Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Manca Žerovnik

Kompozicionalni hierarhični model za ocenjevanje osnovnih frekvenc

diplomsko delo

univerzitetni študijski program prve stopnje
računalništvo in informatika

Mentor: doc. dr. Matija Marolt

Ljubljana 2014

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

V diplomski nalogi ocenite zmožnosti kompozicionalnega hierarhičnega modela za pridobivanje informacij iz glasbe pri ocenjevanju osnovnih frekvenc v glasbenih signalih. Model prilagodite za omenjeno nalogo in ga preizkusite na primerni zbirki skladb. Evaluirajte rezultate v odvisnosti od parametrov učenja in ocenjevanja. Z naknadno obdelavo rezultatov poskusite izboljšati natančnost modela.

Izjava o avtorstvu diplomskega dela

Spodaj podpisana Manca Žerovnik, z vpisno številko **63110354**, sem avtor diplomskega dela z naslovom:

Kompozicionalni hierarhični model za ocenjevanje osnovnih frekvenc.

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom doc. dr. Matija Marolta,
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela,
- soglašam z javno objavo elektronske oblike diplomskega dela na svetovnem spletu preko univerzitetnega spletnega arhiva.

V Ljubljani, dne 8. septembra 2014 Podpis avtorja:

Na začetku bi se rada zahvalila vsem, ki so mi pomagali, da je delo nastalo. Zahvaljujem se mentorju, doc. dr. Matiji Maroltu, za vse nasvete in pomoč. Velika zahvala gre asistentu Matevžu Pesku za veliko časa, ki si ga je vzel za nasvete, pomoč, razlage, spodbude, opominjanje in hvala za moralno podporo čez celo leto. Hvala moji družini - staršem ter Jeri, Aleši, Vidu in Poloni - ki me na moji poti tako in drugače podpira že celo življenje. Hvala Primožu za vzpodbude, pomoč in razumevanje. Hvala dragim prijateljem, ki so mi ves čas študija stali ob strani. Hvala vsem, ki ste verjeli vame.

Žerovnikovim.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
MIR	music information retrieval	pridobivanje informacij iz glasbe
AMT	automatic music transcription	samodejna transkripcija glasbe
DBN	deep belief network	mreže globokega zaupanja
PNO	average precision per frame	povprečna natančnost na okvir
PRE	precision	natančnost
REC	recall	priklic
NMF	non-negative matrix factorization	nenegativna matrična faktorizacija
MIDI	musical instrument digital interface	glasbeni instrumentalni digitalni vmesnik

Povzetek

V nalogi se posvečamo področju pridobivanja informacij iz glasbe. Ukvarjamo se z evalvacijo in optimiziranjem delovanja kompozicionalnega hierarhičnega modela na opravilu ocenjevanja osnovnih frekvenc. Model prilagodimo, da vrača rezultate za to opravilo. Ukvarjamo se z interpretacijo nepravilnih hipotez. Delovanje modela s post-procesiranjem rezultatov poskušamo izboljšati. Model preizkusimo na prosto dostopnih podatkovnih zbirkah, da ga lahko primerjamo z ostalimi in na podlagi različnih grafičnih prikazov delovanja iščemo njegove prednosti in slabosti. Na koncu poskusimo model izboljšati z nenegativno matrično faktorizacijo.

Ključne besede: pridobivanje informacij iz glasbe, ocenjevanje osnovnih frekvenc, kompozicionalni hierarhični model, evalvacija.

Abstract

This thesis focuses on the field of music information retrieval. We present the compositional hierarchical model for music information retrieval and evaluate it on the task of multiple fundamental frequency estimation on publicly available MAPS dataset. We evaluate the robustness of the model by varying the parameters of the model and analyse the results through graphical visualizations of model's provided hypotheses. Additionally, we provide improvements to the results through the analysis of the model's output.

Keywords: music information retrieval, multiple fundamental frequency estimation, compositional hierarchical model, evaluation.

Poglavje 1

Uvod

V glasbi je z vidika računalništva še veliko neraziskanega. Področje pridobivanja informacij iz glasbe se je začelo razvijati v osemdesetih letih prejšnjega stoletja in se je do danes že zelo razširilo. Visoke zmogljivosti tehnologije nam ponujajo priložnosti za opravljanje nalog, ki jih do sedaj zaradi omejenosti človeških sposobnosti nismo mogli opravljati. Uporabnost področja se kaže predvsem pri razvijanju sistema kjer se vhodni zvočni posnetek primerja s posnetki v neki podatkovni bazi in se ga na podlagi podobnosti ustrezno poveže s podobnimi posnetki. Na podlagi tega se razvijajo različni uporabni sistemi. Iskanje glasbe na podlagi mrmranja, ki omogoča iskanje glasbe po melodiji in ne kakor smo vajeni do sedaj, po naslovu. Uporaben je tudi sistem za primerjavo notnih zapisov glasbe, ki omogoča avtomatizirano primerjavo del različnih skladateljev in je zanimiv predvsem za akademske eksperte. Zelo razvit in razvijajoč se sistem je sistem za iskanje plagiatov. Za vsemi temi sistemi se skrivajo dobri algoritmi za računalniško zaznavanje in procesiranje zvočnih signalov.

V nalogi se ukvarjamo s kompozicionalnim hierarhičnim modelom, ki so ga iz podobnega sistema za kategorizacijo slik razvili Pesek, Leonardis in Marolt [27]. Model je namenjen procesiranju zvoka, predvsem glasbe. V nalogi se ukvarjamo z evalvacijo delovanja modela na opravilu ocenjevanja osnovnih frekvenc, katerega bomo predstavili v nadaljevanju, in iščemo izboljšave. Ob dovolj dobri optimizaciji tega opravila bi nek sistem zgrajen na modelu lahko deloval kot orodje za samodejno preslikavo zvočnega signala skladbe v glasbene simbole (note). Že samo informacije,

ki bi bile hitro pretvorjene v računalniški zapis, bi lahko koristno služile marsikateremu glasbeniku. Model bi lahko npr. uporabili tudi za zapis improviziranih del ali pa za zapis glasbe, ki se ohranja z ustnim izročilom. Zagotovo lahko ob začetku naloge rečemo, da se je z modelom iz vidika uporabnosti rezultatov vredno ukvarjati.

V nalogi bomo najprej predstavili področje pridobivanja informacij iz glasbe, opravilo ocenjevanja osnovnih frekvenc in kompozicionalni hierarhični model. Nato bomo opisali opravljeno evalvacijo in delovanje modela ter na koncu opisali metode s katerimi smo poskušali delovanje modela na opravilu ocenjevanja osnovnih frekvenc izboljšati.

Poglavje 2

Pregled področja

2.1 Glasba in informacije v njej

Človek že od nekdaj živi z glasbo. Že zelo zgodaj se je začela izražati potreba po prenosu in posledično zapisu tistega dela informacije o glasbi, ki jo zaznava človeški kognitivni sistem. Na začetku so se informacije o glasbi širile od ust do ust, potem pa so se začeli pojavljati razni zapisi. Zgodovinski viri prve najdene zapise uvrščajo v leto 2000 pr. n. št. Od takrat se je v različnih kulturah pojavljalo veliko različnih načinov zapisovanja not, ki so se do danes izoblikovali v sodoben notni zapis. Glavni informaciji, ki nam jih da sodobni notni zapis, sta višina in dolžina note. Določa tudi dinamiko, tempo in druge informacije, del razumevanja pa je še vedno odvisen od interpretacije posameznika.

Osnova zapisa je nota. Določa dolžino - glede na vrsto note in tonsko višino - glede na položaj note v notnem črtovju. Dolžina je posredno odvisna od tempa in takta, ki sta podana na začetku. Tonska višina je odvisna od ključa, ki določa tonaliteto.
Notno črtovje je baza zapisa. Sestavljeno je iz petih črt, ki jim lahko po potrebi dodamo pomožne črte. Note so postavljene na črte ali med njih.
Takt je zapisan na začetku vsake vrste in določa število dob na takt.
Tempo je določen s številom, ki določa število osnovnih dob na minuto. Lahko pa je določen tudi opisno (hitro, hitreje, zelo hitro, ...).

Temelj vseh sistemov, ki se razvijajo na podlagi tega, da znamo iz glasbe avtomatično pridobiti informacije, so osnovne značilnice. Z značilnicami zaobsegamo naslednje glasbene elemente ([23]):

- Tonska višina, ki je v nekem trenutku definirana kot percepcija osnovne frekvence neke glasbene note, povezane s spremembo glede na prejšnjo noto [30].
- Jakost, ki glede na energijo v signalu določa glasnost zvoka v nekem trenutku.
- Barva, ki poslušalcu omogoča razlikovati med različnimi izvori zvoka. Glede na barvo zvoka ljudje razlikujemo med inštrumenti in med glasovi.
- Harmonija, ki se pojavlja, kadar med izvajanjem zvoka nastopa več tonskih višin naenkrat.
- Ritem, ki ni vezan na percepcijo prej omenjenih značilnic. Ta zaporedju not oziroma zvokov določa trajanje.

Področje pridobivanja informacij iz glasbe, del katerega je ta naloga, se želi približati temu, da bi računalnik znal te značilnice prepoznati prav tako dobro kakor človek, ki je ekspert na glasbenem področju. Želimo si, da bi lahko na podlagi tega na primer pretvorili zvočni zapis v notni zapis in da bi s procesiranjem informacij v glasbi ustvarili nove sisteme in nove razsežnosti v interakciji človeka z glasbo.

2.2 Pridobivanje informacij iz glasbe

Pridobivanje informacij iz glasbe (*ang. Music information retrieval - MIR*) je mlado, a hitro rastoče področje znanosti, ki se je začelo razvijati v osemdesetih letih prejšnjega stoletja. Področje je zelo interdisciplinarno saj zajema kombinacijo računalništva, nevroznanosti, procesiranja signalov, muzikologije, glasbe in psihologije. Kakor že ime pove, gre za pridobivanje informacij, ki se nahajajo v glasbi. Področje zaradi podprtosti z računalnikom ponuja nove razsežnosti procesiranja in uporabnosti le-teh. Želi se približati človeškemu zaznavnemu sistemu in ga nadgraditi s tistimi zmogljivostmi v katerih sodobna tehnologija presega človeka. Dober

primer uporabnosti je iskanje pesmi z določenim tempom po zbirki, kjer se nahaja več tisoč pesmi. Računalniška rešitev v tem primeru ponuja nekaj, kar brez računalnika praktično ne bi bilo izvedljivo v zadovoljivem času.

2.3 Glavne smernice razvoja

MIR se ukvarja s kategoriziranjem, manipulacijo in tudi z ustvarjanjem glasbe. Glavne smeri razvoja področja lahko razdelimo na naslednje naloge, ki še vedno ponujajo ogromno prostora za raziskovanje in napredek ([31]):

- iskanje podobnosti in priporočilni sistemi,
- razvrščanje in gručenje,
- transkripcija tonske višine, ritma in transkripcija glasbe,
- transkripcija glasbe in ločevanje virov,
- povpraševanje z mrmranjem,
- obdelava simboličnih zapisov,
- segmentacija, strukturiranje ter poravnavanje posnetkov in simboličnega zapisa,
- zaščita, prepoznavanje melodije ter prepoznavanje priredb,
- povezovanje MIR z drugimi z znanstvenimi disciplinami.

2.4 Samodejna transkripcija glasbe

Samodejna transkripcija glasbe (*ang. automatic music transcription - AMT*) je eno izmed glavnih opravil na področju MIR. Kljub temu je še vedno daleč od zmogljivosti človeških strokovnjakov na glasbenem področju. Gre za proces samodejnega spreminjanja zvočnega signala glasbe v eno izmed oblik notnega zapisa [3]. Glavni doprinosi, ki jih ponuja AMT, so notni zapisi improvizacij in notni zapisi del pri

zvrsteh, kjer se ta večinoma ne zapiše (ljudske pesmi ustnega izročila, pop, jazz ...). Hkrati omogoča hitrejši napredek drugih opravil področja MIR. Problem samodejne transkripcije lahko razdelimo na več nalog: ocenjevanje osnovnih frekvenc (*ang. multiple f0 ali multi-pitch detection*), pridobivanje melodije (*ang. melody transcription*), zaznavanje pojavitve in konca not (*ang. onset/offset detection*), ocena jakosti in kvantizacija (*ang. loudness estimation and quantisation*), prepoznavanje inštrumentov (*ang. instrument recognition*), pridobivanje ritmičnih informacij (*ang. extraction of rhythmic information*) in časovna kvantizacija (*ang. time quantisation*) [3]. Glavni problem in problem s katerim se bomo ukvarjali je problem ocenjevanja osnovnih frekvenc.

