Univerza
*v Ljubljani*

Fakulteta *za računalništvo*
*in informatiko*

Ciril Bohak

# Segmentacija terenskih posnetkov slovenske ljudske pesmi

Magistrska naloga

Mentor: doc. dr. Matija Marolt

Ljubljana, 2011

Univerza
*v Ljubljani*

Fakulteta *za računalništvo*
*in informatiko*

*Tržaška 25*
*1000 Ljubljana, Slovenija*
*telefon: 01 476 84 11*
*faks: 01 426 46 47*
*www.fri.uni-lj.si*
*e-mail: dekanat@fri.uni-lj.si*

Št.: 137-MAG-RI/2011
Datum: 13. 01. 2011

**Ciril BOHAK**, univ. dipl. inž. rač. in inf.

**L j u b l j a n a**

Fakulteta za računalništvo in informatiko Univerze v Ljubljani izdaja naslednjo magistrsko nalogo

Naslov naloge: **Segmentacija terenskih posnetkov slovenske ljudske pesmi**

**Segmentation of field recordings of Slovenian folk songs**

Tematika naloge:
S segmentacijo lahko zvočni posnetek razdelimo na smiselne enote: daljši posnetek na posamezne pesmi, pesmi na strukturne elemente kot so kitica, refren in uvod, le-te pa nadalje na motive in verze. Segmentacija največkrat predstavlja prvi korak pri semantičnem opisovanju posnetkov, saj omogoča, da posnetek smiselno obravnavamo po delih in olajša postopke kot so iskanje melodije, harmonije, ritma in podobnosti med posnetki. Algoritmi za segmentacijo temeljijo na iskanju podobnosti značilk, ki opisujejo lastnosti kratkih odsekov analiziranega zvočnega signala; npr. barvo, harmonijo ali ritem. Večina obstoječih pristopov je namenjena predvsem segmentaciji popularne glasbe, ki je tudi komercialno najbolj zanimiva.
V okviru magistrske naloge preučite področje segmentacije glasbenih posnetkov. Izpostavite probleme obstoječih pristopov za segmentacijo posnetkov ljudske glasbe in razvijte postopek, ki bo posebej prilagojen za segmentacijo terenskih posnetkov ljudskih pesmi z upoštevanjem njihovih specifičnosti. Pri tem se osredotočite na algoritme, ki temeljijo zgolj na analizi zvočnih posnetkov in za segmentacijo ne potrebujejo dodatnih informacij. Algoritem integrirajte v širši kontekst segmentacije terenskih posnetkov ljudske glasbe in ga ovrednotite na realni zbirki terenskih posnetkov.

M e n t o r:

doc. dr. Matija Marolt

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

D e k a n:

prof. dr. Nikolaj Zimic

Rezultati magistrskega dela so intelektualna lastnina Fakultete za računalništvo in informatiko Univerze v Ljubljani. Za objavljanje ali izkoriščanje rezultatov magistrskega dela je potrebno pisno soglasje Fakultete za računalništvo in informatiko ter mentorja.

*Besedilo je oblikovano z urejevalnikom besedil LaTeX.*

Univerza
*v Ljubljani*

Fakulteta *za računalništvo*
*in informatiko*

*Tržaška 25*
*1000 Ljubljana, Slovenija*
*telefon: 01 476 84 11*
*faks: 01 426 46 47*
*www.fri.uni-lj.si*
*e-mail: dekanat@fri.uni-lj.si*

Št.: 137-MAG-RI/2011
Datum: 13. 01. 2011

**Ciril BOHAK**, univ. dipl. inž. rač. in inf.

**L j u b l j a n a**

Fakulteta za računalništvo in informatiko Univerze v Ljubljani izdaja naslednjo magistrsko nalogo

Naslov naloge: **Segmentacija terenskih posnetkov slovenske ljudske pesmi**

**Segmentation of field recordings of Slovenian folk songs**

Tematika naloge:
S segmentacijo lahko zvočni posnetek razdelimo na smiselne enote: daljši posnetek na posamezne pesmi, pesmi na strukturne elemente kot so kitica, refren in uvod, le-te pa nadalje na motive in verze. Segmentacija največkrat predstavlja prvi korak pri semantičnem opisovanju posnetkov, saj omogoča, da posnetek smiselno obravnavamo po delih in olajša postopke kot so iskanje melodije, harmonije, ritma in podobnosti med posnetki. Algoritmi za segmentacijo temeljijo na iskanju podobnosti značilk, ki opisujejo lastnosti kratkih odsekov analiziranega zvočnega signala; npr. barvo, harmonijo ali ritem. Večina obstoječih pristopov je namenjena predvsem segmentaciji popularne glasbe, ki je tudi komercialno najbolj zanimiva.
V okviru magistrske naloge preučite področje segmentacije glasbenih posnetkov. Izpostavite probleme obstoječih pristopov za segmentacijo posnetkov ljudske glasbe in razvijte postopek, ki bo posebej prilagojen za segmentacijo terenskih posnetkov ljudskih pesmi z upoštevanjem njihovih specifičnosti. Pri tem se osredotočite na algoritme, ki temeljijo zgolj na analizi zvočnih posnetkov in za segmentacijo ne potrebujejo dodatnih informacij. Algoritem integrirajte v širši kontekst segmentacije terenskih posnetkov ljudske glasbe in ga ovrednotite na realni zbirki terenskih posnetkov.

M e n t o r:

doc. dr. Matija Marolt

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

D e k a n:

prof. dr. Nikolaj Zimic

# I Z J A V A  O  A V T O R S T V U

## magistrskega dela

Spodaj podpisani Ciril Bohak

z vpisno številko 63020013

sem avtor magistrskega dela z naslovom:

Segmentacija terenskih posnetkov slovenske ljudske pesmi

S svojim podpisom zagotavljam, da:

- sem magistrsko delo izdelal samostojno pod vodstvom mentorja:

  doc. dr. Matije Marolta

- so elektronska oblika magistrskega dela, naslova (slov., angl.), povzetka (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko magistrskega dela

- in soglašam z javno objavo elektronske oblike magistrskega dela v zbirki »Dela FRI«.

V Ljubljani, dne 3.10.2011

Podpis avtorja:

Ciril Bohak

# Zahvala

V času dodiplomskega študija se vsakemu študentu zdi diplomsko delo velik zalogaj, kjer zbere vse moči in znanje, pridobljeno tekom študija na fakulteti in jih uporabi za izdelavo zaključnega dela - diplomske naloge. Tudi sam sem se srečal s to stopničko in jo uspešno prestopil. Še več. Zaradi neizmerne želje po novem znanju in izpopolnjevanju sem svoj študij nadaljeval na podiplomskem programu, ki mi ga je nudila prav ista fakulteta kjer sem diplomiral.

Stvari na podiplomskem študiju so se bistveno razlikovale od dodiplomskih. Veliko manj predavanj, nič vaj, veliko samoiniciativnega dela, vpeljava v raziskovalne in akademske vode, po vrhu vsega pa službene obveznosti. Vse to naredi nadaljnji študij precej drugačen na pogled enostavnejši, v samem bistvu pa dosti težji. Kot rezultat dosežene stopnje lahko štejem pričujoče magistrsko delo.

Pričetki so velikokrat obetajoči in vzpodbudni. Ideje se krešejo in ne veš katere izbrati ter jih poskusiti uresničiti. Vsake toliko pa se stvari povsem obrnejo na glavo, ko ugotoviš, kaj se da in česa se ne da narediti. Primanjkuje časa da bi uresničil želene cilje, o katerih si razmišljal mesece pred tem. Veliko ostalih dejavnikov te odvrača od ciljev in sili k izbiri novih, drugačnih, uresničljivejših.

Kljub vsemu sem uspešno prispel do sem. Do trenutka v času, ko si lahko zopet rečem, da sem prestopil še eno izmed stopnic moje izobraževalne poti. Predvsem v zaključnem finišu je delo na magistrski nalogi terjalo ogromno časa, truda, živcev in dobre volje. Kljub vsem oviram na poti mi je uspelo.

Te poti pa k sreči nisem prehodil sam. Ob strani mi je stalo kar nekaj ljudi, ki jim je ta košček besedila namenjen. Največja zahvala gre mentorju doc. dr. Matiji Maroltu, za izkazano pomoč, potrpežljivost, dobro voljo, nasvete in potrebne brce, ki so me pripeljale do zastavljenih ciljev. Prav tako gre zahvala vsem prijateljem, ki so v zadnjih letih poslušali o uspehih in neuspehih mojega dela, me vzpodbujali in mi stali ob strani. Vem, da sem vas predvsem v času pred zaključkom (pre)večkrat zanemarjal in druženje z vami nadomestil z delom na nalogi. Hvala vsem domačim, ki so mi stali ob strani, me bodrili in hkrati priganjali z nič kolikokrat ponovljenim vprašanjem: "*Kako daleč pa si kaj z nalogo?*" Hvala vam. Zelo hvaležen sem tudi vsem članom Laboratorija za računalniško grafiko in multimedije, ki so mi bili vedno pripravljeni pomagati tako pri malenkostih kot večjih stvareh tako v okviru dela na magisteriju kot izven. Hvala Saša, Marko, Alenka, Jernej in Matevž.

Še enkrat hvala Vam prijatelji, ki ste mi stali ob strani me vzpodbujali in bodrili.

*"A true friend never gets in your way unless you happen to be going down."*
*- Arnold H. Glasow*

*Vos amici mei dilecti.*

*(Vam, dragi prijatelji in najdražji.)*

# Kazalo

**Izjava** **vii**

**Povzetek** **xxiii**

**Abstract** **xxv**

**1 Uvod** **1**
1.1 Motivacija . . . . . 1
1.2 Digitalni arhiv EtnoMuza . . . . . 2
1.3 Terenski zapisniki in posnetki . . . . . 3
1.3.1 Zgradba terenskih posnetkov . . . . . 3
1.4 Segmentacija zvočnih posnetkov . . . . . 3
1.5 Cilji in namen dela . . . . . 4
1.6 Vsebina in zgradba magistrske naloge . . . . . 5

**2 Pregled področja** **7**
2.1 Pridobivanje informacij iz glasbe . . . . . 7
2.2 Predstavitev zvočnega signala . . . . . 8
2.3 Računalniška obdelava digitalnih zvočnih zapisov . . . . . 9
2.4 Struktura terenskih posnetkov . . . . . 10
2.5 Glasbena struktura . . . . . 11
2.6 Segmentacija zvočnih posnetkov . . . . . 12
2.6.1 Mel-frekvenčni kepstralni koeficienti . . . . . 12
2.6.2 Kromatski vektorji . . . . . 14
2.6.3 Mere razdalj . . . . . 15
2.6.4 Samopodobnostna matrika . . . . . 17
2.6.5 Matrika s časovnim zamikom . . . . . 18
2.6.6 Nenegativna matrična faktorizacija . . . . . 19
2.6.7 Dinamično ukrivljanje časa . . . . . 20
2.6.8 Skriti Markovski model . . . . . 22
2.6.9 Kombinirane metode . . . . . 22
2.7 Segmentacija ljudske glasbe . . . . . 24

2.7.1 Segmentacija glede na dihalne pavze z detekcijo tišine . . . . 24
2.7.2 Iskanje kitic v posnetkih z DTW . . . . . 24
2.7.3 Segmentacija posnetkov glede na vsebino z verjetnostnim modelom . . . . . 27

**3 Metoda za večnivojsko segmentacijo terenskih posnetkov 31**
3.1 Opredelitev problema . . . . . 32
3.2 Segmentacija prvega (višjega) nivoja . . . . . 32
3.3 Segmentacija drugega (nižjega) nivoja . . . . . 32
3.3.1 Predpriprava vhodnega signala . . . . . 33
3.3.2 Iskanje vokalnih pavz . . . . . 33
3.3.3 Iskanje kandidatov za začetke kitic in njihova ocena . . . . . 36
3.3.4 Izbira dejanskih začetkov . . . . . 39
3.4 Izboljšava visokonivojske segmentacije z nizkonivojskim pristopom . 40
3.4.1 Izboljšana natančnost ocene začetka pesmi . . . . . 40
3.4.2 Odprava morebitnega zamika začetka pesmi . . . . . 40

**4 Rezultati eksperimentov 43**
4.1 Priprava posnetkov . . . . . 43
4.2 Segmentacija višjega nivoja . . . . . 44
4.3 Rezultati segmentacije nižjega nivoja . . . . . 44
4.3.1 Določanje optimalnih parametrov . . . . . 45
4.3.2 Ocena prvega koraka metode . . . . . 49
4.3.3 Ocena delovanja celotne metode . . . . . 50
4.3.4 Primeri delovanja . . . . . 50
4.4 Primerjava z obstoječimi metodami . . . . . 54
4.5 Rezultati izboljšav višjenivojske segmentacije . . . . . 56
4.5.1 Izboljšana detekcija začetka pesmi . . . . . 56

**5 Zaključek 57**
5.1 Sklepne ugotovitve . . . . . 58
5.2 Nadaljnje delo . . . . . 58
5.3 Zaključna misel . . . . . 59

**6 Dodatki 61**
6.1 Dodatek A . . . . . 61

**Literatura 64**

# Slike


1.1 Slika prikazuje različne pristope pri segmentaciji zvočnih posnetkov. 4

2.1 Primer notnega zapisa glasbe, opremljenega z besedilom. . . . . 9
2.2 Primer valovne krivulje. . . . . 9
2.3 Slika prikazuje večnivojsko segmentacijo terenskih posnetkov. . . . . 11
2.4 Slika prikazuje samopodobnostno matriko izračunano s pomočjo kosinusne podobnosti med značilnicami MFCC posameznih okvirjev signala. V sliki z dobrim primerom so nad diagonalo označene ponovitve, vzorci pa so razvidni pod diagonalo. Oba primera sta iz digitalnega arhiva EtnoMuza. . . . . 13
2.5 Sliki prikazujeta razporeditev tonskih višin v tonske razrede in kromagram. . . . . 15
2.6 Primera samopodobnostnih matrik dveh pesmi iz digitalnega arhiva EtnoMuza, izračunan na podlagi kromagramov. V sliki z dobrim primerom so nad diagonalo označeni vzorci ponovitev, vidni tudi pod diagonalo. . . . . 16
2.7 Primer matrike s časovnim zamikom izračunane iz samopodobnostne matrike predstavljene v Sliki 2.6(a). V sliki so z belimi črticami označene vodoravne črte, ki ponazarjajo ponavljanje v matriki s časovnim zamikom. . . . . 18
2.8 Sliki prikazujeta prva dva koraka pri izračunu DTW. . . . . 20
2.9 Primer matrike s kumulativno cenilno funkcijo in samopodobnostne matrike na katerih je prikazana najcenejša pot. . . . . 21

3.1 Prikaz delovanja metode. . . . . 33
3.2 Iskanje vokalnih pavz . . . . . 37
3.3 Prikaz iskanja kandidatov za začetke kitic. . . . . 38
3.4 Funkcija verjetnosti, da so kandidati za začetke kitic tudi resnični začetki. . . . . 39

4.1 Primerjava mer precision in recall za pristop iskanja vokalnih pavz na podlagi energije signala pri meri natančnosti $\delta = 2$ sekundi, pri izbiri parametra $\xi_1$. . . . . 46
4.2 Primerjava mer precision/recall za pristop iskanja vokalnih pavz na podlagi zaznavnosti tonske višine pri meri natančnosti $\delta = 2$ sekundi za posamezne vrednosti parametra $\xi_5$. . . . . 47
4.3 Primerjava mer precision/recall za pristop iskanja vokalnih pavz na podlagi energije signala pri meri natančnosti $\delta = 2$ sekundi za posamezne vrednosti parametra $\xi_7$. . . . . 48
4.4 Primerjava mer precision, recall in mere F med posameznimi metodami, tako enostavnimi, kot sestavljenimi. Mera natančnosti je $\delta = 2$ sekundi za posamezne vrednosti parametra. . . . . 50
4.5 Sika prikazuje dva primera pesmi iz arhiva EtnoMuza na katerih deluje metoda nižjenivojske segmentacije zelo dobro. V slikah je z modro izrisan zvočni signal, z oranžno verjetnostna funkcija pojavitve nove kitice, izračunana s pomočjo metode nižjenivojske segmentacije, z zelenimi zvezdicami so označeni dejanski začetki kitic v pesmi določeni ročno, z roza plusi pa so označeni začetki zaznani z metodo opisano v prejšnjem poglavju. . . . . 52
4.6 Sika prikazuje dva primera pesmi iz arhiva EtnoMuza na katerih deluje metoda nižjenivojske segmentacije zelo slabo. . . . . 53
4.7 Slika prikazuje podobnostni matriki 10 sekundnih delov posnetka z začetki v začetku pesmi, pravem začetku tretje kitice in zaznanem začetku tretje kitice. V matrikah je z belo barvo označena podobnost med deli zvočnih posnetkov. . . . . 54
4.8 Sika prikazuje dva primera pesmi iz arhiva EtnoMuza na katerih deluje metoda nižjenivojske segmentacije “povprečno” dobro. . . . . 55

