Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet

Dino Malpera

Detekcija tonova u snimci glazbenog djela hibridnim algoritmom nenadziranog učenja

Zagreb, 2011

Ovaj rad izrađen je na Matematičkom odsjeku Prirodoslovno-matematičkog fakulteta pod vodstvom doc. dr. sc. Luke Grubišića i predan je na natječaj za dodjelu Rektorove nagrade u akademskoj godini 2010./2011..

Sadržaj

1 Uvod

Zadatak glazbene transkripcije je iz zvučnog zapisa glazbenog djela producirati simbolički notni zapis. To znači svaki relevantan glazbeni događaj zastupiti pripadnom notom koja reflektira vrijeme nastupa, trajanje nastupa, visinu tona, instrument koji ga je inicirao, glasnoću i ostale parametre tog glazbenog događaja. Glazbena transkripcija je korisna vještina svakom glazbeniku. I dok uspiješnost glazbene transkripcije kod ljudi ovisi o znanju glazbene teorije i iskustvu, u novije vrijeme, napretkom računala, razvijane su metode automatske (računalne) glazbene transkripcije. Problemi automatske glazbene transkripcije su višestruki. Sama priroda glazbe u izvedbi često uvodi odstupanja od temeljnog predloška - izvođači nisu strojevi; vrijeme nastupa, trajanje, frekvencija i glasnoća pojedinog tona je podložna interpretaciji glazbenih umjetnika. Tako parametri glazbenog događaja ne moraju precizno odgovarati zapisu djela, ako isti postoji (ako je isti uopće moguć).

Ako su pak strojevi čak i izveli glazbeno djelo (nije rijedak slučaj danas) i dalje dolazimo do iduće grupe problema; praksa suvremene produkcije glazbe se većim djelom temelji na korištenju stereofonskog zvuka (popularno *stereo*). Iz toga slijedi da će, ako glazbeno djelo uključuje više od dva instrumenta, barem jedan od kanala biti djeljen među barem dva instrumenta. Naravno, mnoge snimke sadrže znatno više od dva instrumenta, pa je jasno da u toj situaciji imamo više izvora zvuka nego kanala. No, čak kada bi tehnički i imali jednak broj kanala i izvora zvuka, i dalje ostaje da praksa suvremene produkcije glazbe nalaže da se instrumenti rasporede u prostoru kako bi snimka zvučala ugodnije. To sve znači da će pojedini izvori zvuka biti međusobno izmješani, mijenjani i konačno reducirani na dva kanala koji su visoko korelirani, te ih je sasvim praktično gledati kao jedan izvor informacije. U tom kontekstu, jasno je da može dolaziti do maskiranja - tiši glazbeni događaji koji koincidiraju s glasnijim glazbenim događajima neće biti čujni na snimci. Takve koincidencije su iznimno česte; praksa je u glazbi, osobito u glazbi s izraženim ritmom, da veći broj izvođača svira tonove u istom trenutku (*na dobu*). Dodamo li još tome i probleme vezane uz šum na snimci i probleme vezane uz udaraljke (čija pojava u ritmičkim glazbenim djelima dolazi upravo tamo gdje nam najmanje treba - *na dobu*, koincidirajući s drugim tonovima), jasno je da gubitak informacije na snimci može biti toliki da pouzdanu cjelovitu transkripciju nije moguće dobiti (pojedini tonovi ili instrumenti se jednostavno ne čuju).

Najuspiješnija metoda glazbene transkripcije, transkripcija rukom i uhom glazbenika, oslanja se na odlične mogućnosti uha i pripadnih centara mozga kako bi se što točnije odredili objektivni parametri snimke. No, to je tek početak; profesionalni glazbenici čes-

to mogu riješiti i probleme maskiranih tonova koristeći kontekst i iskustvo. Dakle, iako možda ne čuju sam ton, ipak iz iskustva i konteksta znaju da bi on tamo trebao biti, jer to traži glazbena misao.

Ovaj rad posvećen je metodama transkripcije koje ne koriste znanje glazbenih stilova ili glazbenih obrazaca. Drugim riječima, zanimaju nas metode koje nastoje deducirati koji su se i kakvi glazbeni događaji dogodili, pritom koristeći isključivo digitalni zapis glazbenog djela, bez više razine znanja. Pritom prednost dajemo inžinjerskom pristupu. Iduće poglavlje bavi se prirodom signala koje generiraju glazbeni instrumenti i njihovom percepcijom. Treće poglavlje opisuje razne pristupe i metode transkripcije glazbe djeleći ih na metode detekcije nastupa tonova i metode određivanja tona. Četvrto poglavlje predstavlja nov pristup problemu određivanja tonova u polifonim snimkama (MFE) temeljen na nenadziranom učenju ton modela iz snimke koju se transkribira. Slijede prezentacija rezultata numeričkih eksperimenata i zaključak. Rad uključuje i kazalo koje sadrži definicije korištenijih pojmova.

2 Akustika glazbe

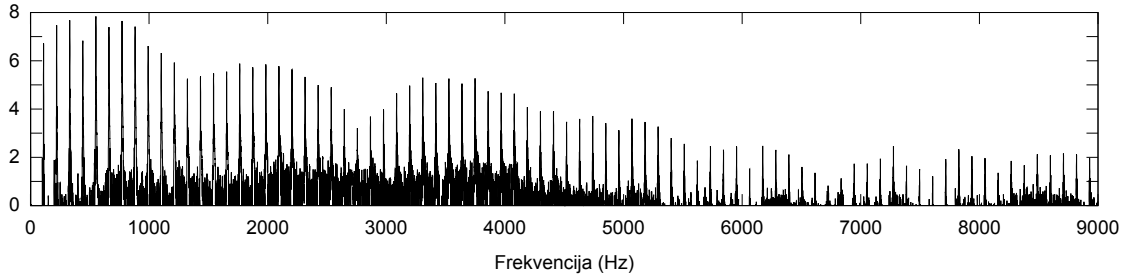
2.1 Parcijali i harmonici

U zvuku glazbenih instrumenata uglavnom dominiraju periodičke komponente, pa mu je razumno pristupati gledajući njegov spektar. Glazbeni instrumenti proizvode tonove koji često sadrže više frekvencijskih komponenti, tj. parcijala. Kod većine instrumenata ti parcijali čine dobro definiran niz frekvencija. U tom slučaju niz čine frekvencije koje su okvirno višekratnici temeljne frekvencije i zovu se harmonici:

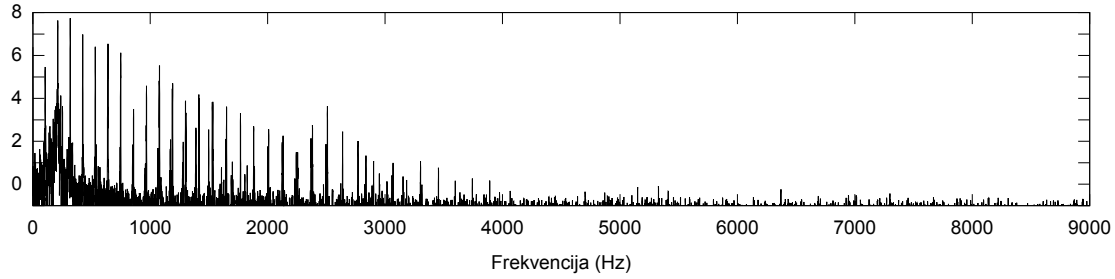
$$f_h = hf_0 + s(h), \quad h > 0. \tag{1}$$

U gornjem izrazu, f_h je frekvencija h-tog harmonika, f_0 fundamentalna frekvencija, a $s(h)$ predstavlja odstupanje harmonika od točnog višekratnika (inharmoničnost). U fiziku generiranja takvih zvukova, tj. zašto napete žice (prilikom trzanja i gudanja) i puhaći instrumenti stvaraju harmoničke signale, nećemo ulaziti. Tonovi ne moraju nužno imati parcijale - i oscilator generira ton. Ako i imaju parcijale, većina ih ne moraju biti harmonici. Udaraljkaški instrumenti spadaju u tu kategoriju; marimba, zvona, timpani imaju frekvenciju, ali malen udio harmonika među parcijalima. No, daleko najzastupljenija i najbitnija kategorija instrumenata je ona bogata harmonicima. U tu kategoriju spadaju gudači žičani instrumenti (klasa violinskih instrumenata), trzalački žičani instrumenti (klasa gitara, čembalo), drveni puhači instrumenti (oboa, klarinet, saksofon, ...), limeni puhači instrumenti (rog, truba, trombon, ...) i neki udaraljkaški instrumenti (klasa klavira). Ovi instrumenti imaju malen udio neharmoničkih parcijala; shodno tome, neharmonički parcijali se često zanemaruju (harmonici su spretniji za modeliranje). Spektri tonova ovih instrumenata liče na češalj (Slika 1).

Tu bi još trebalo spomenuti subharmonike - parcijale čija frekvencija je niža od fundamentalne (djeli fundamentalnu). Subharmonici daju "punoću" zvuku, ali su obično



Slika 1: *Spektar tona saxofona ($F0 = 110Hz$)*



Slika 2: *Spektar tona klavira ($F0 = 105Hz$)*

veoma tihi pa se zanemaruju.

Iako ton može imati mnogo parcijala, ljudi ne čuju svaku pojedinu komponentu (frekvenciju) već percipiraju samo njegovu fundamentalnu frekvenciju. Zašto do toga dolazi je pitanje psihoakustike i u njega nećemo ulaziti. Recimo samo da ljudski auditorni sustav teži pojednostaviti zvučnu sliku i razne akustičke signale predstaviti jednom frekvencijom. To u nekim slučajevima vrijedi čak i za signale koji nisu periodički (određene vrste šumova).

Fundamentalna frekvencija je frekvencija koju je moguće pridružiti tonu u psihoakustičkom eksperimentu. Zamislimo da na raspolaganju imamo izvor zvuka i audio oscilator kojem možemo mijenjati frekvenciju. Ako oscilator možemo ugoditi na frekvenciju koja "odgovara" izvoru zvuka, tada izvor zvuka ima fundamentalnu frekvenciju i to je upravo ona na koju smo ugodili oscilator. Kod harmoničkih tonova fundamentalna frekvencija (ili F0) je često jednaka prvom harmoniku, ali to nije nužno. Ako ton ima dovoljno harmonika, tada odsudstvo prvih nekoliko harmonika, neće utjecati na percepciju fundamentalne frekvencije. Drugim riječima i bez prvih nekoliko harmonika, ton će i dalje biti percipiran kao da ima istu fundamentalnu frekvenciju.

Parcijali tonu daju boju; relativni odnosi amplituda pojedinih harmonika razlikuju se od instrumenta do instrumenta za isti ton i mogu se koristiti za detekciju instrumenata (Slike 1 i 2 prikazuju spektre tonova saksofona i klavira - razlika je očita). No, odnosi harmonika za razne tonove i glasnoće istog instrumenta variraju. Tonovi koji su bliski po fundamentalnoj frekvenciji imat će veoma slične spektre, ali oni koje se razlikuju za nekoliko oktava vjeroatno neće. Također, tiši tonovi često imaju tiše više harmonike, dok porast glasnoće djelom dolazi i iz porasta glasnoća viših harmonika. Izgled spektra također ovisi i o stilu sviranja. Ton koji ima malo harmonika (ili veoma tihe više harmonike) zvučati će "tamno" i "suho". Dodavanje harmonika (ili glasniji viši harmonici) ton čine "svijetlijim" i "punijim". Korisno svojstvo amplituda parcijala jest zaglađenost spektra; amplituda pojedinog harmonika ne odudara značajno u odnosu na susjede (Sli-

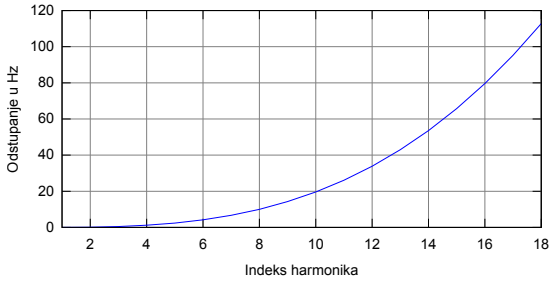
ka 1). Drugim riječima ako povežemo vrhove harmonika, dobit ćemo relativno glatku krivulju. To vrijedi za većinu glazbenih instrumenata i koristi se kao jaka pretpostavka u modelima za analizu glazbe. No, i detalji amplituda harmonika mogu biti od koristi; poznato je da klarinet ima nešto glasnije neparne harmonike od parnih, vidi [15]. To svojstvo koje se često koristi u detekciji instrumenata. Zanimljivo je što se događa kada iz tona uklonimo parne/neparne harmonike; ako uklonimo parne, fundamentalna frekvencija ostaje ista (iako ton naravno zvuči drugačije); ako uklonimo neparne ton postaje za oktavu viši.