2.4.1 Ocenjevanje osnovnih frekvenc

Ocenjevanje osnovnih frekvenc je proces transkripcije polifonične glasbe, ki v določenem časovnem okvirju zazna tone, ki se istočasno pojavijo in so lahko proizvedeni z različnimi inštrumenti. Problem ni enostavno rešljiv. Komponente zvočnih signalov inštrumentov v polifoničnih skladbah se namreč časovno in frekvenčno prekrivajo. Poleg tega probleme povzročajo odmevi in prehodna stanja. Pristope k problemu lahko v grobem razdelimo v dve skupini: pristop s ponavljajočim ocenjevanjem (*ang. iterative estimation approach*) in pristop s skupnim ocenjevanjem (*ang. joint estimation approach*) [7]. Prvi se uporablja tako, da se v vsaki ponovitvi v določenem časovnem okvirju poišče prevladujoča frekvenca, ki se potem ob koncu ponovitve odstrani iz spektra. Pristop s ponavljanjem je manj računsko zahteven a tudi manj natančen. Drugi pristop poišče vse prevladujoče frekvence naenkrat, je večinoma natančnejši a računsko bolj zahteven. Do danes je zaradi večje natančnosti že skoraj povsem prevladal drugi pristop skupnega ocenjevanja. Tudi pri tem pristopu lahko ločimo tehnike reševanja na zaznavanje na podlagi značilnic, zaznavanje s pomočjo statističnih modelov in zaznavanje na podlagi razcepa spektrograma [3]. V nalogi bomo uporabili pristop s kompozicionalnim hierarhičnim modelom, ki bi ga lahko uvrstili med metode, ki opravljajo zaznavanje na podlagi razcepa spektrograma. Naš pristop predstavlja alternativo pristopom, ki se problema lotevajo z globokimi arhitekturami [16], [24].

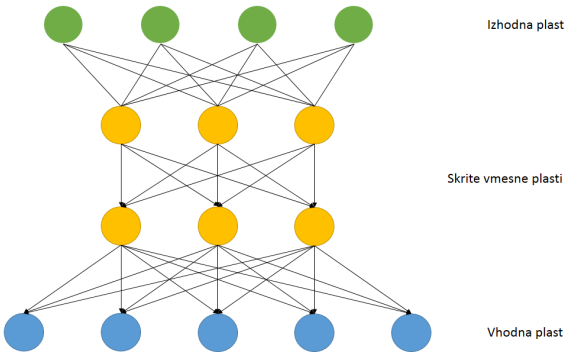
Poglavje 3

Kompozicionalni hierarhični model za pridobivanje informacij iz glasbe

Kompozicionalni hierarhični model je nov biološko navdahnjen pristop na področju pridobivanja informacij iz glasbe, ki predstavlja alternativo globokim arhitekturam, ki delujejo na pristopu z nevronskimi mrežami. Model, ki ga bomo uporabljali, je nastal kot prevedba takšnega modela, ki deluje na področju računalniškega vida, in sta ga zasnovala Leonardis in Fidler [12]. Naš model s podobnim delovanjem, ki se ukvarja s procesiranjem zvočnih informacij, predvsem glasbe, so razvili Pesek, Leonardis in Marolt [27]. Ker je naš model podoben pristopom, ki se problema lotevajo z globokimi arhitekturami, bomo najprej opisali tak pristop.

3.1 Globoke arhitekture na področju MIR

Mreže globokega zaupanja (*ang. deep belief network - DBN*) so vrsta nevronskih mrež, ki so najpomembnejše na področju MIR. Osnovna ideja DBN je, da lahko velike procesirane strukture sestavimo iz več manjših delov. V osnovi so nevronske mreže računalniški modeli, ki se želijo približati delovanju centralnega živčnega sistema. Osnovna zgradba DBN je prikazana na Sliki 4.2. Sestavljene so iz gradnikov,



Slika 3.1: Zgradba DBN

ki imajo več uteženih vhodov in izhodov. Ob določenem vhodu se po povezavah razširi signal, ki ob določenih utežeh in kombinacijah z drugimi vhodi vrne ustrezen izhod. DBN so zgrajene iz več povezanih nivojev. Vmesni nivoji so skriti. Mreža se zgradi ob učenju. Vsak nivo mreže je naučen nenadzorovano in neodvisno od ostalih nivojev. Ob postopku učenja torej mreža sama ugotavlja pravila tako, da bo informacija ob izhodu najbolj optimalna. Takrat se določajo uteži, povezave in prag, ki določa ali bo vhodni signal vrnil izhod ali ne. Pri DBN povezave med vozlišči na istem nivoju ne obstajajo. Za uporabo vzamemo zgornji nivo - izhod, katerega rezultati se glede na opravilo MIR ustrezno nadalje procesirajo. Na področju MIR se DBN uporabljajo kot generator značilnic in za klasifikacijo glasbenih elementov.

Globoke arhitekture se zadnje čase uporabljajo na različnih področjih MIR. Predvsem zaradi možnosti nenadzorovanega učenja in generativnega modeliranja prinašajo marsikatere izboljšave. Mreže globokega zaupanja so bile prvič uspešno uporabljene pri zaznavanju pojavitve in konca not [20]. Kasneje so bile uporabljene tudi pri naslednjih opravilih MIR: prepoznavanje inštrumentov [19, 15], avtomatično prepoznavanje akordov [18], klavirska transkripcija [24], prepoznavanje govora [21], prepoznavanje žanra [22, 14], prepoznavanje žanra, avtorja in ključa [10], modeliranje ritma in melodije [28], prepoznavanje čustev v glasbi [28] in pri analizi vzorca tolkal [2].

3.2 Kompozicionalni hierarhični model

Namen kompozicionalnega hierarhičnega modela je, da kompozicionalno hierarhično modeliranje, ki se uporablja tudi na drugih področjih kognitivne znanosti, uspešno uporabimo za modeliranje človeškega slušnega sistema. Model je podoben drugim globokim arhitekturam, ki temeljijo na pristopu z nevronskimi mrežami, vendar se od slednjih razlikuje po svoji transparentnosti. To bi se lahko izkazalo kot prednost pri razreševanju opravil področja MIR. Zaradi množice informacij, ki jih je možno iz modela razbrati, bi model lahko prilagodili tako, da bi deloval na več opravilih in bi na njem lahko zgradili celosten sistem za analizo glasbe.

Model temelji na kategoriziranju, ker tako deluje tudi človeški zaznavni sistem in hkrati na hierarhičnosti, ker človekov živčni sistem na celičnem nivoju deluje hierarhično. Učenje in povezovanje objektov po gradnikih je prostorsko veliko manj zahtevno od direktnega pristopa. Poleg tega intuitivno nakazuje na hierarhično združevanje elementov glede na vsebino informacije, ki sčasoma z dodajanjem bolj specifičnih elementov določa bolj specifične objekte. Posamezen del določenega nivoja je sestavljen iz pod-delov na nižjem nivoju. Deli na najnižjem nivoju predstavljajo komponente vhodnega signala, na višjih nivojih pa bolj kompleksne informacije. Na vsakem nivoju so informacije dostopne. Hierarhičnost in razvidnost modela na vseh nivojih nam ponujata nove informacije, ki jih z ustrezno interpretacijo lahko uporabimo za temeljitejšo analizo glasbenih elementov vhodnega signala. Model, podobno kot človeški avditorni sistem, vsebuje mehanizma za halucinacijo in inhibicijo informacij.

3.2.1 Zgradba kompozicionalnega hierarhičnega modela

Model je po principu globokih arhitektur zgrajen iz več nivojev, ki jih sestavljajo posamezni gradniki - *deli*. Kompleksnost informacije posameznih delov je zaradi hierarhične zgradbe manjša na nižjih nivojih in večja na višjih. Vsak del je, razen na začetnem nivoju, zgrajen iz delov nižjih nivojev. Spodnji ali vhodni nivo predstavlja vhodni zvočni signal transformiran v časovno frekvenčno domeno. Ta nivo označujemo z $\mathcal{L}_0$ in ga imenujemo tudi ničti nivo hierarhije. Sestavljen je iz gradnikov, ki predstavljajo vse kanale frekvenčnega spektra signala v izbranem časovnem

okvirju.

Zgradba modela je shematično prikazana na sliki 3.2. Deli so sestavi entitet, ki opisujejo signal. Del lahko opisuje posamezne frekvence v signalu, njihove kombinacije, poleg tega pa tudi tone, akorde in časovne vzorce [27]. Posameznim delom pripišemo aktivacijo, ki jo definirata lokacija $\mathcal{L}_p$ in magnituda $\mathcal{A}_p$. Lokacija predstavlja frekvenco, magnituda pa moč aktivacije. Glede na moč aktivacije se na podlagi izbrane pragovne vrednosti določijo deli, ki so aktivni.

Dele na višjih nivojih, imenovane tudi kompozicije, lahko definiramo kot sestav osrednjega dela C in sekundarnih sestavnih delov S z nižjega nivoja, ki se izberejo glede na statistiko sopojavitev. Novi deli nastanejo iz pogostejših sopojavitev. Naslednji nivoji modela $\mathcal{L}_n$ so torej sestavljeni iz kompozicij i, ki so zgrajene iz delov nižjega nivoja $\mathcal{L}_{n-1}$ po enačbi 3.1:

$$P_{n,i} = \{C_{n-1,j}, S_{n-1,k}, (\mu_{n,i}, \sigma_{n,i})\}, \tag{3.1}$$

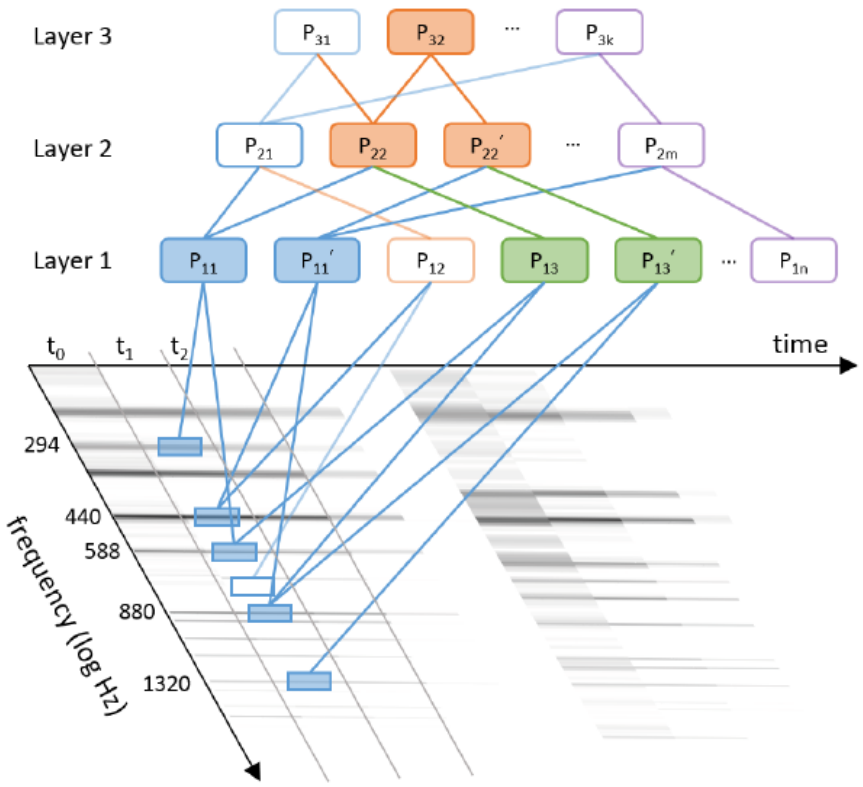
kjer $\mu_{n,i}$ in $\sigma_{n,i}$ predstavljata Gaussovo razporeditev, ki omejuje razdaljo med lokacijami aktivacij delov nižjega nivoja. Vsak del je lahko gradnik poljubnega števila kompozicij višjega nivoja in prav tako lahko vsako kompozicijo sestavlja poljubno število delov nižjega nivoja.

Jakost aktivacije $\mathcal{A}_p$ in lokacijo $\mathcal{L}_p$ določene kompozicije izračunamo glede na vrednosti aktivacij njenih sestavnih delov. $\mathcal{L}_p$ dela na višjih nivojih se določi tako, da vzamemo lokacijo osrednjega dela $\mathcal{L}_c$. Primer za določanje lokacije kompozicije i na nivoju n je prikazan v enačbi 3.2:

$$L_{Pn,i} = L_{Cn-1,j} \tag{3.2}$$

Jakost aktivacije $\mathcal{A}_p$ in lokacijo $\mathcal{L}_p$ določenega dela na višjih nivojih pa določimo po enačbi 3.3:

$$A_p = tanh[G(L_c - L_s, \mu, \sigma)(A_c + A_s)][27], \tag{3.3}$$



Slika 3.2: Slika prikazuje zgradbo kompozicionalnega hierarhičnega modela. Abscisa predstavlja čas. Na sliki je prikazan model zgrajen iz treh nivojev označenih z Layer 1, Layer 2 in Layer 3 ter vhodnega nivoja v časovnem okvirju t1. Povezave med posameznimi deli predstavljajo kompozicije, na primer $\mathcal{P}_{11}$ na nivoju $\mathcal{L}_1$ je del kompozicije $\mathcal{P}_{21}$ in $\mathcal{P}_{22}$ na nivoju $\mathcal{L}_1$. Obarvani deli predstavljajo aktivne dele na vsakem izmed nivojev. Vir slike: [27].

kjer $\mathcal{A}_c$ predstavlja jakost aktivacije osrednjega dela in $\mathcal{A}_s$ jakost aktivacije sekundarnega dela kompozicije, *tanh* pa označuje hiperbolični tangens, ki jakost aktivacije preslika na interval [0,1). Kompozicija je aktivna, kadar so aktivni vsi njeni sestavni deli na nižjih nivojih. V vsakem časovnem okvirju nam model vrača množico aktivnih delov iz katerih lahko razberemo razne informacije o procesiranem signalu.

