Nemanja Milanović

Ekstrakcija melodije iz polifonih zvučnih izvora

U ovom radu istraživana je ekstrakcija melodije iz polifonih zvučnih izvora. To je sprovedeno pomoću analize spektrograma dobijenih Furijeovom transformacijom. Podaci dobijeni iz spektrograma se zatim obrađuju sa ciljem uklanjanja grešaka pri detekciji nota. Rezultati su pokazali da je zadovoljavajuća ekstrakcija melodije moguća kada se za ulaz koriste jednostavniji uzorci sintetičke muzike. Tačnost rezultata kod klavirske muzike je zadovoljavajuća, ali niža u odnosu na sintetičke uzorke. Za komercijalnu muziku, rezultati nisu bili zadovoljavajući.

Uvod

Proces ekstrakcije melodije kao svoj cilj ima dobavljanje niza nota iz polifonog signala koje predstavljaju osnovnu melodiju. Polifoni signal je vrsta signala koja ima više melodijskih linija, harmonija, šumova i drugih frekvencija. Nakon obrade, generiše se monofoni audio signal koji oponaša ono što bi u tradicionalnom notnom zapisu bilo označeno kao osnovna melodija. Glavne primene ovog postupka jesu generisanje notnog zapisa na osnovu pesme, automatska klasifikacija žanrova muzike i Query by humming sistemi. U jednom radu poređeno je više različitih metoda rešavanja ovog problema, od strane različitih autora, i svi su imali tačnost od oko 50% do 70% (Salamon 2008). Međutim, još uvek nije pronađeno rešenje za ovaj problem koje daje rezultate koji su dovoljno pouzdani da bi bili primenljivi. Ekstrakcija melodije se ostvaruje u nekoliko faza. Ove faze zavise od pristupa rešavanju, ali uopšteno rečeno, postupak se sastoji iz pripreme signala, primene Furijeove transformacije i samog algoritma koji obrađuje podatke dobijene Furijeovom transformacijom. U ovom radu, prva faza je priprema izvornog signala primenom različitih filtera. Potom se signal pretvara u spektrogram (grafik koji prikazuje intenzitet date frekvencije u vremenu) pomoću Furijeove transformacije, a zatim se na dobijenom spektrogramu vrši analiza i pronalaženje konture melodije, koja se na kraju dodatno obrađuje radi uklanjanja različitih šumova i smetnji. Krajnji produkt ovog procesa jeste MIDI fajl koji sadrži zapis osnovne melodije koji se može reprodukovati.

Metod

Za implementaciju osmišljenog rešenja korišćen je jezik C++ sa JUCE API-em (ROLI 2018) koji pruža komponente za obradu zvuka kao i GUI elemente. Finalni rezultati dati su u formi MIDI fajlova.

Priprema signala. Obrađuje se digitalni signal dobijen uzorkovanjem muzike koja može biti analogne (muzički instrumenti) i digitalne prirode (digitalni sintisajzeri). Signal je zapisan u nekom od standardnih formata za zapis zvuka (mp3, wav). Da bi se uklonili šumovi koji bi mogli da smetaju ekstrakciji melodije, pre Furijeove transformacije nad signalom se vrši filtriranje različitim vrstama filtera sa beskonačnim impulsnim karakteristikama (infinite impulse response). Korišćena su tri filtera: high pass, low pass i peak filter. High pass filterom omogućava se uklanjanje niskofrekventnih šumova poput onih koji potiču od bubnjeva ili pojedinih nižih harmonija instrumenata poput klavira. Low pass filterom uklanjaju se visoke frekvencije na kojima se uglavnom nalaze intenzivni šumovi i visoki harmonici melodije koji nisu bitni, i prave smetnje pri daljoj kalkulaciji jer često mogu da se

Nemanja Milanović (2001), Petrovac na Mlavi, Srpskih vladara 56, 12300, učenik 2. razreda Požarevačke gimnazije