# Tabele


4.1 Tabela statistike tipov enot v terenskih posnetkih. . . . . 43
4.2 Tabela statistike ponavljajočih delov v enotah solo in večglasnega petja. 44
4.3 Nabor vrednosti parametra $\xi_3$ in vrednosti posameznih mer. . . . . 46
4.4 Primerjava metod z različno dolgimi primerjanimi odseki $\xi_8$ za mero natančnosti $\delta = 2$. . . . . 48
4.5 Primerjava metod po meri F za mero natančnosti $\delta = 2$. . . . . 49
4.6 Primerjava osnovne metode višjenivojske segmentacije in predstavljene izboljšave. . . . . 50
4.7 Primerjava rezultatov delovanja naše, Müellerjeve in Kranenburgove metode segmentacije za mero natančnosti $\delta = 2$. . . . . 56
4.8 Primerjava osnovne metode višjenivojske segmentacije in predstavljene izboljšave. . . . . 56

6.1 Nabor vrednosti parametra $\xi_1$ in vrednosti posameznih mer. . . . . 61
6.2 Nabor vrednosti parametrov $\xi_4$ in $\xi_5$ in vrednosti posameznih mer. . 62
6.3 Nabor vrednosti parametrov $\xi_6$ in $\xi_7$ in vrednosti posameznih mer. . 63

# Seznam uporabljenih kratic in simbolov

**CQT** Constant Q Transform - Konstantna-Q Transformacija

**DTW** Dynamic Time Warping - Dinamično ukrivljanje časa

**GNI** Glasbenonarodopisni inštitut

**HMM** - Hidden Markov Model - Skriti Markovski model

**IEC** International Electrotechnical Commission - Mednarodna elektrotehnična komisija

**ISO** International Organization for Standardization - Mednarodna organizacija za standardizacijo

**MFCC** Mel-frequency cepstral coefficients - Mel-frekvenčni kepstralni koeficienti

**MIR** Music Information Retrieval - Pridobivanje informacij iz glasbe

**NaN** Not-a-Number - Nedefinirana vrednost

**NMF** Non-negative matrix factorisation - Nenegativna matrična faktorizacija

**P** Precision - natančnost

**R** Recall - priklic

**RMS** Root mean Square - koren povprečja kvadratov

**SAZU** Slovenska akademija znanosti in umetnosti

**SI-PLCA** Shift-invariant Probabilistic Latent Component Analysis - Verjetnostna latentna komponentna analiza neodvisna od zamika

**STFT** Short-time Fourier transform - Kratko-časovna Fouriereva transformacija

**ZRC** Znanstvenoraziskovalni center

# Povzetek

S segmentacijo lahko zvočni posnetek razdelimo na smiselne enote. Daljši posnetek lahko razdelimo na posamezne pesmi, intervjuje, komentarje ... pesmi na strukturne elemente kot so kitica, refren, uvod ... le-te pa nadalje na motive in verze. Segmentacija največkrat predstavlja prvi korak pri semantičnem opisovanju posnetkov, saj omogoča, da posnetek smiselno obravnavamo po delih in olajša postopke kot so iskanje melodije, harmonije, ritma in podobnosti med posnetki. Algoritmi za segmentacijo temeljijo na iskanju podobnosti značilnic, ki opisujejo lastnosti kratkih odsekov analiziranega zvočnega signala; npr. barvo, harmonijo ali ritem. Večina obstoječih pristopov je namenjenih predvsem segmentaciji popularne glasbe, ki je tudi komercialno najbolj zanimiva. Segmentacija glasbe je vedno bolj zanimiva tudi za ostale glasbene zvrsti.

V okviru naloge smo preučili področje segmentacije glasbenih posnetkov. Izpostavljeni so problemi obstoječih pristopov, kot so: zmanjšana uspešnost pri ljudski glasbi zaradi slabe kvalitete posnetkov, netočnega petja amaterskih pevcev, zunanjih dejavnikov, ki vplivajo na prekinitev ali slabšo izvedbo pesmi. V nalogi je predstavljen razvoj lastne metode za segmentacijo terenskih posnetkov ljudske glasbe na dveh nivojih, ki deluje zgolj na podlagi analize zvočnih posnetkov in ne uporablja morebitnega dodatnega predznanja podanega s strani strokovnjaka. Na višjem nivoju metoda terenski posnetek razdeli v dele z enako vrsto zvoka (govor, solo petje, večglasno petje ...), na nižjem nivoju pa metoda nadalje razdeli dele s solo petjem in večglasnim petjem v ponavljajoče dele. Metoda združuje že razvit pristop segmentacije terenskih posnetkov na višjem nivoju z novo metodo za segmentacijo na nižjem nivoju. S pomočjo informacij pridobljenih z nižjenivojsko segmentacijo izboljša rezultate metode višjenivojske segmentacije.

Kot del naloge smo delovanje razvite metode preverili na zbirki terenskih posnetkov slovenske ljudske glasbe in rezultate ovrednotili.

**Ključne besede:**

ljudska glasba, glasbena struktura, segmentacija glasbe, pridobivanje informacij iz glasbe

# Abstract

With segmentation we can split audio recordings into meaningful parts. Longer recordings can be split into individual songs, interviews, comments ... songs can be split into structural elements such as chorus and verse ... these can be further split into motives and stanzas. Segmentation is usually the first step towards semantic description of recordings, since it enables us to take into account individual parts and simplifies procedures, such as discovering melody, harmony, rhythm and finding similarity of different parts. Segmentation algorithms are based on finding similar features in short parts of the analyzed audio signal, e.g. timbre, harmony or rhythm. Current approaches mostly deal with popular music, which is more commercially interesting. Music segmentation is lately reaching out to other genres as well.

In this thesis we provide an overview of research in music segmentation. We have pointed out the problems of current approaches, e.g. lower performance on folk music because of low quality sound recordings, inaccurate singing of amateur singers, distractions during recording or poor performance. We present a novel approach for two-level segmentation of field recordings of folk music. The method uses audio recordings only and does not rely on prior knowledge given by an expert. On the higher level, method splits field recordings into individual units of the same type (speech, solo singing, choral singing ...). On the lower level, the method further splits parts with solo and choral singing into individual repeating parts. The proposed algorithm combines an existing approach for higher-level segmentation of field recordings with a newly developed low-level segmentation method. Low-level segmentation method is also used to improve high-level segmentation.

As part of this thesis we have tested and evaluated our method on collection of field recordings of Slovenian folk music.

**Keywords:**

folk music, musical structure, music segmentation, music information retrieval

# Poglavje 1

# Uvod

## 1.1 Motivacija

Glasba je del človekovega življenja vsaj že iz časov kamene dobe, kar pričajo najdeni inštrumenti stari preko 50 000 let. Del človekovega vsakdana se je tako prepletal z glasbo skozi celotno zgodovino in je še vedno (po nekaterih ocenah celo vedno bolj) pomemben del človekovega življenja. Človek z glasbo izraža svoja občutja in čustva tako, da glasbo ustvarja (poje ali jo igra) ali pa glasbo posluša, pri čemer je zelo dobrodošla možnost izbire zvrsti, avtorja, pesmi, izbranega posnetka, ali celo zgolj dela posnetka. Takšne možnosti so nam vedno bolj na razpolago, predvsem zaradi razvoja tehnologij, ki omogočajo hitro brskanje med zapisi in neposreden dostop do posameznih delov. Še posebej se to zadnja odraža leta s porastom digitalizacije in razvitih metod za iskanje po digitalni glasbi.

Skozi zgodovino se je glasba prenašala na bodoče rodove preko ljudskega izročila. Prvi zapisi glasbe so temeljili na opisu melodije z označevanjem načina prehodov melodije v samem besedilu pesmi. Kasneje so se začeli pojavljati prvi simbolični zapisi melodije s pomočjo not. Prvi zvočni posnetki, so se pojavili šele v drugi polovici 19. stoletja. Z razvojem tehnologije se je močno spremenil tako način snemanja zvočnih posnetkov, kot tudi njihova kvaliteta ter obstojnost nosilcev zapisa. Zvočni posnetek praviloma vsebuje veliko več informacij o glasbi, kot jih lahko zapišemo v simbolični predstavitvi, saj vsebuje tudi lastnosti konkretne izvedbe glasbe, ki se lahko med seboj močno razlikujejo glede na izvajalce, prostor izvajanja, snemalne pogoje in druge dejavnike.

Ljudsko glasbo so ustvarjali vsakdanji ljudje in tako večinoma ni bila nikoli zapisana ali objavljena, ampak se je med generacijami prenašala zgolj z ustnim izročilom. Pri prenosu pesmi med generacijami je velikokrat prišlo do manjših sprememb v besedilih ali melodiji, kar se odraža v mnogih variantah istih ljudskih pesmi v različnih (bolj ali manj oddaljenih) okoljih. Spremembe večinoma niso namerne in se dogajajo predvsem zaradi pozabljenega besedila ali melodije, z zamenjavo dela besedila

z besedilom druge pesmi ali dodanega besedila.

Tako je bilo vse dokler niso pričeli etnomuzikologi sestavljati zbirk ljudske glasbe, ki so vsebovali transkripcije (notne zapise), besedila in ostale metapodatke o pesmih odkritih na terenu. Z leti so te zbirke postali arhivi, ki se še vedno dopolnjujejo, hkrati pa uporabljajo za najrazličnejše raziskave na področju etnomuzikologije.

V svetu je dostopnih vedno več arhivov ljudske glasbe posameznih geografskih in etničnih področij; verjetno najbolj znana med njimi je zbirka pesmi Essen Folksong Database [42], ki vsebuje preko 20 000 pesmi, večinoma iz Nemčije, Poljske in Kitajske ter manjše zbirke pesmi nekaterih drugih (predvsem evropskih) držav. Zbirko je leta 1995 zbral Schaffrath in predstavil v ***kern* formatu[1]. Dobro znan je tudi digitalni arhiv Finnish Folk Tunes [12] v obsegu 9 000 ljudskih melodij, ki so bile izdane kot zbirka knjig med leti 1898 in 1933 in so bile digitalizirane v letih 2002 in 2003. V svetu je še veliko zbirk ljudske glasbe. Omenimo naj zgolj The American Folk Song Collection [2], in Australian Folk Songs [1].

Raziskave v okviru magistrskega dela so bile izvedene nad podatki iz arhiva EtnoMuza [44], ki vsebuje zbirke ljudske glasbe in plesov Glasbenonarodopisnega inštituta (GNI) Znanstvenoraziskovalnega centra (ZRC) Slovenske akademije znanosti in umetnosti (SAZU).

## 1.2 Digitalni arhiv EtnoMuza

V okviru projekta EtnoMuza: digitalna multimedijska shramba slovenske ljudske glasbene in plesne kulture je med letoma 2006 in 2008 nastal digitalni arhiv predstavljen v [44]. Glasbeni arhiv je nastal v sodelovanju Laboratorija za računalniško grafiko in multimedije in GNI ZRC SAZU. Vse od postavitve se v digitalni arhiv dodaja digitalizirane vsebine obstoječega arhiva, kot tudi nove vsebine.

Digitalni arhiv vsebuje rokopise, zvočne posnetke, plese in slikovno gradivo. Nekatere vsebine so med seboj povezane. Tako je na primer vsak zvočni posnetek povezan s terenskim zapisnikom. V digitalnem arhivu je trenutno preko 13 000 rokopisov, preko 1000 plesov, slikovno gradivo in preko 300 terenskih zapisnikov. Namensko razvit vmesnik omogoča enostavno dodajanje novega gradiva, dodajanje terenskih zapisnikov in povezovanje z gradivom iz arhiva, pregled posameznih vnosov (rokopisov, posnetkov, slik) ali pregled terenskih zapisnikov in z njimi povezanega gradiva. Vmesnik omogoča tudi iskanje med podatki v arhivu na različne načine; iskanje lahko omejimo na posamezno vrsto gradiva, ali ga še dodatno omejimo s pomočjo različnih filtrov, ki jih prilagodimo glede na želene lastnosti iskanih gradiv in si s tem olajšamo iskanje želenih posnetkov.

[1] http://www.lib.virginia.edu/artsandmedia/dmmc/Music/Humdrum/kern_hlp.html

## 1.3 Terenski zapisniki in posnetki

Terensko delo etnomuzikologov obsega iskanje še nepopisanih ljudskih pesmi, ki jih odkrijejo s pomočjo intervjujev s posamezniki na terenu, večinoma starejšimi ljudmi, ki ohranjajo tradicijo ljudske pesmi. Skozi intervjuje spoznajo katere pesmi so se prenesle na današnje generacije. Terenski zapisnik je dokument, ki vsebuje informacije o terenskih posnetkih etnomuzikologov. Etnomuzikologi v dokument zapišejo podatke o tem kdaj in kje je posnetek nastal, kaj posnetek vsebuje, kdo je avtor terenskega posnetka (zapisovalec), kdo vse so osebe na terenskem posnetku (informatorji), s kakšno napravo je bil posnet, ipd.

Terenski posnetek je zvočni posnetek posnet na terenu in vsebuje različne vsebine; na posnetkih so lahko intervjuji s posamezniki ali skupinami ljudi, posnete ljudske pesmi, ki so odpete v solo, večglasni izvedbi ali s spremljavo, inštrumentalna ljudska glasba (zaigrana na eno ali več glasbil), pritrkavanje zvonov ... Intervjuji vsebujejo informacije o izvoru ljudskih pesmi odpetih v terenskem posnetku, opis priložnosti ob katerih so se določene ljudske pesmi pele, pojasnitve morebitnih nejasnosti ...

### 1.3.1 Zgradba terenskih posnetkov

Pri poslušanju zvočnega posnetka lahko človek brez posebnega napora določi, kateri so (glede na zahtevane kriterije) smiselni deli na katere lahko razdelimo nek zvočni posnetek. Tako ni posebno težko določiti v katerem delu posnetka je govor, solo petje, večglasno petje, inštrumentalna glasba ali kakšna izmed kombinacij naštetega; primer takšne segmentacije je prikazan na Sliki 1.1(a).

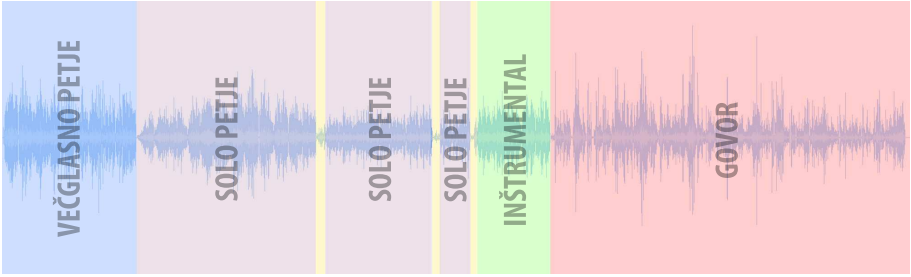
Človek lahko hitro ugotovi ali se v nekem zvočnem posnetku začnejo ponavljati deli s podobnimi značilnostmi; to so lahko deli s podobno ali enako melodijo, ritmom, barvo zvoka, harmonijo ali kombinacijo le-teh; takšni deli lahko predstavljajo posamezne dele pesmi; primer takšne segmentacije je prikazan na Sliki 1.1(b).

Težka naloga pa je za zgoraj opisane procese izdelati formalni postopek, ki bi opisal kako lahko do podobnih rezultatov pridemo avtomatizirano s pomočjo računalnika, kar je edina racionalna možnost uporabe na velikih zbirkah glasbe, ki so danes na voljo.

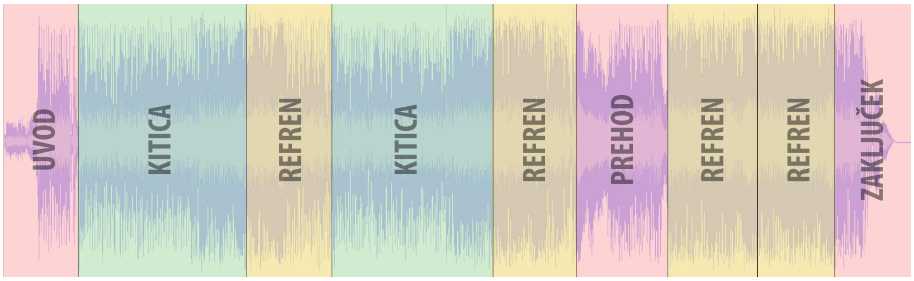
## 1.4 Segmentacija zvočnih posnetkov

V prejšnjem razdelku je predstavljena zgradba terenskih posnetkov na katero lahko gledamo na več nivojih. Posnetek lahko razdelimo na vsebinsko smiselne dele (govor, petje ...) na višjem nivoju in na ponavljajoče ali neodvisne dele na nivoju posamezne pesmi (kitice, verzi ...).