2.2 Inharmonici

Još jedno svojstvo koje utječe na boju tona je inharmoničnost; ranije smo spomenuli da su harmonici okvirno višekratnici temeljne frekvencije. No, harmonici svoje teoretske pozicije rijetko slijede savršeno precizno (što je dobro, jer ton inače zvuči "siromašno"). To odstupanje harmonika (ili inharmoničnost) ovisi o indeksu harmonika (kao što je sugerirano funkcijom s u formuli (1)); viši harmonici mogu odstupati više nego niži. Također, pokazano je da su dublji tonovi osjetljiviji na inharmoničnost nego viši [17]; veći broj parcijala pridonosi inharmoničnosti tonova s niskom fundamentalnom frekvencijom i njihova odstupanja moraju biti manja. Slijedi ilustrativna formula koja dobro opisuje inharmoničnost žičanih instrumenata, a uporište ima u fizici napete žice [16]:

$$f_h = h f_0 \sqrt{1 + \beta(h^2 - 1)}, \tag{2}$$

gdje je f_h frekvencija h-tog harmonika, f_0 fundamentalna frekvencija, a β koeficijent inharmoničnosti (tipična vrijednost je $\beta = 0.0004$). Pitanje koje je zanimljivo za razvoj modela harmoničkih tonova je koliko inharmoničnosti je previše, tj. koliko najviše može parcijal odstupati od svoje teoretske pozicije prije nego ga se počinje percipirati kao zaseban ton; zanima nas $\max |s(h)|$. Odgovor na to pitanje ovisi o trajanju tona. Veća odstupanja parcijala u veoma kratkim tonovima nećemo primjetiti, dok za duže tonove hoćemo. Korisnije saznanje jest da prag inharmoničnosti ne ovisi značajno o frekvenciji i iznosi oko 4Hz, vidi [19]. No, to se odnosi samo na pojedini harmonik; ako veći broj harmonika pomaknemo zajedno i tako nagnemo krivulju inharmoničnosti, što je slučaj za napete žice kao što sugerira forumla (2) i Slika 3, dopustiva odstupanja harmonika postaju veća. Ipak, iako harmonici mogu značajno odstupati od teoretske pozicije (iz Slike 3 je vidljivo da se uz $\beta = 0.0004$ 14-ti harmonik nalazi na pola puta do teoretske pozicije slijedećeg harmonika, dok se 18-ti harmonik nalazi na poziciji idućeg harmonika, a u stvarnosti odstupanja mogu biti i znatno veća), nešto stabilnije svojstvo je



Slika 3: *Inharmoničnost (Hz) prvih 18 harmonika tona* ($F0 = 100Hz$)

odstupanje od svojih susjeda; razlika frekvencija dva susjedna harmonika biti će veoma blizu fundamentalne frekvencije, čak i ako su harmonici veoma daleko od svojih teoretskih pozicija. Ako očekujemo da se harmonici nalaze na svojim teoretskim pozicijama, koristimo pretpostavku o spektralnim pozicijama harmonika, a ako pak računamo na stabilnost relativnih pozicija sukcesivnih harmonika oslanjamo se na stabilnost spektralnih intervala. Obje pretpostavke se koriste prilikom modeliranja tonova.

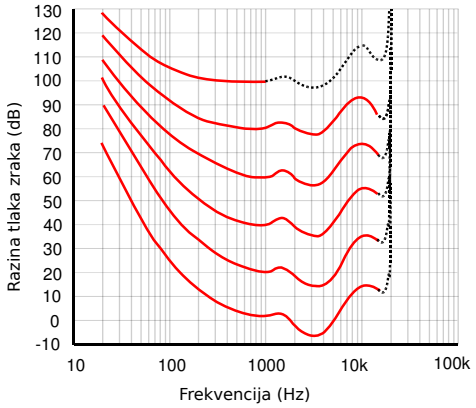
Zanimljivo je uočiti da, ukoliko je aritmetička sredina razlika teoretskih i stvarnih frekvencija harmonika bitno veća od nule (drugim riječima ako su stvarne pozicije svih harmonika veće od teoretskih), kao što prethodna formula sugerira za žičane instrumente, ton će zvučiti "više" nego prvi harmonik sugerira; tu postaje jasna razlika između fundamentalne frekvencije i prvog harmonika - prvi harmonik je prvi opaženi harmonik, dok je fundamentalna frekvencija stvar percepcije auditornog sustava.

Dodatan uzrok prividne inharmoničnosti može biti i ograničena rezolucija sučelja; ako koristimo DFT sa 23ms prozorom bez interpolacije, razmak između frekvencijskih koeficijenata bit će 43Hz, što za tonove s niskom fundamentalnom frekvencijom (malim razmakom među harmonicima) može uvesti značajna odstupanja od očekivanih pozicija.

Iako, zvuk većine instrumenata sadržava neku razinu inharmoničnosti, ona može nestati u slučaju mode-locking fenomena, vidi [18]. U slučaju gudačih instrumenata do toga će doći kad se žica gudi (harmonici će postati točni višekratnici fundamentalne frekvencije); ako se žica trzne ton će biti inharmoničan.

2.3 Ostala razmatranja

Za razliku od znanstvenika i inžinjera, glazbenici umjesto o frekvencijama govore o tonovima u glazbenoj skali. Skala je podjeljena na oktave, gdje razmak od jedne oktave među dva tona govori da je kvocijent fundamentalnih frekvencija tih tonova dva. To će značiti da, dok oktave na glazbenoj skali dolaze jednako razmaknute (za 12 polutonova), razlike među frekvencijama tonova rastu eksponencijalno; ako ton A3 ima frekvenciju



Slika 4: *Krivulja jednake glasnoće*

220Hz, njegova viša oktava (A4) imati će frekvenciju 440Hz, a njemu (A4) viša oktava (A5) 880Hz (duplo od A4, a ne triput od A3). Zato o razmaku među tonovima ne možemo govoriti u hercima, radije u polutonovima i oktavama. Spomenuli smo da se oktave dijele na polutonove i tonove; ton je dva polutona, a poluton je dvanaestina oktave - omjer fundamentalnih frekvencija tonova u odnosu polutona je $2^{1/12}$. Raspon ljudskog sluha prostire se kroz 10-ak oktava (20Hz-20kHz), pri ćemu je uho osjetljivije u višem djelu intervala (više frekvencije).

Što se relativne glasnoće zvuka tiče, o njoj se često govori u decibelima. Decibel je definiran kao logaritm omjera vrijednosti; to ga čini zanimljivim za uspoređivanje vrijednosti koje se mogu značajno razlikovati u veličini, primjerice snage:

$$L_{dB} = 10 \log_{10} \frac{P_1}{P_0}.$$

U akustici se decibelima izražava odnos glasnoće (tlaka zraka) dva zvuka. Iako je decibel mjera odnosa vrijednosti, često je spretno govoriti o odnosu prema nekoj referentnoj točki. U akustici se kao referentna točka uzima prag čujnosti ljudskog uha. No, osjetljivost uha ovisi o frekvenciji; uho je osjetljivije u srednjem djelu spektra (oko 2kHz) koji odgovara rasponu frekvencija ljudskog glasa. Ovisnost osjetljivosti uha na glasnoću o frekvenciji ilustrira krivulja jednake glasnoće (Slika 4) koja povezuje točke koje ljudsko uho percipira jednako glasnima.

Osim oblika spektra i inharmoničnosti, na boju tona utječe mnogo faktora. Neki od njih su vezani uz nastup tona. Nastup tona je početni dio (vremenski) tona, u kojem energija tona raste od nule do maksimalne vrijednosti koju će dosegnuti. Trajanje nastupa i način na koji se spektralna krivulja mijenja tokom nastupa bitno utječu na

sposobnost identifikacije instrumenata; ukoliko zvuku klavira maknemo nastup, teško ćemo ga identificirati - klavir je udaraljkaški instrument koji zvuk stvara time što batić udara žicu, što daje specifičan nastup. Bez nastupa ton klavira ćemo lako zamjeniti za neki drugi žičani instrument. Još jedno svojstvo koje utječe na boju zvuka su blage oscilacije frekvencije ili amplitude tona. Te oscilacije su kod većine instrumenata jedva primjetne (ako se na njih ne obrati pozornost), ali tonu daju "živost". Kod nekih instrumenata, oscilacije su izraženije; orgulje imaju veoma izražene oscilacije u amplitudi i lako ih je prema tome identificirati.

S obzirom da se neke metode automatske transkripcije glazbe oslanjaju na auditorne modele, red je nešto reći i o njima. Auditorni sustav čovijeka počinje uhom, gdje se vibracije zraka bubnjićem i trima košćicama prenose do pužnice. Pužnica je ispunjena tekućinom koja vrši stanovitu kompresiju signala; većina metoda koristi neku vrst kompresije o ćemu će više riječi biti u Odjeljku 3.1.1. Osjetilne dlačice, koje se nalaze na unutarnjoj stjenki pužnice i koje animira kretanje tekućine, generiraju električne signale koji se auditornim živcem prenose do mozga. Zahvaljujući različitoj poziciji raznih grupa osjetilnih dlačica u pužnici, neke grupe aktiviraju samo neke frekvencije. To opažanje mnogi autori vide kao temelj za korištenje niza pojasno propusnih filtera u sučelju o ćemu će više riječi biti u Odjeljku 3.1.1. Što se sa signalom dalje događa (kad dođe do mozga) manje je jasno; studije na pacijentima koji imaju oštećenja određenih djelova mozga indiciraju da se posebno vrši temporalna organizacija zvuka (određivanje ritma), a posebno određivanje tonova, vidi [20].

Veoma specifični instrumenti koje treba izdvojiti su udaraljke. Udaraljke možemo podjeliti na one koje imaju fundamentalnu frekvenciju i one koje nemaju. U one koje nemaju spadaju činele, snare i bas bubnjevi, koji su obilno korišteni u zabavnoj i jazz glazbi. Iako u početnim milisekundama ovi instrumenti imaju neke periodičke osobine, veoma brzo prelaze u kaotični režim što za poslijedicu ima zvuk bogat tranzijentima, vidi [21].

Još jedan instrument koji zaslužuje biti izdvojen jest glas - iznimna raznolikost u svim aspektima boje, čine glas teškim za modeliranje čak i kad se modeli treniraju za određeni glas. To se osobito odnosi na skladbe koje uključuju tekst; vokali imaju različite spektre, a suglasnici su bogati tranzijentima. No, iako većina autora ne koristi posebne modele za glas, ljudski auditorni sustav je posebno vješt u radu s glasom, osobito kad zuk sadrži tekst. No, to već zalazi u područje prepoznavanja govora i time se nećemo baviti.

3 Pristupi transkripciji glazbe

3.1 Uvod

Da bi uopće mogli govoriti o glazbenoj transkripciji, nužno trebamo snimku glazbenog djela. Snimka treba sadržavati dovoljno informacije (biti dovoljno kvalitetna) kako bi zadatak glazbene transkripcije bio sto lakši. Važan parametar digitalnog audio zapisa je frekvencija uzorkovanja. Ukoliko je frekvencija uzorkovanja suviše niska, više frekvencije neće biti dobro pohranjene. Naime Nyquist-Shannon-ov teorem o uzorkovanju kaže da je signal moguće u potpunosti rekonstruirati kad je frekvencija uzorkovanja barem dvostruko veća od najveće frekvencije iz tog signala. Iako u praksi postoje snimke frekvencije uzorkovanja od 20kHz pa i niže, velika većina snimaka koristi frekvenciju uzorkovanja od 44.1kHz. Iz toga slijedi da te snimke ne bi smjele sadržavati frekvencije više od 22.05kHz.

3.1.1 Sučelje

Digitalni zapis zvuka sastoji se od niza uzoraka i možemo ga interpretirati kao funkciju amplitude u ovisnosti o vremenu. Ta funkcija nam daje reprezentaciju signala u vremenskoj domeni. Takva reprezentacija često nije spretna, te se signal transformira u druge domene kako bi informacija koju nosi bila pristupačnija. To je jedan od zadataka sučelja. Jedna od korištenijih reprezentacija signala je u frekvencijskoj domeni. Orginalni signal se Fourierovom transformacijom rastavlja u sumu frekvencija. Redovito se gleda samo jedan (odabrani) dio signala - okvir. Heisenbergov princip neodređenosti za harmoničku analizu opisuje odnos između rezolucije u vremenu i frekvenciji; što je okvir manji imamo jasniju sliku događaja u vremenu i manje jasnu u frekvenciji i obratno. Tu valja napomenuti da veliki okviri jasnoću u frekvenciji daju samo za spore (stacionarne) događaje. Ako želimo dobiti što bolji uvid u informacije koje signal sadrži, zanimat će nas njegove stacionarne komponente kao i njegove tranzientne komponente. Stacionarne komponente su intuitivno one koje imaju stabilnu frekvenciju, dok su tranzienti brzi kratki naleti energije koji nemaju periodička svojstva. To nas vodi na metode simultane reprezentacije signala u vremenu i frekvenciji.

Gradeći na metodi Fourierove transformacije dolazimo do metode STFT (Short Time Fourier Transform): orginalni signal se segmentira u relativno kratke segmente (okvire) na koje se primjeni Fourierova transformacija. Iako Fourierova transformacija daje naoko intuitivan pogled na signal (kao sumu frekvencija), to je istina samo za stacionarne signale; frekvencijski prikaz signala čije se komponente brzo mijenjaju kroz vrijeme ili sadrže tranzijente, zahtijevaju interpretaciju. Tu dolazimo do pretpostavke o kvazipe-

riodičnosti; računamo s tim da su komponente signala za dovoljno kratke vremenske intervale (male okvire) dovoljno slične stacionarnim valovima:

$$s(t) = r(t) + \sum_{k=0}^{N-1} A_k \cos(k\omega t + \phi_k),$$

gdje je $r(t)$ rezidual. Fourierova transformacija signal prikazuje u ortogonalnoj bazi funkcija. U slučaju diskretne Fourierove transformacije bazne funkcije su sinusoide. Takav izbor baznih funkcija daje odličnu lokalizaciju u frekvenciji (za stacionarne komponente), ali slabu u vremenu (za tranziente).