Model za izgradnjo uporablja nenadzorovano učenje. Proces učenja na podlagi statistične metode sestavlja kompozicije iz sopojavitev delov, ki so med seboj čimbolj disjunktni in se pogosto aktivirajo na podobnih razdaljah. Za zmanjšanje števila delov se na koncu uporabi še požrešen algoritem. Učenje je natančneje opisano v poglavju 3.2.5.

Ko je model zgrajen, ga lahko uporabimo za klasifikacijo in prepoznavanje značilnic vhodnega zvočnega signala za vsak posamezen časovni okvir.

Model za povečanje robustnosti uporablja tri biološko navdahnjene mehanizme: halucinacijo, inhibicijo ter mehanizem za samodejno uravnavanje jakosti (*ang. automatic gain control - AGC*).

Poleg teh mehanizmov je potrebno omeniti še dve lastnosti, ki model ločita od običajnih pristopov z DBN. Gre za relativnost in deljivost delov (*ang. relativity and shareability of parts*), ki jih bomo opisali v nadaljevanju.

3.2.2 Mehanizmi

Preden začnemo opisovati mehanizme, moramo opisati pojem pokritja (*ang. coverage*). Pokritje dela P na lokaciji L_P opišemo z naslednjo enačbo 3.4:

$$c(P, L_P) = \bigcup \{c(C, L_P), c(S, L_P + \mu)\} \tag{3.4}$$

Pokritje dela P predstavlja množico vseh informacij vhoda, ki jih pokriva nek del in vsi njegovi pod-deli v drevesni strukturi. Pokritje se računa od zgoraj navzdol od nekega aktivnega dela P. Na ničtem nivoju je ta množica enaka vsem delom, katerih aktivacija večja od nič:

$$A_p > 0, \tag{3.5}$$

na tem nivoju gre torej za množico frekvenčnih komponent, ki se pojavijo. Primer pokritja si lahko pogledamo na primeru: na sliki 3.2 je pokritje dela P_{22} na lokaciji 440 Hz, kar lahko zapišemo kot c(P_{22},440), množica frekvenc {588 Hz, 880 Hz, 1320 Hz}.

Halucinacija

Vloga halucinacije je, da nadomesti manjkajočo informacijo, ki lahko nastane zaradi različnih razlogov, predvsem napak v signalu. Halucinacija ponazarja človeško percepcijo, ki glede na vsebino informacije logično dopolni pomanjkljivosti. Mehanizem glede na ostalo vsebino lahko delu, ki sicer ni aktiven, določi moč aktivacije.

Kakor smo povedali v poglavju 3.2.1 je zato, da je nek del aktiven, potrebno, da so aktivni vsi njegovi pod-deli. Halucinacija pa na podlagi parametra τ_1 to omogoči tudi nekaterim drugim delom. Pogoj, ki je veljal prej, je ob delovanju halucinacije spremenjen, tako da je del aktiven, če je število frekvenčnih komponent v pokritju tega dela ($|c(P,L_P)|$), deljeno z maksimalnim številom frekvenčnih komponent, ki bi jih lahko pokrival, večje od τ_1. Če je τ_1=1, je torej delovanje modela enako navadnemu delovanju brez halucinacije. Parameter τ_1 lahko določamo za vsak nivo posebej in ga med delovanjem tudi dinamično spreminjamo.

Inhibicija

Inhibicija je mehanizem, ki posnema delovanje človeškega avditornega sistema in odstranjuje odvečne informacije ter zmanjšuje šum, ki je lahko prisoten v signalu. Glede na moč najmočnejših hipotez mehanizem izbere tiste, ki imajo v primerjavi z njimi zelo nizko jakost aktivacije in jih odstrani. Inhibicija deluje tako, da če se del P na lokaciji L_P in del Q na lokaciji L_Q pojavita na istem nivoju L_n in pokrivata iste dele informacije, se potem tisti del, ki ima nižjo aktivacijo, odstrani. Pri tem je vse odvisno še od parametra τ_2. Pogoj za obstoj nekega dela P ob delovanju inhibicije na našem modelu lahko opišemo z enačbo 3.6, ki je bila predstavljena v

[26]:

$$\forall P \exists Q : \frac{|c(P, L_P) \backslash c(Q, L_Q)|}{|c(P, L_P)|} < \tau \wedge A_Q > A_P. \tag{3.6}$$

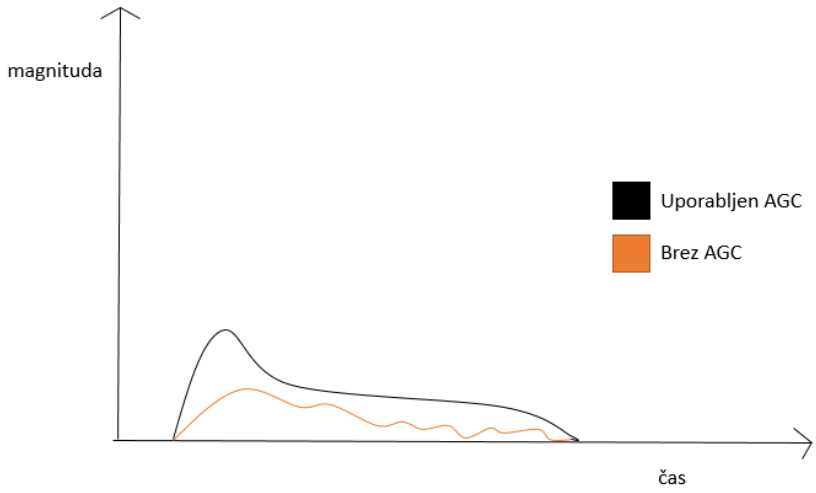
Parameter τ_2 pri inhibiciji določa jakost vpliva mehanizma. Vrednost določa mejno vrednost, ki nam v odstotkih pove, kako podobna morata biti dela, da bo določen del odstranjen. Če je vrednost τ_2 enaka 0.6, to pomeni, da bodo vsi deli, katerih množica pokritja ima več ali enako kot 60 odstotkov elementov enakih množici pokritja drugega delu na isti plasti z višjo aktivacijo, odstranjen.

Samodejno uravnavanje jakosti

Tudi mehanizem za samodejno uravnavanje jakosti odpravlja nepravilnosti, ki se pojavljajo v signalu. Z upoštevanjem časovne komponente glede na hipoteze prejšnjih časovnih okvirov s poudarjanjem pojavitev in stabilizacijo aktivacij ob manjših nihanjih uravnava aktivacije. Mehanizem modelu dodaja dodatno dimenzijo, dimenzijo časa, ki prej ni bila nikjer upoštevana. Deluje tako, da ko se nekje pojavi nova aktivacija in traja nekaj časa, jo mehanizem ojača, da poudari pojavitev, potem pa jo malo zmanjša in ohranja stabilno vrednost tudi v pojemanju, kjer ponavadi v signalu pride do raznih manjših nihanj. Delovanje mehanizma je prikazano na sliki 3.3.

3.2.3 Relativnost in deljivost delov

Relativnost delov omogoča, da se nek del, kakor je prikazano tudi na sliki 3.2, lahko aktivira na večih lokacijah. To pa zato, ker deli na višjem niso definirani z absolutno lokacijo delov na ničtem nivoju iz katerih so sestavljeni, ampak so definirani z odmikom v razdalji med dvema sopojavljajočima deloma na nižjem nivoju. Nek del, ki je na višjem nivoju in je sestavljen iz pod-delov nižjega nivoja na razdalji ene oktave, se tako lahko v istem časovnem okvirju aktivira večkrat, odvisno od tega kakšno je število delov z istim odmikom lokacijama na spodnji plasti. Relativnost delov se ohranja na vseh nivojih modela. Relativnost omogoča visoko nivojsko abstraktno



Slika 3.3: Prikaz aktivacije nekega dela z uporabo in brez uporabe mehanizma za samodejno uravnavanje jakosti. Po abscisi teče čas, po ordinati pa magnituda aktivacije dela. Vidimo, da AGC poudari pojavitev aktivacije in odstranja šum pri manjših aktivacijah, kjer se ta ponavadi začne pojavljati.

predstavitev signala, ne glede na lokacijo dela na ničtem nivoju, na primer predstavitev intervala. Zaradi relativnosti lahko ne glede na lokacijo pojavitev en del pokrije veliko delov spodnje plasti, ki odražajo podobno strukturo. Deljivost delov pomeni, da je nek del nivoja $\mathcal{L}_{n-1}$ sestavni del mnogim delom nivoja $\mathcal{L}_n$. Zaradi tega nam ni potrebno hraniti večih primerkov iste kompozicije.

Relativnost in deljivost si lahko ogledamo na sliki 3.2, kjer se del P_{22} aktivira dvakrat, ker se dva dela na nižjem nivoju pojavita na razdalji ene oktave in sicer se v istem okvirju ta del aktivira na lokacijah 294 Hz in 440 Hz. Tak pojav omogoča relativnost delov. Funkcijo deljivosti delov pa lahko na sliki opazimo pri delu P_{11}, ki je gradnik večih delov višjega nivoja.

3.2.4 Konstantna Q transformacija

Zvočni signal, ki služi kakor vhod modela, najprej transformiramo s konstantno Q transformacijo. Konstantna Q transformacija je pretvorba signala iz časovnega

prostora v frekvenčnega. Sorodna je Fourierevi transformaciji. Prednost konstantne Q transformacije je to, da se časovna ločljivost ob višjih frekvencah povečuje, ker je frekvenčna lestvica predstavljena logaritmično. To je podobno delovanju človeškega slušnega sistema.

Konstantna Q transformacija vsebuje množico logaritemsko razporejenih filtrov, kjer ima k-ti filter pasovno širino, ki je večkratnik pasovne širine predhodnega filtra *k-1*, kakor je predstavljeno v enačbi 3.8.

$$\delta f_k = 2^{\frac{1}{n}} * \delta f_{k-1} \tag{3.7}$$

$$= (2^{1/n})^k * \delta f_{min}, \tag{3.8}$$

kjer δf_k predstavlja pasovno širino k-tega filtra, δf_{min} centralno frekvenco najnižjega filtra, n pa predstavlja število filtrov na oktavo.

Konstantno Q transformacijo lahko predstavimo z enačbo 3.9:

$$X[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} W[k,n]x[n]e^{\frac{-j2\pi Qn}{N}}, \tag{3.9}$$

kjer je x[n] vhodni signal. Q je faktor "kakovosti", izračunamo pa ga tako:

$$Q = \frac{f_k}{\delta f_k}, \tag{3.10}$$

V 3.10 f_k predstavlja centralno frekvenco k-tega filtra, $\delta\ f_k$ pa enako kot zgoraj pasovno širino k-tega filtra.

N[k] v enačbi 3.9 predstavlja dolžino okna za k-ti frekvenčni kanal. Izračun dolžine okna za k-ti frekvenčni kanal je predstavljen v enačbi 3.11:

$$N[k] = (\frac{f_s}{\delta f_k}) = (\frac{S}{f_k})Q. \tag{3.11}$$

Katerokoli okensko funkcijo bomo pri Q transformaciji uporabili, bo ta odvisna od dolžine okna, ker bomo z dolžino okna normalizirali relativno moč. Pri višjih frekvenčnih kanalih se namreč pri konstantni transformaciji relativna moč zmanjšuje.

W[k,n] v enačbi 3.9 predstavlja okensko funkcijo. Pri Q transformaciji, ki se uporablja na našem modelu, se uporablja Hammingovo okno:

$$W[k,n] = \alpha - (1-\alpha)cos(\frac{2\pi n}{N[k]}), \alpha = 25/46, 0 \leq n \leq N[k]-1. \tag{3.12}$$

Pri konstantni Q transformaciji za predstavitev izhodnega signala potrebujemo manj frekvenčnih kanalov za pokritje nekega signala kakor pri Fourierevi transformaciji.

3.2.5 Učenje modela

Učenje modela poteka nenadzorovano, uči se plast za plastjo. Učenje je odvisno predvsem od statistike aktivacij posameznih delov.

Učenje prvega nivoja

Prvi nivo se zaradi drugačne strukture uči malo drugače od višjih nivojev. Na ničtem nivoju, ki je predstavljen v obliki frekvenčnih kanalov in njihovih magnitud, si izberemo nek frekvenčni kanal, ki ima dovolj visoko amplitudo in bo predstavljal centralni del, potem pa temu delu dodamo še enega z višjo frekvenco, ki bo predstavljal sekundarni del dela. Razliko v frekvenčnih kanalih shranimo kot odmik med tema dvema deloma. Dodamo jih v množico kandidatov za prvi nivo, če tak del tam še ne obstaja. Želimo si, da bi bila množica delov nekega nivoja čim bolj disjunktna in da bo pokritje učne množice ob tem čim večje. Število kandidatov na podlagi teh dveh kriterijev zmanjšamo s požrešnim algoritmom, ki bo opisan spodaj.