Avtomatsko ugotavljanje oz. razpoznavanje takšne večnivojske zgradbe posnetkov je cilj te naloge. Pri tem bomo najprej preizkusili že obstoječe metode segmentacije na pesmih slovenske ljudske glasbe, ter predstavili lastno metodo za segmen-




(a) glede na tip vsebine (govor, solo petje, večglasno petje, inštrumentalna glasba).




(b) glede na strukturo pesmi (uvod, kitica, refren, prehod, zaključek).

Slika 1.1: Slika prikazuje različne pristope pri segmentaciji zvočnih posnetkov.

tacijo. V nadaljevanju smo predstavljeno nizkonivojsko metodo združili z metodo visokonivojske segmentacije v terenskih posnetkih [30].

## 1.5 Cilji in namen dela

Delo v okviru naloge obsega pregled in opis glavnih pristopov za iskanje smiselnih delov v terenskih posnetkih ljudske glasbe. Smiselni so deli z isto vrsto zvoka (govor, petje ...) in tudi ponavljajoči deli v nekem zvočnem posnetku. V pregledani literaturi opisani v pregledu področja nismo zasledili metode, ki bi reševala opisane probleme na zvočnih posnetjih ljudske glasbe. Obstoječe metode so namenjene segmentaciji zgolj na posameznem nivoju. Razvita metoda je po našem vedenju prva metoda namenjena segmentaciji terenskih posnetkov ljudske glasbe na več nivojih, ki poleg ločenih segmentacij na posameznem nivoju uporabi tudi informacije pridobljene na nižjem nivoju za izboljšanje delovanja na višjem nivoju. V okviru naloge smo razvili nov pristop za iskanje smiselnih delov v zvočnih posnetkih slovenske ljudske glasbe na dveh nivojih. Pristop na prvem nivoju terenski posnetek razdeli glede na vrsto zvoka v posnetku. Na drugem nivoju iščemo dele s ponavljajočimi lastnostmi v delih iz segmentacije prvega nivoja. V nalogi je razvit algoritem predstavljen, analiziran, testiran in primerjan s podobnimi že razvitimi pristopi. Cilji

dela v okviru magistrske naloge so:

- razviti metodo za segmentacijo ljudskih pesmi, ki upošteva specifičnosti tovrstnih posnetkov;
- integrirati razvito metodo z višjenivojsko segmentacijo terenskih posnetkov [30];
- ovrednotiti metodo na zbirki terenskih posnetkov Glasbenonarodopisnega inštituta ZRC SAZU in izvesti primerjavo z obstoječimi segmentacijskimi metodami.

## 1.6 Vsebina in zgradba magistrske naloge

Naloga je sestavljena iz več poglavij, ki smiselno predstavijo posamezni del naloge. Prvo poglavje obsega uvod in predstavitev teme magistrske naloge skupaj s predstavitvijo problemov, ki smo si jih v okviru ciljev naloge zadali rešiti.

Poglavje 2 obsega pregled področja razdeljen v opis splošnih pristopov pri reševanju problema segmentacije zvočnih posnetkov in opis metod pomembnih pri razvoju lasnega pristopa za večnivojsko segmentacijo terenskih posnetkov ljudske glasbe v okviru magistrske naloge. V poglavju je predstavljeno področje pridobivanja informacij iz glasbe, definicija zvočnega signala, računalniška obdelava digitalnih zvočnih zapisov, zgradba zvočnih posnetkov, segmentacija glasbenih posnetkov skupaj z značilnicami in metodami ter iskanje smiselnih delov v zvočnem zapisu.

V poglavju 3 je podrobneje je opredeljen problem, ki ga magistrska naloga rešuje, predstavljena je metoda segmentacije posnetkov na višjem in nižjem nivoju razvita v okviru magistrske naloge ter izboljšava višjenivojske segmentacije z nizkonivojskim pristopom.

Poglavje 4 opisuje eksperimente izvedene v okviru naloge, s katerimi smo ovrednotili delovanje naše metode in jo primerjali z že obstoječimi metodami segmentacije. Tukaj smo lahko z ostalimi metodami primerjali zgolj segmentacijo na posameznem nivoju in ne tudi delovanja celotne razvite metode, saj je v okviru naloge razvita metoda prva, ki združuje segmentacijo na dveh nivojih.

V zadnjem poglavju 5 so predstavljeni zaključki do katerih smo prišli med izdelavo magistrske naloge, možne izboljšave in možnosti za nadaljnje delo na sami metodi ali njeni uporabi.

Na kratko lahko opise poglavij strnemo v spodnje alineje:

- Uvod in predstavitev ciljev magistrske naloge (poglavje 1).
- Pregled področja in predstavitev pomembnih metod za segmentacijo zvočnih posnetkov (poglavje 2).
- Predstavitev metode za večnivojsko segmentacijo terenskih posnetkov ljudske glasbe razvite v okviru magistrske naloge (poglavje 3).

- Rezultati testov na zbirki posnetkov iz arhiva EtnoMuza, primerjava z obstoječimi metodami, vrednotenje, prednosti in slabosti predstavljenega pristopa (poglavje 4).
- Možne izboljšave, nadaljnje delo in zaključek (poglavje 5).

# Poglavje 2

# Pregled področja

S porastom digitalizacije in vedno večjim številom digitalnih zbirk na različnih področjih znanosti in umetnosti, tudi arhivi ljudske glasbe vstopajo v digitalno dobo. Digitalizacija zbirk je dodatno vzpodbujena s strani Evropske unije, ki podpira projekte za ohranjanje kulturne dediščine. Večino novih posnetkov in gradiv se danes zajema in zbira v digitalni obliki, s čimer odpade korak digitalizacije in so vsebine veliko hitreje dostopne končnim uporabnikom.

Dostopnost glasbe v takšnih arhivih je zelo dobrodošla, saj je to dodatna vzpodbuda za raziskovalno delo, ki vključuje vsebino takšnih arhivov. Prav tako se z dostopnostjo zbirk tudi promovira kulturna dediščina, v našem primeru pomeni ljudsko glasbo. Ljudska glasba je v odprtih glasbenih zbirkah postala dostopna širši javnosti in povečuje možnosti po "ponovnem" odkritju pozabljene glasbe.

Poleg popularizacije, javno dostopne zbirke odpirajo nova vrata raziskovalcem, tako etnomuzikologom, kot raziskovalcem s področja pridobivanja informacij iz glasbe, saj jim dostop do digitalnih zbirk omogoči hitro zbiranje glasbenih posnetkov.

## 2.1 Pridobivanje informacij iz glasbe

Področje pridobivanja informacij iz glasbe (ang. Music Information Retrieval - MIR) [21], je področje računalništva, ki se ukvarja z načini obdelave glasbe v digitalni obliki, za kar so potrebne dobro definirane metode za reševanje zadanih problemov. MIR med drugim zajema tudi:

- računske metode za klasifikacijo, združevanje, pridobivanje značilnic, računanje podobnosti, iskanja podobnih vzorcev;
- metode za avtomatsko identifikacijo in prepoznavo glasbe, kot so sledenje notam, avtomatska spremljava, iskanje na podlagi poizvedbe;
- prepoznavanje sestavov in inštrumentov v glasbenih posnetkih, informacije o ritmu, tempu, melodiji, harmoniji, barvi zvoka ...

Večina MIR raziskav v zadnjih letih je bila na področju popularne glasbe, kjer so raziskave tudi tržno zelo zanimive. Prav tako so največje prav zbirke popularne glasbe, kar omogoča raziskovalcem lažje razvijanje in testiranje novih metod in pristopov za odkrivanje informacij.

Tega ne moremo trditi za dosti manj popularno, a iz kulturnega vidika zelo pomembno ljudsko glasbo. Ciljna množica uporabnikov je tukaj veliko manjša, vsekakor pa je raziskovanje na tem področju pomembno za ohranjanje kulturne dediščine. Prav tako je bilo v zadnjih letih digitaliziranih več zbirk ljudske glasbe, kar tudi vzpodbuja raziskave na področju ljudske glasbe.

Iz glasbe pridobljene informacije uporabimo kadar želimo na podlagi lastnosti in značilnosti glasbe sprejemati odločitve. Na podlagi takšnih informacij se lahko na primer odločamo ali želimo neko glasbo uvrstiti na seznam za predvajanje ali ne. Iz lastnosti pesmi nekega seznama za predvajanje lahko v neki zbirki poiščemo pesmi, ki imajo podobne lastnosti in jih dodamo v seznam predvajanja. Prav tako si lahko na podlagi lastnosti glasbe, ki jo nekdo posluša, zgradimo model posameznikovega glasbenega okusa in mu na podlagi pridobljenih informacij predlagamo novo, njemu nepoznano glasbo. Podobnih scenarijev si lahko zamislimo še veliko, prav tako so sistemi, ki opisane stvari počno že v uporabi.

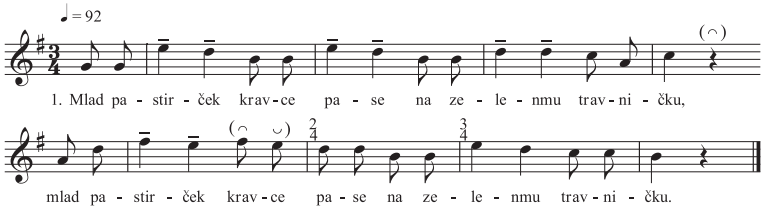
S pridobivanjem informacij iz glasbe lahko pomagamo raziskovalcem na področju glasbe in glasbenim ustvarjalcem. Tako lahko na podlagi pridobljenih informacij v neki zbirki poiščemo vse pesmi ali dele pesmi z iskanimi lastnostmi.

## 2.2 Predstavitev zvočnega signala

Glasba človeka spremlja že od pradavnine. Povsem drugače je z njenim zapisovanjem. Podobno kot poznamo za zapis jezika pisavo, za glasbo dolgo ni obstajalo širše sprejete oblike zapisa. Prvi zapisi glasbe na tablicah iz obdobja 2000 let pred našim štetjem so že povzemali osnovne glasbene lastnosti. Zapis se je preko antike in bizantinskega obdobja spreminjal in je v srednjem veku dobil danes vsem poznano standardno notno črtovje, ki se je iz začetne oblike s štirimi črtami spremenilo v pet črtno obliko, kakršno je v uporabi še danes. Do modernega zapisa glasbe so se spremenile še oblike not, dodatno uveljavljeni simboli za opis pavz in takta. Še dandanes se zapis glasbe v notnem črtovju spreminja, saj skladatelji v črtovje vpeljujejo nove simbole, s katerimi predstavijo do sedaj neopisljive lastnosti glasbe.

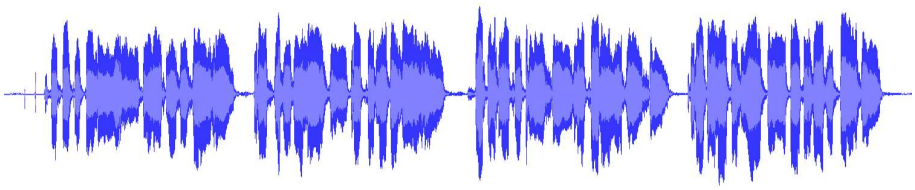
Na papirju je glasba največkrat predstavljena na zgoraj omenjen način s pomočjo notnega zapisa, opremljeno še z dodatnimi oznakami in morebiti besedilom pesmi. Primer takšnega zapisa glasbe iz digitalnega arhiva EtnoMuza je na Sliki 2.1. Takšna oblika zapisa je zelo kompaktna, kar omogoča enostavno ponovljivo izvajanje glasbe, še vedno pa dopušča, da izvajalec simbole interpretira na sebi lasten način.

Glasba je v digitalni obliki predstavljena v simbolični obliki ali v obliki zvočnega posnetka. Simbolična oblika nam omogoča kompakten zapis pomembnih lastnosti




Slika 2.1: Primer notnega zapisa glasbe, opremljenega z besedilom.

glasbe, ne omogoča pa zapisa vseh lastnosti in podrobnosti vsebovanih v nekem zvočnem posnetku. Glasbo v simbolični obliki lahko večkrat ponovno reproduciramo, njene lastnosti so odvisne od kvalitete in načina reprodukcije. Primer predstavitve glasbe v simbolični obliki je notni zapis. Zvočni posnetek nam poleg osnovnih glasbenih lastnosti omogoča hranjenje konkretnih akustičnih lastnosti snemalnega okolja, inštrumentov in opreme. Kvaliteta in lastnosti zvočnega posnetka so odvisne od uporabljene snemalne opreme in od okolja v katerem je zajemanje izvedeno. Pretvorba simbolične oblike glasbe v zvočni posnetek je enostavna in se vrši z izvedbo posameznega dela, enako pa ne velja za obratni proces t.i. transkripcije. Zvočni posnetek je v računalniku predstavljen v obliki valovne krivulje, ki opisuje spreminjanje pritiska v snovi skozi katero se zvok širi. Primer takšne krivulje je predstavljen na Sliki 2.2.



Slika 2.2: Primer valovne krivulje.

## 2.3 Računalniška obdelava digitalnih zvočnih zapisov

Kadar se srečujemo z velikimi zbirkami podatkov, je določena dela smiselno avtomatizirati. S tem prihranimo čas in denar, zmanjšamo pa tudi možnost človeške napake, v primeru dobro definiranih postopkov in algoritmov. Pri vedno večjih glasbenih zbirkah zelo narašča tudi potreba po zmogljivejših sistemih za hranjenje in pridobivanje informacij. Primer takšne zbirke je digitalni arhiv EtnoMuza, ki

vsebuje veliko količini gradiva. Arhiv je dobro okolje za razvoj metod pridobivanja informacij iz glasbe. Večina raziskav je opravljenih na zbirkah popularne in klasične glasbe. Izvajalci popularne in klasične glasbe so večinoma šolani glasbeniki, kar ne drži za ljudsko glasbo. Zaradi velikih razlik v kvaliteti posnetkov in izvedbi glasbe predstavlja ljudska glasba izziv za raziskovalce v področju MIR.

## 2.4 Struktura terenskih posnetkov

Zaradi velike količine podatkov v glasbenih arhivih se povečuje potreba po vpeljavi avtomatskih metod. V etnomuzikoloških arhivih to velja na primer za segmentacijo terenskih posnetkov in skladb. Zanimiv primer uporabe algoritmov za avtomatsko obdelavo arhiva je tudi ugotavljanje vrste zvoka nekega dela terenskega posnetka in možne razdelitve terenskega posnetka v vsebinsko enotne dele. Terenski posnetek je lahko sestavljen iz več delov. Deli posnetka vsebujejo intervjuje s posamezniki, kar predstavlja govorno vsebino. Med samim intervjujem lahko posameznik ali skupina ljudi zapoje ali zaigra v intervjuju opisano pesem. Posnetek je lahko sestavljen iz več krajših odsekov posnetih v različnih časovnih obdobjih, kar posledično pomeni, da lahko posnetek vsebuje vsebinsko različne, včasih povsem nepovezane stvari.

Vprašanje: *"Ali je mogoče razviti metodo za avtomatsko segmentacijo terenskih posnetkov v vsebinsko smiselne dele?"* je pri obdelovanju velikih arhivov zvočnih posnetkov povsem na mestu. V zvočnih posnetkih lahko iščemo:

- dele posnetka (višji nivo), ki vsebujejo isto vrsto zvoka (govor, solo petje, večglasno petje, inštrumental, ali katero izmed kombinacij naštetega.),
- v delih posnetka z glasbo še nadalje iščemo dele (nižji nivo), ki predstavljajo smiselno strukturo glasbe in imajo enake lastnosti (npr. deli z enako melodijo, barvo, ritmom ipd.)

Pri organizaciji arhiva nam pomaga segmentacija obeh nivojev, saj posnetke hierarhično razčleni na smiselne dele in s tem omogoča enostavnejše brskanje ali iskanje po arhivu. Takšen pristop je lahko osnova za vizualizacijo vsebine arhiva. Primer segmentacije terenskega posnetka v smiselne dele glede na vrsto zvoka in ponavljajoče dele je prikazan v Sliki 2.3. Predstavljena struktura v je značilna tudi za zvočne posnetke digitalnega arhiva EtnoMuza.

Na področju segmentacije zvočnih posnetkov je bilo narejenih veliko raziskav, ki pa so večinoma zajemale zbirke popularne glasbe. Struktura pri popularni glasbi je zelo dobro definirana, glasba je točno odigrana in odpeta, prav tako je večja kvaliteta zvočnih posnetkov. Ljudska glasba je bolj problematična saj je netočno zaigrana in/ali zapeta, posneta pri slabih snemalnih pogojih v vsakdanjem okolju z visoko stopnjo šuma iz okolice. Problem segmentacije ljudske glasbe v terenskih posnetkih je kompleksnejši problem, zaradi potrebe po upoštevanju večje količine

Slika 2.3: Slika prikazuje večnivojsko segmentacijo terenskih posnetkov.

