAS test dataset, u kome imamo neke instrumente koji su zastupljeni isključivo kao prateći instrumenti (violončelo i klarinet) i koje klasifikator teže prepoznaje, takođe važi i suprotno, da klasifikator dobro prepoznaje one instrumente koji su u datasetu prisutni većinom kao glavni instrumenti (obična i električna) gitara. Najbolji rezultat pokazuju klasifikacije uz pomoć običnog i Melovog spektrograma (*F1 Score = 0.6*), dok MFCC, kao i u slučaju monofonog sadržaja, pokazuju nešto slabije performanse (*F1 Score = 0.55*).

a)

	precision	recall	f1-score	support
cel	0.24	0.51	0.33	74
cla	0.16	0.22	0.19	59
flu	0.52	0.52	0.52	153
gac	0.69	0.71	0.70	359
gel	0.68	0.62	0.65	529
org	0.47	0.47	0.47	229
pia	0.79	0.46	0.58	827
sax	0.57	0.89	0.70	317
tru	0.55	0.55	0.55	159
vio	0.56	0.80	0.66	174
micro avg	0.60	0.59	0.60	2880

b)

	precision	recall	f1-score	support
cel	0.25	0.46	0.32	74
cla	0.10	0.29	0.15	59
flu	0.49	0.58	0.53	153
gac	0.77	0.69	0.73	359
gel	0.70	0.64	0.67	529
org	0.60	0.44	0.50	229
pia	0.81	0.45	0.58	827
sax	0.54	0.92	0.68	317
tru	0.42	0.58	0.49	159
vio	0.53	0.78	0.63	174
micro avg	0.58	0.60	0.59	2880

c)

	precision	recall	f1-score	support
cel	0.18	0.58	0.28	74
cla	0.07	0.10	0.08	59
flu	0.47	0.57	0.51	153
gac	0.54	0.66	0.60	359
gel	0.74	0.55	0.63	529
org	0.39	0.34	0.36	229
pia	0.79	0.57	0.66	827
sax	0.50	0.85	0.63	317
tru	0.36	0.67	0.47	159
vio	0.34	0.79	0.48	174
micro avg	0.51	0.60	0.55	2880

Slika 16 Rezultati polifone klasifikacije uz pomoć: a) Energetskog spektrograma 2) Mel Spektrograma 3)MFCC-a

4. Zaključak

U ovom radu je pokušana klasifikacija muzičkog sadržaja na osnovu predominantnog instrumenta koji se javlja u istom pri čemu je za klasifikaciju upotrebljena vizuelizacija audio signala tj. spektrogram. Ova karakteristika predstavlja promenu energetske zastupljenosti frekvencija od kojih se sastoji muzički signal u vremenu. U radu je eksperimentisano sa običnim energetskim, kao i sa Melovim spektrogramom koji se dobija nakon primena nelinearnih transformacija (Melovi filtri) nad frekvencijskim spektrom signala, što podržava način na koje ljudsko uho percipira različite frekvencije. Eksperiment je pokazao da su obe ove karakteristike jednako uspešne kod klasifikacije instrumenata. Takođe je eksperimentisano i sa MFCC karakteristikom audio signala, koja se često koristi u prepoznavanju govora. Ova karakteristika pokazuje nešto slabije karakteristike u odnosu na spektrograme, iz razloga što ne sadrži u sebi potrebne informacije koje se tiču boje zvuka, na osnovu kojih se pravi razlika između instrumenata. Kao klasifikator korišćena je konvoluciona neuronska mreža, koji je jedan od najefikasnijih alata za klasifikaciju slika u kompjuterskoj viziji, kao i za prepoznavanje vizuelnih objekata. Klasifikator je isproban na IRMAS Datasetu koji obezbeđuje monofone muzičke materijale za treniranje klasifikatora, kao i znatan broj polifonih sadržaja za testiranje. Najbolji rezultat klasifikator je pokazao kod testiranja monofonog sadržaja i to uz pomoć običnog spektrograma, tačnost ove klasifikacije je 86%. Kada je reč o klasifikaciji polifonog muzičkog sadržaja, sistem je postigao najbolji rezultat takođe korišćenjem običnog spektrograma, gde je dobijeni F1 Score 0.6.

Ostavljeno je dosta prostora za dalje poboljšanje klasifikacije. Za ovako nešto, moglo bi se pokušati sa kombinovanjem više vrsta klasifikacija, gde bi npr. rezultat klasifikacije CNN-a opisan u ovom radu, zajedno sa nekim drugim karakteristikama, bio ulaz u novi klasifikator (npr obična neuronska mreža). Takođe, mogle bi se poboljšati perfomasne klasifikacije polifonih sadržaja na način što bi se mreži obezbedio polifoni sadržaj za fazu treninaranja. Ovo bi bilo moguće pronalaženjem novog, polifonog, dataseta, ili pokušajem augmentacije korišćenog IRMAS dataseta, tj kombinovanjem postojećeg monofonog, da bi se dobio polifoni sadržaj.

5. Literatura

[1] Markus Schedl, Emilia Gómez, and Julián Urbano. 2014. Music Information Retrieval: Recent Developments and Applications.

[2] https://scholarworks.montana.edu/xmlui/bitstream/handle/1/12732/DonnellyP1215.pdf?sequence=4&isAllowed=y

[3] http://www.thefouriertransform.com/

[4] https://medium.com/sho-jp/fourier-transform-101-part-3-fourier-transform-6def0bd2ca9b

[5] Müller, Meinard. (2015). The Fourier Transform in a Nutshell.

[6] https://www.dsp.etfbl.net/multimediji/2017/08a_audio_spektralna_analiza.pdf

[7] https://en.wikipedia.org/wiki/Window_function

[8] https://www.audiolabs-erlangen.de/content/05-fau/professor/00-mueller/02-teaching/2016s_apl/LabCourse_STFT.pdf

[9] https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53

[10] https://www.researchgate.net/figure/The-mel-scale-used-to-map-the-linear-frequency-scale-to-a-logarithmic-one_fig3_259479391

[11] http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[12] https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

[13] https://link.springer.com/content/pdf/bbm%3A978-3-319-03116-3%2F1.pdf

[14] https://wiki.aalto.fi/display/ITSP/Cepstrum+and+MFCC

[15] Aggarwal, C. C. (2018), *Neural Networks and Deep Learning* , Springer , Cham .

[16] Solanki, Arun & Pandey, Sachin. (2019). Music instrument recognition using deep convolutional neural networks. International Journal of Information Technology. 10.1007/s41870-019-00285-y.

[17] Yoonchang Han, Jaehun Kim, Kyogu Lee, Yoonchang Han, Jaehun Kim, and Kyogu Lee. 2017. Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music.