Učenje višjih nivojev

Učenje na višjih nivojih poteka zelo podobno kakor učenje na prvem nivoju, le gradniki so nekoliko drugačni. Nivo L_n je zgrajen iz kompozicij, ki so sestavljene iz

delov nivoja L_{n-1}. Kompozicije so sestavljene iz delov, ki se najpogosteje pojavljajo skupaj na podobnih razdaljah. Take kompozicije dodamo v množico kandidatov. Tisti del v kompoziciji, ki ima nižjo lokacijo, predstavlja centralni del, μ in σ pa sta ocenjeni glede na vse sopojavitve delov.

Ko se izbere množica kandidatov za kompozicije na določenem nivoju, se ta naknadno uredi in sicer tako, da nekatere kandidate izločimo, tako da je množica kandidatov na novem nivoju čim manjša in hkrati pokriva dovolj informacije za ustrezno delovanje modela.

Izbira kandidatov poteka s požrešnim algoritmom, kjer je v vsaki iteraciji izbrana kompozicija, ki doprinese in pokrije največ informacije o vhodnem signalu. Tako kompozicijo dodamo kot nov del na novem nivoju. Algoritem se zaključi, ko je pokrita zadostna količina informacij, kar je določeno s parametrom τ_3, ali pa ko v množici kandidatov ni več takih kandidatov, ki bi povečali količino informacije, ki jo izbrani kandidati pokrivajo. Algoritem je predstavljen v [26] in je prikazan s psevdokodo 1 spodaj:

Algorithm 1 Prikazan je požrešen algoritem za izbiro kandidatov iz množice kompozicij $\mathcal{P}$. Funkcija *perc* izračuna odstotek informacije, ki ga nek del v učni množici pokrije. $\mathcal{L}n$ predstavlja nivo, ki ga gradimo (vir: [27]).

procedure REFINE($\mathcal{P}$)
  $prevCov \leftarrow 0$
  $coverages \leftarrow \emptyset$
  $\mathcal{L}_n \leftarrow \emptyset$
  repeat
    for $P \in \mathcal{P}$ **do**
      $coverages[P] \leftarrow perc(\mathcal{L}_n \cup P)$
    $Chosen \leftarrow \underset{P}{\text{argmax}}(coverages)$
    $\mathcal{L}_n \leftarrow \mathcal{L}_n \cup Chosen$
    $\mathcal{P} \leftarrow \mathcal{P} \setminus Chosen$
    if $coverages[Chosen] = prevCov$ **then**
      break *//No added coverage - finish*
    $prevCov \leftarrow coverages[Chosen]$
  until $prevCov > \tau_3 \vee \mathcal{P} = \emptyset$

Poglavje 4

Uporaba modela za ocenjevanje osnovnih frekvenc

Model smo prilagodili za ocenjevanje osnovnih frekvenc. To smo izvedli tako, da smo modelu dodali dodatno funkcijo, ki vrača hipoteze. Te hipoteze so predstavljene s frekvencami in so določene za vsak časovni okvir zvočnega signala, na katere vhodni signal ob začetku razdelimo. Lahko izbiramo katero plast modela bi radi opazovali kakor izhod.

Nato smo se lotili iskanja podatkovnih zbirk za preizkušanje modela, kjer smo zato, da bo možna evalvacija, iskali tiste, ki jih za isto opravilo uporabljajo tudi ostali, ki delujejo na področju MIR. Model smo potem naučili in ga uporabili. Glede na dobljene rezultate in primerjave z ročnimi transkripcijami smo najprej ocenili robustnost modela glede na različne parametre, ki jih pri delovanju modela lahko spreminjamo. Po analizi smo model prilagodili za najboljše delovanje in rezultate, ki nam jih vrne, evalvirali. Model smo ocenili tako, da smo primerjali MIDI vrednosti naših hipotez in ročno transkribiranega posnetka.

MIDI (Musical Instrument Digital Interface) je elektronski protokol za komuniciranje med elektronskimi glasbenimi napravami. Ko bomo v nalogi govorili o MIDI vrednostih, bomo s tem mislili na parameter protokola v angleščini imenovan Note Number, ki nam s številko pove, katera tipka klavirja je v uporabi - gre za preslikavo klavirskih tipk v številke, kjer ton A4 (440 Hz) preslikamo v vrednost 69.

4.1 Priprava modela

Za ocenjevanje osnovnih frekvenc smo zgradili model, kjer je na ničtem nivoju s konstantno Q transformacijo signal pretvorjen v 345 frekvenčnih kanalov med 55 in 8000 Hz. Časovni okvir je dolg 50 ms.

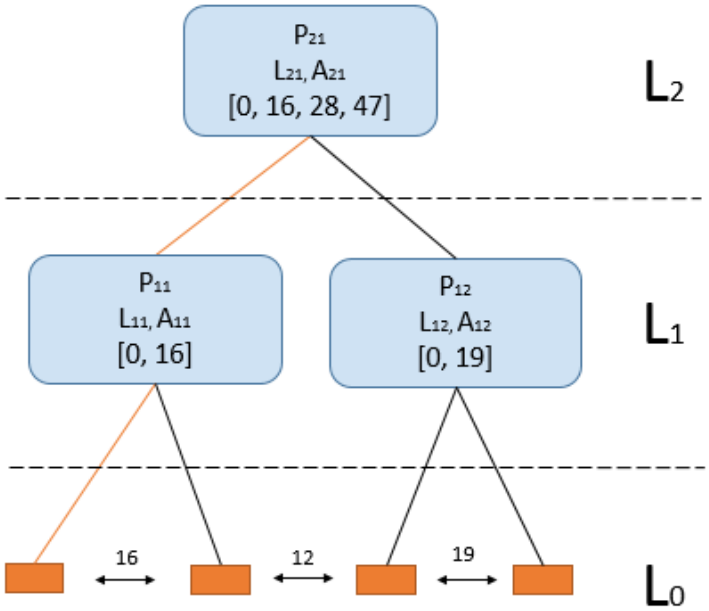
Naučili smo dva nivoja $\mathcal{L}_1$ in $\mathcal{L}_2$, ki vsebujeta 23 in 12 delov. Za ocenjevanje osnovnih frekvenc smo uporabili nivo $\mathcal{L}_2$, tako da smo model preuredili, da nam vrača informacijo o frekvencah, ki jih model zazna v posameznem časovnem okvirju. Informacija, ki jo pridobimo iz modela, je neposredna in ne rabi dodatnega procesiranja. Za učno množico smo uporabili le majhno bazo oseminosemdesetih posnetkov tipk klavirja. Informacije, ki jo dobimo iz modela, ni potrebno dodatno procesirati.

Za celoten potek izvajanja najprej z določenimi parametri naučimo hierarhijo, potem pa z drugim naborom parametrov model deluje. V nadaljevanju bomo najprej analizirali različno naučene hierarhije potem pa še samo delovanje modela. Obe fazi bomo obravnavali in optimizirali ločeno.

4.1.1 Pridobivanje informacij iz modela

Opisali bomo na kakšen način so pridobljene informacije, ki jih kot rezultat pri opravilu ocenjevanja osnovnih frekvenc preberemo iz zgornje plasti. Med delovanjem modela *(ang. inference)*, dobimo kot rezultat dele, ki so aktivni. Vsak tak del ima aktivacijo in lokacijo, iz katerih pridobimo ton, za katerega je najbolj verjetno, da ga del opisuje. Če je nek del aktiven večkrat, izračunamo ton za vsako izmed lokacij, kjer se pojavi. Če se ista lokacija v nekem okvirju pojavi večkrat, vzamemo tisto, ki ima najvišjo aktivacijo. Natančen opis delovanja modela v praksi in opis pridobivanja informacije, ki jo vzamemo za našo hipotezo, je predstavljen na sliki 4.1. Prikazujemo primer aktiviranega dela P_{21} na drugem nivoju in njegovo poddrevo. Aktiven je, ker so aktivni vsi njegovi pod-deli. Na ničtem nivoju L_0 štirje oglati pravokotniki predstavljajo frekvenčne kanale, ki se aktivirajo v tem okvirju. V spodnji vrstici vsakega dela na naučenih nivojih so v oglatih oklepajih predstavljeni relativni odmiki, ki jih hrani vsak del. Vsak del na višjem nivoju se aktivira, ko je kompozicija, ki se pojavi v nekem okvirju sestavljena iz delov z enakim odmikom kakor jih hrani del. Pri tem lahko omenimo, da nivoja L_0 in L_1 predstavljata polni

dvodelni graf, vendar smo na sliki prikazali le aktivne povezave. Del na drugem nivoju L_2 je prav tako aktiven, ko se aktivirajo kompozicije z enakimi odmiki, kakor jih ta del hrani v svoji množici relativnih odmikov. Za vsak tak del potem uporabimo funkcijo, ki na podlagi znanja o harmonikih, ki se ob določenih tonih tipično pojavijo, izračuna relativen odmik verjetnega tona od lokacije aktivnega, ki ga sopojavitev frekvenčnih kanalov s takimi odmiki najverjetneje predstavlja. Izračun lokacije nekega dela smo že opisali. Lokacija v tem primeru "potuje" po oranžnih povezavah in je enaka frekvenčnemu kanalu prvega gradnika nivoja L_0. Hipoteza je torej ton, ki ga izračunamo iz relativnega odmika verjetnega tona od lokacije. Aktivacijo dobljene hipoteze dobimo iz aktivacije aktivnega dela, ki ga opazujemo, in jo izračunamo po formuli, opisani v poglavju 3.



Slika 4.1: Prikazano je delovanje modela v enem okvirju. L_n prikazuje številko nivoja, oranžni pravokotniki predstavljajo aktivacije na ničtem nivoju, številke med njimi pa relativni odmik med njimi v frekvenčnih kanalih. Deli na višjih nivojih so modre barve. V oglatem oklepaju so predstavljeni relativni odmiki kompozicij, ki sestavljajo del.

4.2 Opis zbirke

Delovanje modela smo najprej preizkusili na dvanajstih pesmih, ki jih bomo v nadaljevanju imenovali zbirka 12 skladb. Gre za skladbe formata .wav s pripadajočimi ročno transkribiranimi posnetki. Zbirko sestavljajo klavirski posnetki skladb različnih zvrsti. Dodaten opis posameznih skladb se nahaja v tabeli 4.1. Skladbe so pridobljene iz različnih virov in sicer smo med skladbami, za katere obstaja ročna transkripcija, iskali in izbrali dvanajst takih, ki so si med seboj čim bolj različne in ki niso dolge, da jih bomo lahko opazovali ročno in da bo postopek analize hiter.

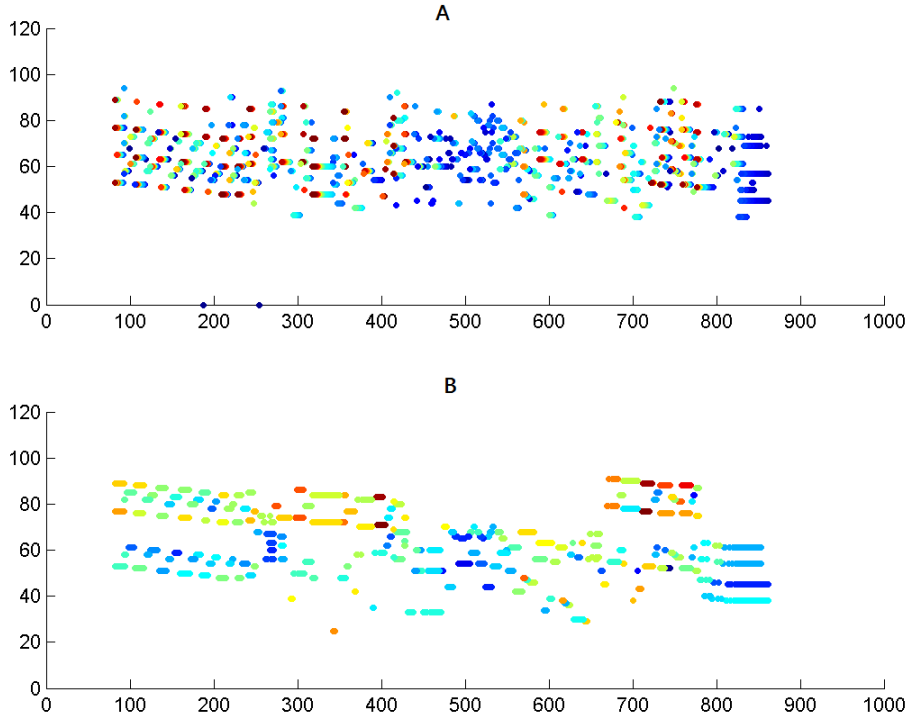
Najprej smo delovanje modela analizirali grafično. Slika 4.2 prikazuje delovanje modela na eni pesmi iz zbirke. Zaradi boljše preglednosti je prikazanih le prvih 1000 okvirjev.