šuma v različnih oblikah in netočne izvedbe. To za popularno glasbo posneto v snemalnih studiih z večkratnim ponovnim snemanjem ne velja.

## 2.5 Glasbena struktura

Pod pojmom glasbena struktura si predstavljamo hierarhično strukturo, ki opisuje kako je sestavljena glasba. Strukturo glasbe si najlaže predstavljamo za glasbo v simbolični obliki, kjer med zaporedji simbolov iščemo podobne strukture (npr. ponavljajoča zaporedja not, akordov, ipd.). Podobno velja tudi za glasbo v zvočnih posnetkih. Glasbeno strukturo lahko iščemo na različnih nivojih [9]. Avtor analizira razlike in povezave med *površinsko strukturo* in *globoko strukturo* glasbe. Na časovni skali se lahko ponavljajo krajši odseki (nekaj not, ali nekaj sekund zvočnega posnetka) ali pa daljši odseki (daljša zaporedja not oz. nekaj 10 sekund dolgi deli zvočnega posnetka). Pesmi popularne glasbe so sestavljene na podoben način. Največkrat vsebovani deli popularne skladbe so: uvod, kitica, refren, prehod, instrumentalni solo in zaključek. Izjemoma popularna glasba vsebuje tudi druge dele. Ljudska glasba ima večinoma še enostavnejšo zgradbo. V večini primerov je posamezna skladba sestavljena iz kitic s ponavljajočo melodijo.

Glasbena struktura pripomore k večji prepoznavnosti neke skladbe, saj si človek glasbo zapomni tudi po zaporedju in ponovitvah posameznih delov. Še vedno ni znano zakaj človek v glasbi uživa, veliko teorij zagovarja stališče, da ima struktura glasbe pri tem pomembno vlogo.

> *Večina odraslih ima otroške fantazije o izdelavi in urejanju večjih struktur iz manjših. Ena vrsta razumevanja glasbe vključuje izgradnjo velikih mentalnih struktur iz manjših, delov glasbe. Morda je pogon, ki gradi te*

*mentalne glasbene strukture isti, ki nam poskuša pomagati razumeti svet.*
- Minsky, 1981

Analiza glasbene strukture (ang. Music structural analysis) je obširno področje, ki zajema analizo glasbe v simbolični obliki in obliki digitalnih posnetkov. Magistrska naloga obsega področje segmentacije zvočnih posnetkov glasbe v smiselne dele, na različnih nivojih, kot je predstavljeno v poglavju 2.4. Posamezne dele, ki predstavljajo strukturo nekega glasbenega posnetka, združujemo glede na njihovo podobnost. Podobnost v našem primeru pomeni ujemanje lastnosti kot so ritem, melodija, itd. Načini pridobivanja lastnosti so predstavljeni v naslednjem poglavju o segmentaciji zvočnih posnetkov.

## 2.6 Segmentacija zvočnih posnetkov

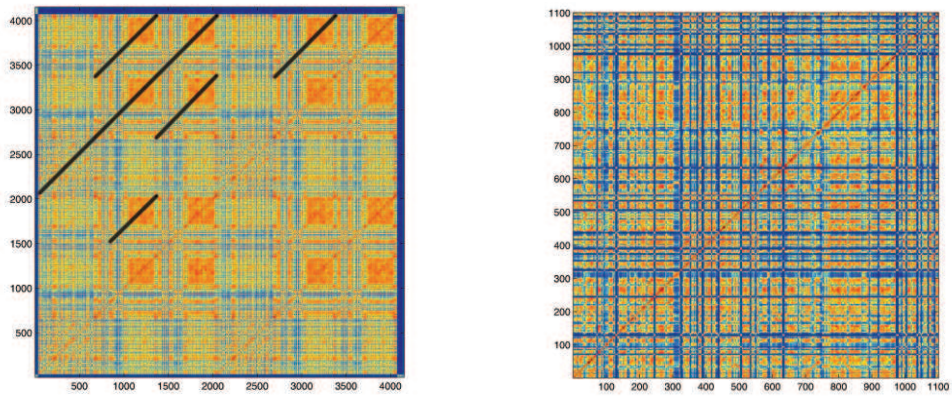
Obstoječi pristopi za segmentacijo glasbenih posnetkov so večinoma narejeni za segmentacijo popularne glasbe. Dober pregled področja je podan v [36], kjer avtorji povzemajo najnovejše rezultate raziskav o segmentaciji in iskanju struktur v glasbenih posnetkih. Tipični pristopi za segmentacijo temeljijo na izračunu zvočnih značilnic, na podlagi katerih z uporabo mer podobnosti izračunajo podobnostno predstavitev, ki je podlaga algoritma za segmentacijo. Pregled najpogosteje uporabljenih značilnic, mer podobnosti in segmentacijskih algoritmov je podan v nadaljevanju.

### 2.6.1 Mel-frekvenčni kepstralni koeficienti

Mel-frekvenčni kepstralni koeficienti (ang. Mel-frequency cepstral coefficients - MFCC) [41], so značilnice, ki predstavljajo obliko frekvenčnega spektra. Mel-frekvenčni kepster je kompaktna predstavitev zvočnega spektra ustrezajočega človekovi nelinearni percepciji tonske višine, opisane z Mel skalo. Mel skala [43] je zaznavna skala tonskih višin, ki so po občutku za poslušalca med seboj enakomerno oddaljene. Velikokrat uporabljena formula za preračun $f$ Hertzov v $m$ mel-ov je:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) \tag{2.1}$$

Pri iskanju glasbene strukture je uporaba koeficientov MFCC predstavljena v delih [14, 7, 15, 37]. Največkrat se uporabljajo kot značilnice (ang. feature) za izračun mere podobnosti (na primer v samopodobnostnih matrikah 2.6.4). Ker prvih nekaj koeficientov MFCC predstavlja globalno barvo zvoka (povezano s sestavom oz. inštrumenti, ki v glasbi nastopajo), lahko z merami podobnosti, ki temeljijo na teh koeficientih, najdemo dele glasbe, ki so si med seboj podobni po barvi. Primer uporabe pri izračunu samopodobnostne matrike je prikazan na Sliki 2.4. V sliki je za izračun mere podobnosti uporabljena kosinusna razdalja.




(a) primer, ko metoda deluje uspešno (pesem 'Abraham ma sedem sinu')

(b) primer, ko metoda odpove (pesem 248-23(GNI M))

Slika 2.4: Slika prikazuje samopodobnostno matriko izračunano s pomočjo kosinusne podobnosti med značilnicami MFCC posameznih okvirjev signala. V sliki z dobrim primerom so nad diagonalo označene ponovitve, vzorci pa so razvidni pod diagonalo. Oba primera sta iz digitalnega arhiva EtnoMuza.

Značilnice MFCC se uporabljajo za glasbo v kateri je prisotnih več vokalov in/ali inštrumentov z različnimi barvami zvoka. Izračunamo jih po naslednjih korakih predstavljenih v [41]:

1. izvorni zvočni signal $\mathbf{x}(n)$ razdelimo v kratke, praviloma prekrivajoče, časovne odseke (okvirje);

2. izračunamo diskretno Fourierevo transformacijo nad vsemi okvirji izvornega zvočnega signala dolžine $N$:

$$X(k) = \sum_{n=0}^{N-1} w(n)x(n)e^{-\frac{j2\pi kn}{N}}, \tag{2.2}$$

   za $k = 0, 1 \ldots N-1$, kjer $k$ predstavlja frekvence $f(k) = \frac{kf_s}{N}$ in je $f_s$ frekvenca vzorčenja v Hertzih, $w(n)$ pa časovno okno (na primer Hammingovo okno podano z $w(n) = 0,54 - 0.46\cos(\frac{\pi n}{N})$;

3. magnitudni spekter pretvorimo v Mel skalo z uporabo zbirke Mel filtrov $H(k,m)$, nato z logaritmiranjem še v logaritemsko:

$$X'(m) = \ln(\sum_{k=0}^{N-1} |X(k)| \cdot H(k,m)), \tag{2.3}$$

za $m = 1, 2 \ldots M$, kjer je $M$ število filtrov in velja $M \ll N$, $|X(k)|$ pa predstavlja absolutno vrednost $X(k)$. Zbirka Mel filtrov je zbirka trikotnih filtrov definiranih s srednjimi frekvencami $f_c(m)$:

$$H(k,m) = \begin{cases} 0 & for \quad f(k) < f_c(m-1) \\ \frac{f(k)-f_c(m-1)}{f_c(m)-f_c(m-1)} & for \quad f_c(m-1) \leq f(k) < f_{c(m)} \\ \frac{f(k)-f_c(m+1)}{f_c(m)-f_c(m+1)} & for \quad f_c(m) \leq f(k) < f_c(m+1) \\ 0 & for \quad f(k) \geq f_c(m+1) \end{cases} \tag{2.4}$$

Srednje frekvence so izračunane z aproksimacijo Mel skale (glej enačbo 2.1). Nadalje izračunamo fiksno resolucijo Mel skale z uporabo $\Delta\phi = \frac{\phi_{max}-\phi_{min}}{M+1}$, kjer je $\phi_{max}$ najvišja frekvenca v zbirki filtrov po Mel skali izračunana iz $f_{max}$ s pomočjo enačbe 2.1, $\phi_{min}$. Najnižja frekvenca v Mel skali ustreza frekvenci $f_{min}$, $M$ je število filtrov v zbirki. Uporabimo lahko na primer vrednosti $f_{max} = 11025$ Hz, $f_{min} = 0$ Hz in $M = 30$. Srednje frekvence v Mel skali so podane kot $\phi_c(m) = m \cdot \Delta\phi$ za $m = 1, 2 \ldots M$. Za izračun srednjih frekvenc v Hertzih uporabimo inverz enačbe 2.1 podan z $f_c(m) = 700(10^{\phi_c(m)/2595} - 1)$, ki jo vstavimo v enačbo 2.4.

4. izračunamo diskretno kosinusno transformacijo $X'(m)$ z uporabo:

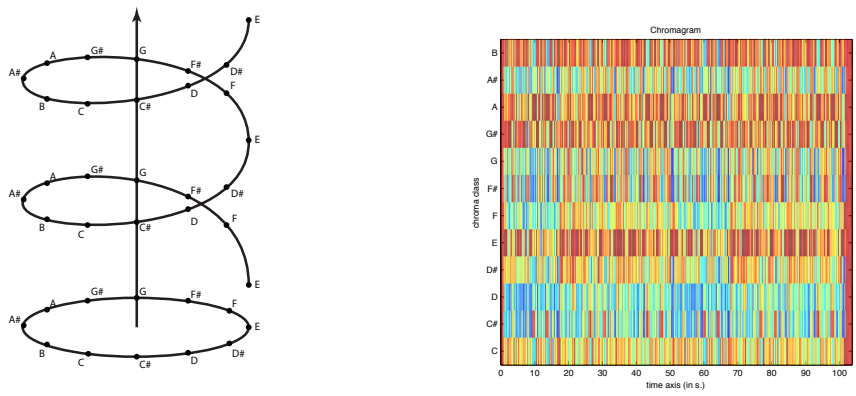
$$c(l) = \sum_{m=1}^{M} X'(m) \cos(l\frac{\pi}{M}(m - \frac{1}{2})). \tag{2.5}$$

Metoda deluje dobro na posnetkih v katerih je prisotnih več zvočnih virov z različno barvo zvoka. Primer pesmi v kateri je prisotnih več zvokov različnih barv (več vokalov in inštrument) je predstavljen na Sliki 2.4(a)) so dobro vidne diagonale, ki predstavljajo ponavljanja. V Sliki 2.4(b) so razlike v barvi zelo izrazite. Do izraza pridejo razlike v barvi samoglasnikov in soglasnikov. Ponavljajoči deli v posnetku ne pridejo do izraza saj je vsak vokal podobne barve. V digitalnem arhivu večina posnetkov sodi v drugo skupino, za kar uporaba značilnic MFCC ni ustrezna.

### 2.6.2 Kromatski vektorji

Človek zaznava višino tona periodično med različnimi oktavami glede na tonski razred. Toni v razmaku ene ali več oktav so toni istega tonskega razreda. Ta lastnost nam omogoča razdelitev zvoka glede na tonsko višino in tonski razred [18]. Tonskih razredov je 12 in so C, C#, D, D#, E, F, F#, G, G#, A, A#, B. Kako se toni uvrščajo v posamezen razred glede na višino je prikazano na Sliki 2.5(a), kjer so tonske višine nanizane na spiralo, kot na spirali pa predstavlja tonski razred.

Kromatski vektor (ang. Chroma vector) je predstavitev v obliki 12-dimenzionalnega vektorja, kjer vrednost na posameznem mestu vektorja predstavlja seštevek frekvenčnih komponent posameznega tonskega razreda v opazovanem časovnem okvirju zvočnega posnetka.




(a) prikaz razporeditve tonskih višin v tonske razrede na spirali.

(b) primer kromagrama za pesem 'Abraham ma sedem sinu'

Slika 2.5: Sliki prikazujeta razporeditev tonskih višin v tonske razrede in kromagram.

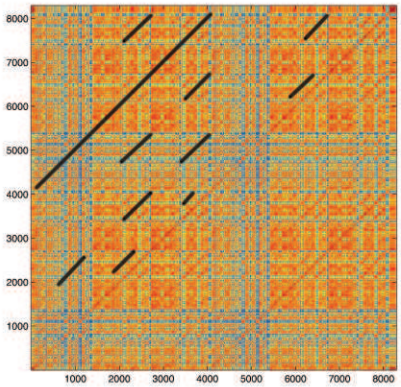
Kromatske značilnice iz zvočnega posnetka lahko pridobimo tako, da zvočni signal razdelimo na 88 tonskih pasov. Za vsak pas izračunamo kratko-časovno povprečno energijo (ang. short-time mean-sqare power - STMSP) in seštejemo vse vrednosti STMSP, ki ustrezajo istemu tonskemu razredu [31]. Na ta način dobimo 12-dimenzionalni kromatski vektor za posamezni okvir signala.

Na Sliki 2.5(b) je prikazan kromagram, sestavljen iz kromatskih vektorjev za okvirje dolžine 100 ms, z delnim prekrivanjem okvirjev, za pesem iz digitalnega arhiva EtnoMuza. Kromagram lahko nadalje uporabimo pri izračunu samopodobnostne matrike za iskanje ponavljajočih delov pesmi glede na tonske značilnosti zajete v kromagramu. Pristopi [4, 17, 18] uporabljajo kromatske vektorje pri izračunu samopodobnostne matrike ali matrike s časovnim zamikom.

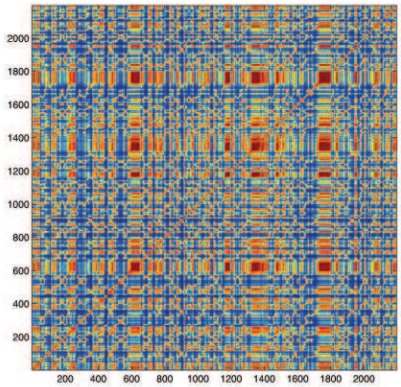
V Sliki 2.6 sta primera samopodobnostnih matrik izračunana z mero podobnosti $d_{RMS}$ (glej 2.6.3) za pesmi iz digitalnega arhiva EtnoMuza, kot v [18]. Metoda dobro deluje na primeru iz Slike 2.6(a), kjer so v sliki dobro razvidne ponovitve in slabo na primeru iz Slike 2.6(b), kjer ponovitve v samopodobnostni matriki niso izražene. Razlog je netočno petje amaterskih pevcev v posnetkih, ki različnih kitic v pesmi ne zapojejo z istimi tonskimi višinami. To se odraža v zamaknjenih vrednostih med tonskimi razredi. Probleme z zamaknjenimi tonskimi višinami rešuje metoda predstavljena v [32], ki jo opisujemo v 2.6.7.