Alternativni pristup daje transformacija valića. Valić je dio periodičkog vala koji je lokaliziran u vremenu. Diskretna transformacija valića problemu međusobno suprotnih zahtjeva između lokalnosti u vremenu i frekvenciji pristupa tako što su bazne funkcije nižih frekvencija manje lokalizirane u vremenu nego one viših frekvencija i obratno. Time kod viših frekvencija dobivamo veću lokalnost u vremenu i manju rezoluciju u frekvenciji. Kod transformacije valića govorimo o reprezentaciji u vremenskoj skali, s obzirom da je lokalnost vezana uz frekvenciju kroz parametar skaliranja. Diskretna transformacija valića također vrši rastav u ortogonalnu bazu; idući korak je proučiti rastave koji nisu ortogonalni.

Tako dolazimo do rijetko popunjenih aproksimacija; signal se prikazuje u takvoj bazi u kojoj je broj koeficijenata različitih od nule znatno manji od broja koeficijenata jednakih nuli. Metode koje koriste rijetko popunjene aproksimacije rastav signala vrše na članove (atome) predefiniranih riječnika. Riječnici sadrže razne funkcije (radi se o parametriziranim klasama) različitih lokalnosti, frekvencija, skaliranja. Zadatak postaje optimizacijski problem; naći onaj podskup članova riječnika koji najbolje opisuje signal. Uspiješnost ovog pristupa ovisi o dva izbora; izboru riječnika - želimo da riječnik sadrži funkcije koje odgovaraju komponentama koje očekujemo u signalu, i izboru metode optimizacije - želimo da riješenje bude što bolje, a algoritam što brži. Metoda uparivanja (*matching pursuit*) [5] koristi pohlepni pristup, u svakom koraku birajući onaj atom iz riječnika koji najbolje opisuje rezidual prethodnog koraka s obzirom na neku funkciju cijene, i zatim uklanjajući taj atom iz signala. Metoda uparivanja naslijeđuje svojstva pohlepnih alogritama; brza je u odnosu na druge metode rijetko popunjenih aproksimacija, ali loš izbor u ranijim koracima vodi u suboptimalno riješenje. Iako je primjena metoda rijetko popunjenih aproksimacija u transkripciji glazbe rijetka, metoda uparivanja je našla svoje mjesto [6] najvjeroatnije zahvaljujući brzini. Za razliku od metode uparivanja, basis pursuit [7] koristi konveksnu optimizaciju; iako daje bolje rezultate, me-

toda je prespora za mnoge primjene. Razvijena su mnoga proširenja metode uparivanja; HRMP (*high resolution matching pursuit*) - metoda uparivanja visoke rezolucije [8] koja koristi drugačiju funkciju cijene (orginalna metoda uparivanja koristi skalarni produkt, što može dovesti do toga da metoda "stvara" energiju tamo gdje je ranije nije bilo) kako bi izbjegla suboptimalna riješenja u nekim situacijama; molekularna metoda uparivanja [9] i harmonička metoda uparivanja [10] koriste činjenicu da neki signali imaju strukturu (primjerice harmonički zvuk), pa umjesto atoma u svakom koraku iz signala pokušavaju ukloniti molekulu (sumu atoma) koja bi predstavljala harmonijski niz.

U praksi ipak najkorišteniji pristupi bazirani su na STFTu, vjeroatno zahvaljujući egzaktnosti rastava i velikoj brzini FFT algoritma. FFT ima vremensku složenost $O(n \log n)$ i veoma je efikasno implementiran na računalima. STFT u svakom koraku izračunava DFT signala s nadodanom funkcijom prozora:

$$X_k = \sum_{n=0}^{N-1} w(n) x_n e^{-\frac{2\pi i}{N} kn}, \quad k = 0, \ldots, N-1.$$

gdje je x signal od N uzoraka, w funkcija prozora, a X spektar signala (pri ćemu je X_k k-ti Fourierov koeficijent). Izbor funkcije prozora ovisi o širini okvira, tj. o željenom odnosu rezolucije u frekvenciji i amplitudi. Za širine okvira preferiraju se potencije od 2, jer takve N-ove FFT algoritam najbrže računa. Okviri iz sukcesivnih koraka STFTa razmaknuti su za fiksnu vrijednost - širinu koraka. To uvodi uređaj na okvire STFTa.

Kako bi dobili što točniju frekvenciju sinusoidalnih komponenti signala, često se koriste metode interpolacije. Najjednostavnija metoda iz ove grupe je interpolacija nadjevanjem nulama u kojoj se ulazni vektor DSPa proširuje nulama. Metoda kvadratične interpolacije spektralnih vrhova traži parametre parabole koja najbolje opisuje vrh. Iz tih parametara tada procjenjuje točnu amplitudu i frekvenciju vrha.

Nestacionarne komponente signala mogu čitanje spektra učiniti problematičnim, što zahtjeva posebne metode za detekciju i procjenu sinusoida. Primjerice, čest slučaj su tonovi koji tokom trajanja mijenjaju amplitudu (AM - amplitude modulation) i frekvenciju (FM - frequency modulation), te su razvijene metode koji procjenjuju parametre takvih komponenti signala [11]. Vrijedna spomena je i time-frequency reassignment metoda koja grupira energiju signala u lokalne centre mase i time "pooštruje" spektogram signala u vremenu i frekvenciji. Tako se procjenjuje točna pozicija spektralnih vrhova u frekvenciji (koristeći procjenu instantne frekvencije) i vremenu (koristeći procjenu grupnog kašnjenja).

Korištenija transformacija iz vremenske u frekvencijsku domenu je *constant Q tran-*

sform [12]. Q iz imena predstavlja kvocijent frekvencije i rezolucije. Transformacija taj omjer čuva konstantnim, što znači da rezolucija ovisi o frekvenciji (viši frekvencijski pojasevi imat će manju rezoluciju). Tako su frekvencijski koeficijenti logaritamski razmaknuti, pa ova transformacija ima sličnosti s diskretnom transformacijom valića. Prednost logaritamske distribucije frekvencijskih koeficijenata kod metoda koje koriste modele izvora zvuka [13] (u ovom slučaju glazbenih instrumenata) leži u tome što su tada relativne pozicije harmonika konstantne neovisno o fundamentalnoj frekvenciji.

U prethodnom poglavlju (2.3) spomenuo sam da se prema auditornom modelu signal rastavlja na frekvencijske pojaseve, te neka sučelja imitiraju taj pristup [2]; ulazni signal se provlači kroz niz (često preklapajućih) pojasno propusnih filtera i tako dobiva velik broj kanala koje se obrađuje neovisno, a rezultati obrade zatim kombiniraju u idućem koraku.

Nakon što se signal transformirao u frekvencijsku domenu, često se primjenjuje neka vrsta kompresije; amplitude se promjene nekom monotonom funkcijom, izbor koje ovisi o tome što nas zanima. Logaritamska kompresija dati će prednost detaljima (ali i šumu), $\log(x+1)$ će također dati prednost detaljima, ali će ograničiti kodomenu odozdo (valja naglasiti da ovo ima više smisla kad je $x \gg 1$). Kvadratna funkcija će bolje reprezentirati energiju signala i dati prednost glasnijim komponentama. Također se koriste eksponencijalne kompresije s raznim eksponentima; 0.67 se smatra optimalnim, a 0.5 se često bira jer je efikasno implementiran na računalima.

3.1.2 Uklanjanje šuma

Snimke često sadrže više ili manje šuma. Neki autori koriste uklanjanje šuma kao poseban korak [1, 2], dok drugi koriste prisutstvo šuma kao radnu pretpostavku metoda transkripcije [3, 4]. Uklanjanje šuma je osobito važno u metodama koje koriste iterativnu procjenu F0 kandidata, što će detaljnije biti objašnjeno u Odjeljku 3.3.2. Specifičan problem glazbene transkripcije je šum udaraljki koji je tranzijentne prirode, nije bijel, kratak je i često glasan. Ipak, većina autora ne pridaje posebnu pažnju takvoj vrsti šuma. Pristupi uklanjanju šuma su razni; od izbjeljivanja šuma (*noise whitening*) [14] do generativnih noise modela [1]. Nijedna od metoda nije posebno razvijena za glazbenu transkripciju, već se koriste standardne metode uklanjanja šuma.

3.1.3 Automatska transkripcija glazbe

Automatska transkripcija glazbe je složeni zadatak koji se može podjeliti na više jednostavnijih podzadataka:

1. detekcija glazbenih događaja
2. određivanje tonova
3. detekcija instrumenta
4. određivanja tempa

Iz popisa je jasno da se podzadaci grupiraju u dvije grupe; podzadatke vezane uz rad s vremenom i podzadatke vezane uz rad s tonovima. Iako neki pristupi MFE-u koriste detekciju nastupa tonova kao posebni korak [31], intervali aktivnosti pojedinih tonova se uglavnom induciraju iz MFE-a, o ćemu će biti više riječi u Odjeljku 3.4. Idućа dva podpoglavlja bave se prvom i drugom točkom s prethodnog popisa; zadnje dvije točke s popisa nisu toliko vezane uz prve dvije i njima se nećemo baviti.

3.2 Pristupi detekciji nastupa tonova

Cilj detekcije nastupa tonova je naći vremena početaka tonova u glazbenoj snimci. Takva informacija se kasnije može koristiti u razne svrhe - kompresiji, indeksiranju, analizi, procjeni tempa. Pod vremenom početaka tonova se uglavnom misli na početak početka; uzmimo da inicijalni dio razvoja tona u vremenu predstavlja porast energije tona do maksimuma. U tom kontekstu nastup tona predstavlja točku u vremenu u kojoj je razvoj tona počeo, tj. u kojoj je energija počela rasti.

Većina detektora nastupa tonova imaju dvije komponente: funkciju detekcije i biranje vrhova iz funkcije detekcije. Funkcija detekcije je funkcija vremena i predstavlja alternativnu domenu u koju je snimka prebačena, a u kojoj su lakše vidljive promjene koje indiciraju pojave nastupa (Slika 5 - srednja). Funkcija će imati stabilne vrijednosti u dijelovima snimke u kojima nema nastupa, dok će dijelovi u kojima se nalaze nastupi biti predstavljeni vrhovima u funkciji detekcije. Idealna funkcija detekcije ima oštre vrhove u točkama nastupa, a nulu drugdje. Nakon izračuna funkcije detekcije preostaje iz nje odabrati nastupe po nekom kriteriju (Slika 5 - donja).

Izbori funkcije detekcije su raznoliki; najjednostavnije funkcije detekcije prate promjenu amplitude ili energije signala uz konvoluciju u vremenu kako bi se prednost dala globalnijim promjenama [22]. Konvolucijskom transformacijom funkcije f smatramo:

$$(f * \psi)(t) = \int_{-\infty}^{\infty} f(\tau)\psi(t-\tau)\mathrm{d}\tau,$$

gdje je $\psi(t)$ konvolucijska jezgra. Takvi pristupi dobre rezultate daju samo u rijetkim prilikama kad su nastupi veoma istaknuti - primjerice nastupi udaraljki. Idući pristup

bio bi koristiti prve derivacije energije po vremenu, kako bi se prednost dala naglim promjenama (porastima) energije.

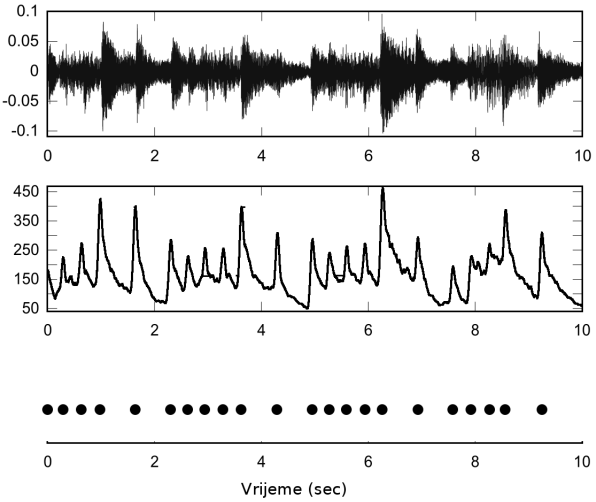
Kako se većina energije signala nalazi u donjem djelu spektra, svi frekvencijski pojasi nisu ravnopravno zastupljeni kada gledamo promjenu energije signala; mala promjena energije može biti značajna za viši dio spektra, ali mala u odnosu na ukupnu energiju signala (većina koje dolazi iz nižih frekvencija). Zato mnogi autori analiziraju pojedine frekvencijske pojase odvojeno i zatim objedine rezultate kako bi dobili funkciju detekcije. Drugo rješenje istog problema je koristiti težinsku funkciju, koja višim frekvencijama daje veću težinu, vidi [23].

Zanimljiv pristup je pratiti promjenu spektra kroz vrijeme; gledaju se razlike spektara među sukcesivnim koracima STFT-a na način da se spektar jednog koraka smatra vektorom. Zatim se vektori sukcesivnih koraka uspoređuju - pristupi uključuju: L_1 normu razlika vektora, L_2 normu korigirane razlike [28]:

$$DM(t) = \sum_{k=0}^{\frac{N}{2}-1} d(|X_k(t)| - |X_k(t-1)|)^2,$$

gdje se korekcija sastoji u tome da se gledaju samo frekvencije koje imaju porast energije, time dajući prednost nastupima (funkcija d vraća 0 za negativne argumente).