MENTOR: Igor Šikuljak, student Fakulteta tehničkih nauka Univerziteta u Novom Sadu

poklope sa harmonicima melodije, dajući im nepotrebno pojačanje intenziteta. Peak filter ne smanjuje ni jednu frekvenciju, već pojačava frekvencije u određenom opsegu. U ovom slučaju to pojačavanje nije jakog intenziteta i ima ulogu u dodatnom isticanju frekvencija tamo gde bi se tražena melodija najverovatnije našla, što je uglavnom oko 600 Hz. Ovi filteri obrađuju izvorni zvuk uzorak po uzorak. Pojam muzičkog uzorka podrazumeva vremenski najmanji mogući zapisiv deo signala, koji se uglavnom zapisuje kao ceo broj ili broj sa decimalnim zarezom u intervalu od 0 do 1.

Furijeova transformacija. U ovom koraku se primenom brze Furijeove transformacije (fast Fourier transform, FFT) audio podaci pretvaraju u oblik spektrograma koji je pogodniji za dalju obradu. Zbog toga su odluke koje se donesu u implementaciji FFT-a ključne za dalju obradu podataka. Prva takva odluka jeste rezolucija FFT-a. Naime, svaka Furijeova transformacija izvršava se na određenom broju uzoraka koji mora biti oblika $2n$, gde je n prirodan broj. Tako se dobija $2n - 1$ frekvencija koje su ravnomerno raspoređene u opsegu od 0 Hz do 22 kHz. Što je veća vrednost rezolucije FFT-a, to je moguće preciznije određivanje frekvencija. Međutim, ovde se javlja problem: povećanjem ove vrednosti, povećava se i broj uzoraka koji se obrađuje. To znači da se gubi na vremenskoj preciznosti, jer se trajanje svakog bloka uzoraka drastično povećava. Drugim rečima, kada bi se kao ulaz koristio snimak na kome se note jako brzo menjaju, sa većom FFT rezolucijom došlo bi do mnogo grešaka pri njegovoj obradi. Ako se uzme u obzir da je sample rate (stopa, odnosno frekvencija uzorkovanja) većine audio fajlova 44.1 kHz (44100 uzoraka po sekundi), a FFT rezolucija 4096, tada jedan blok traje oko 0.093 s. To je približno jednako trajanju šesnaestine na 128 bpm (128 beats per minute, allegro), što je najčešće tempo moderne pop muzike. U suprotnom slučaju, ako je FFT rezolucija jako mala, vremensko trajanje blokova je manje, ali preciznost određivanja frekvencija takođe pada.

Ekstrakcija tona. Nakon obrade uzoraka FFT-om, iz svakog obrađenog bloka se odmah dobija melodijska kontura, tj. dominantna frekvencija u tom bloku. Pošto je rezultat FFT-a samo niz brojeva koji predstavljaju intenzitete određenih frekvencija, potrebno je prvo dovesti u međusobnu vezu indekse niza i frekvencije predstavljene u hercima, što je moguće uraditi sledećom formulom: $\nu = \frac{i \times \mathrm{SR}}{F_S}$, gde je ν frekvencija, i indeks u nizu rezultata, SR frekvencija uzorkovanja (sample rate), a F_S FFT rezolucija. Nakon toga traži se dominantna frekvencija koja je definisana kao maksimalna parcijalna suma harmonika frekvencija koje su iznad određenog intenziteta. Drugim rečima, za blok koji se obrađuje, uzima se maksimalni i minimalni intenzitet i na osnovu toga se stavlja relativna donja granica (u rasponu od 0 do 1). Samo frekvencije čiji je intenzitet iznad date donje granice ulaze u dalju obradu. Nakon odabira frekvencija sledi sumiranje harmonika. Harmonici tona se dobijaju na osnovu istoimenog matematičkog reda, gde talasna dužina n-tog harmonika odgovara n-tom članu reda – ako prvi harmonik ima frekvenciju od 100 Hz, drugi će imati frekvenciju od 200 Hz, treći 300 Hz i tako dalje:

$$\sum_{n=1}^{\infty} \frac{1}{n} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