### 2.6.3 Mere razdalj

Mera razdalje pove kako podobni sta med seboj dve vrednosti. Naj bo $F$ zaloga vrednosti, ki jih značilnice lahko zavzemajo. Za primerjavo dveh vrednosti $x, y \in F$




(a) primer, ko se ponovitve v samopodobnostni matriki lepo izrazijo z linijami vzporednimi diagonali (pesem 'Abraham ma sedem sinu')




(b) primer, ko se ponovitve v samopodobnostni matriki ne izrazijo (pesem 248-23(GNI M))

Slika 2.6: Primera samopodobnostnih matrik dveh pesmi iz digitalnega arhiva EtnoMuza, izračunan na podlagi kromagramov. V sliki z dobrim primerom so nad diagonalo označeni vzorci ponovitev, vidni tudi pod diagonalo.

definiramo mero razdalje kot:

$$d : F \times F \to \mathbb{R}_{\geq 0}. \tag{2.6}$$

Vrednost $d(x, y)$ je nizka, če sta $x$ in $y$ podobna v nasprotnem primeru, pa je vrednost $d(x, y)$ visoka. Primer mere razdalje je lahko evklidska razdalja med vektorjema $\mathbf{p}$ in $\mathbf{q}$ v $n$-dimenzionalnem prostoru predstavljena z:

$$d_E(\mathbf{p}, \mathbf{q}) = d_E(\mathbf{q}, \mathbf{p}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{2.7}$$

Na področju MIR se velikokrat uporablja tudi kosinusna razdalja. Kosinusna razdalja med vektorjema $\mathbf{p}$ in $\mathbf{q}$ je definirana:

$$d_{cos} = 1 - \frac{\|\mathbf{p}\|\|\mathbf{q}\|}{\mathbf{p} \cdot \mathbf{q}} = 1 - \frac{\sqrt{\sum_{i=1}^{n} (p_i)^2} \sqrt{\sum_{i=1}^{n} (q_i)^2}}{\sum_{i=1}^{n} p_i q_i}, \tag{2.8}$$

kjer $\|\mathbf{p}\|$ predstavlja drugo normo vektorja $p$. Kot mera razdalje se v MIR uporablja tudi koren povprečja kvadratov razdalje (ang. Root Mean Square - RMS). RMS je za vektorja $\mathbf{p}$ in $\mathbf{q}$ definirana kot:

$$d_{RMS}(\mathbf{p}, \mathbf{q}) = \frac{\| \frac{\mathbf{p}}{\max_i p_i} - \frac{\mathbf{q}}{\max_i q_i} \|}{\sqrt{l}}, \tag{2.9}$$

kjer $\|_\|$ predstavlja drugo normo, $l$ pa dolžino vektroja.

Mere razdalj so osnova za izračun podobnosti med različnimi deli zvočnega signala in so v nadaljevanju uporabljene v predstavljenih metodah za odkrivanje podobnosti v zvočnih posnetkih.

### 2.6.4 Samopodobnostna matrika

Samopodobnostna matrika (ang. Self-similarity matrix) je grafična predstavitev podobnosti delov zaporedja (podzaporedij). Pri zvočnih posnetkih to pomeni prikaz medsebojno podobnih časovnih okvirjev posnetka. V dvodimenzionalnem prikazu obe osi predstavljata čas. Intenziteta točke v matriki predstavlja stopnjo podobnosti med dvema okvirjema signala glede na izbrano mero podobnosti. Podobnosti med posameznimi deli posnetka se v samopodobnostni matriki izražajo kot črte z visoko intenziteto. Iskanje delov z enakimi lastnostmi v samopodobnostni matriki se prevede na iskanje diagonalno poudarjenih sledi v matriki, kjer glavna diagonala predstavlja osnovni signal, vzporedne sledi pa med seboj povezujejo dele s podobnimi lastnostmi glede na izbrano mero podobnosti.

Samopodobnostna matrika je poseben primer metode izrisa ponovitev (ang. Recurrence plot), znane statistične metode, za odkrivanje ponavljajočih dogodkov v časovnih zaporedjih [11], za katere velja, da sta vektorja $\mathbf{x}(i)$ in $\mathbf{x}(j)$ podobna glede na izbrano mero podobnosti. Podobnost zapišemo tudi kot:

$$\mathbf{x}(i) \approx \mathbf{x}(j), \tag{2.10}$$

kjer so vrednosti predstavljene z $i$ prikazane na vodoravni, vrednosti predstavljene z $j$ pa na navpični osi. Metoda je bila razvita za prikaz podobnosti v kaotičnih sistemih. Posamezno vrednost v samopodobnostni matriki lahko izračunamo kot:

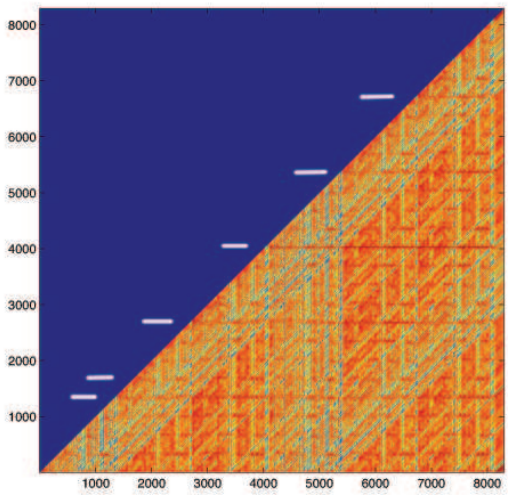
$$\mathbf{R}(i,j) = d(\mathbf{x}(i), \mathbf{x}(j)), \tag{2.11}$$

kjer je $d(\mathbf{a}, \mathbf{b})$ mera podobnosti (glej poglavje 2.6.3), $\mathbf{x}(i)$ in $\mathbf{x}(j)$ sta elementa primerjanih vektorjev (na primer kromatska vektorja dveh okvirjev zvočnega posnetka).

Samopodobnostno matriko lahko v glasbi uporabimo za prikaz ponavljajočih lastnosti (kot je barva, če so značilnice MFCC oz. harmonija ali melodija, če so značilnice kromatske ipd.) v glasbi skozi čas. Začetki uporabe samopodobnostnih matrik na področju segmentacije so predstavljeni v [14], kjer avtor predstavi možno uporabo za iskanje ponavljajočih delov v glasbi na podlagi akustičnih lastnosti posnetka. Predstavljena je uporaba avtokorelacijske metode za odkrivanje podobnih delov v samopodobnostni matriki na primeru posnetka bobnanja. Z avtokorelacijo so odkriti ponavljajoči udarci v celotnem zvočnem posnetku.

Na Sliki 2.7 je prikazan primer matrike s časovnim zamikom matrike izračunane iz samopodobnostne matrike prikazane v Sliki 2.6(a). Pristopi [4, 17, 18] uporabljajo

samopodobnostno matriko zgolj za izgradnjo matrike s časovnim zamikom predstavljene v nadaljevanju. Pristopi segmentacije ne izvajajo neposredno z uporabo samopodobnostne matrike ampak posredno preko matrike s časovnim zamikom.




Slika 2.7: Primer matrike s časovnim zamikom izračunane iz samopodobnostne matrike predstavljene v Sliki 2.6(a). V sliki so z belimi črticami označene vodoravne črte, ki ponazarjajo ponavljanje v matriki s časovnim zamikom.

### 2.6.5 Matrika s časovnim zamikom

Ekvivalentna predstavitev predstavitvi s samopodobnostno matriko je matrika s časovnim zamikom (ang. Time-lag Matrix). Tudi matrika s časovnim zamikom je namenjena predstavitvi podobnih delov v časovnem zaporedju. Za razliko od samopodobnostne matrike se ponavljajoči deli v matriki s časovnim zamikom ne izražajo kot diagonalno ampak kot vodoravno poudarjene sledi vzporedne s časovno osjo. Intenziteta točk v sledi tudi pri matriki s časovnim zamikom določa stopnjo podobnosti delov zvočnega zapisa glede na izbrano mero podobnosti. Na obeh oseh je še vedno časovna skala, na eni (običajno vodoravni) je absolutni čas v posnetku, na drugi (običajno navpični), pa je časovni odmik dela od začetka zaporedja.

Matematično lahko zapišemo, da za matriko s časovnim zamikom velja:

$$\mathbf{x}(t) \approx \mathbf{x}(t - \tau), \tag{2.12}$$

kjer je $\tau$ časovni zamik ob katerem se del posnetka s podobnimi lastnostmi ponovi. Tako se ponovitve odražajo z odmikom od glavne diagonale. Matrika je definirana

kot:

$$\mathbf{R}(t,l) = d(\mathbf{x}_t, \mathbf{x}_{t-l}),\ t = 1, 2 \ldots N,\ l = t \ldots N \tag{2.13}$$

kjer je $t$ časovno zaporedje, $l$ časovni zamik, $N$ končni čas, $d$ pa izbrana mera podobnosti odvisna od posameznega pristopa. Tako je na primer v [18] kot $d$ uporabljena mera podobnosti $d_{RMS}$ (glej 2.6.3) med vektorjema $\mathbf{x}_t$ in $\mathbf{x}_{t-l}$.

Matrika s časovnim zamikom je za iskanje sledi primernejša in enostavnejša od samopodobnostne matrike, saj se ponavljajoči deli izražajo kot sledi vzporedne z vodoravno osjo (ki označuje absolutni čas v zaporedju), kar olajša njihovo odkrivanje. Matrike s časovnim zamikom je na področje MIR vpeljal Goto, ki jih uporabi v [17], za iskanje ponavljajočih delov v popularni glasbi, kar uporabi tudi v izdelku SmartMusicKIOSK[1]

### 2.6.6 Nenegativna matrična faktorizacija

Pri iskanju ponavljajočih delov v samopodobnostni matriki si lahko pomagamo z nenegativno matrično faktorizacijo (ang. Non-negative Matrix Factorisation - NMF). NMF je postopek, ki matriko $\mathbf{X}$ razčleni v produkt dveh matrik $\mathbf{W}$ in $\mathbf{H}$, tako da velja:

$$\mathbf{X} \approx \mathbf{WH}, \tag{2.14}$$

kjer je $\mathbf{X}$ vhodna matrika, $\mathbf{W}$ in $\mathbf{H}$ pa sta matriki na kateri razčlenimo originalno matriko. Postopkov za iskanje matrik $\mathbf{W}$ in $\mathbf{H}$ je več in so odvisne od namena uporabe. Pomembno je omeniti, da je zmnožek matrik $\mathbf{W}$ in $\mathbf{H}$ zgolj približek originalne matrike.

V [26] predstavljen pristop uporabi NMF za dekompozicijo samopodobnostne matrike. V delu kot mero podobnosti uporabijo kosinusno podobnost nad značilnicami, ki opisujejo barvo zvoka. Na podlagi izbrane mere podobnosti je sestavljena samopodobnostna matrika, ki predstavlja matriko $\mathbf{X}$. NMF je uporabljena za oceno nenegativnih faktorjev v matrikah $\mathbf{W}$ in $\mathbf{H}$. Matrika $\mathbf{W}$ vsebuje bazne vektorje, matrika $\mathbf{H}$ pa koeficiente kodiranja. Vsak element samopodobnostne matrike je predstavljen kot:

$$s_{i,j} = \sum_{k=1}^{\tau} A_k(i,j), \tag{2.15}$$

kjer velja:

$$A_k = W(:,k)H(k,:), \tag{2.16}$$

kjer je $\tau$ stopnja dekompozicije. Glede na stopnjo dekompozicije lahko ugotovimo število strukturnih elementov. Po opravljeni dekompoziciji uporabijo postopek gručenja (ang. Clustering) za povezovanje delov z enakimi lastnostmi.

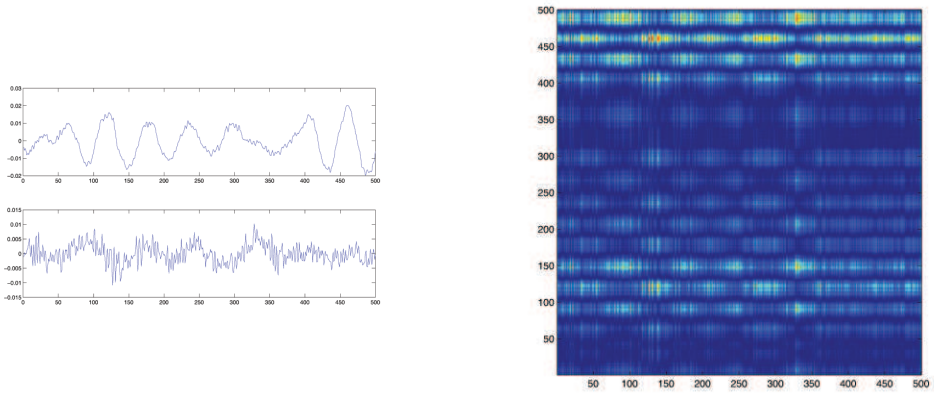
Drugačen pristop z uporabo izvedenke NMF, verjetnostne latentne komponentne analize neodvisne od zamika (SI-PLCA - ang. Shift-invariant Probabilistic Latent

[1] http://staff.aist.go.jp/m.goto/SmartMusicKIOSK/

Component Analysis) [48] za iskanje ponavljajočih vzorcev v posamezni komponenti dekompozicije kromagrama. Pristop uporabi SI-PLCA za dekompozicijo časovno-frekvenčne matrike zvočnega signala.

### 2.6.7 Dinamično ukrivljanje časa

Dinamično ukrivljanje časa (ang. Dynamic Time Warpping - DTW), prilagojena oblika dinamičnega programiranja, je metoda za odkrivanje optimalne poravnave med dvema podanima (časovno odvisnima) zaporedjema. V osnovi je bila metoda razvita za prepoznavanje govora v avtomatskih transkripcijskih sistemih. Na področju MIR se metoda uporablja za premoščanje časovnih transformacij in razlike v hitrosti pri časovno odvisnih zaporedjih.




(a) primer delov zvočnega signala iz arhiva

(b) primer samopodobnostne matrike dveh signalov z leve slike

Slika 2.8: Sliki prikazujeta prva dva koraka pri izračunu DTW.

Metoda kot vhod vzame dve zaporedji predstavljeni z vektorjema $\mathbf{X}$ in $\mathbf{Y}$:

$$\mathbf{X} = (x_1, x_2 \dots x_N), \mathbf{Y} = (y_1, y_2 \dots y_M); M, N \in \mathbb{N}, \tag{2.17}$$

Izračunamo samopodobnostno matriko $\mathbf{C}$, kjer elementi matrike predstavljajo ceno med elementi vhodnih vektorjev. Element matrike predstavlja ceno med dvema elementoma, izračunano s pomočjo mere razdalje:

$$c = 1 - d, \tag{2.18}$$

kjer je $c$ element matrike in je $d$ izračunan kot RMS razdalja med primerjanima elementoma. Razdalja $d$ zavzema vrednosti na intervalu $[0, 1]$ in je predstavljena v 2.6.3. Cilj je najti poravnavo med $\mathbf{X}$ in $\mathbf{Y}$ z največjo podobnostjo. Intuitivno takšna

pot poteka po "dolini" z majhnimi cenami v samopodobnostni matriki $\mathbf{C}$. Rezultat je viden v matriki kumulativne cenilne funkcije prikazane na sliki 2.9(a). Ukrivljena pot (ang. wrapping path) je zaporedje $\mathbf{p} = (p_1 \dots p_L); p_l = (n_l, m_l) \in [1 : N] \times [1 : M]; l \in [1 : L]$, ki ustreza:

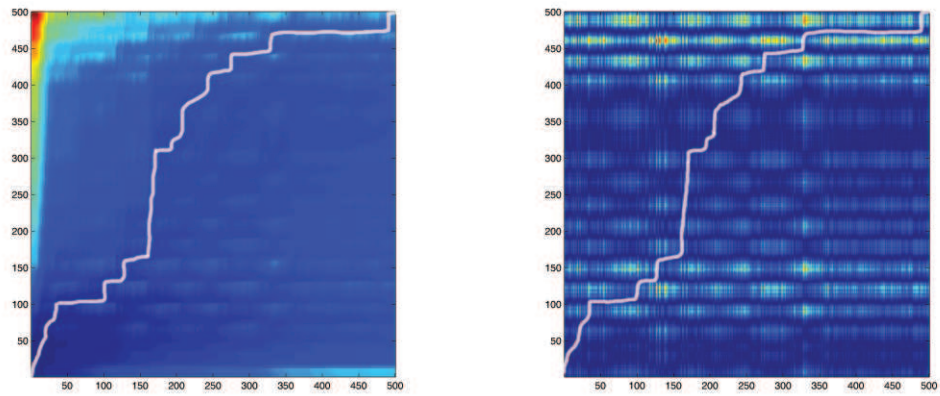
- *robnim pogojem*: $p_1 = (1, 1)$ in $p_L = (N; M)$;
- *monotonosti*: $n_1 \leq n_2 \leq \dots \leq n_L$ in $m_1 \leq m_2 \leq \dots \leq m_L$;
- *pogoju koraka*: $p_{l+1} - p_l \in \{(1, 0), (0, 1), (1, 1)\}$ za $l \in [1 : L - 1]$.

Ceno $c_p(\mathbf{x}, \mathbf{y})$ celotne ukrivljene poti $\mathbf{p}$ med $\mathbf{x}$ in $\mathbf{y}$ glede na lokalno cenilno funkcijo $c$ izračunamo kot:

$$c_p(\mathbf{x}, \mathbf{y}) = \sum_{l=1}^{L} c(x_{n_l}, y_{m_l}). \tag{2.19}$$

Optimalna ukrivljena pot $\mathbf{p}^*$ je tista pot, med vsemi možnimi ukrivljenimi potmi, katere skupna cena je najmanjša. Takšno pot $\mathbf{p}^*$ imenujemo tudi DTW in je definirana kot:

$$\text{DTW} = c_{p^*}(\mathbf{x}, \mathbf{y}) = \min \{c_p(\mathbf{x}, \mathbf{y}) | p \text{ je } (N, M) - \text{ukrivljena pot}\}. \tag{2.20}$$




(a) primer kumulativne cenilne matrike za zaporedji s slike 2.8(a), z označeno najcenejšo potjo

(b) primer samopodobnostne matrike z označeno najcenejšo potjo med zaporedjema s slike 2.8(a)

Slika 2.9: Primer matrike s kumulativno cenilno funkcijo in samopodobnostne matrike na katerih je prikazana najcenejša pot.