Tokom "mirnih" djelova signala u kojima nema nastupa, sinusoidalne komponente bi trebale biti gotovo stacionarne, što posebno znači da bi faza svake takve komponente



Slika 5: *Česti koraci u detekciji nastupa tonova: funkcija detekcije (srednja slika) i biranje vrhova (donja slika). Gornja slika predstavlja snimku (orginalni signal).*

trebala biti stabilna, krećući se na predvidljiv način (derivacija razmotane faze po vremenu je konstantna, ako je instantna frekvencija kostantna). No, u trenutku nastupa, stacionarnost signala bila bi poremećena imajući za posljedicu izmijenjen hod faze. To je intuicija iza metoda koje koriste informaciju o fazi za detekciju nastupa. O fazi nije bilo mnogo riječi do sada iz jednostavnog razloga; većina metoda transkripcije glazbe ne koristi fazu, već samo amplitudu frekvencija. No, neke metode detekcije nastupa koriste fazu. Neki pristupi fazu gledaju kao izoliranu kategoriju; uglavnom se sumira druga derivacija faze po svim frekvencijama i eventualno primjeni neka vrsta zaglađivanja, vidi [24]. Takvi pristupi su veoma osjetljivi na šum u signalu, jer on može značajno utjecati na fazu. Zanimljivija riješenja gledaju kompleksni spektar; Dexbury [25] gleda razliku između kompleksnog spektra okvira i njegove procjene na temelju predhodnog okvira. Ovaj pristup daje bolje rezultate od metoda koje koriste samo fazu ili samo promjenu energije.

Izbor vrhova funkcije detekcije je najjednostavnije vršiti koristeći fiksni prag; svi vrhovi koji pređu prag se klasificiraju kao nastupi tonova. Očekivano, ovaj pristup daje nezavidne rezultate; nemaju svi vrhovi koji predstavljaju nastupe tonova sličnu vrijednost funkcije detekcije. Glasniji djelovi glazbenog djela imat će veće vrhove, dok će tiši imati manje. To nas vodi do adaptivnih pragova; prag je funkcija vremena dobivena zaglađivanjem funkcije detekcije. Obzirom da je prag ovisan o funkciji detekcije, glasnije regije imat će viši prag, a tiše niži.

3.3 Pristupi detekciji fundamentalne frekvencije

3.3.1 SFE

Obzirom na složenost problema počeci glazbene transkripcije vežu se uz monofone skladbe. Tu se pretpostavlja da se u jednom trenutku čuje samo jedan ton. No, u realnosti to je varljiva teza; zahvaljujući učinku reverberacije i u monofonim snimkama često se u jednom trenutku mogu čuti dva i više tonova (ton koji se trenutno svira i reverb tona koji se svirao ranije).

Pristupi procjene fundamentalne frekvencije monofonih skladbi (SFE) djele se u dvije grupe ovisno o domeni u kojoj djeluju: pristupi vremenske domene i pristupi spektralne domene. Pristupi vremenske domene koriste pretpostavku periodičnosti signala uspoređujući signal sa svojom odgođenom verzijom, tražeći valnu duljinu. Najjednostavnija

metoda iz ove grupe je zasigurno autokorelacija:

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau},$$

gdje je x signal, W širina okvira, t vrijeme u kojem računamo autokorelaciju, a τ period kojeg tražimo. Ova jednostavna metoda otporna je na šum u signalu, ali je spora i loše se nosi s većom inharmoničnošću. Neke modifikacije autokorelacijske funkcije umjesto produkta gledaju razliku ili kvadrat razlike:

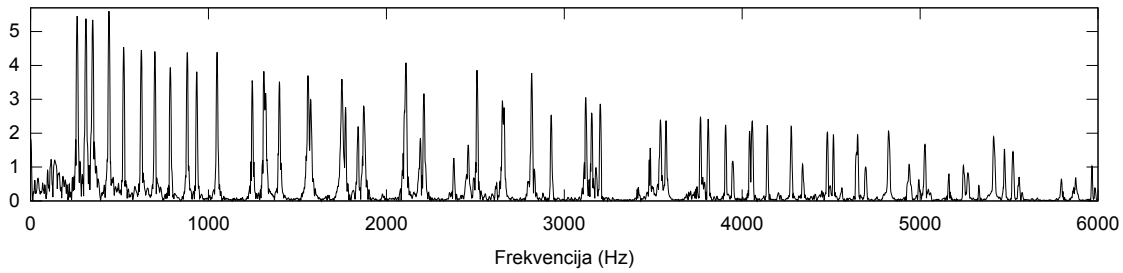
$$r_t(\tau) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau})^2.$$

Nakon što se izračuna autokorelacija, fundamentalni period određuje se tražeći najmanji lokalni minimum. Više harmonike signala se ne detektira posebno - pretpostavlja se da im period djeli fundamentalni; tu se oslanja na pretpostavku o spektralnim pozicijama harmonika opisanu u Odjeljku 2.2. Jasno je da ove metode lako rade sub-harmoničke greške, tj. potcjenjuju fundamentalnu frekvenciju vraćajući vrijednosti koje su djeljitelji stvarne fundamentalne frekvencije, jer frekvencija perioda τ ima i period 2τ.

Značajna izvedenica autokorelacijske funkcije je YIN algoritam [26]; autori kroz šest koraka nadograđuju autokorelacijsku funkciju na razne načine, konačno dosegnuvši poboljšanje od 20 puta.

Veoma zanimljiva metoda je metoda dvosmjernog neslaganja [27] koja radi harmonički *pattern matching* u frekvencijskoj domeni. Imamo dva niza podataka, opaženi niz frekvencija, za koje želimo utvrditi čine li harmonički niz, i hipotetičan niz frekvencija. Računamo dvije vrste greške; grešku prvog niza prema drugom i grešku drugog niza prema prvom. Greška se računa tako da za svaku frekvenciju jednog niza tražimo najbližu frekvenciju u drugom nizu. Zatim izračunamo razliku tih dviju frekvencija. Sve takve razlike svih frekvencija oba niza se konačno kombiniraju u završnu vrijednost na temelju koje se procjenjuje odgovara li hipotetičan niz opaženom stanju, tj. je li hipoteza prihvaćena.

Pristupi spektralne domene oslanjaju se na pretpostavku o spektralnim intervalima harmonika opisanu u Odjeljku 2.2 i počivaju na slijedećem razmatranju: ako je zvuk harmoničan, tada je razmak među njegovim sukcesivnim harmonicima relativno konstantan - spektar signala ima periodička obilježja. Tada ima smisla računati autokorelaciju spektra; dobiveni period odgovarat će razmaku među harmonicima, dakle fundamentalnoj frekvenciji. Alternativna metoda je umjesto autokorelacije spektra koristiti cepstrum,



Slika 6: *Spektar mješavine četiri tona jednakih amplituda*

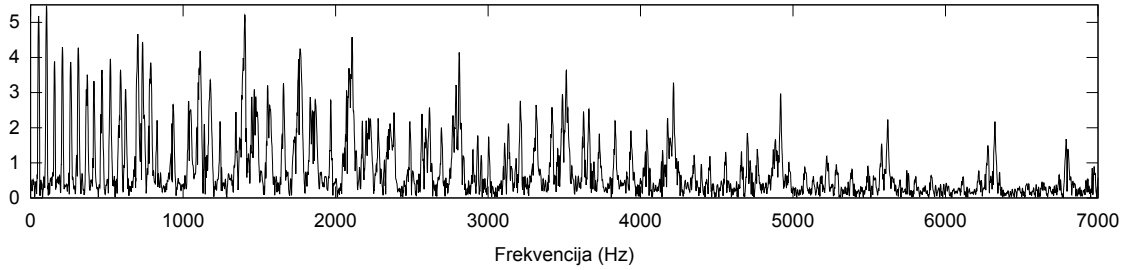
Fourierovu transformaciju spektra logaritmiranih vrijednosti. Ove metode rade superharmoničke greške, tj. precjenjuju fundamentalnu frekvenciju iz istih razloga iz kojih su ju pristupi vremenske domene potcjenjivali, uvažavajući da u spektru veća perioda (veći razmak među harmonicima) znači višu fundamentalnu frekvenciju.

Iako problem SFE nije toliko zanimljiv, obzirom da je većina glazbe ipak polifona, primjenu će uvijek naći u evaluaciji MFE metoda. Kako bi evaluirali uspiješnost transkripcije polifone snimke, treba nam referentna transkripcija iste. Problem nije jednostavan, a jedno riješenje je snimiti sve instrumente monofono i zatim ih miješati zajedno. Obzirom da su SFE metode značajno pouzdanije od MFE metoda, svaku monofonu dionicu moguće je transkribirati SFE metodom (redovito se koristi YIN), te zatim rezultate transkripcije svih dionica sintetizirati kako bi se dobila referentna transkripcija.

3.3.2 MFE

Procjena višestrukih fundamentalnih frekvencija (MFE) ne može se riješiti nadogradnjom metoda za SFE iz razloga što kod polifonih skladbi dolazi do preklapanja parcijala pojedinih tonova. SFE metode računaju na to da većina opaženih vrhova spektra pripada jednom jedinom tonu dok u polifoniji nije uvijek jasno koji vrh pripada kojem tonu (pod vrh "pripada" tonu misli se da je vrh harmonik tona). Problem ilustrira Slika 6 koja prikazuje spektar mješavine četiri tona jednakih glasnoća; u višem djelu spektra je gotovo nemoguće odrediti koji vrh pripada kojem tonu. Nadalje, u polifonim snimkama veći broj tonova mogu (i često jesu) u odnosu oktave; kako su fundamentalne frekvencije tonova takvog niza međusobno djeljive, takvi tonovi će djeliti većinu harmonika pa se, sa stajališta parcijala, može činiti da se radi o jednom tonu. Dodatno pitanje koje MFE uvodi je pitanje broja izvora zvuka (polifonije), tj. koliko tonova je prisutno u signalu.

Metode MFE-a oslanjaju se na razne pretpostavke prilikom modeliranja zvučne slike. Često korištene su harmoničnost tonova, glatkoća spektralne krivulje tonova i sinkrona evolucija harmonika tona u vremenu. Svojstvo harmoničnosti tonova objašnjeno je u



Slika 7: *Spektar kompleksne mješavine tonova raznih amplituda*

Odjeljcima 2.1 i 2.2; računa se s tim da su parcijali višekratnici fundamentalne frekvencije ili da su razmaknuti za fundamentalnu frekvenciju. Svojstvo glatkoće spektralne krivulje (također objašnjeno u Odjeljku 2.1) igra centralnu ulogu u MFE metodama; koristi se za bodovanje raznih hipoteza tonova i određivanje koji ton je vjeroatniji, a nužno je za detekciju tonova koji su u oktavnim odnosima (to je jedini način da se takvi tonovi detektiraju). Slika 7 prikazuje spektar signala kompleksne mješavine tonova pri ćemu, unatoč velikom broju vrhova, možemo relativno lagano odrediti nekoliko tonova tražeći sličnosti u amplitudi pravilno raspoređenih kandidata koje stvaraju glatku krivulju tona. Svojstvo sinkrone evolucije tonova u vremenu polazi od činjenice da će parcijali koje generira isti izvor zvuka zajedno rasti i padati. Sigurno će imati isti nastup, a jamačno i veoma sličan pad (iako viši harmonici često padaju nešto brže).

Kako bi predočili veličinu problema, uzmimo u obzir da imamo K ton kandidata za signal polifonije P. Time dolazimo do $\binom{K}{P}$ kombinacija koje bi trebalo ispitati što daje golemi broj kombinacija čak i za mali broj kandidata i male polifonije (to sve uz pretopstavku da znamo točnu polifoniju).

Prvi korak većine MFE metoda je izbor ton kandidata. Ton kandidat je niz spektralnih vrhova koji čine niz harmonika. Većina pristupa u obzir uzima samo ton kandidate koji imaju određeni broj vrhova (iako ne svi [32]); time reduciraju broj kandidata na one za koje postoji dovoljno svjedočanstva. No, neki instrumenti kao marimba ili zvona imaju mali broj harmonika (iako oni imaju veliku energiju), pa ako je limit na broj vrhova postavljen prenisko, tonovi tih instrumenata neće biti detektirani. Pristupi se razlikuju i po tome koliku inharmoničnost dopuštaju. Tipično, izbor ton kandidata počinje odabirom vrha iz odabranog frekvencijskog pojasa; većina metoda zanemaruje veoma niske ($< 50Hz$) i veoma visoke ($> 2100Hz$) frekvencije. Ukoliko postoje vrhovi na višekratnicima frekvencije odabranog vrha, oni se dodaju u niz. Pritom se uglavnom u obzir uzima mala okolina oko višekratnika (redovito pola polutona ili fiksni interval). Tu se postavlja pitanje što ako se u toj okolini nalazi više vrhova; neki pristupi tada

uzimaju najbliži vrh [33], a neki maksimalni [2]. Ipak, Yeh [1] u obzir uzima i vrh čija je amplituda najbliža prosječnoj amplitudi prethodna tri odabrana parcijala koristeći pretpostavku o glatkoći spektralne krivulje. Još jedna zanimljiva stvar koju Yeh radi je da za više harmonike ograničava okolinu oko teoretske pozicije harmonika koju se uzima u obzir; niži harmonici imaju okolinu od pola polutona, a više ograničava s:

$$\alpha \leq \frac{0.3}{2h+1},$$

gdje je h harmonik, a α koeficijent koji određuje širinu okoline.