4.3 Evalvacija naučenih hierarhij

Za izboljšavo rezultatov smo najprej ovrednotili delovanje modela z različno zgrajenimi hierarhijami. To smo naredili, da smo ocenili robustnost modela in ugotovili kakšen vpliv ima različno naučena hierarhija na rezultate.

4.3.1 Ocenjevanje robustnosti parametrov

Za opazovanje delovanja modela smo najprej ocenili različno zgrajene hierarhije, da bi videli kako to vpliva na rezultate modela. Poleg tega smo želeli oceniti robustnost modela na parametre. Eksperiment smo izvedli tako, da smo zgenerirali 24 hierarhij, ki so bile naučene z različnimi vrednostmi vpliva inhibicije, halucinacije in mehanizma za samodejno uravnavanje jakosti. Potem smo za vsako posamezno hierarhijo pogledali delovanje modela na zbirki dvanajstih pesmi.



Slika 4.2: Grafični prikaz delovanja modela na skladbi Rubalcaba. Na grafu označenem s črko A so z vizualizacijo predstavljene hipoteze našega modela, s tem da smo prikazali enako število hipotez na časovni okvir kakor jih je prisotnih v osnovnem posnetku. Vzeli smo hipoteze z najmočnejšo aktivacijo. Graf označen s črko B prikazuje ročno transkribiran posnetek. Pri prikazu rezultata modela smo z vizualizacijo v Jet barvnem prostoru prikazali tudi moč aktivacije posamezne hipoteze. Pri vseh grafih nam abscisa predstavlja časovne okvirje, ordinata pa MIDI vrednosti označene točke.

Ko smo najprej primerjali število delov na posameznem nivoju, se to na prvem nivoju ne glede na parametre ni spreminjalo (v našem primeru jih je bilo na prvem nivoju povsod 22), na drugem nivoju pa je število delov pri različnih hierarhijah nihalo med 12 in 20 vendar brez kakšnega posebnega pravila, da bi lahko rekli kateri parameter ima vpliv na to. Tudi korelacije med uspešnostjo delovanja in številom

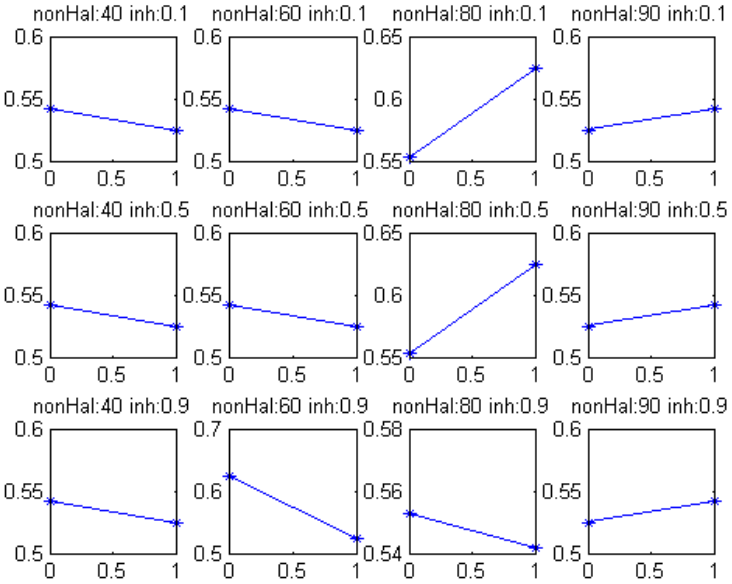
delov, ki se pojavijo, nismo zaznali.

Tabela 4.1: V spodnji tabeli je opis dvanajstih skladb, ki smo jih uporabili za začetno testiranje modela. V prvem stolpcu so indeksi skladb, ki jih na nekaterih mestih v nadaljevanju uporabljamo namesto imen. V drugem stolpcu so imena skladb, ki jih uporabljamo v nalogi. V tretjem stolpcu se nahaja natančnejši opis skladb. V četrtem stolpcu je dolžina posameznih skladb v sekundah, v petem stolpcu je povprečna stopnja polifonije v skladbi, v zadnjem stolpcu pa je odstotek skladbe, v katerem je stopnja polifonije večja od 2.

Indeks	Ime	Opis	Dolžina	Povprečna polifonija	Poli > 2
1	988-v20	Bach: The Goldberg variations 988-V20	114,9	1,70	0,07
2	mz_333_2	Mozart: Piano Sonata No. 13 in B flat major, K333, 2. stavek	522,28	2,71	66,42
3	mz_333_3	Mozart: Piano Sonata No. 13 in B flat major, K333, 3. stavek	347,23	2,22	36,82
4	Aria	Bach: Air	123,60	3,45	89,07
5	rubalcaba	sodobna glasba	43,59	2,35	43,40
6	Cal_drea	California dreaming piano instrumental	123,78	3,32	72,98
7	Smokegtz	Smoke gets in your eyes	86,37	2,52	49,62
8	Invert6	piano inversion	62,35	4,32	75,12
9	woods3	standard blues form	33,63	1,65	14,58
10	Bp054	standard blues form	166,55	2,27	38,07
11	Blu4pia2	standard blues form	139,45	2,08	31,01
12	Bwv780	Bach BWV 780	103,13	1,96	0,05

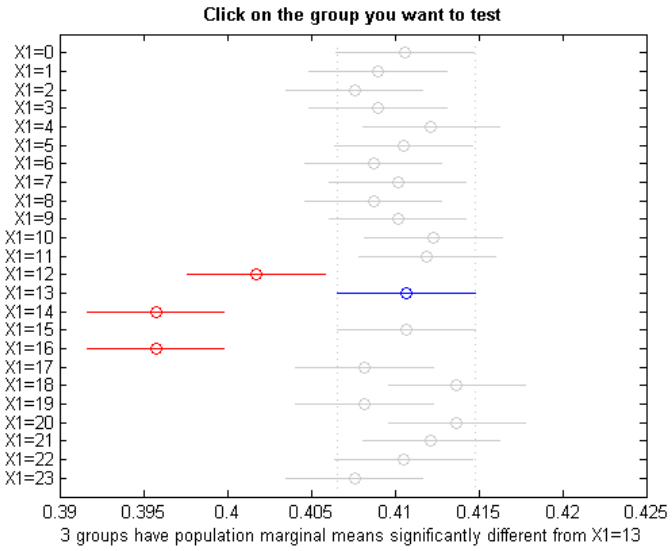
Potem smo se osredotočili še na redundantnost delov pri različnih parametrih, kar je prikazano na sliki 4.3. Redundantnost delov označuje odstotek podobnosti delov

istega nivoja. Prišli smo do zaključka, da je redundantnost delov v hierarhiji nižja, če je mehanizem za samodejno uravnavanje jakosti vklopljen. Inhibicija in halucinacija pa na redundantnost delov v hierarhiji ne vplivata bistveno. Predvidevamo, da AGC pozitivno vpliva na nižjo redundantnost, ker so aktivacije posameznih delov skozi čas bolj stabilne in je manj nihanj in s tem odvečnih informacij. Predvidevamo, da manjša količina informacij, ki jih moramo pokriti, omogoča boljšo optimizacijo pri algoritmu za izbiro delov pri učenju za želen odstotek pokritja informacije, ki temelji na tem, da bi bili deli med seboj čim bolj disjunktni. Zagotovo pa lahko rečemo, da vklopljen AGC pozitivno deluje na samo delovanje in izboljšuje delovanje modela na opravilu ocenjevanja osnovnih frekvenc, ker manj redundantni deli privedejo do manj redundantnih hipotez, ker istih informacij ne upoštevamo večkrat.



Slika 4.3: Redundantnost hierarhij glede na vpliv mehanizma za samodejno uravnavanje jakosti (AGC). Na abscisni osi je vrednost AGC (vklopljen ali izklopljen), na ordinatni osi pa je predstavljena vrednost redundance. Vpliv AGC-ja je predstavljen na štiriindvajsetih različnih hierarhijah.

Analizo učinka parametrov pri učenju hierarhij smo naredili tudi tako, da smo za posamezne pesmi ocenjevali povprečno natančnost naših hipotez na okvir, ki jih ob delovanju z istimi parametri model vrne. Vpliv parametrov smo ocenili z analizo variance, kar je prikazano na sliki 4.4. Povprečno natančnost smo v tem primeru računali kot odstotek pravilnih hipotez na časovni okvir, ki smo jih izbrali tako, da smo med vsemi hipotezami, ki jih vrne model, vzeli določeno število tistih z najmočnejšo aktivacijo. To število je bilo enako številu aktivacij za isti okvir v ročno transkribiranem posnetku.



Slika 4.4: Slika prikazuje analizo variance ocene povprečne natančnosti na okvir na zbirki dvanajstih skladb. Na ordinatni osi so prikazane številke hierarhij, ki jih ocenjujemo, na abscisni osi pa je ocena povprečne natančnosti na okvir.

Na sliki 4.4 vidimo, da glede na pomembnost razlike med hierarhijami jih lahko ločimo na dva dela. Hierarhije, ki so očitno slabše od ostalih, so hierarhije 12, 14 in 16. Vrednost parametra inhibicije pri teh treh hierarhijah je različen za vsako izmed njih, kar pomeni, da vpliv inhibicije pri gradnji hierarhije na natančnost delovanja ni vplival. Vse tri hierarhije pa so bile zgrajene z enako vrednostjo drugih dveh parametrov in sicer je bil AGC izklopljen, parameter nonHalucinatedPartPercent

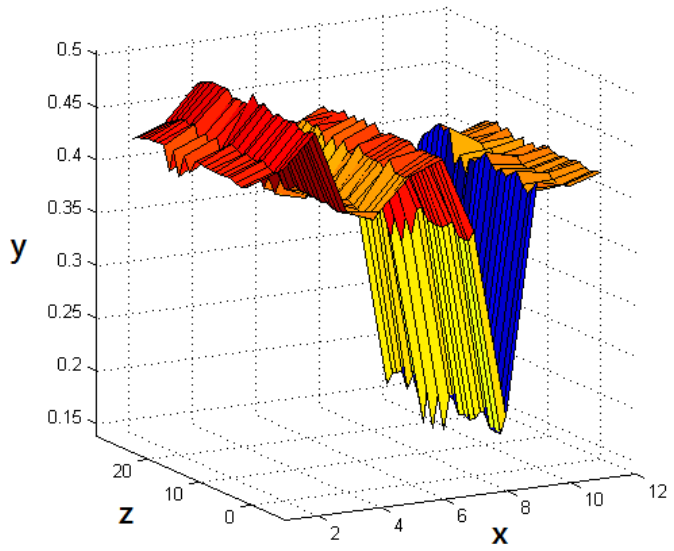
pa je imel vrednost 0,8. Številke hierarhije, ki jih vidimo na ordinatni osi, po vrsti določajo hierarhije, ki so bile zgenerirane v trojni zanki, kjer je zunanja zanka določala vrednost parametra nonHalucinatedPartPercent, ki smo ga po vrsti vzeli iz množice števil 40, 60, 80, 90, ugnezdena zanka je določala vrednost vpliva inhibicije in sicer po vrsti 0,1, 0,5 in 0,9, najbolj ugnezdena zanka pa je določala ali je AGC izklopljen ali vklopljen v takem vrstnem redu. Številka hierarhije 0 torej pomeni, da je bila hierarhija zgrajena s parametri nonHalucinatedPartPercent: 40, vpliv inhibicije: 0,1 in AGC je bil izklopljen.

Hierarhije smo na ta način evalvirali na bazi zgoraj omenjenih dvanajstih pesmi. Kakor je opisano tudi pod sliko, smo ugotovili, da je za točnost delovanja pri učenju hierarhij najbolj bistven parameter halucinacije, ki ob previsoki vrednosti poslabša delovanje. Parameter, ki ga opazujemo pri halucinaciji (nonHalucinatedPartPercent), predstavlja odstotek pod-delov v pod-drevesu nekega dela, ki morajo biti aktivni, da je ta del označen kot aktiven.

Prišli smo do sklepa, da izklopljen AGC v kombinaciji s parametrom nonHalucinatedPartPercent nastavljenim na 80 pri gradnji hierarhije vrne hierarhije, ki delujejo slabše kakor hierarhije zgrajene z ostalimi vrednostmi parametrov. Pričakovali smo, da bo ob nižji vrednosti parametra halucinacije model deloval bolje, ker bo zapolnil več izgubljenih informacij, prišli pa smo do rezultata, da dobro deluje tudi pri zelo visoki vrednosti. Nenavadno je, da pri vrednosti 80 deluje slabo, pri vrednostih 90 in 60 pa bolje. Sklepamo, da ravno takrat (pri vrednosti 80) pridemo do meje, ko se nepravilnosti, ki jih vklopljen AGC sicer odpravi, zaradi že dovolj visoke stopnje halucinacije upoštevajo v naših končnih rezultatih, medtem ko nepravilnosti, ki jih ne odpravimo z AGC-jem, s halucinacijo zaradi prenizkega vpliva še niso odstranjene. Model pri omenjenih parametrih ne deluje veliko slabše, smo pa potrdili domnevo, da vklopljen AGC vpliva na izboljšanje povprečne natančnosti hipotez.