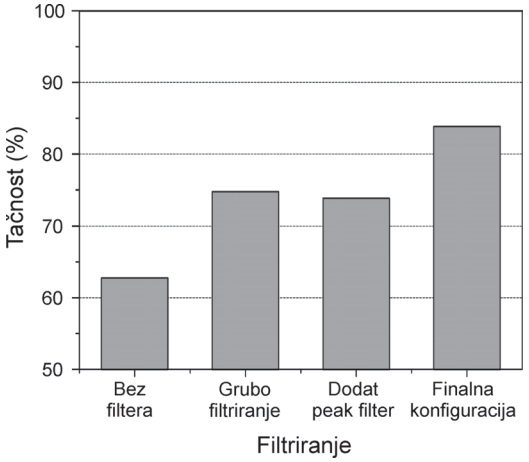
Ako se uzme u obzir formula koja povezuje indekse niza i frekvencije, zaključuje se da element na indeksu $n \cdot x$ ima n puta veću frekvenciju od elementa na indeksu x, što omogućava lako sumiranje harmonika na osnovu indeksa frekvencije. Nakon dobijenih suma za svaku odabranu frekvenciju nađe se frekvencija sa maksimalnom sumom, i ona se uzima kao dominantan ton. Značajna prepreka ovoj metodi jeste suzbijanje dinamike, ili kompresija muzike, koja je često prisutna u komercijalnoj muzici. Kompresija predstavlja suzbijanje dinamičkog opsega pesme (razlike između najtiše i najglasnije tačke u pesmi), a to se postiže pojačavanjem slabijih frekvencija i smanjivanjem onih koje su najistaknutije. Zatim se jačina celokupnog zvuka dodatno pojačava, i stvara se tzv. komercijalna glasnoća. Zbog ovog procesa, spektrogram postaje „zamućeniji" i visoke frekvencije na kojima se nalaze harmonici većine instrumenata se osetno pojačavaju, što znatno otežava ekstrakciju frekvencija. Proces kompresije nemoguće je efikasno i precizno obrnuti.

Završna obrada. Nakon ekstrakcije tonova note se stavljaju u niz, a osim tonalne vrednosti, čuva se i njihovo trajanje. Zatim se, brisanjem određenih tonova, vrši uklanjanje verovatnih grešaka. Ovaj postupak sadrži više faza. Prvo se vrši analiza celokupne melodije, pravi se histo-

gram nota, a na osnovu histograma vrši uklanjanje nota koje su verovatne greške. Prvi kriterijum za ukljanjanje jeste koliko se često nota javlja, prebrojano po FFT blokovima. Note koje se javljaju ređe od unapred određene granice se uklanjaju. Nakon toga se vrši korekcija oktava: kada je nota pre ili posle date note za oktavu viša od iste, spušta se za oktavu i pravi se kontinuirani ton. Ovakve greške dešavaju se zbog načina na koji harmonici funkcionišu. Pošto je prvi harmonik frekvencije duplo veće od osnovne, dobija se da je prvi harmonik oktava, koja po svojoj prirodi često ume da bude intenzivnija od osnovnog tona. Na kraju se vrši uklanjanje grešaka bazirano na trajanju nota. Note koje su veoma kratke tretiraju se kao šumovi ili smetnje i zanemaruju se. Nakon obrade, rezultati se čuvaju u formi MIDI fajla.