Bitna podjela ton kandidata je na one koji su harmonički vezani i one koji nisu. Potonji, neharmonički vezani tonovi (NHRF0) zadavoljavaju slijedeći izraz:

$$F_1 = \frac{m}{n} F_0,$$

gdje su F_1 i F_0 fundamentalne frekvencije, a m i n prirodni brojevi koji nemaju zajedničkih faktora. Od dva harmonički vezana tona (HRF0), jedan je višekratnik drugog, što znači da onaj veće fundamentalne frekvencije vjeroatno djeli većinu svojih parcijala s onim manje fundamentalne frekvencije. Razlikovanje ove dvije klase ton kandidata omogućava specifićan pristup svakoj grupi; NHRF0 je lako detektirati i njihova problematika su parcijali koje djele s drugim NHRF0 tonovima. S druge strane HRF0 je veoma teško detektirati, jer većinu svojih harmonika posuđuju od svog (temeljnog) NHRF0.

Metode MFE-a djele se u dvije grupe prema redoslijedu u kojem procjenjuju fundamentalne frekvencije: metode simultane procjene i metode iterativne procjene. Prve nastoje sve tonove procjeniti odjednom, dok druge zadatak rastavljaju u niz koraka, pri ćemu u svakom koraku procjenjuju točno jednu fundamentalnu frekvenciju.

3.3.3 Metode simultane procjene

Može se činiti da je simultana procjena bolji pristup jer obećava da će uzeti u obzir interferenciju raznih tonova u razmatranje, simultano promatrajući njihovu kombinaciju (što uključuje i eventualno poklapanje parcijala i konkuriranje za razne parcijale). No, kako smo ranije uočili, broj mogućih kombinacija je prevelik za izravno provjeravanje te se uglavnom koristi neka vrsta optimizacije.

Mnogi autori reduciraju broj ton kandidata prije glavnog algoritma. Redukcija se vrši po raznim kriterijima; najčešći je zasigurno ukupna energija svih parcijala ton kandidata; što je ukupna energija veća, vjeroatnije je da se radi o tonu, a ne o slučajnom nizu parcijala koji su se našli na pravim mjestima. Yeh također koristi i inharmoničnost,

te zaglađenost spektra [1], a Emiya udaljenost fundamentalne frekvencija kandidata od frekvencije tonova dobro ugođene glazbene skale [31]. Izbor ton kandidata može se vršiti koristeći fiksni prag ili fiksni broj ton kandidata [32][33] (bira se najboljih F).

Najjednostavniji pristupi koriste veoma grube parametre redukcije (izbor desetak najboljih kandidata) i zatim vrše iscrpnu provjeru svim mogućih kombinacija. Konačno se bira kombinacija koja ostvari najbolji rezultat. Kao mjera uspiješnosti se često uzima udio ukupne energije spektra koji je kombinacija opisala.

Zanimljiviji pristupi uključuju statističke modele; Goto [3] koristi MAP procjenitelj pomoću EM algoritma kako bi ustanovio najvjeroatnije tonove. On zvučnu sliku modelira kao težinski model miješanja ton modela, pri ćemu je ton model težinska suma normalnih distribucija čiji raspored i varijanca reflektiraju harmoničku strukturu tonova (položene su na teoretskim pozicijama harmonika i dopuštaju inharmoničnost). Formalno, ton model je:

$$p(x|F,\mu(F)) = \sum_{h=1}^{H} p(x,h|F,\mu(F))$$

$$p(x,h|F,\mu(F)) = c(h|F)G(x,Fh,\psi)$$

$$G(x,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

gdje je F fundamentalna frekvencija, h harmonik, H broj harmonika koji se uzimaju u razmatranje, $G(x,\mu,\sigma^2)$ normalna distribucija, p uvjetne vjeroatnosti, ψ inharmoničnost, a c definira krivulju ton modela (relativne odnose amplituda) i vrijedi $\sum_{h=1}^{H} c(h|F) = 1$. Paramtear $\mu(F) = \{c(h|F)|h = 1,\ldots,H\}$ predstavlja oblik ton modela. Opaženi spektar je tada težinska suma svih mogućih ton modela:

$$p(x|\theta) = \int_F \omega(F)p(x|F,\mu(F)),$$

gdje je $\theta = \{\omega,\mu\}$ parametar modela, a suma ide po svim dopustivim fundamentalnim frekvencijama F.

Popularna metoda je i NMF - nenegativna faktorizacija matrice; uzmimo da znamo dva ton modela (spektra tonova) m_1 i m_2 za dva tona i imamo spektar x koji sadrži neku mješavinu tih tonova. Pišemo:

$$x = \begin{bmatrix} m_1 & m_2 \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \end{bmatrix},$$

gdje su ω_1 i ω_2 težine pojedinog modela u sumi. Spektar x je nenegativan, ton modeli su spektri pa kao takvi također nenegativni, i težine ω_1 i ω_2 su opet po definiciji nenegativne. Ako sad uzmemo spektre sukcesivnih okvira $x^{(t)}$ u matrici X, očekujemo rastav:

$$\begin{bmatrix} x^{(t)} & \dots & x^{(t+L)} \end{bmatrix} = \begin{bmatrix} m_1 & \dots & m_k \end{bmatrix} \begin{bmatrix} \omega_1^{(t)} & \dots & \omega_1^{(t+L)} \\ \vdots & \vdots & \vdots \\ \omega_k^{(t)} & \dots & \omega_k^{(t+L)} \end{bmatrix}$$

$$X = MW$$

gdje je k broj ton modela, a L broj sukcesivnih okvira. Općenito ovakav rastav ne postoji, ali metodom NMF možemo naći optimalnu aproksimaciju koja uvažava restrikciju na nenegativnost ton modela i težina (matrica M i W). Na bazne funkcije M možemo postaviti razna ograničenja, pri ćemu je za aplikacije u automatskoj transkripciji glazbe zanimljivo ograničenje na harmoničnost, vidi [34] [35]. Privlačna karakteristika NMF-a leži i u tome što za razliku od drugih metoda, koje djeluju na jednom okviru i time u potpunosti ignoriraju vremensku komponentu, NMF u obzir uzima vremenski razvoj zvučne slike (težine W ovise o vremenu).

3.3.4 Metode iterativne procjene

Metode iterativne procjene su znatno brže od metoda simultane procjene što je posljedica toga da ne uzimaju u obzir kombinacije tonova, već u svakom koraku izdvajaju jedan ton. Korak ovih metoda ima dvije faze; u prvoj iz spektra biraju najistaknutiji ton, a u drugoj ga uklanjaju kako ne bi smetao idućoj iteraciji metode. Kriteriji biranja najistaknutijeg tona su razni: energija ton kandidata, harmoničnost, broj opaženih parcijala.

Zanimljivija pitanja su pitanje uklanjanja odabranog tona iz spektra i pitanje kriterija zaustavljanja. Pristupe uklanjanju odabranog tona iz spektra djelimo u dvije grupe: izravno uklanjanje i uklanjanje uz korištenje spektralnih modela. Izravno uklanjanje je trivijalan pristup koji iz spektra uklanja sve vrhove koji pripadaju odabranom tonu. Jasno je da ovaj pristup nije prikladan za signale više polifonije. Uklanjanje uz korištenje spektralnih modela, nakon uklanjanja vrhova koji su pripadali odabranom tonu, istima pridružuje nove vrijednosti s ciljem da aproksimira izgled spektra kad u njemu ne bi bilo odabranog tona. Uočimo da preklapani parcijal nekog tona nema točnu vrijednost sa stajališta ni jednog ton modela tonova kojima pripada. To znači da taj parcijal potencijalno narušava glatkoću krivulje spektra tog tona. Zato metoda uklanjanja uz

korištenje spektralnih modela, zaglađuje spektar odabranog tona te ga zatim ukloni iz spektra ostavljajući pritom razliku na mjestima svojih parcijala.

Procjena polifonije

Kriterij zaustavljanja metoda iterativne procjene i dimenzija prostora kod metoda simultane procjene ovise o procjeni polifonije. Kod iterativnih metoda, kriterij zaustavljanja može biti vezan uz veličinu reziduala, tj. preostalog spektra; ako smo u svakom koraku uklanjali točne tonove, u konačnici bi nam trebao ostati samo šum koji ima malu energiju. Naravno, ovaj pristup ima smisla samo ako je SN-omjer dovoljno velik; SN-omjer predstavlja odnos energije signala prema energiji šuma. Dobar kriterij zaustavljanja je i minimalna energija tona; ako ton ima premalu energiju, vjeroatno se ne čuje, a ako smo došli do takvog tona, znači da smo u ranijim iteracijama uhvatili sve tonove koji se čuju. Za metode simultane procjene, možemo postepeno povećavati hipotezu polifonije, sve dok cijena modela ne počne stagnirati. Dok je hipoteza polifonije manja od stvarne polifonije, model će sadržavati premalo tonova da bi uspiješno opisao signal, pa će dodavanje novih tonova (povećavanje hipoteze polifonije) davati bolje rezultate. S druge strane, dodavanje novih tonova u model nakon što se dosegne prava polifonija, neće značajno poboljšavati cijenu modela. Zato ima smisla za procjenu polifonije uzeti prvu vrijednost nakon koje poboljšanje cijene modela postane zanemarivo.

3.3.5 Kombinirane metode

Yeh [1] koristi iterativnu metodu kao sredstvo poboljšanja rada simultane metode. Poboljšanje se sastoji u reduciranju broja ton kandidata, kako bi simultana metoda imala manje kombinacija za provjeriti. Inesta [33] pak koristi iterativnu metodu kako bi preciznije bodovao kombinacije tonova.

3.4 Praćenje nota

Sve ranije navedene metode detekcije fundamentalne frekvencije operiraju na razini jednog okvira, tj. otkrivaju tonove prisutne u veoma kratkom vremenskom intervalu. Idući korak je primjeniti analizu vremenske evolucije; detektirati gdje tonovi počinju, gdje završavaju, gdje je detekcija tona u nekom okviru bila izolirana greška. Prirodni pristup problemu daje HMM; skrivena stanja predstavljaju tonove u svim okvirima, a opzervacije predstavljaju procjenjene fundamentalnih frekvencija u svim okvirima. Ovdje se na "sve okvire" misli samo teoretski; realno se stanja drže u modelu dok god to ima smisla, nakon čega se uklanjaju. U HMM se često ugrađuju posebna ograničenja vezana uz

prirodu problema. Tako Chang razlikuje stanja nastupa i *sustain*-a [30], a Klapuri je u HMM ugradio muzikološki model [29].

Metode koje za MFE koriste NMF nemaju posebne metode za praćenje tonova, nego je analiza vremenske evolucije dio NMF metoda, kao što je opisano u Odjeljku 3.3.3.

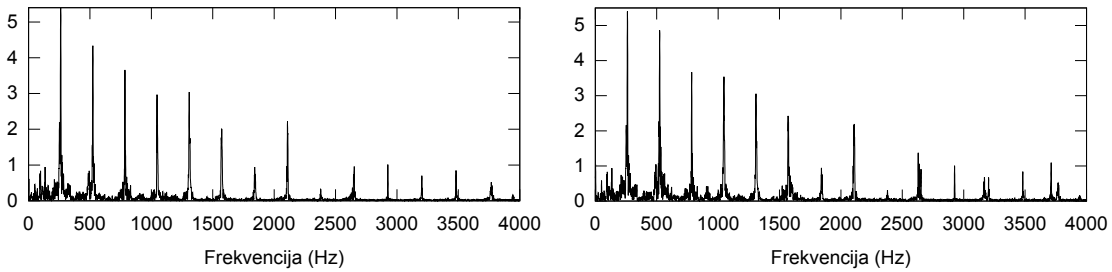
4 Predloženi pristup

4.1 Opis pristupa

Iako pretpostavka o glatkoći spektralnih krivulja ton modela daje dobru heuristiku za razlikovanje "točnih" od "krivih" ton kandidata (krivulje zvukova većine instrumenata doista izgledaju zaglađeno), nju ipak treba uzeti s rezervom. Krivulje nisu baš sasvim glatke. Nadalje, ako želimo odgovoriti na pitanje, sadrži li neki spektar jedan ili dva harmonički vezana tona, glatkoća spektra nam neće biti dovoljni kriterij; Slika 8 prikazuje dva spektra čija glatkoća nije bitno različita. Ipak, lijevi sadrži samo jedan ton, dok desni sadrži dva. Pouzdanije razlikovanje dvaju tonova u desnom slučaju imali bi kada bi nam bio poznat jedan od tonova; tada bismo mogli računati u kojoj mjeri se njegova spektralna krivulja razlikuje u odnosu na referencu. Veća razlika ukazivala bi na prisutnost nekog drugog tona koji interferira s harmonicima polaznog tona i uzrokuje odstupanja.

Jedan korak u ovom smjeru bio bi naučiti ton modele iz baze zvukova instrumenata i zatim koristiti tu bazu ton modela za modeliranje spektralnih krivulja u programu za automatsku transkripciju [1]. No, ton modeli naučeni na ovaj način neće nužno biti slični onima u konkretnoj snimci koju želimo transkribirati. Shodno tome, njihova korisnost je upitna; ton model iz baze najsličniji opaženom spektru daje veći argument tezi da je opaženi spektar gladak, nego što govori da opaženi spektar predstavlja samo jedan ton.