Za iskanje optimalne hierarhije in interpretacijo delovanja različnih hierarhij smo natančnost hierarhij za primerjavo med sabo prikazali grafično, kar je prikazano na sliki 4.5.

Po kratki evalvaciji različno zgrajenih hierarhij so rezultati pokazali, da je za najbolj optimalno delovanje modela potrebno pri gradnji vključiti mehanizem za



Slika 4.5: Abscisna os predstavlja posamezne pesmi iz zbirke, ordinata predstavlja povprečno natančnost naših hipotez na okvir, aplikata pa posamezne hierarhije. Na podlagi prikaza smo v nadaljevanju iskali nepravilnosti delovanja in tudi nepravilnosti v bazi pesmi, ki jih uporabljamo kot testno množico.

samodejno uravnavanje jakosti, prameter nonHalucinatedPartPercent pa nastaviti na vrednost okrog 0,4.

Kot zaključek evalvacije različno naučenih hierarhij moramo poudariti, da rezultat ni neodvisen od hierarhije, saj se že samo med temi dvanajstimi skladbami odstotek povprečne natančnosti na okvir pri isti skladbi, ki rezultate pridobi z različnimi hierarhijami, razlikuje tudi za 8,5 odstotka. Tak rezultat potrjuje sklepanja o pomembnosti tega kako so zgrajeni deli, ki jih bomo uporabljali pri delovanju modela. Čeprav smo boljše hierarhije iskali že kar na podlagi delovanja modela, je pomembna sama struktura hierarhije, ki pa smo jo tako najlažje ovrednotili.

4.4 Evalvacija rezultatov modela

4.4.1 Opisi mer

Mere, ki smo jih izbrali, so standardne mere za analizo uspešnosti razvrščanja in so uveljavljene tudi na področju MIR, zato jih lahko primerjamo z ostalimi rezultati na istem opravilu. Izbrali smo jih tudi zato, ker so vsaka po svoje zelo informativne glede delovanja modela, kar je opisano v nadaljevanju. Izračunali smo jih za vsako skladbo posebej in pri primerjavah celotnih podatkovnih zbirk vzeli povprečje mer za vse skladbe skupaj. Model smo ovrednotili s tremi merami: natančnost (*ang. precission - PRE*) (4.1), priklic (*ang. recall - REC*) (4.2) in povprečna natančnost na okvir (PNO).

$$PRE = \frac{TP}{TP + FP} \tag{4.1}$$

$$REC = \frac{TP}{TP + FN} \tag{4.2}$$

V enačbah 4.1 in 4.2 oznaka *TP* - pravi pozitivi (*ang. true positive*) predstavlja število hipotez, ki jih pravilno napovemo, torej tistih, ki se nahajajo v ročno transkribiranem posnetku in jih tudi mi napovemo. Oznaka *FP* - lažni pozitivi (*ang. false positive*) predstavlja število tistih, ki so v naših hipotezah, a jih ni v ročno transkribiranem posnetku. Oznaka *FN* - lažni negativi (*ang. false negative*) predstavlja število tistih vrednosti, ki so v ročno transkribiranem posnetku, a jih ni med našimi hipotezami.

Natančnost

Natančnost si lahko v našem primeru razlagamo kakor odstotek verjetnosti, da bo naključno izbrana hipoteza izmed hipotez, ki jih model vrne, pravilna.

Priklic

Priklic si v našem modelu lahko predstavljamo kakor odstotek verjetnosti, da bo, če si med tistimi toni, ki se dejansko pojavijo v nekem časovnem okvirju skladbe,

izberemo en ton, ta tudi med hipotezami našega modela.

Povprečna natančnost na okvir

Natančnost in priklic smo merili nad celo skladbo naenkrat. Povprečna natančnost na okvir pa je mera, ki vzame natančnost vsakega časovnega okvirja in izračuna povprečje. Predstavlja povprečno natančnost časovnega okvirja v skladbi.

4.4.2 Vrednotenje

Za zbirko dvanajstih pesmi smo mere izračunali tako, da smo med hipotezami, ki nam jih model vrne, glede na moč aktivacije upoštevali enako število hipotez na okvir kakor se jih nahaja v ročno transkribiranem posnetku. Na začetku smo dobili rezultate, prikazane v tabeli 4.2.

Tabela 4.2: Rezultati dvanajstih pesmi, kakršne smo dobili na začetku. V drugem stolpcu je za vsako pesem podana povprečna natančnost na okvir.

Ime pesmi	PNO
Aria	0,4187
988-v20	0,4265
Blu4pia2	0,3445
Bp054	0,3481
Bwv780	0,3945
Cal_drea	0,4761
Invert6	0,3389
mz_333_2	0,4631
mz_333_3	0,4638
rubalcaba	0,3129
Smokegtz	0,4675
woods3	0,1240

Z iskanjem najboljših parametrov smo nadaljevali optimizacijo. Spreminjali smo enake parametre kot prej (vpliv inhibicije, halucinacije in mehanizma za uravnavanje jakosti). Tokrat smo to spreminjali pri samem delovanju modela, tako da smo vzeli

najboljše parametre za učenje, ki smo jih definirali v prejšnji fazi, in jih fiksirali. Rezultate nam je uspelo nekoliko izboljšati. Pri tem pa nismo opazovali le zgoraj omenjenih mer, ampak smo bili pozorni na dodatne mere, ki kažejo na opažene nepravilnosti pri delovanju modela.

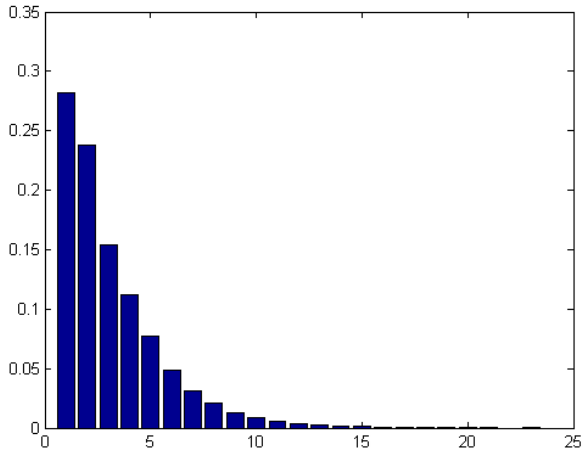
Po prvem vrednotenju modela smo hipoteze, ki nam jih vrne model, preučili s pomočjo grafičnih prikazov in ročno. Ugotovili smo, da se pojavlja izredno veliko oktavnih napak. Ob osnovnem tonu, ki ga proizvede nek vir, se namreč ponavadi vedno pojavijo višji harmoniki, to so toni, ki zvoku določajo višino tona in barvo. Človeško uho vse skupaj zazna kakor en ton, umetno zaznavanje pa višje harmonike težko loči od osnovnega tona. Prvi harmonik ima dvakrat višjo frekvenco od osnovnega tona. To razdaljo v glasbeni terminologiji imenujemo oktava, napačne hipoteze, ki so od pravilnih oddaljene za oktavo, smo zato poimenovali oktavne napake. Zaznali smo tudi, da so poleg odvečnih hipotez večji problem tudi manjkajoče hipoteze. Primer delovanja modela in ročne analize rezultatov, ki jih vrne model, je prikazan v tabeli 4.3. V prvem stolpcu tabele je številka okvirja, v drugem so MIDI vrednosti, ki se v tem okvirju pojavijo v ročno transkribiranem posnetku, v zadnjem stolpcu tabele pa so MIDI vrednosti hipotez, ki nam jih vrne model. V zadnjem stolpcu so okrepljene vrednosti tiste, ki se sploh ne pojavijo pa bi se morale. Vrednosti, ki se pojavijo kot hipoteze in se v ročni transkripciji ne pojavijo, so razporejene po oklepajih med pravilne hipoteze v oklepajih za njimi, in sicer ob tistih vrednostih zaradi katerih predvidevamo, da se zaradi medsebojne harmonične sorodnosti sploh pojavijo. Pri ročni analizi smo potrdili pričakovanja, da so največji problem harmoniki, ki se pojavljajo ob tonih. Skoraj vse hipoteze, ki se pojavijo v določenem okvirju, namreč lahko na podlagi harmonične sorodnosti povežemo z nekim osnovnim tonom, ki se tisti trenutek pojavi. To je lastnost, ki jo je pri analizi večglasnega vhoda težje obravnavati, ker je pojavljajoče harmonike potrebno ločiti od tonov, ki dejansko nastopajo v skladbi. Gre za pričakovano nepravilnost delovanja pri opravilu ocenjevanja osnovnih frekvenc, ne glede na vrsto pristopa, ker se v spektru vhodnega signala pred obdelavo hipotez z modelom pojavi veliko frekvenc.

Tabela 4.3: V tabeli je prikazano delovanje modela in ročna analiza na primeru nekaj okvirjev skladbe mz_333_3 ob izbiri hierarhije, ki nam je vrnila najboljše rezultate.

Okvir	Rešitev	Rezultat
1	78	78 (90 oktava)
2	78	78 (90 oktava; 97 oktava+kvinta)
3	78	78 (90 oktava; 97 oktava+kvinta)
4	78	78 (90 oktava; 97 oktava+kvinta)
5	78	78 (90 oktava)
6	78	78 (90 oktava; 97 oktava+kvinta)
7	78	78 (90 oktava; 71 oktava+kvinta↓)
8	78; 59	78 (90 oktava; 97 oktava+kvinta); 59 (70 kvinta+terca; 83 2 oktavi; 87 2 oktavi+terca;)
9	78; 59	78 (90 oktava; 97 oktava+kvinta); 59 (83 2 oktavi; 87 2 oktavi+terca;)
10	78; 59	78 (90 oktava; 97 oktava+kvinta); 59 (83 2 oktavi; 87 2 oktavi+terca;)
11	78; 59; 75	59 (71 oktava; 83 2 oktavi; 55 terca↓); 75 (87 oktava; 99 2 oktavi; 63 oktava↓); **78** (90 oktava; 97 oktava+kvinta)
12	59; 75	59 (83 2 oktavi; 87 2 oktavi+terca;); 75 (71 terca↓; 99 2 oktavi); 97
200	54; 70; 61	54 (66 oktava; 78 2 oktavi 94 3 oktave+terca); 70 (82 oktava; 89 oktava+kvinta; 101 oktava+kvinta+oktava); 61 (73 oktava; 85 2 oktavi; 92 2 oktavi+kvinta; 97 3 oktave); 87; 95
875	83	83 (95 oktava; 102 oktava+kvinta) 70; 78; 85; 97; 62
6484	80	80 (68 oktava↓; 92 oktava; 64 oktava+terca↓; 88 terca+terca; 72 terca+terca↓); 99; 76; 52; 70; 58; 65
5910	42; 54	42;54 (66 oktava; 78 2 oktavi; 61 kvinta; 85 2 oktavi + kvinta) 92; 86

Glede na stanje rezultatov smo torej uvedli nove mere za opazovanje obnašanja modela in vpliva parametrov na le-te. Prvo mero, ki smo jo uvedli, smo poimenovali odmik od okvirja in predstavlja za posamezen okvir povprečni odmik iskanih hipotez, ki jih model vrne, a jih izgubimo pri izbiri kandidatov, kjer vzamemo le zgornjih n

hipotez razvrščenih po aktivaciji. Želimo si imeti omejitev kandidatov, a to takšno, ki bi večala priklic in natančnost, ne pa le eno od teh dveh mer. Povprečje mere "odmik od okvirja"je za dvanajst pesmi znašalo 1,9686 in se je večinoma gibalo med 1,5 in 2,5. V prihodnosti bi mero lahko uporabili za izboljšanje algoritma za izbiro kandidatov. Na splošno je odmik pravilnih hipotez od izbranega okvirja največkrat enak 1 ali 2, kar prikazuje tudi porazdelitev odmikov na sliki 4.6.

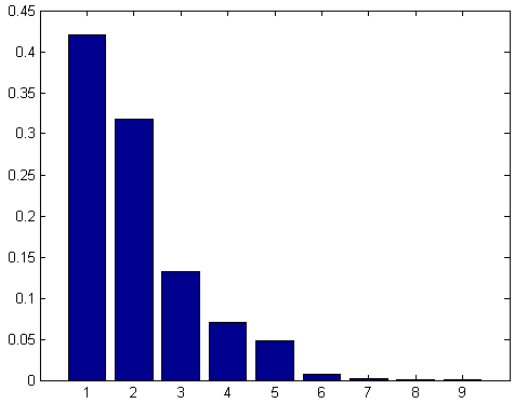


Slika 4.6: Slika prikazuje porazdelitev odmikov pravilnih hipotez od okvirja izbranih kandidatov, ko za izbiro kandidatov gledamo n kandidatov z najvišjo aktivacijo, kjer je n enak številu hipotez na okvir v ročno transkribiranem posnetku. Abscisna os nam predstavlja vrednost odmika, ordinatna os pa odstotek takih odmikov glede na vse odmike, ki jih opazujemo.