Za iskanje optimalne ukrivljene poti $\mathbf{p}^*$ bi lahko preverili vse možne ukrivljene poti kar je časovno zelo potratno in bi vodilo v eksponentno časovno odvisnost od velikosti $N$ in $M$. Optimalno ukrivljeno pot lahko poiščemo tudi s pomočjo dinamične

programiranja, pri čemer sestavimo matriko $\mathbf{D}$, imenovano tudi kumulativna cenilna matrika, katere primer je na sliki 2.9(a) in je definirana kot:

$$\mathbf{D}(n, m) = \mathrm{DTW}(\mathbf{x}(1:n), \mathbf{y}(1:m)). \tag{2.21}$$

DTW predstavlja možnost ugotavljanja podobnosti v zvočnem posnetku kljub neenakemu tempu igranja ali petja. Uporabo dinamičnega ukrivljanja časa na področju MIR je Müeller dobro predstavil v delu [31]. V delu so predstavljene tudi variacije metode in njihova uporaba.

### 2.6.8 Skriti Markovski model

Skriti Markovski model (ang. Hidden Markov model - HMM) je statistični model, ki predvideva, da je modelirani sistem Markovski proces s skritimi stanji. Pri uporabi HMM se želimo modela naučiti iz samih podatkov in s tem iščemo statistični model, ki odraža strukturo v podatkih. Vsako stanje v HMM ima verjetnostno porazdelitev preko vseh izhodnih simbolov, parametri sistema pa so verjetnosti prehodov med stanji, povprečja in variance za posamezna stanja. Pri uporabi HMM je zelo pomembno učenje samega modela, za kar lahko uporabimo različne pristope nenadzorovanega učenja. Opis HMM, postopka učenja parametrov in ugotavljanja najbolj verjetnega stanja za posamezni okvir zvočnega signala pri razpoznavi govora Rabiner predstavi v [41, 40].

Pristop [29], uporablja 48 HMM-jev za detekcijo različnih akordov na podlagi česar zgradijo model razmejevanja delov zvočnega posnetka z vokalom in brez. V [27] avtor predstavi pristop za iskanje podobnih delov v pesmih. Z uporabo nenadzorovanega učenja ergodičnih HMM-jev, je posamezni HMM naučen za posamezno pesem. Stanje v katerem se HMM nahaja odraža skupino podobnih delov pesmi. Uporabljen Viterbijev algoritem [41] poišče najbolj verjetno zaporedje stanj pri danih podatkih in modelu. Rezultati Viterbijevega algoritma odražajo zgradbo pesmi. Zaporedni deli posnetka z istim stanjem so združeni v enotne segmente.

### 2.6.9 Kombinirane metode

V prejšnjih poglavjih so predstavljeni pristopi, ki za segmentacijo glasbenih posnetkov uporabljajo posamezno metodo in eno vrsto značilnic. Seveda pa obstajajo tudi pristopi, ki uporabljajo kombiniran pristop več metod in/ali značilnic izmed katerih je nekaj predstavljenih v nadaljevanju. Avtorji v [3] predstavijo pristop kjer glasbo modelirajo kot dinamično teksturo. Metoda uporablja kombinacijo značilnic MFCC in kromatskih vektorjev kot vhod za linearni dinamični sistem, ki na podlagi vhodnih značilnic vrne množico časovnih vrst znano tudi pod imenom dinamična tekstura (ang. dynamic texture) iz področja računalniškega vida [10]. Avtorji so na področke MIR vpeljali novo tehniko, prej uporabljeno na drugem raziskovalnem področju.

Kombinacijo značilnic MFCC in kromatskih vektorjev uporablja tudi pristop [13], kjer je predstavljena računsko učinkovita metoda za detekcijo ponavljajočih delov v popularni in rock glasbi. Značilnice MFCC in kromatski vektorji so uporabljene za izgradnjo seštevka matrik razdalj (samopodobnostni matriki). Prednost metode je uporaba ločenih matrik, kjer lahko za vsako uporabimo druge pristope za izboljšavo. Enaka kombinacija značilnic je uporabljena tudi v [34]. Opisan pristop je prav tako namenjen iskanju strukture v popularni glasbi. Pristop uporablja značilnice MFCC za "surovo segmentacijo", kromatske vektorje pa za primerjavo ujemanja posameznih segmentov. Pristop nato najdene segmente združuje in iz njih sestavi strukturo celotnega glasbenega posnetka.

Kombinacija uporabe značilnic MFCC in kromatskih vektorjev skupaj s komponento glasnosti (ang. loudness) je predstavljena v [22]. Uporabo kombinacije značilnic MFCC in kromatskih vektorjev skupaj z novo značilnico ritmogramom (ang. rhythmogram), predstavi avtor v [23, 24]. Ritmogram je prikaz ritmične informacije in razvoja ritma skozi čas s pomočjo avtokorelacije. Enaka kombinacijo značilnic je uporabljena tudi v pristopu [35]. Metoda posnetek najprej razdeli na veliko potencialnih segmentov, med katerimi paroma izračuna mero razdalje in jo pretvori v verjetnost pojavitve istega segmenta. Metodo za avtomatsko oceno glasbene strukture v glasbenih posnetkih, ki uporablja značilnice MFCC in kromatske vektorje, namesto ritmograma pa tehniko spektralnega kontrasta (ang. spectral contrast) je predstavljena v [38]. Predstavitev koeficientov spektralnega kontrasta, je podrobneje podana v [25].

Pristop [6] predstavi uporabo značilnic spektra in metodo združevanja segmentov za karakterizacijo ponavljajočih struktur v popularni glasbi. V prvem koraku je izračunana samopodobnostna matrika na podlagi spektralne podobnosti med okvirji. Meje med segmenti kitica-refren poiščemo s koreliranjem jedra vzdolž diagonale matrike. Po opravljeni segmentaciji izračunamo statistiko spektra vsakega okvirja. V drugem koraku združimo segmente glede na medsebojno podobnost glede na statistiko spektra z uporabo matrične dekompozicije.

Delo [28] predstavi uporabo metode konstantne-Q transformacije za izračun časovno-frekcvenčne predstavitve nad katero je izvedeno iskanje ponavljajočih vzorcev in strukture v zvočnih posnetkih. Konstantna-Q transformacija (ang. constant Q transform - CQT) je podrobno predstavljena v delu [5]. CQT je večločljivostna transformacija, ki predstavi spekter na logaritmični osi s slabšo časovno ločljivostjo pri nizkih in boljšo pri visokih frekvencah. Logaritmična skala je uporabljena, ker bolje aproksimira dejansko percepcijo višine tona.

Omenimo naj še pristop [38], ki kot nabor značilnic uporabi opis barve glasu (značilnice MFCC in spektralnega kontrasta) uporabljene v formatu MPEG-7[2]. Standard MPEG-7 je namenjen opisu vsebine večpredstavitvenih datotek sprejet

[2] http://mpeg.chiariglione.org/standards/mpeg-7/mpeg-7.htm

in potrjen s strani organizacij ISO[3] in IEC[4] leta 2006.

## 2.7 Segmentacija ljudske glasbe

### 2.7.1 Segmentacija glede na dihalne pavze z detekcijo tišine

Rezultati objavljeni v članku [47] in doktorski disertaciji [46] pričajo o tem, da lahko enoglasno ljudsko glasbo posneto na terenu v šumnih okoliščinah, z amaterskimi pevci, zelo dobro razdelimo v smiselne dele že z enostavno detekcijo vokalnih pavz v zvočnem posnetku. Smiselni deli ljudskih pesmi so večinoma prav kitice in morebitni refren.

Vokalna pavza, včasih poimenovana tudi dihalna pavza, je del zvočnega posnetka kjer pevec ne poje. Avtor v omenjenih delih vokalne pavze išče tako, da v kratkih delih posnetkov, dolgih nekaj 100 ms, zazna tonsko višino. Če tonska višina ne presega nekega definiranega praga, predpostavi, da v trenutnem delu posnetka ni prisotne tonske višine in s tem posledično ni prisotnega vokala.

Avtor pri zaznavanju prisotnosti tonske višine v zvoku uporablja algoritem YIN predstavljen v [8], ki je namenjen detekciji tonske višine v zvočnih posnetkih glasbe in govora. Algoritem YIN je bil primarno razvit za detekcijo tonske višine v govoru in enoglasnem petju, za kar ga avtor tudi uporablja. Problemi se pojavijo, ko želimo enak pristop uporabiti pri večglasnem petju ali petju z inštrumentalno spremljavo, ko algoritem YIN ne da dobrih rezultatov.

Metoda predstavljena v [47] zvočni posnetek razdeli na 100 ms dolge okvirje za katere z algoritmom YIN izračuna tonske višine. Dobljene tonske višine zgladi na podlagi petih predhodnih in petih sledečih tonskih višin. Dele posnetka, kjer algoritem veliko bližnjih okvirjev označi kot "brez tonske višine", označi kot vokalne pavze, na podlagi katerih nato izvede nadaljnjo segmentacijo. Pri tem avtor postavi tudi eksperimentalno določene velikosti smiselnih delov. Tako je najkrajši del posnetka brez tonske višine tisti, ki vsebuje vsaj 10 zaporednih okvirjev (233 ms) zvočnega posnetka brez zaznane tonske višine, najdaljši smiselni del posnetka pa ne sme presegati 360 okvirjev, kar pomeni 8,4 sekunde posnetka.

Pristop je prilagojen zvočnim posnetkom zbirke Nizozemskih ljudskih pesmi [20], kjer je uporabljen na posnetkih enoglasnega petja brez instrumentalne spremljave.

### 2.7.2 Iskanje kitic v posnetkih z DTW

Iskanje kitic v pesmih je eden izmed možnih pristopov k segmentaciji glasbe v smiselne dele. Takšen pristop je še posebej primeren pri ljudski glasbi, kjer je struktura zelo preprosta in so pesmi sestavljene zgolj iz kitic. Ker je kitica del pesmi kate-

[3]International Organization for Standardization
[4]International Electrotechnical Commission

rega melodija se skozi celotno pesem ponavlja, se iskanje le-teh, v neki pesmi, lahko prevede na problem iskanja delov s ponavljajočo melodijo.

Za reševanje tovrstnega problema obstaja več pristopov [32, 33]. Avtorji v delih predstavijo pristop, pri katerem na podlagi predloge melodije v simbolični obliki poiščejo ustrezne ponavljajoče dele glasbe, ki ustrezajo melodiji iz predloge. Sledi podrobnejši opis pristopa [32].

Kot vhod algoritem sprejme zvočni posnetek ljudske pesmi in MIDI[5] predstavitev predloge melodije ene kitice, ki se v tej ljudski pesmi ponavlja. Za iskanje ponavljajočih delov predloge v celotni pesmi so v pristopu uporabljeni kromatski vektorji (predstavljeni v 2.6.2), ki predstavljajo močno srednje-nivojsko predstavitev za povezovanje harmoničnosti v glasbi. Pretvorba iz MIDI zapisa v obliko s kromatskimi vektorji je enostavna, saj lahko iz podatkov o času dogodkov in višin tonov določimo v katerega izmed kromatskih razredov sodi del posnetka za okno fiksne dolžine [19]. Nadalje avtorji uporabijo izvedene kroma značilnice CENS [31].

Zaporedje kromatskih značilnic MIDI predloge dolžine $K$ imenujemo $\mathbf{X}$, zaporedje značilnic zvočnega signala dolžine $L$ pa $\mathbf{Y}$. Značilnice $x_k, k \in [1 : K]$ in $y_l, l \in [1 : L]$ so normirani 12-dimenzionalni vektorji. Cenilna funkcija med kromagramom MIDI predloge in ustreznim delom zglajenega kromagrama zvočnega posnetka, v odvisnosti od $\mathbf{X}$ in $\mathbf{Y}$, z uporabo izvedenke dinamičnega ukrivljanja časa (DTW, glej 2.6.7), je definirana kot:

$$\Delta(l) := \frac{1}{K} \underset{a \in [1:l]}{min} (\mathrm{DTW}(\mathbf{X}, \mathbf{Y}_{a:l})), \tag{2.22}$$

kjer $\mathbf{Y}_{a:l}$ predstavlja podzaporedje $\mathbf{Y}$ z začetnim indeksom $a$ in končnim indeksom $l \in [1, L]$ in $\mathrm{DTW}(\mathbf{X}, \mathbf{Y}_{a:l})$ predstavlja DTW ceno med $\mathbf{X}$ in $\mathbf{Y}_{a:l}$ glede na lokalno mero podobnosti (v konkretnem primeru kosinusno mero podobnosti glej 2.6.3). Cenilna funkcija $\Delta$ je izračunana zelo učinkovito s pomočjo dinamičnega programiranja. Funkcijo $\Delta$ lahko razložimo kot: majhna vrednost $\Delta(l)$ za nek $l \in [1 : L]$ odraža, da je podzaporedje $\mathbf{Y}$ z začetnim indeksom $a_l$ (kjer je $a_l \in [1 : l]$ minimalni indeks v enačbi 2.22) in končni indeks $l$, podobno zaporedju $\mathbf{X}$. Indeks $a_l$ lahko pridobimo z enostavnim algoritmom sledenja znotraj izračuna DTW.

Sama segmentacija sledi preprosti požrešni segmentacijski metodi. Z uporabo $\Delta$ funkcije poišče indeks $l \in [1 : L]$ z minimizacijo $\Delta$ in izračunom začetnega indeksa $a_l$. Tako interval $S_1 := [a_l : l]$ predstavlja prvi segment. Na vrednost mere $\Delta(l)$ gledamo kot na *ceno* segmenta. V izogib prevelikim prekrivanjem različnih segmentov je izločena okolica $[L_l : R_l] \subset [1 : L]$ okoli indeksa $l$ iz nadaljnje obravnave. V pristopu sta uporabljeni vrednosti $L_l := max(1, l - \frac{2}{3}K)$ in $R_l := min(L, l + \frac{2}{3}K)$, kar predstavlja izključitev dveh tretjin dolžine predloge na levo in desno od indeksa $l$, kar se odraža v prilagojeni $\Delta$ funkciji z $\Delta(m) := \infty$ za $m \in [L_l : R_l]$. Za detekcijo

[5]Musical Instrument Digital Interface - opis formata je dostopen na spletnem naslovu: http://www.midi.org/

naslednjega dela $S_2$ je uporabljen enak pristop z modificirano cenilno funkcijo, in tako naprej za ostale dele. Postopek je ponovljen dokler vrednosti $\Delta$ ne ležijo nad ustrezno izbranim *kvalitetnim pragom* $\tau > 0$.

Predstavljene razširitve metode so:

- prva izboljšava je namenjena povečanju kvalitete kromatskih značilnic [16]. Pristop predstavi izbiro prevladujočih spektralnih koeficientov, kar se odraža s signifikantnim dušenjem šumnih komponent. Zaradi uporabe na enoglasni glasbi je mogoče izvesti korak več, z izbiro spektralnih komponent signala, ki ustrezajo osnovni frekvenci ($F0$). Korak je izveden tako, da je s pomočjo algoritma YIN [8] izračunana osnovna frekvenca vsakega okvirja, kateri se poišče ustrezna tonska višina enakomerno uglašene tonske lestvice. Srednja frekvenca tonske višine enakomerno uglašene tonske lestvice je najbližja izračunani $F0$. Pri izračunu kromatskih značilnic je energija dodeljena zgolj delu spektra, ki ustreza najdeni tonske višine enakomerno uglašene tonske lestvice, ostali deli spektra pa se ne upoštevajo (njihova energija je postavljena na 0). Nazadnje je dobljena *redka* predstavitev tonske višine projicirana na kromatsko predstavitev in zglajena. Avtorji takšni značilnici pravijo $F0$-izboljšan kromagram. Izboljšava odpravlja slabosti šumnih snemalnih okolij, tipičnih pri snemanju ljudske glasbe.