Predloženo riješenje uči ton modele iz snimke koju transkribira, koristeći dva prolaza kroz snimku. U prvom prolazu, metoda uočava djelove snimke koji su dovoljno "čisti" i na njima uči ton modele. Ovdje "čisti" dio snimke predstavlja onaj koji ima nisku polifoniju, visok SN-omjer i slična svojstva koja ukazuju na to da su tonovi na tom djelu snimke minimalno onečišćeni interferencijama. U drugom prolazu se vrši sama transkripcija iterativnom metodom koristeći naučene ton modele.



Slika 8: *Spektar tona C4 (lijevo) i spektar mješavine tonova C4 i C5 (desno)*

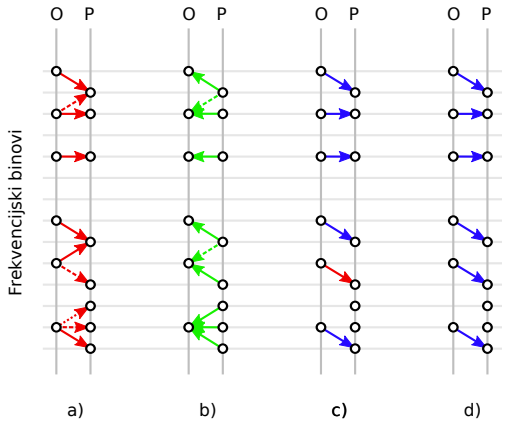
Kako se ton modeli uče iz same snimke, postavlja se pitanje njihove vjerodostojnosti; možda su oni onečišćeni šumom ili interferencijom. Problem rješavamo koristeći prikladno sučelje koji procjenjuje pouzdanost svakog spektralnog vrha korištenjem praćenja vrhova (*peak tracking*). Ta informacija u vidu funkcije pouzdanosti daje mjeru koliko se nekom vrhu, harmoniku, a samim time i tonu, može vjerovati, te se metoda usmjerava na pouzdanije harmonike.

Uobičajeni pristupi računanju razlike između ton modela i ton kandidata uglavnom nisu fleksibilni. Greška se računa kao suma grešaka po harmonicima, pri ćemu se ne uzima u obzir okolina pojedinog harmonika. Kako bismo povećali moć generalizacije naučenih ton modela i tako ih maksimalno iskoristili, korisimo posebnu metodu računanja grešaka, koja daje prednost lokalnoj aproksimaciji. Povrh toga, spomenuta metoda nastoji pronači onu aproksimaciju koja maksimalno griješi na harmonicima koji konstituiraju uzorak (parni harmonici, harmonici djeljivi s tri, harmonici djeljivi s dva i tri), a minimalno na svim drugima, na taj način isključujući interferenciju iz računanja greške.

Zadnja novina vezana je uz način na koji se ton kandidati formiraju. Gotovo svi iterativni pristupi (a i mnogi simultani) koriste neku heuristiku za izbor harmonika ton kandidata. Idući harmonik biraju prema tome koliko se od prethodnog razlikuje u frekvenciji, amplitudi ili odstupanju od teoretske pozicije. Kako izbor točnih vrhova, tj. formiranje što boljih ton kandidata ima presudan utjecaj za daljnji rad svake metode MFEa, zaključujemo da je u ovoj fazi prerano uvesti takve heuristike. S druge strane, s obzirom da se broj svih mogućih ton kandidata redovito mjeri u stotinama tisuća, ton kandidate reprezentiramo grafovima koristeći relaciju uvjetovanosti.

4.2 Sučelje

Praćenje vrhova za primjene u analizi zvuka istraživao je Serra [36]. Njegov model promatra vrhove između dva sukcesivna okvira i pronalazi koje pridruživanje je najvjeroatnije (koji vrh iz početnog okvira odgovara kojem u idućem). Predloženi model dopušta nestajanje vrhova na nekoliko okvira (što je čest slučaj, pogotovo ako je prisutno udaranje parcijala - *partial beating*), a vrhove među okvirima prati na osnovu sličnosti u frekvenciji i amplitudi. Rad algoritma ilustrira Slika 9. Algoritam prvo (podslika *a*) za svaki opaženi vrh (linija označena s O) traži kandidate među vrhovima koji su praćeni u ranijim koracima (linija označena s P), pri ćemu je dopuštena razlika jedan frekvencijski koeficijent (maksimalna promjena u frekvenciji koju vrh može napraviti između dva sukcesivna okvira). Razlog za tako jako ograničenje leži u tome što bi veća margina, uz dopuštanje nestajanja vrhova među okvirima, dovela do težeg optimizacijskog problema



Slika 9: *Rad algoritma za praćenje vrhova; poželjnost kandidata je sugerirana iscrtkanošću (poželjniji su manje iscrtkani)*

i manje pouzdanosti rješenja. Pronađeni kandidati se zatim rangiraju po poželjnosti; bliži po frekvenciji i amplitudi su poželjniji. Zatim svaki od vrhova praćenih u ranijim koracima formira svoju listu kandidata među opaženim vrhovima, koristeći isti kriterij (podslika *b*). Ukoliko prvi kandidat opaženog vrha ima upravo njega kao najpoželjnijeg kandidata (opaženi i praćeni vrh su se međusobno odabrali), taj opaženi vrh postaje dio putanje tog praćenog vrha (plave linije, podslika *c*). Nepridruženi opaženi i pamćeni vrhovi traže idućeg kandidata; proces se nastavlja dok svi vrhovi nisu ili pridruženi ili terminirani ili preskočeni (podslika *d*).

Glavna motivacija za praćenje vrhova leži u statistikama koje se o njima računaju, a iz kojih je moguće odrediti pouzdanost pojedinog vrha. Za svaki vrh računamo težinsku aritmetičku težinu vrijednosti vrha i težinsku aritmetičku sredinu derivacije vrijednosti vrha po vremenu, u okolini okvira, za svaki okvir i svaki vrh. Okolina je definirana promjerom jezgre, a težine reflektiraju udaljenost od njegovog središta. Jezgra je opisana sljedećom funkcijom:

$$\omega(k) = \sqrt{\cos\left(\frac{x}{r}1.515\right)},$$

gdje je r radius jezgre. U radu je korišten $r = 4$. Pouzdanost vrha $c_p(t)$ računamo kao:

$$c_p(t) = \sqrt{\frac{\log(\psi_p(t)\beta(m))}{n_c}},$$

$$\psi_p(t) = \frac{m_p(t)}{s_p(t) + z_c}\alpha_p(t) + u_c(1 - \alpha_p(t)),$$

$$m_p(t) = \frac{1}{\omega_p(t)} \sum_{k=-r}^{r} \kappa_p(t+k)\omega(k)a_p(t+k),$$

$$s_p(t) = \frac{1}{\omega_p(t)} \sum_{k=-r}^{r} \kappa_p(t+k)\omega(k)a_p(t+k)',$$

$$\omega_p(t) = \sum_{k=-r}^{r} \kappa_p(t+k)\omega(k),$$

$$\alpha_p(t) = \frac{\omega_p(t) - 1}{\Omega - 1},$$

$$\beta(m) = \begin{cases} 1 & : \quad m \geq l_c \\ \frac{m}{l_c} & : \quad m < l_c \end{cases},$$

gdje je p oznaka vrha, t oznaka okvira, $m_p(t)$ težinska aritmetička sredina vrijednosti vrha p u okviru t, $s_p(t)$ težinska aritmetička sredina derivacije po vremenu, $\omega_p(t)$ ukupna suma težina, $\kappa_p(t)$ indikator prisutnosti vrha p u okviru t, $a_p(t)$ vrijednost vrha p u okviru t, Ω maksimalna teoretska suma težina ($\int_{-\infty}^{\infty} \omega(k)\mathrm{d}k$), α funkcija koja vrednuje kontinuiranost prisutstva vrha u vremenu (vrh koji je prisutan svaki drugi okvir je manje pouzdan od onog koji je prisutan svaki okvir), β funkcija koja reducira pouzdanost vrhova niske amplitude, a n_c, z_c, l_c i u_c konstante. Uloga omjera $\frac{m_p(t)}{s_p(t)}$ jest da predoči SN-omjer signala; signal čija derivacija se malo mijenja u odnosu na amplitudu je stabilan. Konstanta z_c izbjegava djeljenje s nulom. Smisao funkcije $\alpha_p(t)$ i konstante u_c leži u tome da uvede kontrolu nad vrhovima koji nisu kontinuirani u vremenu; želimo reducirati značaj takvih vrhova, ali ne pretjerano - u nekim situacijama (partial beating) nekontinuiranost vrha nije poslijedica njegove nepouzdanosti. Konstanta l_c je odabrana tako da bude nešto veća od prve bočne latice (*sidelobe*-a) korištenog prozora. Logaritam u funkciji pouzdanosti $c_p(t)$ reducira raspon vrijednosti; vrh s omjerom $\frac{m}{s}$ od 2 i vrh s omjerom tih vrijednosti od 4 se razlikuju 2 puta i to je razlika koja će nas zanimati, jer je drugi vrh značajno pouzdaniji od prvog (iako su oba nepouzdana). S druge strane vrh s spomenutim omjerom od 100 i drugi s omjerom 200 se također razlikuju duplo, ali ta razlika nas neće zanimati; oba vrha su veoma pouzdana. Konačno, praktično je pouzdanost ograničiti odozgo, osobito na način da pada u interval $[0, 1]$; tomu služi konstanta n_c.

Za izračun spektra koristimo STFT s Blackman prozorom:

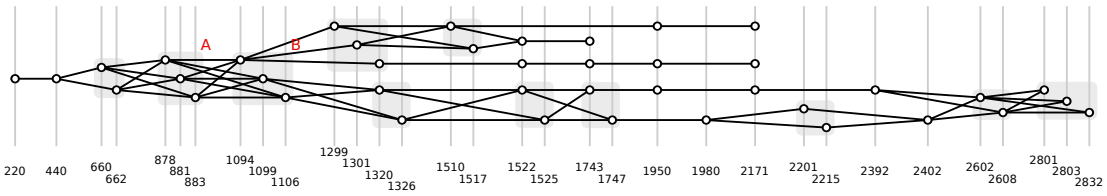
$$b(n) = \frac{1-\alpha}{2} - \frac{1}{2}\cos\left(\frac{2\pi n}{N-1}\right) + \frac{\alpha}{2}\cos\left(\frac{4\pi n}{N-1}\right),$$

pri ćemu je $\alpha = 0.16$, a N je zadan širinom okvira koja iznosi 93ms. Korak STFTa je širine 3ms; ovako kratak korak uvjetuje metoda praćenja vrhova. Promjer jezgre je tada 28% širine prozora; u tom intervalu se računaju statistike vrhova. Kao mjeru interpolacije, koristimo nadjevanje nulama s faktorom osam. Preprocesiranje uključuje prijenos snimke sa stereofonskog formata u monofonski korištenjem aritmetičke sredine kompleksnog spektra i normalizaciju signala. Za izlaz sučelja, broj okvira se reducira četiri puta, tako da efektivni korak u ostatku algoritma iznosi 12ms. Na spektar primjenjujemo $\log(x + 1)$ kompresiju.

4.3 Ton kandidati

Raspon frekvencija u kojima tražimo ton kandidate je 50-2100Hz. Koristimo pretpostavku o harmoničnosti. Uz to da dopuštamo da prva tri harmonika ne budu prisutna, pri ćemu će takvi (krnji) kandidati biti dodani samo ako ne postoje regularni kandidati s istom fundamentalnom frekvencijom. Što se biranja harmonika tiče, koristimo Yehov kriterij (prezentiran u Odjeljku 3.3.2); oslanjamo se na stabilnost intervala između harmonika. Tu dopuštamo preskakanje do dva harmonika odjednom. No, koristimo još jedan kriterij; dok stabilnost intervala mjeri relativno odstupanje sukcesivnih vrhova, malo toga govori o globalnoj slici krivulje inharmoničnosti. Tu krivulju ograničavamo s formulom (2) pri ćemu za koeficijent β uzimamo 0.0004. Za taj kriterij ne dopuštamo odstupanja od dva i više sukcesivnih vrhova - tada bi krivulja inharmoničnosti postala previše iskrivljena.

Za opis svih ton kandidata za jednu fundamentalnu frekvenciju koristimo graf. Graf ima jedan početni čvor (koji može odgovarati prvom harmoniku, ako je prisutan), i mnogo vrhova koji odgovaraju raznim kandidatima za harmonike. Put kroz graf, od početnog čvora do nekog završnog čvora predstavlja jedan ton kandidat. Slika 10 ilustrira jedan takav graf. Tu vidimo sve fenomene koji se u ovakvom grafu mogu dogoditi; putevi se mogu križati, razdvajati u neovisne puteve i ponovno spajati. Uočimo da izbor nekih vrhova ne uvjetuje izbor idućih vrhova; na Slici 10, uz izbor bilo kojeg vrha iz grupe



Slika 10: *Graf svih ton kandidata fundamentalne frekvencije F0=220Hz*

vrhova označene s **A** možemo doći u bilo koji vrh iz grupe vrhova označene s **B**. S druge strane, izbor vrha frekvencije 1094, uvjetuje izbor vrha frekvencije 1299. To nas vodi na relaciju uvjetovanosti:

Definicija 1. *Dva vrha p_1, p_2, koji su kandidati za harmonike redom h_1 i h_2, $h_1 < h_2$ su u relaciji uvjetovanosti ako svaki ton kandidat koji prolazi kroz p_1, mora prolaziti i kroz p_2.*

Definicija se prirodno proširuje na grupe vrhova. Nezavisnom grupom vrhova smatramo skup vrhova koji konkuriraju za isti harmonik. Jasno je da broj ton kandidata koje neki graf sadrži općenito ovisi eksponencijalno o broju vrhova (pogotovo uz veći broj neuvjetovanih vrhova), što nije prikladno za pohranu u memoriji računala. Također, obilazak svakog puta posebno nema smisla ni sa stajališta vremenske složenosti. Uz relaciju uvjetovanosti moguće je primjeniti ideju dinamičkog programiranja i segmente između dviju neuvjetovanih grupa gledati kao neovisnu cjelinu. Ovakav pristup osigurava iznimne uštede u vremenskoj i prostornoj složenosti uz zadržavanje fleksiblinosti do samog trenutka klasifikacije ton kandidata.