Taki rezultati, ki jih moramo obravnavati z ozirom na stopnjo polifonije, so po ročni analizi delovanja pričakovani. Povprečna stopnja večglasnosti na zbirki dvanajstih pesmi znaša 2,55. Oktavne napake, ki smo jih zaznali, pa se pri tistem tonu, ki ga model zazna kot najmočnejšega, skoraj vedno pojavijo z enako aktivacijo kakor osnovni ton. To pomeni, da je skoraj vedno en ton v okviru nepravilen, kar pomeni da bo ena izmed pravilnih hipotez izpadla iz okvirja (odmik bo enak 1). Pri večglasju se isti pojav ponavadi ponovi še pri enem tonu, kar pomeni, da se odmik

naslednjega pravilnega tona poveča še za ena. Po opazovanju te mere sklepamo, da model za vsak ton ponavadi vrne še en sorodni harmonik, ki ima zelo podobno aktivacijo kakor osnovni ton ali pa kar isto. Med množico izbranih hipotez ni mogoče ločiti, katere hipoteze so harmoniki in katere hipoteze so pravilne. Tako delovanje smo pričakovali, kljub temu pa to ni dobro. Želeli smo si, da bi model čim več informacije o sorodnih harmonikih nekega tona zajel v delu, ki določa pravi ton, in da te informacije model ne bi še enkrat uporabil.

Druga mera, ki smo jo opazovali, je število manjkajočih hipotez na okvir, torej tistih hipotez, ki se v rezultatu modela sploh ne pojavijo, čeprav se v skladbi frekvence pojavijo. Porazdelitev števila izgubljenih na okvir je prikazana na sliki 4.7.

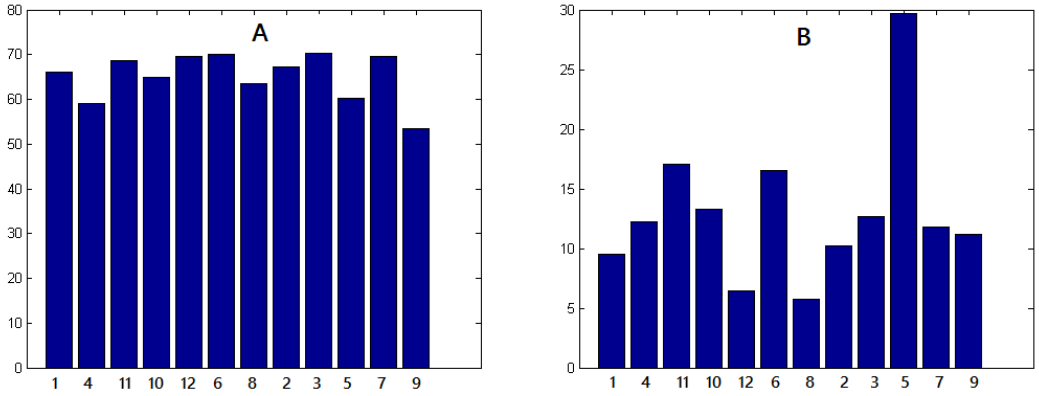


Slika 4.7: Slika prikazuje porazdelitev hipotez na okvir, ki manjkajo. Abscisna os predstavlja število izgubljenih na okvir, ordinatna os pa odstotek takega števila izgubljenih na primeru dvanajstih skladb.

Na zbirki dvanajstih pesmi smo izračunali, da se povprečno pojavlja 1,02 manjkajočih hipotez na okvir. S tem mislimo na število tistih tonov, ki se pojavijo v transkribiranem posnetku, med našimi hipotezami pa ne. Porazdelitev kaže, da je nekaj tudi takih okvirjev, kjer je to število večje. Pričakovali smo, da bo en del rezultatov tak, predvsem zato, ker se nek ton lahko včasih pojavi hkrati kakor višjih harmonik nekega drugega tona v večglasju in bi informacija o njem lahko zaradi

delovanja našega modela bila uporabljena drugje. Prav tako se take stvari dogajajo zaradi nepravilnosti v signalu, ko zaradi takih stvari nek del ni izbran kot aktiven, ker ima aktivnih premalo pod-delov.

Tretja mera, ki smo jo analizirali in pogledali, je število oktavnih napak navzgor in navzdol, število smo definirali tako, da smo šteli vse hipoteze, ki se od ročno transkribiranih vrednosti razlikujejo za eno oktavo in število ločili na vrsto oktavne napake, glede na to, ali je bila napačna frekvenca nižja (oktavna napaka navzdol) ali višja (oktavna napaka navzgor) od vrednosti ročno transkribiranega posnetka. Odstotek oktavnih napak, ki se pojavljajo na zbirki dvanajstih skladb, je grafično prikazan na sliki 4.8.



Slika 4.8: Slika prikazuje odstotek oktavnih napak navzgor (graf A) in navzdol (graf B). Abscisna os označuje indeks skladb, ordinatna os pa odstotek pojavitve oktavnih napak za vse aktivacije ročno transkribiranega posnetka.

Ker so skladbe same po sebi ponavadi sestavljene iz tonov, ki so si harmonično sorodni, se potem harmoniki takih tonov v skladbi razporedijo tako, da njihovo zaporedje v funkciji, ki določa verjetni ton določene razporeditve harmonikov, zelo pogosto vzorec prepozna kot drug ton. Največkrat se pojavi problem pri zaznavanju prvega harmonika kakor originalnega.

Ob veliki večini aktivacij se pojavi oktavna napaka navzgor, kar pomeni, da bi ustrezen algoritem, ki bi te napake zaznaval in izločil iz posnetka, lahko izboljšal

delovanje modela. Poleg odstranitve sorodnih harmonikov, ki se pojavljajo med hipotezami, bo očitno potrebno urediti tudi problem hipotez, ki jih model sploh ne zazna. To bomo v nadaljevanju poskušali rešiti z dodatnim post-procesiranjem rezultatov modela.

Ob opazovanju spreminjanja mer smo poskušali poiskati optimalne parametre za delovanje modela, preučevanje delovanja je tudi pomagalo k odpravi nekaterih nepravilnosti pri delovanju modela in k odkrivanju slabo nastavljenih parametrov. Rezultate nam je uspelo nekoliko izboljšati, kar je prikazano v tabeli 4.4.

Ugotovili smo, da je tudi v drugi fazi, torej pri delovanju modela na vhodnem signalu bolje, če je AGC vklopljen. Optimalna vrednost parametra za uravnavanje halucinacije je 40, vrednost parametra, ki uravnava inhibicijo pa 0,9.

Največji vpliv na vse mere je imel parameter inhibicije, ki je zmanjšal mero "odmik od okvirja", zmanjšal je število oktavnih napak tako navzgor kot navzdol, kar smo pričakovali in je dobro.

Tabela 4.4: V tabeli je prikazana povprečna natančnost na okvir v prvem in natančnost v drugem stolpcu za dvanajst skladb, na katerih smo opazovali delovanje modela. Prikazano je izboljšanje delovanja po optimiziranju parametrov.

	P. natančnost na okvir	Natančnost
Pred izboljšavo	0,30	0,32
Po izboljšavi	0,39	0,40

4.5 Primerjava

Model smo za ovrednotenje in primerjavo rezultatov s trenutnim stanjem istega opravila na področju MIR preizkusili na prosto dostopni zbirki MAPS (MIDI Aligned Piano Sounds) [11], ki vsebuje klavirske posnetke, MIDI zapise teh posnetkov ter tekstovne datoteke z informacijami o posnetkih. Zbirka MAPS je razdeljena na več delov, glede na vrsto klavirja s katerim so proizvedene skladbe. Vsak del vsebuje posnetke posameznih tonov, akordov, naključnih kombinacij tonov in skladb. Mi

smo model testirali na skladbah vsakega dela, ki so v vsakem delu zbirke zbrane v direktoriju MUS.

Za vsako skladbo smo naredili tri izračune. Prvi je bil normalen izračun točnosti glede na pridobljene hipoteze, drugi, ki ga v tabeli 4.5 označujemo s pripono 12 ob imenu dela zbirke, spregleda oktavne napake navzgor in navzdol. Tretji način računanja točnosti, ki smo ga uporabili, spregleda vse oktavne napake in njihove večkratnike. Ta način primerja le tonske razrede. V tabeli 4.5 so rezultati tretjega načina prikazani v vrstici, kjer je imenu dela zbirke dodana pripona mod. Za MIDI vrednosti hipotez in transkribiranega posnetka smo namreč povsod vzeli originalne vrednosti po modulu 12.

Model smo preizkusili na vseh devetih delih zbirke MAPS.

Neposredna primerjava nakazuje, da so predstavljeni rezultati slabši od trenutno najboljših na tem področju. Wenninger [32] predlaga metodo, ki na isti bazi deluje s povprečno natančnostjo na okvir 77.1 %, Böck [4] pa predlaga metodo, ki ima povprečno natančnostjo na okvir 68.7 %. Nam [24] predstavi metodo z globokimi arhitekturami, ki ima oceno F na isti bazi 74.4 %, vendar je rešitev namenjena zgolj transkripciji klavirske glasbe.

Tabela 4.5: Rezultati prvih dveh delov zbirke, ki so izračunani kot povprečje vseh skladb posameznega dela. Rezultate smo pridobivali na celotnih skladbah in odražajo trenutno delovanje modela. Za vsak del smo izračunali natančnost na tri načine, ki so opisani zgoraj.

Ime dela zbirke	PNO	PRE	REC
AkPnBcht	0,5797	0,2243	0,5682
AkPnBcht_12	0,7026	0,4263	0,6824
AkPnBcht_mod	0,8753	0,3185	0,6799
AkPnBsdf	0,6321	0,2193	0,6337
AkPnBsdf_12	0,7267	0,5010	0,7078
AkPnBsdf_mod	0,8929	0,3830	0,7219
AkPnCGdD	0,6630	0,2338	0,6644
AkPnCGdD_12	0,7193	0,4148	0,6979
AkPnCGdD_mod	0,8960	0,3706	0,7261
AkPnStgb	0,2834	0,1210	0,2774
AkPnStgb_12	0,4647	0,2752	0,4410
AkPnStgb_mod	0,6280	0,4244	0,6329
ENSTDkAm	0,5250	0,2067	0,5288
ENSTDkAm_12	0,6300	0,3809	0,6128
ENSTDkAm_mod	0,8326	0,3572	0,6510
ENSTDkCl	0,5648	0,2201	0,5638
ENSTDkCl_12	0,6436	0,3934	0,6243
ENSTDkCl_mod	0,8395	0,3752	0,6388
SptkBGAm	0,6581	0,2963	0,6593
SptkBGAm_12	0,6958	0,5122	0,6725
SptkBGAm_mod	0,8666	0,4618	0,7015
SptkBGCl	0,6940	0,3021	0,6987
SptkBGCl_12	0,7132	0,5040	0,6953
SptkBGCl_mod	0,8647	0,4543	0,6852
StbgTGd2	0,6161	0,3356	0,6158
StbgTGd2_12	0,6334	0,5170	0,6074
StbgTGd2_mod	0,8058	0,4856	0,6011

Poglavje 5

Izboljšave

Model do sedaj pri opravilu ocenjevanja osnovnih frekvenc ni uporabljal nobenega post-procesiranja pridobljenih hipotez. Prostor za izboljšave nam ponuja klasifikacija, kajti model je zaenkrat sam deloval kot klasifikator. Z boljšimi klasifikacijskimi metodami bi točnost klasifikacije utegnili izboljšati. Opazovanje modela nas je prav tako privedlo do zaključka, da bi bilo ugodno poiskati ali razviti algoritem, ki bo odstranjeval oktavne napake. Najprej smo rezultate poskusili izboljšati z nenegativno matrično faktorizacijo, ker se je ta prav na opravilu ocenjevanja osnovnih frekvenc že izkazala za uspešno [1, 5, 8, 9, 29].

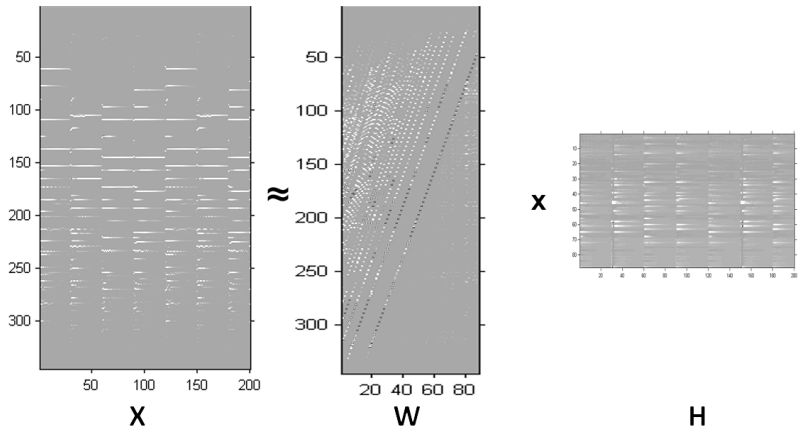
5.1 Nenegativna matrična faktorizacija

Nenegativna matrična faktorizacija (*ang. non-negative matrix factorization - NMF*) je postopek, kjer neko matriko X zapišemo kot produkt dveh nenegativnih matrik W in H. Ponavadi točna rešitev ne obstaja in je zato rezultat približek, pridobljen z numeričnimi metodami, tako da velja:

$$X \approx W \cdot H. \tag{5.1}$$

Metoda se uporablja na večih področjih, na primer pri računalniškem vidu, kemometriji in pri priporočilnih sistemih.