- druga izboljšava se spopade z netočnim petjem izvajalcev, ki velikokrat zgrešijo želeno tonsko višino za nekaj poltonov. Izboljšava se zanaša na opažanja predstavljena v [17], da dvanajst rotacij 12-dimenzionalnega kromatskega vektorja ustreza dvanajstim naravnim možnim transpozicijam. Rotacija 12-dimenzionalnega kromatskega vektorja odraža naravno transpozicijo melodije. Z izboljšavo je ugotovljen indeks rotacije, ki minimizira razliko med tonsko višino MIDI predloge in pridobljene melodije in za ta indeks rotira kromatsko predstavitev. Minimizacija indeksa je izvedena z 12-kratno rotacijo kromatske predstavitve in izračunom dvanajstih funkcij razdalje (za vsak zasuk). Med njimi nato izberemo tisti indeks zasuka, za katerega dobimo minimalno mero razdalje. To je predstavljeno z minimizacijo $\Delta^0 \ldots \Delta^{11}$ poimenovano z $\Delta^{trans}; [1 : L] \rightarrow \mathbb{R} \cap \{\infty\}$:
$$\Delta^{trans}(l) := \underset{i \in [0:11]}{min} (\Delta^i(l)). \tag{2.23}$$

- zadnja izboljšava predstavljenega algoritma obravnava uporabo zamaknjenih filtrov za izračun kromatskih značilnic, pri čemer so vse množice uporabljenih filtrov za izračun teh značilnic zamaknjene za polovico poltona. Tako skupaj z novo množico filtrov dobimo s skupaj s starimi cenilnimi funkcijami, 24 cenilnih funkcij $\Delta$. Minimizacijo preko vseh 24 $\Delta$ funkcij imenujemo $\Delta^{fluc}$ poimenovano tudi *mera razdalje neodvisna na nihanja*. Izboljšava upošteva nihanje pevčevega glasu med petjem s čimer se poveča natančnost cenilne funkcije.

Metoda zvočne posnetke monofonične ljudske glasbe segmentira v ponavljajoče dele z visoko natančnostjo. Eksperimenti izvedeni v okviru predstavljenega dela so zajemali 47 zvočnih posnetkov ljudske glasbe, ki skupaj predstavljajo 156 minut posnetkov. Pesmi so sestavljene iz 4 do 34 ponavljajočih delov. V zbirki so bili ročno označeni začetki in konci ponavljajočih delov. Kot tolerančni faktor $\delta$ v eksperimentu je razdalja meje določene z metodo od ročno postavljene meje v sekundah. Za tolerančni faktor $\delta = 2$ so vrednosti mer točnosti sledeče: natančnost (ang. precision) $P = 0,912$, priklic (and. recall) $R = 0,940$ in mera F1 (ang. F1-measure) $F = 0.926$, kar predstavlja visoko uspešnost metode pri segmentaciji zvočnih posnetkov. Rezultati so precej slabši za tolerančni faktor $\delta = 1$, kjer so vrednosti $P = 0,637$, $R = 0,639$ in $F = 0,638$.

Metoda za svoje delovanje uporablja predlogo melodije v simbolični obliki. S tem je do neke mere olajšano iskanje segmentacije, saj že s samo predlogo dobimo približno dolžino iskanih segmentov in njihovo melodijo. Za podatke uporabljene v okviru magistrske naloge takšne predloge nimamo in je potrebno lastnosti ponavljajočih segmentov pridobiti iz samih zvočnih posnetkov ljudske glasbe, kar predstavlja dodaten izziv.

### 2.7.3 Segmentacija posnetkov glede na vsebino z verjetnostnim modelom

V predhodnih poglavjih predstavljeni pristopi so namenjeni segmentaciji vsebinsko zaključenih zvočnih posnetkov, ki vsebujejo glasbo. Pristopi uporabljajo različne definicije smiselnosti posameznih delov, saj so prirejeni za uporabo na raznovrstni glasbi z različno sestavo. Med opisanimi pristopi ni takšnega, ki zajema višjenivojsko segmentacijo. Pristop, ki rešuje takšen problem predstavi avtor v [30]. V delu je predstavljen pristop, ki terenski zvočni posnetek, razdeli na dele glede na vrsto zvoka (govor, solo petje ...). Poleg tega, da pristop terenski posnetek razdeli v dele glede na vsebino, jim tudi pripiše oznake. Predstavljen pristop terenske posnetke razdeli na dele, ki vsebujejo: govor, solo petje, večglasno petje (vsi nastopi z dvema ali več pevci), inštrumentalna glasba (vključuje inštrumentalno glasbo s petjem) in pritrkavanje zvonov. Oznake so določene glede na vsebino terenskih posnetkov slovenske ljudske glasbe iz digitalnega arhiva EtnoMuza. Pristop določi oznako dela posnetka glede na 9 izračunanih značilnic:

- kvocient variance RMS energije preko kvadratnega povprečja RMS energije, kjer je RMS energija definirana kot:

$$r = \sqrt{\frac{1}{W} \sum_{i=0}^{W-1} x_i^2}, \tag{2.24}$$

  kjer $x$ predstavlja zvočni signal in $W$ velikost okna. Značilnica predstavlja

vrednost nihanj energije signala, ki je tipično večja za govor kot ostale vrste signalov;

- povprečna spektralna entropija kot je definirana v [39]. Entropija predstavlja nestabilnost energije signala izračunano na več spektralnih pasovih, ki je tipično nizka za dele s pritrkavanjem zvonov, nekaj višja za glasbo in visoka za ostalo. Izračunamo jo kot:

$$H = -\sum_{i=0}^{L-1} \frac{X_i}{\sum_{j=0}^{L-1} X_j} log_2 \frac{X_i}{\sum_{j=0}^{L-1} X_j}, \tag{2.25}$$

  kjer $L$ predstavlja število spektralnih pasov in $X_i$ energijo $i$-tega pasu (bolj podrobno opisano v [39]);

- varianca delt spektralne entropije, kjer so delte izračunane kot linearni trend preko 5 zaporednih oken;

- variance prvih treh MFCC koeficientov (brez ničtega), ki definirajo obliko spektra signala,

- variance delt prvih treh MFCC koeficientov (brez ničtega), kjer so delte izračunane kot linearni trend preko pet zaporednih oken.

Na podlagi predstavljenih značilnic je sestavljen klasifikator, naučen na množici kratkih, ročno označenih posnetkov. Vse značilnice so izračunane na signalu nad katerim je uporabljeno Hammingovo okno dolžine 46ms s 23ms prekrivanjem. Povprečja in variance so bile izračunane na celotni dolžini 3-sekundnih delov. Za klasifikacijo je uporabljen multinomski logistični regresijski kvalifikator zaradi enostavnosti in dobrih rezultatov. Nadalje je lahko rezultat klasifikatorja obravnavan kot verjetnostna porazdelitev med posameznimi razredi. Klasifikacija ima 78% klasifikacijsko točnost. Do napak prihaja predvsem pri govoru in solo petju (kar lahko enostavno razložimo z dejstvom, da gre za amaterske pevce), prav tako prihaja tudi do mešanja med solo petjem in večglasnim petjem, kjer vsi pevci pojejo isto melodijo.

Sledi postopek segmentacije za katerega se najprej poišče množica kandidatov za meje za katere se izračuna verjetnost razmejitve posnetka. Glede na združeno verjetnost vseh delov ob upoštevanju predznanja o dolžini delov posameznega razreda. Kandidati za meje med posameznimi deli so pridobljeni glede na dva kriterija. Prvi je energija signala (posamezni smiselni deli so ločeni s tišino) drugi kriterij pa so meje med deli z različnimi oznakami (solo petje, ki sledi govoru ipd.). Formalno lahko pristop predstavimo kot:

$$\mathcal{B} = \left\{ (b_l, b_r) | \forall t \in [b_l, b_r] : \left[ \begin{array}{l} e_t < max(e_t^f E_l, E_0) or \\ d_t > max(d_t^f + D_l, D_0) \end{array} \right] \right\}, \tag{2.26}$$

kjer sta $E_0$ in $E_l$ globalni in relativni prag za izbiro kandidata na podlagi energije in sta $D_0$ in $D_l$ globalni in relativni prag za izbiro mej med deli glede na vsebino. Namesto iskanja najboljših vrednosti za $D_l$ in $E_l$ pristop obravnava izbiro mej kot klasifikacijski problem, ki ga reši z dvema regresijskima klasifikatorjema (eden temelječ na energiji, drugi pa na KL divergenci) za napovedovanje verjetnosti delitve posnetka na določeni kandidatni meji. Značilnici uporabljeni pri klasifikaciji pragu sta $s_e$ za oceno signala pod pragom in $m_e$ za oceno maksimalne razlike med vsebino signala levo in desno od meje. Izračunata se:

$$s_e = \sqrt{\sum_{t=b_l}^{b_r} max(e_t^f E_l, E_0) - e_t} \text{ in} \tag{2.27}$$

$$m_c = \frac{1}{N+1} \underset{c}{max} \left| \sum_{t=b_l-N}^{b_l} P(c_t = c) - \sum_{t=b_r}^{b_r+N} P(c_t = c) \right|, \tag{2.28}$$

kjer $P(c_t = c)$ predstavlja verjetnost, da $c_t$ v času $t$ pripada razredu $c$ in $N$ število upoštevanih okvirjev na levo in desno stran meje. Za najboljša kvalifikatorja sta se izkazala količina divergence nad pragom ($s_d$) in skupna količina divergence v okviru meje ($t_d$):

$$s_d = log \left( 1 + \sum_{t=b_l}^{b_r} d_t - max(d_t^f + D_l, D_0) \right), \tag{2.29}$$

$$t_d = log \left( 1 + \sum_{t=b_l}^{b_r} d_t \right). \tag{2.30}$$

Za razliko od izbire fiksnih pragov, klasifikatorja izboljšata natančnost izbire. Dodatna prednost pa je, da lahko na rezultate klasifikatorjev gledamo kot na verjetnost razdelitve posnetka na kandidatni meji; dejstvo, ki je izkoriščeno pri segmentaciji.

Sama segmentacija sledi logičnemu Bayesovskemu modeliranju tako, da maksimirajo:

$$P(seg|data) \propto P(data|seg)P(seg) \tag{2.31}$$

Segmentacija je predstavljena kot zaporedje segmentov $S_{i1}, S_{i2} \ldots S_{iN}$, $0 < i1 < i2 < \ldots < iN$, kjer se $S_{i1}$ prične s časom 0 in konča z $B_{i1}$, $S_{i2}$ začne z $B_{i1}$ in konča z $B_{i2}$ in tako naprej. Vsako od mej $B_t \in \mathcal{B}$ je diskretna naključna spremenljivka z dvema možnima rezultatoma; ali kandidatna meja predstavlja dejansko mejo med dvema deloma in tako posnetek razdeli na dva dela, ali pa ne. Verjetnostna funkcija spremenlivke je definirana z rezultatoma klasifikatorjev temelječih na energiji ($P_e$) in KL divergenci ($P_{kl}$):

$$P(B_t = true) = max(P_e(B_t), P_{kl}(B_t)) \tag{2.32}$$

V pristopu je verjetnost vsakega segmenta odvisna zgolj od lokacije prejšnjega segmenta:

$$P(S_{i1})P(S_{i2}|S_{i1})P(S_{i3}|S_{i2})\dots P(S_{iN|S_{iN-1}}). \tag{2.33}$$

Za izračun segmenta $S_i$ pri podanem segmentu $S_j$ moramo upoštevati vse kandidatne meje znotraj segmenta, kot tudi njegovo trajanje. Če naj bi se segment začel v času $j$ in končal v času $i$, morajo biti vse meje znotraj segmenta napačne ($false$), vrednost v času $j$ pa more biti resnična ($true$). Segmentacija je nadalje omejena s predznanjem o tipični dolžini segmentov glede na njihov razred, kar opišemo kot:

$$P(S_i|S_j) = P(D_i|S_i, S_j)P(B_i = true)\prod_{j<k<i} P(B_k = false). \tag{2.34}$$

S tem se enačba 2.33 spremeni v:

$$\prod_{(i,j)\in\mathfrak{G}} (P(B_i = true)P(D_i|S_i, S_j)) \times \prod_{j\in\mathfrak{G}} P(B_k = false), \tag{2.35}$$

kjer je $\mathfrak{G}$ množica vseh indeksov segmentov in $(i, j)$ par zaporednih indeksov iz te množice.

Verjetnost trajanja segmenta glede na podane meje je odvisna od razreda segmenta, ki je pridobljena z že predstavljenim klasifikatorjem, in imajo, glede na analizirano zbirko terenskih posnetkov, povprečne vrednosti in standardne deviacije $(\mu_c, \sigma_c)$. Z dodatno vsiljenim najkrajšim trajanjem segmenta $D_{min}$ dobimo:

$$P(D_i|S_i, S_j) = \begin{cases} \sum_c P(C_i = c|S_i, S_j)G(i-j, \mu_c, \sigma_c) \\ o, i-j < D_{min} \end{cases}, \tag{2.36}$$

kjer $P(C_i = c|S_i, S_j)$ predstavlja verjetnost, da segment $S_i$ pripada razredu $c$ in je izračunan kot povprečna verjetnost klasifikacije vseh okvirjev v segmentu v razred $c$. $G$ je osnovna Gaussova funkcija. Za iskanje zaporedja, ki maksimira enačbo 2.35 in s tem predstavlja optimalno rešitev je uporabljen pristop dinamičnega programiranja, ki privede do enostavne rešitve. Za vsak segment $S_i$, ki se konča s kandidatno mejo $B_i$, izračunamo najbolj verjetno segmentacijo, ki se konča z mejo $d(S_i)$ po naslednjih pravilih:

$$d(S_i) = \begin{cases} 0,5 & i = 0, \\ P(B_i = true)\max\limits_{0\le j<i}(d(S_j), c(i,j)) & i > 0 \end{cases}, \tag{2.37}$$

in je $c(i, j)$ definiran kot:

$$c(i,j) = P(D_i|S_i, S_j)\prod_{j<k<i} P(B_k = false), \tag{2.38}$$

kjer $S_0$ predstavlja mejo pri času 0. $S_0$ je meja, če se v času 0 tudi začne sama vsebina v posnetku ali če je šum ali tišina, zaradi tega mu pripišemo verjetnost $0,5$. Tako dobljeni deli so označeni z razredom $c$, ki maksimira $P(C_i = c|S_i, S_j)$.

# Poglavje 3

# Metoda za večnivojsko segmentacijo terenskih posnetkov

V predhodnem poglavju smo predstavili dosedanje pristope k reševanju problema segmentacije zvočnih posnetkov na različnih nivojih. Med predstavljenimi pristopi so vsi omejeni zgolj na segmentacijo posnetkov na posameznem nivoju, ni pa pristopa, ki bi združeval segmentacijo na več nivojih. V nadaljevanju opisan pristop vključuje nekatere predstavljene ideje in jih poveže v enotni pristop za večnivojsko segmentacijo terenskih posnetkov.

Terenski posnetek na dveh nivojih razdelimo v smiselne dele, kot jih pri svojem delu iščejo etnomuzikologi pri ročnem delu. Ta različna nivoja segmentacije zajemata:

- na prvem (višjem) zvočni posnetek razdelimo na vsebinsko smiselne dele [30] in tako dobimo dele posnetka označene glede na vsebino;

- na drugem (nižjem) nivoju razdelimo dele pridobljene iz segmentacije prvega (višjega) nivoja označene kot *solo* in *večglasno petje* in jih nadalje razdelimo na smiselne dele. Kot rezultat segmentacije drugega nivoja dobimo posamezne kitice, kar v primeru ljudskih pesmi pomeni dele s ponavljajočo melodijo.

V nadaljevanju uporabimo rezultate segmentacije drugega nivoja za boljšo segmentacijo prvega nivoja pri delih označenih kot *solo petje* in *večglasno petje*. V tem delu poizkušamo izboljšati oceno začetkov pesmi izračunanih pri segmentaciji prvega nivoja, kot tudi poiskati meje v delih prvega nivoja, kjer se zaporedoma pojavljajo pesmi, a jih segmentacija prvega nivoja med seboj ni uspela ločiti.

## 3.1 Opredelitev problema

Algoritem predstavljen v tem poglavju kot vhod dobi digitaliziran terenski posnetek, ki ga z dvonivojsko segmentacijo razdeli na posamezne enote, enote pa se na kitice.

Rezultat segmentacije višjega nivoja so časi začetkov in koncev enot posnetkov s posamezno vrsto zvoka (*solo petje*, *večglasno petje*, *inštrumentalno glasbo*, *pritrkavanjem zvonov*, *govor*). Enote so opremljene z ustreznimi oznakami. Enote z oznakami solo petje in večglasno petje uporabimo kot vhod v segmentacijo na nižjem nivoju. Rezultati segmentacije nižjega nivoja so časi začetkov ponavljajočih kitic znotraj enot posnetkov pridobljenih z višjenivojsko segmentacijo.

Vsak del metode je v nadaljevanju opisan v svojem poglavju. Najprej segmentacija na višjem nivoju (3.2), nato segmentacija na nižjem nivoju (3.3) in na koncu še izboljšave višjenivojske segmentacije z rezultati nizkonivojskega pristopa (3.4).