4.4 Metoda klasifikacije

Zadatak klasifikacije je iz skupa ton modela pronaći onaj koji je najbliži ton kandidatu ili zaključiti da takvog nema. Tri su osnovna razloga zašto klasifikacija može ne uspijeti. Prvi i najmanje zanimljiv je preveliki šum i/ili prevelika interferencija. Drugi, zanimljiviji počiva na opažanju da spektar tona ovisi o glasnoći, stilu sviranja i raznim drugim faktorima. Neke od tih varijacija (primjerice glasnoća) ne mijenjaju lokalnu strukturu tona (detalje), ali mijenjaju globalni izgled spektra. Zato ima smisla, umjesto uspoređivanja pojedinih harmonika ton modela s ekvivalentnim harmonicima ton kandidata, uspoređivati relativne odnose harmonika ton modela s relativnim odnosima istih harmonika ton kandidata (primjerice, razlika drugog i prvog harmonika ton kandidata prema razlici drugog i prvog harmonika ton modela). Treći i najzanimljiviji razlog zašto klasifikacija može ne uspijeti je interferencija s drugim harmonički vezanim ili harmonički nevezanim tonovima. Interferencija harmonički nevezanih tonova može se manifestirati na harmonicima višekratnicima najmanjeg zajedničkog višekratnika fundamentalnih frekvencija tonova. U tom slučaju situacija je istovjetna interferenciji s harmonički vezanim tonom pripadne fundamentalne frekvencije. Ako se pak ne manifestira na taj način (zbog razlike u inharmoničnosti ili nekog drugog razloga), greška će biti slučajno distribuirana. Vratimo se stoga interferenciji harmonički vezanih kandidata. Interferencija će uzrokovati da se ton kandidat razlikuje od svog točnog ton modela na točno određenim

mjestima - višekratnicima omjera frekvencija. Drugim riječima ako je omjer frekvencija dva (ton interferira sa svojom višom oktavom), greške će se javljati na parnim harmonicima. Ovo svojstvo nam otvara zanimljivu perspektivu; kad bi mogli prepoznati uzorak u kojem je greška distribuirana, mogli bi računanje razlike između ton modela i ton kandidata ograničiti na one harmonike u kojima nema interferencije. U protivnom ne računamo samo grešku između ton kandidata i ton modela, nego i razinu interferencije.

Jedan pristup tom problemu je tražiti najmanje relativne greške. Za prvi harmonik ton kandidata gledamo njegove relativne razlike prema idućih nekoliko harmonika. Isto radimo i za ton model. Harmonik u kojem je greška tih razlika između ton kandidata i ton modela najmanja, označimo kao najpouzdaniji u okolini prvog harmonika i koristimo kao polaznu točku za iduću usporedbu. Označene harmonike zvat ćemo markeri. Ako je ton model veoma sličan ton kandidatu, markeri će biti svi harmonici. Ako ton kandidat ima interferenciju s višom oktavom, markeri će biti na neparnim harmonicima. Idući korak je odrediti formiraju li greške nekakav uzorak. U tu svrhu više ne možemo koristiti greške relativnih odnosa harmonika, nego trebamo grešku po harmoniku. Prvi korak računanja greške harmonika je pronaći najbliži lijevi i desni marker. Zatim markere proglasimo točnima vrhovima i spektar ton kandidata nagnemo (*skew*) tako da vrijednosti vrhova na markerima odgovaraju vrijednostima istih harmonika ton modela. Zatim računamo apsolutne greške između vrhova ton modela i modificiranih vrhova ton kandidata. Kad imamo sve greške svih harmonika, uzorke tražimo uspoređujući aritmetičke sredine vrijednosti prikladnih skupova; za otkrivanje HRF0 dvostruke fundamentalne frekvencije, usporedit ćemo aritmetičke sredine parnih i neparnih harmonika. Za otkrivanje HRF0 trostruke fundamentalne frekvencije, usporedit ćemo aritmetičku sredinu harmonika djeljivih s tri, sa aritmetičkom sredinom harmonika koji nisu djeljivi s tri. Tu valja uočiti da će i HRF0 sa četverostrukom fundamentalnom frekvencijom uzrokovati velike razlike među aritmetičkim sredinama parnih i neparnih harmonika. Da bi izbjegli pogrešnu klasifikaciju, za određivanje HRF0 s faktorom k, osim omjera aritmetičkih sredina harmonika djeljivih s k i nedjeljivih s k, koristimo i omjer aritmetičkih sredina harmonika koji se nalaze u skupu $\mathcal{S} = \{k + 2ki \mid i = \{1, \ldots, n\}\}$, s onima koji se nalaze u skupu $\mathcal{S}^{\mathsf{c}} \setminus \{2ki \mid i = \{1, \ldots, n\}\}$. Kombinacije hipoteza o interferencijama ne razmatramo posebno, jer su međusobni utjecaji uzoraka raznih HRF0-a relativno mali.

Uočimo da je ovakav pristup klasificiranju nekonzistentan za ton kandidate koji se značajno razlikuju od ton modela zbog relativnog pristupa odabiru markera. No, takve slučajeve je lako prepoznati, jer su tada greške na markerima velike; u tom slučaju traženje uzoraka nema ni smisla, pa se izračun prekida, a ton model označuje kao neprikladan. Ukoliko su pak greške na markerima male, ima smisla proučiti distribuciju

grešaka na ostalim harmonicima.

Pouzdanost vrha igra veliku ulogu u svim izračunima. Kod računanja aritmetičkih sredina harmonika (prilikom traženja uzoraka), pouzdanosti vrhova se koriste kao težinski faktori, pa se u stvari radi o težinskim aritmetičkim sumama, pri ćemu je težina greške harmonika minimum pouzdanosti harmonika ton modela i harmonika ton kandidata. Kod traženja markera, greške relativnih razlika se djele minimumom pouzdanosti svih četriju vrhova koji sudjeluju u izračunu.

4.5 Ton modeli

U prvom prolazu kroz snimku, iz snimke se izvlače ton kandidati, te se krajem prvog prolaza orginalni zapis snimke napušta; drugi prolaz operira samo na ton kandidatima i ne koristi orginalnu snimku. Nakon što se iz pojedinog okvira izvuku svi mogući ton kandidati, oni prolaze kroz nekoliko eliminatornih koraka, u kojima se odbacuju oni koji imaju premalu energiju, premali broj vrhova ili premalu prosječnu pouzdanost. Preostali ton kandidati se čuvaju za drugi prolaz, a iz njih se biraju i kandidati za učenje ton modela. Izbor kandidata za učenje ton modela donosi još dvije provjere. Prva se sastoji od četiri kriterija: prosječna pouzdanost, omjer opaženih harmonika od teoretski mogućih, udio energije u ukupnoj energiji okvira, omjer broja opaženih harmonika u odnosu na broj deset. Ovi kriteriji prednost daju ton kandidatima koji imaju jak SN-omjer, koji pripadaju niskoj polifoniji, koji imaju pouzdane vrhove i velik broj opaženih parcijala.

Drugi kriterij se bavi odnosom (interferencijom) ton kandidata; iako u ovom trenutku ne možemo pouzdano izbjeći ton kandidate koji interferiraju s HRF0-ima, ipak možemo primjeniti jeftinu heuristiku kako bi izbjegli barem očite slučajeve. Situacije koje želimo izbjeći su one u kojima interferirajući HRF0 ima znatno veću energiju od našeg ton kandidata. Takve je lako prepoznati gledajući prosječnu amplitudu harmonika; uzimamo prvih deset harmonika ton kandidata koji nas zanima i proučavamo ton kandidate čija fundamentalna frekvencija je višekratnik one odabranog ton kandidata i počinju vrhom koji pripada grafu odabranog ton kandidata. Tada za svaki takav ton kandidat računamo prosječnu amplitudu prvih onoliko harmonika koliko ih ulazi u raspon prvih deset harmonika odabranog ton kandidata (intervali moraju biti isti, inače vrijednosti nema smisla uspoređivati).

Ton kandidati koji su prošli oba kriterija se prosljeđuju klasifikatoru. Ako je sličan ton model već naučen, ton kandidat se može koristiti za poboljšanje ton modela; ukoliko neki harmonici ton kandidata imaju veću pouzdanost od pripadnih harmonika ton modela, vrijednosti se kopiraju. Ako klasifikacija nije uspijela (sličan ton model ne postoji),

ton kandidat se uči.

Razmotrimo sljedeću situaciju; klasifikator u nekom okviru nauči ton model koji sadrži interferirajući HRF0. Interferirajući HRF0 nije veoma dominantan, pa ga ranije opisana preselekcija ne odbaci. Pitanje je kako se riješiti tog ton modela, koji je očito kriv. Jasno, da smo prvo naučili ton model i kasnije pokušali naučiti taj sličan ton model, ali onečišćen interferencijom, on bi bio odbačen. No, kako je redosljed sada obrnut, pitanje je kako prepoznati i ukloniti sporni ton model. Odgovor je jednostavan; razumno je očekivati da će se kad tad kao kandidat za ton model pojaviti pripadan neinterferirani ton model. Tokom postupka klasifikacije, on će neminovno biti uspoređen s pogrešno naučenim ton modelom. Greška među tim modelim imat će uzorak pripadne višekratnosti, ali s negativnim predznakom, jer ton model griješi u odnosu na ton kandidat. U tom slučaju, ton model se briše, a ton kandidat dodaje u bazu. Tu je jasno da se baza ton modela kontinuirano konsolidira (dobri ton modeli se poboljšavaju, loši odbacuju), te je izgledno da će u konačnici sadržavati samo dobre ton modele.

U drugom prolazu kroz snimku, ton kandidati se klasificiraju koristeći naučene ton modele. Pri tome se nakon klasificiranja svih tonova u okviru primjenjuju još dva reduktivna koraka; ton čija prosječna energija markera je znatno veća od energije prvog harmonika se odbacuje - time se izbjegavaju subharmoničke greške. Također se uspoređuju prosječne energije markera ton kandidata koji su harmonički vezani; ako ne postoje značajne razlike, viši se odbacuje; ako su razlike jako velike, tiši se odbacuje.

4.6 Praćenje tonova

Zadnja faza je praćenje tonova među okvirima. Koristimo transformaciju signala konvolucijom s pravokutnom jezgrom promjera sedam okvira. Ukoliko je ton opažen u barem četiri, smatramo ga prisutnim. Ukoliko je prisutan u tri i manje, prethodni okvir predstavlja kraj tog tona. Još jedan kriterij koji može uzrokovati kraj tona je promjena u energiji tona; ako ton u nekom okviru ima znatno veću energiju nego u zadnjem opaženom okviru, zadnji opaženi okvir predstavlja kraj tog tona, a ovaj okvir novi početak tog tona. Tu energija tona djeluje kao harmonička funkcija detekcije.

Da bi konvolucija imala smisla, frekvencije tonova zaokružujem na dobro ugođenu glazbenu skalu s referencom $A4 = 440Hz$.

5 Rezultati

Glavni problem vezan uz evaluaciju metoda glazbene transkripcije je nalaženje pouzdane referentne transkripcije. Iako problem ima više složenih riješenja, najjednostavnije je krenuti od dobre referentne transkripcije. U tu svrhu se redovito koriste MIDI datoteke koje sadrže simbolički glazbeni zapis, a mogu se naći na internetu (*http://www.free-scores.com/*). Idući korak je od simboličkog zapisa dobiti audio zapis (snimku); postupak se zove generiranje audio zapisa (*rendering*), a uključuje dodjeljivanje zvukova pojedinim MIDI događajima (tonovima). Zvukovi mogu biti sintetizirani (izbor je širok i uključuje razne sintetizatore - *synthesizer*) ili reproducirani koristeći uzorke snimki stvarnih instrumenata (*patch*evi). U tu svrhu koristimo glazbene fontove (*fluid soundfonts*).

Računamo preciznost (p), odziv (o), F-mjeru (F) i točnost rezultata (a):

$$p = \frac{TP}{TP + FP}$$

$$o = \frac{TP}{TP + FN}$$

$$F = 2\frac{p\,o}{p + o}$$

$$a = \frac{TP}{TP + FP + FN},$$

pri ćemu je TP broj točno klasificiranih tonova, FP broj tonova koji su detektirani iako ih referentna transkripcija ne sadrži, te FN broj tonova koji nisu detektirani, a prisutni su u referentnoj transkripciji. Konačne vrijednosti dobivene su aritmetičkom sredinom vrijednosti po okvirima, pri ćemu okvire u kojima nema tonova (ni detektiranih ni u referentnoj transkripciji) zanemarujemo. Osim ukupnih rezultata, zanimali su nas i rezultati prema polifoniji. Tablica 1 prikazuje rezultate po glazbenim djelima.