Ideja metode je, da se zaradi nenegativnosti matrik W in H osnovno matriko zapiše kakor kombinacijo posameznih stolpcev matrike W, z utežmi ki jih določa matrika H. Na področju transkripcije glasbe se ta metoda uporablja predvsem zato, ker lahko eno izmed matrik obravnavamo kot slovar, drugo pa kot aktivacije zapisov slovarja v določenem časovnem okvirju. Pri ocenjevanju osnovnih frekvenc se uporablja več različnih načinov za iskanje pojavljajočih se osnovnih frekvenc. Nekateri uporabljajo metode za iskanje tona v slovarju, drugi pa uporabljajo že vnaprej definiran slovar, kjer za vsak stolpec uporabimo spekter posameznega tona. Mi smo se lotili faktorizacije na slednji način. Model smo najprej preuredili, da nam namesto



Slika 5.1: Na sliki vidimo primer delovanja NMF na skladbi, ki predstavlja enostavno zaporedje akordov zaigranih na klavir. Matrika X predstavlja hipoteze, ki jih vrne naš model, slovar pa je bil zgrajen iz spektra posnetih tonov klavirja. Rezultat je matrika aktivacij H, kjer po abscisni osi teče čas, po ordinatni pa stolpci slovarja W. Temnejše barve predstavljajo močnejše aktivacije.

MIDI vrednosti hipoteze vrne v obliki številke frekvenčnega kanala (0-345). Matrika X ima v našem primeru dimenzijo *št.okvirjev* x *345*. Stolpci predstavljajo časovne okvirje, vsako polje v stolpcu pa predstavlja moč aktivacije, ki jo naš model določi za posamezen frekvenčni kanal v tistem časovnem okvirju. Slovar oz. matrika W

ima dimenzijo *88* x *345*. Vsak stolpec predstavlja povprečne aktivacije frekvenčnih kanalov za posnetek klavirskega tona, ki jih vrne naš model. Slovar smo zgradili iz posnetkov 88 klavirskih tipk. Kasneje smo ga zamenjali s podobno zgrajenim slovarjem, le da je bil ta zgrajen iz povprečnih aktivacij frekvenc v spektru skladbe in ne iz hipotez našega modela, ker se je izkazalo, da tak slovar deluje bolje. Matrika aktivacij H z dimenzijo *št. okvirjev* x *88* v stolpcih določa moč aktivacije za vsako komponento slovarja. To je matrika, ki jo z NMF-jem določimo in nas zanima. Za boljše razumevanje zgradbe matrike H je tu opis: v stolpcu 1 so za časovni okvir 1 podane uteži, ki določajo s kakšnimi aktivacijami se stolpci slovarja pojavijo v časovnem okvirju. V naši terminologiji: posamezno polje matrike H določa, s kakšno aktivacijo se nek ton izmed 88 pojavi v časovnem okvirju enakemu številki stolpca. Tista polja, ki imajo aktivacijo večjo od nič, torej določajo nove hipoteze. Grafični prikaz delovanja je prikazan na sliki 5.1.

Metoda, ki smo jo uporabili, je primerna predvsem zato, ker pri iskanju rešitve stremi k čim višji stopnji redkosti matrike H (*ang. sparsity*). Gre za to, da za rekonstrukcijo signala porabimo čim manj elementov slovarja (čim manj elementov matrike H je večjih od 0). S tem se poskušamo znebiti vseh harmonično sorodnih frekvenc, ki se pojavljajo ob osnovnih tonih in nam vračajo nepravilne hipoteze.

Algoritem, ki smo ga uporabili, se rešitvi približuje z metodo padajočega gradienta. Ob tem upoštevamo omejitve redkosti, tako da ima vsak vektor (okvir) na koncu želeno ℓ_2 in ℓ_ϵ normo. Predstavljen je bil v [8] in je prikazan z našo kodo:

```
stepsize=0.1;
epsilonnorm=0.5;
steps=500;
for i=1:steps
    % metoda padajocega gradienta
    H = H-stepsize*dictionary'*(dictionary*H-X);
    % projekcija vektorja na l_epsilon hiperravnino
    s=H+(epsilonnorm-sum(tanh(H.^2)))/size(H,1);
    m=epsilonnorm/size(H,1);
    % resevanje kvadratne enacbe, tako da
    % ima projekcija vektorja zeljeno l_2 normo
```

```
        alfa=(-(s-m)'*m+sqrt(((s-m)'*m).^2-sum(s-m)^2*sum(m.^2-norm(X)^2)));
        alfa = alfa /(sum(s-m)^2+eps);
        s=m+alfa'.*(s-m);
        s(s<0)=0;
        H=s;
end
```

Delovanje NMF-ja smo preizkusili na zbirki dvanajstih pesmi, da bi najprej pogledali primerjavo rezultatov modela z NMF-jem in brez NMF-ja. Pričakovali smo, da bi bile možne izboljšave, prišli pa smo do ugotovitve, da na našem modelu zaenkrat NMF ne deluje dobro, kar je prikazano v tabeli 5.1.

Tabela 5.1: V tabeli so opisani rezultati, ki smo jih pridobili z našim algoritmom NMF in so izračunane z merami, ki smo jih izračunali enako kakor v poglavju 4.5. Prikazano je povprečje rezultatov na zbirki dvanajstih pesmi.

	povprečna natančnost na okvir	natančnost	priklic
Brez uporabe NMF	0.5192	0.1867	0.5170
Brez uporabe NMF_mod12	0.7834	0.3156	0.6181
Z uporabo NMF	0.0922	0.0267	0.0877
Z uporabo NMF_mod12	0.4871	0.1492	0.3268

Kakor so pokazali rezultati, NMF zaenkrat sploh ni uporaben za post-procesiranje rezultatov našega modela. Očitno povsem pokvari naše hipoteze. Iz tega lahko sklepamo, da tega algoritma na našem modelu ne moremo uporabiti. Možna razlaga za to je, da je problem v neaditivnosti modela. Aktivacije se v modelu ne seštevajo tako, da bi lahko končen rezultat modela preprosto razčlenili kot seštevek posameznih tonov. V modelu delujejo razne funkcije, ki spreminjajo aktivacije, tako da te niso le posledica preprostega seštevanja. Predvidevamo, da je ena izmed ovir že sam izračun aktivacij, kjer na vsaki plasti hiperbolični tangens preslika aktivacije na interval [0,1).

Možno je tudi, da algoritma NMF nismo dovolj optimizirali, ker je kar nekaj parametrov, ki jih lahko spreminjamo in pomembno vplivajo na rezultat modela. Pomembna je tudi metoda, ki meri divergenco med V in WH. Glede na to obstaja več

različnih NMF postopkov, ki delujejo različno. Tu je prostor za nadaljnje raziskave in izboljšanje.

Poglavje 6

Zaključek

V nalogi smo spoznali področje pridobivanja informacij iz glasbe in se bolj posvetili opravilu ocenjevanja osnovnih frekvenc. Kompozicionalni hierarhični model, ki se je prej na področju MIR uporabljal za prepoznavanje akordov, je bil preoblikovan tako, da smo pridobili hipoteze, ki ocenjujejo frekvence, ki se pojavljajo v določenem časovnem okvirju vhodnega zvoka. Dokazali smo, da se model da uporabljati za to opravilo. Delovanje modela smo ovrednotili, optimizirali in primerjali z rezultati, ki jih na isti podatkovni zbirki dosegajo drugačne metode. Naši rezultati niso primerljivi z najboljšimi, so pa obetavni in predvidevamo, da jih bomo z dodatnimi nadgradnjami delovanja izboljšali. Trenutno se ukvarjamo z izboljšavo nenegativne matrične faktorizacije. V prihodnosti bomo najprej implementirali metodo podpornih vektorjev, kar bi z boljšo klasifikacijo utegnilo izboljšati natančnost hipotez. Potem bomo z istimi metodami poskusili izvesti transkripcijo melodije, tako da bomo na obstoječih hipotezah uporabili strojno učenje.

Literatura

[1] Samer A. Abdallah and Mark D. Plumbley. Unsupervised analysis of polyphonic music by sparse coding. *IEEE Transactions on Neural Networks*, 17(1):179–196, 2006.

[2] Eric Battenberg and David Wessel. Analyzing Drum Patterns Using Conditional Deep Belief Networks. In *Proceedings of ISMIR*, pages 37–42. Citeseer, 2012.

[3] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 41(3):407–434, 2013.

[4] S. Bock and Markus Schedl. Polyphonic piano note transcription with recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–124. IEEE, 2012.

[5] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Discriminative Non-negative Matrix Factorization for Multiple Pitch Estimation. In *Proceedings of ISMIR*, pages 205–210. Citeseer, 2012.

[6] Judith C Brown and Miller S Puckette. An efficient algorithm for the calculation of a constant Q transform. *The Journal of the Acoustical Society of America*, 92(5):2698–2701, 1992.

[7] Y E H Chunghsin. *Multiple fundamental frequency estimation of polyphonic recordings*. PhD thesis, Universit{é} Lille 1, 2008.

[8] Arshia Cont. Realtime multiple pitch observation using sparse non-negative constraints. In *Proceedings of ISMIR*, 2006.

[9] Arnaud Dessein, Arshia Cont, Guillaume Lemaitre, and Others. Real-time polyphonic music transcription with non-negative matrix factorization and beta-divergence. In *Proceedings of ISMIR*, pages 489–494, 2010.

[10] Sander Dieleman, Philémon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In *Proceedings of ISMIR*, pages 669–674. University of Miami, 2011.

[11] V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. 2010.

[12] Sanja Fidler and Ales Leonardis. Towards scalable representations of object categories: Learning a hierarchy of parts. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[13] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC Music Database: Popular, Classical and Jazz Music Databases. In *Proceedings of ISMIR*, volume 2, pages 287–288, 2002.

[14] Philippe Hamel and Douglas Eck. Learning Features from Music Audio with Deep Belief Networks. In *Proceedings of ISMIR*, pages 339–344. Utrecht, The Netherlands, 2010.

[15] Philippe Hamel, Sean Wood, and Douglas Eck. Automatic Identification of Instrument Classes in Polyphonic and Poly-Instrument Audio. In *Proceedings of ISMIR*, pages 399–404. Citeseer, 2009.

[16] Eric J. Humphrey, Juan P. Bello, and Yann LeCun. Feature learning and deep architectures: new directions for music informatics. *Journal of Intelligent Information Systems*, 41(3):461–481, 2013.

[17] Eric J. Humphrey, Juan Pablo Bello, and Yann LeCun. Moving Beyond Feature Design: Deep Architectures and Automatic Feature Learning in Music Informatics. In *Proceedings of ISMIR*, pages 403–408. Citeseer, 2012.

[18] Eric J. Humphrey, Taemin Cho, and Juan Pablo Bello. Learning a robust tonnetz-space transform for automatic chord recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 453–456. IEEE, 2012.

[19] Eric J. Humphrey, Aron P. Glennon, and Juan Pablo Bello. Non-linear semantic embedding for organizing large instrument sample libraries. In *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*, volume 2, pages 142–147. IEEE, 2011.

[20] Alexandre Lacoste and Douglas Eck. A supervised classification algorithm for note onset detection. *EURASIP Journal on Advances in Signal Processing*, 2007, 2006.

[21] Honglak Lee, Peter Pham, Yan Largman, and Andrew Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems*, pages 1096–1104, 2009.

[22] Tom Li, Antoni B. Chan, and A. Chun. Automatic musical pattern feature extraction using convolutional neural network. In *Proc. Int. Conf. Data Mining and Applications*, 2010.

[23] L. S. Lloyd and H. Boyle. *Intervals, Scales and Temperaments.* St. Martin's Press, 1963.

[24] Juhan Nam, Jiquan Ngiam, Honglak Lee, and Malcolm Slaney. A Classification-Based Polyphonic Piano Transcription Approach Using Learned Feature Representations. In *ISMIR*, pages 175–180, 2011.

[25] Nicola Orio. *Music retrieval: A tutorial and review.* now publishers Inc, 2006.

[26] Matevž Pesek, Ales Leonardis, and Matija Marolt. Compositional Hierarchical Model for Music Information Retrieval. *Master thesis.*

[27] Matevž Pesek, Aleš Leonardis, and Matija Marolt. Compositional Hierarchical Model for Music Information Retrieval. *Proceedings of ISMIR, In Press*, 2014.

[28] Erik M. Schmidt and Youngmoo E. Kim. Learning emotion-based acoustic features with deep belief networks. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 65–68. IEEE, 2011.

[29] Paris Smaragdis. Polyphonic pitch tracking by example. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pages 125–128. IEEE, 2011.

[30] Rainer Typke, Frans Wiering, and Remco C. Veltkamp. A survey of music information retrieval systems. 2005.

[31] George Tzanetakis. Music Information Retrieval. *In Press*, 2014.

[32] F. Weninger, C. Kirst, B. Schuller, and H.-J. Bungartz. A discriminative approach to polyphonic piano note transcription using supervised non-negative matrix factorization. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6–10, 2013.