Rezultati i diskusija

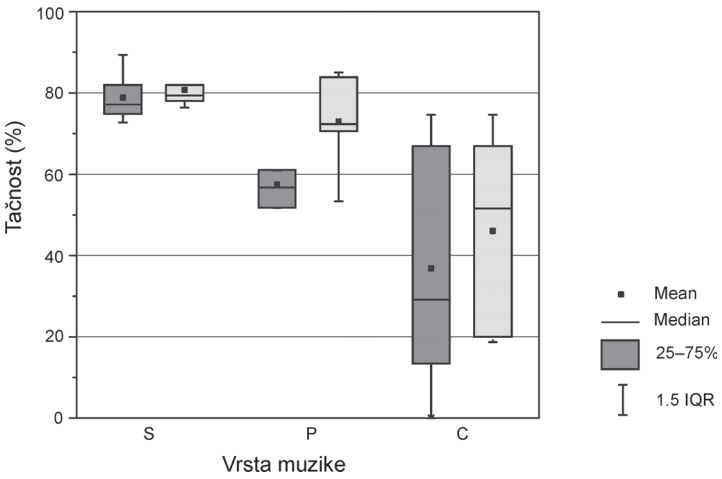
Zbog prirode grešaka, rezulati su razvrstani u dve grupe: na rezultate sa i bez oktavnih grešaka, jer oktavne greške predstavljaju posebnu kategoriju grešaka kod kojih se rezultati mogu donekle smatrati ispravnim u muzičkom smislu. Samo merenje zasnovano je na ispitivanju da li se u originalnom notnom zapisu dela i u dobijenom notnom zapisu (tj. MIDI fajlovima) u svakom MIDI ticku (najmanjoj jedinici vremena koju MIDI standard podržava) poklapaju note. Na osnovu toga, rezultat se prevodi u procente. Kao ulazni podaci programu su davani audio klipovi u trajanju od oko 15-20 sekundi. Filteri su konfigurisani na sledeći način: low pass: 4500 Hz, high pass: 509 Hz i peak filter: 590 Hz sa intenzitetom od 1.5. Ovakva konfiguracija filtera daje najbolje rezultate. Pri prvom merenju filteri nisu bili uključeni, pri drugom low pass i high pass filteri su podešeni na vrednosti bliske njihovim graničnim vrednostima, pri trećem je ubačen i peak filter i svi filteri su dovedeni bliže optimalnim vrednostima, a poslednje merenje predstavlja merenje sa optimalno podešenim filterima (slika 1). Treba imati na umu da efekat ovih filtera varira od primera do primera, ali ove vrednosti su nakon testiranja dobijene kao optimalne za većinu primera. Test sa grafika na slici 1 je rađen na sintetičkom izvoru. Na slici 2 prikazana je uspešnost algoritma u zavisnosti od vrste muzike.



Slika 1. Uticaj filtriranja na rezultate pri obradi sintetičkog izvora. Pri prvom merenju filteri nisu bili uključeni, pri drugom low pass i high pass filteri su podešeni na vrednosti bliske njihovim graničnim vrednostima, pri trećem je ubačen i peak filter i svi filteri su dovedeni bliže optimalnim vrednostima, a u poslednjem merenju filteri su optimalno podešeni.

Figure 1. The effect of filters on results when processing a synthetic source. In the first case filters were turned off, in the second case low pass and high pass filters were set to values close to their limits, in the third case the peak filter was added and all filters were set closer to optimal values, and in the last case all filters were optimally set up.

Rezultati su pokazali da kod sintetičkih izvora ima najmanje grešaka, da je kod klavirske muzike udeo grešaka veći, dok je kod komercijalne muzike broj grešaka najveći. Kod sintetičkih izvora ima manjih grešaka u detekciji vremena početka i kraja tona. U slučaju klavirske muzike ima znatno više grešaka i u detekciji samih tonova, a i u detekciji njihovih početaka i krajeva. Kod komercijalne muzike, detekcija tonova je loša, a počeci i krajevi nota u dosta slučajeva nisu ni približno tačni. U pojedinim primerima umesto melodije kao rezultat dobijen je šum koji program tumači kao nasumičnu notu. Za muzičara je u ovim rezultatima moguće pronaći korisnih informacija za dalju obradu pesme, kao što su pojedine pogođene note koje otkrivaju u kojoj skali je delo pisano. Ipak, ovakav nivo kvaliteta ne može se smatrati ni približno dovoljnim da bi se moglo reći da ostvaruju prvobitni cilj ovog istraživanja.