Tablica 1: Rezultati prema glazbenim djelima

Glazbeno djelo	**sastav**	**preciznost**	**odziv**	**F-mjera**	**točnost**
J. S. Bach, Prelude, BWV 997	klavir	0.656	0.845	0.738	0.604
F. Devienne, Devienne sonata, opus 71, no 3	klarinet, klavir	0.806	0.776	0.79	0.653
W. A. Mozart, Ave Verum Corpus	gudački kvartet	0.752	0.812	0.781	0.637
T. Albinoni, Adagio	orgulje, violina	0.517	0.66	0.58	0.415
J. S. Bach, Brandenburg Concerto No.2 u F-duru, drugi stavak	violina, oboa, flauta, kontrabas	0.797	0.78	0.789	0.665
prosječno		0.706	0.775	0.736	0.595

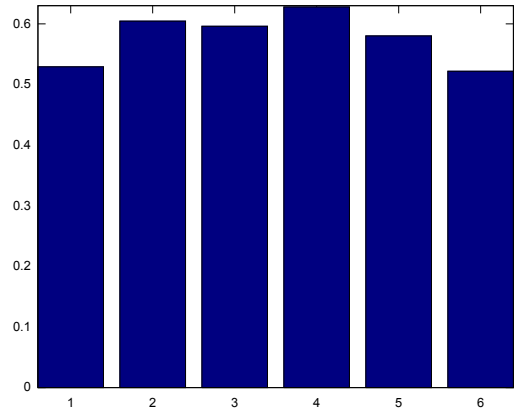
Nešto slabiji rezultati na snimci Albinonijevog Adagia, poslijedica su izražene modulacije u amplitudi koju uključeni instrumenti posjeduju.

Usporedba s rezultatima **MIREX** natjecanja iz 2010. godine, danih u Tablici 2, ilustrira uspiješnost predložene metode; predloženi pristup je usporediv sa suvremenim metodama u području detekcije tonova iz polifonih snimki.

Tablica 2: Rezultati MIREX 2010

Oznaka kandidata	**preciznost**	**odziv**	**točnost**
AR1	0.721	0.799	0.654
AR2	0.722	0.799	0.655
AR3	0.752	0.829	0.692
AR4	0.753	0.828	0.692
BD1	0.716	0.485	0.468
CRVRC1	0.638	0.556	0.490
CRVRC3	0.515	0.637	0.457
DCL1	0.499	0.676	0.457
DHP1	0.712	0.632	0.553
JW1	0.369	0.579	0.354
JW2	0.384	0.541	0.361
LYLC1	0.584	0.399	0.373
NNTOS1	0.142	0.061	0.060

Slika 11 prikazuje rezultate po polifoniji; vidimo da se metoda dobro nosi s većim polifonijama.



Slika 11: *Točnost metode prema polifoniji*

6 Zaključak

Glazbena transkripcija je složen problem koji se zadovoljavajuće može riješiti jedino koristeći kontekst i iskustvo. Zbog mnogobrojnih primjena glazbene transkripcije, razvijene su metode automatske glazbene transkripcije koje nastoje riješavati pojedine probleme kao što su detekcija nastupa tonva i detekcija tonova. Temeljni problemi glazbene transkripcije uključuju šum i interferenciju među tonovima.

Predložena metoda daje obečavajuće rezultate koristeći nenadzirano učenje ton modela kako bi minimizirala pogreške koje uzrokuje interferencija.

7 Literatura

[1] Chunghsin Yeh, Multiple fundamental frequency estimation of polyphonic recordings, PhD thesis, Université Paris VI - Pierre et Marie Curie, 2008

[2] Anssi Klapuri, Multiple Fundamental Frequency Estimation Based on Harmonicity and Spectral Smoothness, IEEE transactions on speech and audio processing, vol. 11, No. 6, november 2003

[3] Masataka Goto, A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals, Speech Communication 43, 311–329, 2004

[4] Emmanuel Vincent and Xavier Rodet, Music Transcription with ISA and HMM, IRCAM, Analysis-Synthesis Group, 2004

[5] S. G. Mallat and Z. Zhang, Matching Pursuits with Time-Frequency Dictionaries, IEEE Transactions on Signal Processing, December 1993, pp. 3397-3415

[6] Pierre Leveau, David Sodoyer and Laurent Daudet, Automatic instrument recognition in a polyphonic mixture using sparse representations, Proceedings of the International Symposium on Music Information Retrieval (ISMIR'07), Vienna (2007)

[7] Scott Shaobing Chen, David L. Donoho and Michael A. Saunders, Atomic decomposition by basis pursuit, SIAM Journal on Scientific Computing, vol. 20, 33-61, 1998

[8] R. Gribonval, Ph. Depalle, X. Rodet, E. Bacry and S. Mallat, Sound signals decomposition using a high resolution matching pursuit, Proc. International Computer Music Conference (ICMC'96), p 293-296, 1996

[9] Laurent Daudet, Sparse and Structured Decompositions of Signals With the Molecular Matching Pursuit, IEEE transactions on audio, speech, and language processing, vol. 14, no. 5, september 2006

[10] Rémi Gribonval and Emmanuel Bacry, Harmonic Decomposition of Audio Signals With Matching Pursuit, IEEE transactions on signal processing, vol. 51, no. 1, january 2003

[11] M. Abe and J. O. Smith, AM/FM rate estimation for time-varying sinusoidal modeling, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05), volume 3, iii/201–iii/204, Philadelphia, PA, USA., (2005)

[12] Judith C. Brown, Calculation of a constant Q spectral transform, J. Acoust. Soc. Am., 89(1):425–434, 1991.

[13] Hirokazu Kameoka, Takuya Nishimoto and Shigeki Sagayama, A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering, IEEE transactions on audio, speech and language processing, vol. 15, no. 3, march 2007

[14] E. Benetos and S. Dixon, Multiple-F0 estimation of piano sounds exploiting spectral structure and temporal evolution, in Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition, pp. 13-18, Sep. 2010.

[15] Antti Eronen, Automatic musical instrument recognition, Master of science thesis, Tampere university of technology, Department of Information Technology

[16] H. Fletcher, E.D. Blackham and R. Stratton, Quality of piano tones, Journal of the Acoustical Society of America, vol. 34, no. 6, pp. 749-761, 1962

[17] Davide Rocchesso and Francesco Scalcon, Bandwidth of Perceived Inharmonicity for Physical Modeling of Dispersive Strings, IEEE Transactions on Signal Processing 7(5), pp. 597-601, 1999

[18] Fletcher and Rossing, The Physics of Musical Instruments (2nd edition), Springer-Verlag New York, Inc., 1998.

[19] B. C. J. Moore, R. W. Peters and B. C. Glasberg, Thresholds for the detection of inharmonicity in complex tones, Journal of the Acoustical Society of America, vol. 77, no. 5, pp. 1861-1867, 1985

[20] I. Peretz and M. Coltheart, Modularity of music processing, Nature Neuroscience, 6(7), July 2003.

[21] Thomas D. Rossing, Acoustics of percussion instruments: Recent progress, Acoustical Science and Technology, Vol. 22 (2001) , No. 3 pp.177-188

[22] A. W. Schloss, On the Automatic Transcription of Percussive Music From Acoustic Signal to High-Level Analysis, Ph.D. dissertation, Tech. Rep. STAN-M-27, Dept. Hearing and Speech, Stanford Univ., Stanford, CA, 1985.

[23] P. Masri, Computer Modeling of Sound for Transformation and Synthesis of Musical Signal, Ph.D. dissertation, Univ. of Bristol, Bristol, U.K., 1996.

[24] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, A combined phase and amplitude based approach to onset detection for audio segmentation, in Proc. 4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03), London, U.K., Apr. 2003, pp. 275–280.

[25] Chris Duxbury, Juan Pablo Bello, Mike Davies and Mark Sandler, Complex domain onset detection for musical signals, Proc. of the 6th Int. Conference on Digital Audio Effects (DAFx-03), London, UK, September 8-11, 2003

[26] A. de Cheveigné and H. Kawahara, YIN, a fundamental frequency estimator for speech and music, Journal of the Acoustical Society of America 111, 1917–1930., (2002)

[27] R. C. Maher and J. W. Beauchamp, Fundamental frequency estimation of musical signals using a two-way mismatch procedure, J. Acoust. Soc. Am. 95 (4), 2254–2263, 1994.

[28] C. Duxbury, M. Sandler and M. Davies, A hybrid approach to musical note onset detection, in Proc. Digital Audio Effects Conf. (DAFX,'02), Hamburg, Germany, 2002, pp. 33–38.

[29] M. Ryynänen and A. Klapuri, Polyphonic music transcription using note event modeling, in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'05), Mohonk, NY, USA. (2005)

[30] Wei-Chen Chang, Alvin W. Y. Su, Chunghsin Yeh, Axel Roebel and Xavier Rodet, Multiple-f0 tracking based on a high-order HMM model, Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland, September 1-4, 2008

[31] Valentin Emiya, Roland Badeau and Bertrand David, Automatic transcription of piano music based on hmm tracking of jointly-estimated pitches, in Proc. Eur. Conf. Sig. Proces. (EUSIPCO), Lausanne, Switzerland, Aug. 2008.

[32] Zhiyao Duan, Bryan Pardo and Changshui Zhang, Multiple Fundamental Frequency Estimation by Modeling Spectral Peaks and Non-peak Regions, Audio, Speech, and Language Processing, IEEE Transactions on vol. 18, no. 8, pp. 2121-2133, Nov. 2010

[33] Antonio Pertusa, José M. Iñesta, Multiple fundamental frequency estimation using gaussian smoothness, Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, no. March 31 2008-April, pp. 105-108

[34] E. Vincent, N. Bertin and R. Badeau. Harmonic and inharmonic nonnegative matrix factorization for polyphonic pitch transcription. In IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008, pages 109–112, 2008.

[35] S. A. Raczynski, N. Ono and S. Sagayama, Multipitch analysis with harmonic nonnegative matrix approximation, in Proc. of 8th International Symposium on Music Information Retrieval. (2007)

[36] X. Serra, A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. Ph.D. thesis.(Stanford University), 1989

Kazalo

auditorni model — model koji opisuje kako auditorni sustav čovijeka obrađuje zvuk

decibel — logaritam kvocijenta vrijednosti koji je spretan za izražavanje odnosa dviju vrijednosti kad se one mogu značajno razlikovati u veličini

DFT — Discrete Fourier Transform - diskretna fourierova transformacija

FFT — Fast Fourier Transform - algoritam koji računa diskretnu fourierovu transformaciju signala

frekvencija uzorkovanja — učestalost kojom se iz kontinuiranog signala uzimaju uzorci u svrhu dobivanja diskretnog signala

frekvencijska domena — prikaz signala kao funkcije frekvencije

frekvencijski pojas — dio spektra koji promatramo

fundamentalna frekvencija — frekvencija koja se u psihoakustičkom eksperimentu pridružuje tonu

harmonik — parcijal tona koji je okvirno višekratnik fundamentalne frekvencije

inharmoničnost — odstupanje harmonika od teoretske pozicije

MFE — Multiple F0 Estimation - procjena višestrukih fundamentalnih frekvencija, zadatak izračuna fundamentelnih frekvencija tonova prisutnih u signalu, pri čemu se dopušta veći broj simultanih tonova (polifonija)

monofono — u jednom trenutku se može čuti najviše jedan ton. Monofonija je prirodno vezana uz snimku jednog instrumenta

nadjevanje nulama — metoda interpolacije spektra koja se sastoji u proširivanju ulaznog vektora DFTa nulama

nastup tona — *onset*, početni dio glazbenog događaja tokom kojeg energija raste od nule do maksimuma

NMF	Non-negative matrix factorisation - nenegativna matrična faktorizacija, faktorizira matricu u produkt nenegativnih matrica
oktava	razlika među tonovima čije fundamentalne frekvencije se razlikuju duplo
okvir	dio signala koji promatramo
parcijal	frekvencija kompleksnog tona koja je različita od fundamentalne
pojasno propusni filter	filter koji propušta samo jedan dio spektra, dok ostatak odbacuje
prozor	funkcija koja se koristi za kontrolu omjera rezolucije spektra u vremenu i frekvenciji
SFE	Single F0 Estimation - procjena jedne fundamentalne frekvencije, zadatak izračuna fundamentalne frekvencije tona, uz pretpostavku da je to jedini ton u signalu
SN-omjer	SN-ratio, signal to noise ratio - odnos signala i šuma
spektralni vrh	lokalan vrh spektra
stacionaran signal	signal čija se statistička svojstva ne mijenjaju kroz vrijeme - intuitivno, signal koji je periodičan
STFT	Short Time Fourier Transform - signal se segmentira u kratke segmente, te se Fourierova transformacija primjenjuje na svaki od njih
sučelje	modul koji prethodi glavnim metodama i koji snimku iz vremenske domene transformira u ciljnu domenu koju traže glavne metode
ton	zvuk kojem je u psihoakustičkom eksperimentu moguće pridružiti frekvenciju
transformacija valića	transformacija signala u sumu valića
tranzient	kratak nalet energije koji nema periodička svojstva
uklanjanje šuma	zadatak reduciranja šuma u signalu
valić	dio periodičkog vala lokaliziran u vremenu

vremenska domena prikaz signala kao funkcije vremena