Slika 2. Rezultati testiranja u odnosu na vrstu muzike. S – uzorci sintetičke muzike, P – klavirske, C – komercijalne. Tamnijim nijansama označeni su rezultati sa oktavnim greškama, svetlijim – bez oktavnih grešaka.

Figure 2. Testing results with different types of music – the first set is synthetic samples (S), the second piano music (P), and the third is commercial music (C). The dark series represents results with octave errors, and the light series disregards the octave errors.

Razlog za ovako nizak stepen uspešnosti ekstrahovanja melodije kod komercijalne muzike može se objasniti prirodom te muzike. U slučaju klavirske muzike ili nekih jednostavnijih primera komercijalne muzike, note na spektrogramu su prilično jasno definisane. Kod komercijalne muzike ovo uglavnom nije slučaj zbog suzbijanja dinamike. Suzbijanjem dinamike dolazi do zamućenja spektrograma što znatno otežava ekstrakciju tonova. Povišenjem donje granice intenziteta iznad koje se tonovi obrađuju prilikom ekstrakcije tonova došlo je do zanemarljivog poboljšanja.

Zaključak

Iz dobijenih rezultata vidi se da, kada je u pitanju komercijalna muzika, preciznost nije ni blizu dovoljno visoka da bi rezultati bili verni stvarnoj osnovnoj melodiji. U pojedinim slučajevima, rezultati uopšte nemaju sličnosti sa pesmom iz koje su generisani. Glavni problem (osim kompresije komercijalnih pesama) jeste to što muzika dolazi u dosta različitih formi koje je prilično teško generalizovati i iz njih izvući opštu definiciju pojma melodije. Različitim poboljšanjima, kao što su odgovarajuće metode filtriranja, uključivanje Melove skale (Rashidul *et al*. 2004), moguće je dobiti bolje rezultate, međutim, za sada nije pronađena metoda koja daje rezultate koji su zadovoljavajući u odnosu na ciljane potrebe.

Dalji rad se može preusmeriti na domene ispitivanja veštačkom inteligencijom. Umesto numeričke analize spektrograma dobijenog Furijeovom transformacijom, ulazni signal bi mogao da se analizira uz pomoć klasifikatora. Kao set podataka za treniranje može se koristiti set pesama i njihovih melodija. Ovaj set se može deliti po žanrovima radi dalje optimizacije. Ova tematika je još uvek nedovoljno ispitana i efikasnost ovakvih sistema na ekstrakciju melodije je još uvek nepoznata.

Literatura

ROLI 2018. JUCE. ROLI Ltd. Dostupno na: https://juce.com/ [Pristupljeno 14. oktobra 2018].

Salamon J. 2008. Melody Extraction from Polyphonic Music Signals. PhD thesis. Universitat Pompeu Fabra, Barcelona, Spain

Rashidul H., Jamil M., Rabbani G., Rahman S. 2004. Speaker Identification Using Mel Frequency Cepstral Coefficients. U *Proceedings of the 3rd International Conference on Electrical & Computer Engineering*, ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh. ICEE, str. 565-8.

Nemanja Milanović

Melody Extraction from Polyphonic Audio Sources

This paper studies melody extraction from polyphonic audio sources. The aim of the research was extracting the main basic melody from commercial songs. This was done by analysis of the spectrogram generated by the fast Fourier transform. Before applying the Fourier transform, the source signal was processed using three infinite impulse response filters. Dominant frequencies were then extracted from the generated spectrogram, which we call the melody contour. The melody contour was then processed in different ways with the intent of removing errors. The precision of the results was measured by checking if the MIDI file of the source audio matched the output MIDI file at each MIDI tick (smallest time interval that the MIDI standard supports).

Results show that it is possible to get a decent output, which is the case with simple synthetic examples. In the case of piano music, there is a drop in precision, and with commercial music, precision is even worse. The main problem with commercial music is compression, which is applied to every song to gain loudness. This process enhances the harmonics of the melody, which presents a big problem with calculating dominant frequencies.