## 3.2 Segmentacija prvega (višjega) nivoja

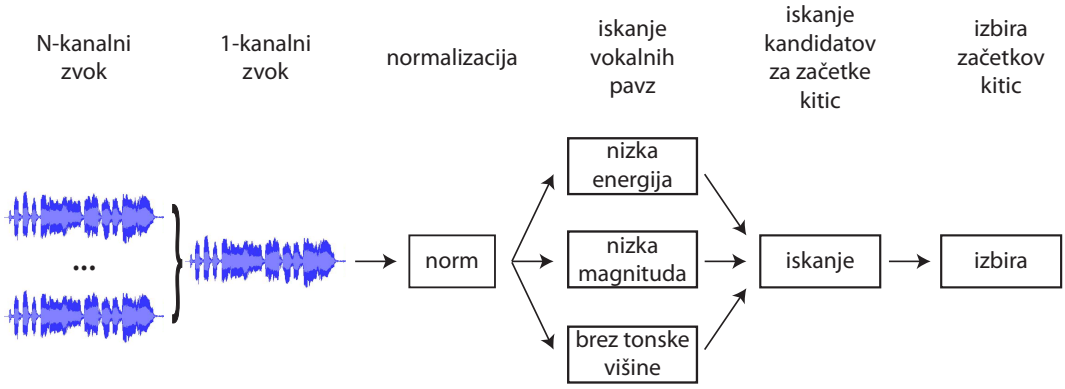
Segmentacija višjega nivoja je izvedena nad celotnimi terenskimi posnetki. Na tem nivoju za segmentacijo uporabimo že predstavljeno metodo, ki jo je razvil Marolt [30] in je podrobno predstavljena v poglavju 2.7.3. Kot smo že omenili metoda za vhod vzame datoteko z zvočnim zapisom celotnega terenskega posnetka ter na izhod vrne čase začetkov in koncev posameznih delov, ki so označeni z ustreznimi oznakami. Dele terenskega posnetka označi z oznakami *govor*, *solo petje*, *večglasno petje*, *inštrumentalna glasba* in *pritrkavanje zvonov*. Pridobljene dele z oznakami *solo petje* in *večglasno petje* neposredno uporabimo v segmentaciji nižjega nivoja opisanega v naslednjem poglavju.

## 3.3 Segmentacija drugega (nižjega) nivoja

Segmentacija nižjega nivoja je izvedena nad segmenti pridobljenimi z visokonivojsko segmentacijo označenih z oznako *solo petje* ali *večglasno petje*. S pridobljeno segmentacijo razrežemo originalni posnetek in nad posameznim delom izvedemo nižjenivojski pristop. Osnovna ideja nižjenivojske segmentacije je poiskati ponavljajoče dele posnetka glede na melodijo. Pri tem upoštevamo dejstvo, da pevci pesmi ne odpojejo točno, kar se odraža z razliko v tonski višini za nekaj poltonov med posameznimi ponavljajočimi deli, in dejstvo, da tempo v odpetih skladbah zelo niha, tako znotraj kitic, kot med kiticami. Glede na to, da so pevci amaterski pevci, ki nimajo treniranega glasu in glasbene izobrazbe, so navedena dejstva pričakovana in jih je nujno upoštevati. To je tudi bistvena razlika s segmentacijo popularne glasbe. Večina algoritmov predstavljenih v prejšnjem poglavju temelji na segmentaciji popularne glasbe, ki predstavlja točno tonsko in ritmično odigrane skladbe brez velikih zunanjih vplivov (sum okolice, slabi snemalni pogoji ipd.). Kot pokažemo

v poglavju 4, tovrstni algoritmi dosegajo slabe rezultate pri segmentaciji ljudske glasbe. To je povod za razvoj lastne metode nizkonivojske segmentacije, ki upošteva lastnosti zvočnih posnetkov ljudske glasbe (amaterski pevci, slabi snemalni pogoji ipd.).

Nižjenivojsko segmentacijo razdelimo na več ločenih korakov: predpriprava vhodnega signala, iskanje vokalnih pavz, iskanje kandidatnih začetkov kitic, ocena kandidatov in izbira dejanskih začetkov kitic. Delovanje metode po korakih je opisano v nadaljevanju in prikazano v Sliki 3.3.




Slika 3.1: Prikaz delovanja metode.

### 3.3.1 Predpriprava vhodnega signala

Za vhodni signal najprej preverimo ali je v enokanalni ali večkanalni obliki. V kolikor je vhodni signal v večkanalni obliki, ga pretvorimo v enokanalno obliko:

$$\mathbf{s} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{s}_i, \tag{3.1}$$

kjer je $\mathbf{s}$ enokanalni signal, izračunan kot povprečna vrednost signalov posameznih kanalov, $\mathbf{s}_i$ pa vektor časovne vrste $i$-tega kanala. Nadalje vhodni signal normaliziramo:

$$\mathbf{s}_{norm} = \frac{\mathbf{s}}{max(|\mathbf{s}|)}. \tag{3.2}$$

Tako pripravljen vhodni signal nato uporabimo v nadaljnjih korakih.

### 3.3.2 Iskanje vokalnih pavz

V posnetkih ljudskih pesmi navadno nastopajo amaterski pevci, ki med izvedbo delajo značilne dihalne pavze. Le-te se v samem posnetku izražajo kot deli tišine.

Tišino v posnetkih zaznavamo na več načinov. V našem delu uporabljamo 3 pristope: iskanje delov posnetkov z nizko energijo signala, iskanje delov posnetkov z nizko glasnostjo ter iskanje delov posnetkov brez zaznavne prevladujoče osnovne frekvence. Vokalne pavze predstavljajo možne začetke kitic, ki jih nato ovrednotimo in s tem dobimo dejanske začetke kitic.

### Iskanje vokalnih pavz glede na energijo

Energijo signala računamo na 200 ms dolgih okvirjih signala, na sledeč način:

$$E_j = \sum_{i=j*w_s}^{(j+1)*w_s-1} |s_i|^2, j = 0, 1 \ldots \lfloor \frac{N}{w_s} \rfloor, \tag{3.3}$$

kjer je $E_j$ energija signala $j$-tega okvirja, $s_i$ je $i$-ti element vhodnega vektorja $\mathbf{s}_{norm}$ in $w_s$ dolžina okvirja za katerega računamo energijo. Pavze so deli signala pod pragom $\xi_1$. Eksperimentalno smo prag za tišino nastavili na $\xi_1 = \frac{\overline{E}}{120}$, kjer je $\overline{E}$ srednja vrednost energij izračunanih za posamezna okna signala. Zaporedne okvirje signala z energijo nižjo od $\xi_1$ združimo v dele, ki določajo vokalne pavze. Konci dobljenih vokalnih pavz ($e_i \in \mathfrak{K}_{le}$) nam predstavljajo množico ($\mathfrak{K}_{le}$) kandidatov možnih začetkov kitic.

Ostane nam zgolj še izključitev neželenih delov pridobljenih zaradi nizkega praga, potrebnega za detekcijo vokalnih pavz. Iz množice izključimo vokalne pavze krajše od $\xi_2 = 0,7$ kratnika povprečne dolžine zaznane vokalne pavze v posnetku. Povprečna dolžina vokalne pavze je število zaporednih okvirjev, ki vokalno pavzo določajo. Krajše vokalne pavze niso tiste, ki jih iščemo in se jim želimo zaradi pohitritve v nadaljevanju metode izogniti. Tudi $\xi_2$ je določen eksperimentalno. Primer zaznanih zvočnih pavz je prikazan v sliki 3.2(a).

### Iskanje vokalnih pavz glede na amplitudno ovojnico

Amplitudno ovojnico zvočnega signala izračunamo s filtriranjem absolutnega signala z Butterworthovim filtrom četrte stopnje in mejno frekvenco določeno s faktorjem 0.01. V logaritmični skali nato iščemo dele signala, kjer je signal pod mejno vrednost določeno s parametrom $\xi_4 = -60$ dB, določeno eksperimentalno. Tako pridobljene dele signala podobno kot v podpoglavju 3.3.2 združimo glede na njihovo bližino, kjer za razliko ne iščemo striktno zaporednih delov signala, ki ustrezajo pogoju magnitude. Med deli signala z nizko amplitudo dovoljujemo maksimalni razmak predstavljen s parametrom $\xi_5 = 0,5$ sekunde, medtem ko zaznane dele krajše od 125 ms ignoriramo.

Metoda izkorišča podoben pristop kot predhodno opisana metoda, le da manj zgladi amplitudno ovojnico. Razlika je v izbiri delov, ki predstavljajo vokalne pavze. Predhodno predstavljena metoda izloča vokalne pavze krajše od deleža povprečno dolge vokalne pavze, opisan pristop pa izloča vokalne pavze fiksne dolžine.

Rezultat vokalnih pavz odkritih s tem pristopom je prikazan v sliki 3.2(b). Konce ($e_i \in \mathfrak{K}_{ll}$) vokalnih pavz združimo v množico $\mathfrak{K}_{ll}$ in nam predstavljajo kandidate možnih začetkov kitic. Iz slik predstavljenih metod je razvidno, da metodi najdeta večino istih vokalnih pavz, vsaka pa najde tudi vokalne pavze, ki jih druga metoda ne odkrije. S tem prispevamo v množico kandidatov za začetke kitic večni nabor kandidatov.

### Iskanje glede na spremembo višine tona

Za detekcijo delov posnetka brez prevladujoče osnovne frekvence uporabljamo pristop na kratko predstavljen v poglavju 2.7.1. Vhodni posnetek najprej prevzorčimo na frekvenco vzorčenja $f_s = 1024$ Hz. Nad tako preoblikovanim signalom poženemo algoritem YIN [8], ki za vsak okvir signala izračuna osnovno frekvenco ($F0$). Okvir je velik 32 zaporednih vzorcev, kar v našem primeru pomeni 31,3 ms. Zaporedni okvirji se med seboj ne prekrivajo.

$$\mathbf{F0} = \mathrm{YIN}(\mathbf{s}), \tag{3.4}$$

kjer je $\mathbf{f}_0$ vektor zaznanih frekvenc v vhodnem posnetku predstavljenim z $\mathbf{s}$. Tako dobljeno zaporedje zaznanih frekvenc v zvočnem posnetku zgladimo s povprečenjem preko 5 predhodnih in 5 sledečih okvirjev. Povprečenje ignorira okvirje brez zaznane frekvence in izračuna povprečje na preostalih okvirjih.

$$\overline{F0}_i = \frac{1}{v} \sum_{i-5}^{i+5} F0_i, \tag{3.5}$$

kjer je $v$ število okvirjev z definirano frekvenco v predpisani okolici $i$-ja. V zaporedju zaznanih frekvenc nato poiščemo okvirje kjer se zaznana frekvenca od srednje frekvence celotnega posnetka ($Fs$) razlikuje več kot ekvivalent $\xi_3 = 20$ poltonov.

$$Fs = \frac{1}{w} \sum_{i=1}^{L} F0_i, \tag{3.6}$$

kjer je $w$ število okvirjev v signalu z zaznano frekvenco in $L$ dolžina vektorja $F0$. Obdržimo torej zgolj okvirje za katere velja:

$$|\overline{F0}_i - Fs| < 20 \text{ poltonov}, \tag{3.7}$$

kjer $|_|$ predstavlja absolutno vrednost. Vrednost $\xi_3$ je določena eksperimentalno. Tukaj postopamo podobno kot v predhodno omenjenih pristopih iskanja vokalnih pavz. Zaporedne okvirje z zaznano frekvenco, katerih število je večje od eksperimentalno določenega parametra *minimalne dolžine brez zaznane frekvence*, združimo v vokalne pavze. Parameter minimalne dolžine brez zaznavne frekvence je v našem

primeru 4 okvirjev oz. $\xi_4 = 4$ ms. Dele krajše od $\xi_4$ brez tonske višine ignoriramo. Upoštevamo tudi parameter *minimalne dolžine segmenta*, ki je v našem primeru 256 okvirjev oz. $\xi_5 = 250$ ms. Tudi ta parameter je določen eksperimentalno. Zopet dobimo vokalne pavze katerih konci ($e_i \in \mathfrak{K}_{\mathfrak{F}0}$) nam predstavljajo množico ($\mathfrak{K}_{F0}$) kandidatov možnih začetkov kitic.

Slika 3.2(c) prikazuje rezultat zaznanih zvočnih pavz s predstavljeno metodo.

### 3.3.3 Iskanje kandidatov za začetke kitic in njihova ocena

Pri iskanju kandidatov za začetke kitic združimo množice vseh pristopov zaznavanja vokalnih pavz $\mathfrak{K} = \mathfrak{K}_{le} \cup \mathfrak{K}_{ll} \cup \mathfrak{K}_{F0}$, kjer so množice $\mathfrak{K}_{le}$, $\mathfrak{K}_{ll}$ in $\mathfrak{K}_{F0}$ rezultat pristopov predstavljenih v predhodnem poglavju (3.3.2). Primer rezultata takšne detekcije vokalnih pavz je razvidna iz slike 3.2(d), kjer so zbrani vsi konci vokalnih pavz (in s tem predvideni začetki kitic). Predstavniki posamezne metode so predstavljeni z različnimi odtenki zelene barve. V kolikor kandidati za začetke, pridobljeni s posameznim pristopom, sovpadajo, se v množici $\mathfrak{K}$ ta kandidat pojavi le enkrat.

Predpostavimo, da je prvi zaznan kandidat $k_0$ iz množice $\mathfrak{K}$ resnični začetek. Za $\xi_8 = 10$ sekundni del posnetka s pričetkom v $k_0$ izračunamo podobnosti z enako dolgimi deli posnetka z začetki v $k_i \in \mathfrak{K}, k > 0$. 10 sekundne dele posnetka smo izbrali eksperimentalno. Ta dolžina predstavlja približno polovico povprečne dolžine kitice eksperimentalne množice pesmi. Povprečna dolžina kitice eksperimentalne množice pesmi meri 20,03 sekund. Mera s katero računamo podobnosti je opisana z naslednjimi koraki in prikazana v Sliki 3.3.3:

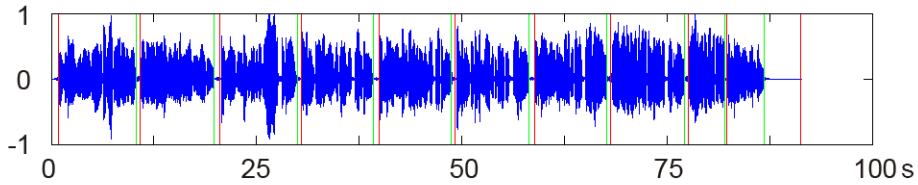
1. Najprej za vsakega izmed delov posnetkov izračunamo kromagrame (glej 2.6.2), s 50 ms dolgim oknom. Kot rezultat dobimo matriko 12 dimenzionalnih vektorjev za vsakega izmed delov posnetkov, ki se začnejo s kandidatom za začetek kitice ($k_i$).

2. Na tem mestu definiramo cenilno funkcijo s pomočjo RMS razdalje med posameznimi vektorji:

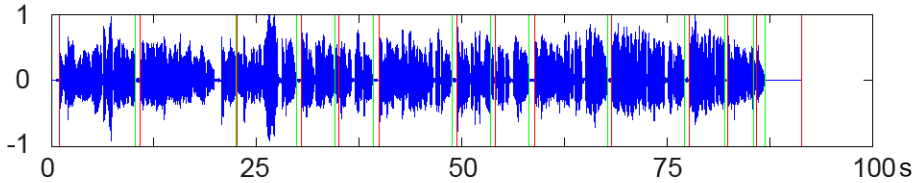
$$c(\mathbf{a}, \mathbf{b}) = \frac{\sqrt{(\sum_i (a_i - b_i)^2}}{\sqrt{12}}, \tag{3.8}$$

kjer je $c$ cenilna funkcija med kromatskima vektorjema $\mathbf{a}$ in $\mathbf{b}$, $a_i$ in $b_i$ pa sta $i$-ta elementa kromatskega vektorja. Majhna vrednost cenilne funkcije pomeni večjo podobnost, večja vrednost pa manjšo podobnost med deloma posnetka. Cenilno funkcijo uporabimo v DTW (glej poglavje 2.6.7) za izračun cene med deloma posnetka z DTW definirane kot:

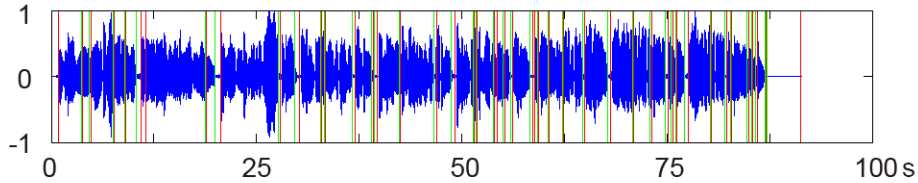
$$c_p(p_1, p_2) = \sum_{l=1}^{L} c(p_1(l), p_2(l)) \tag{3.9}$$



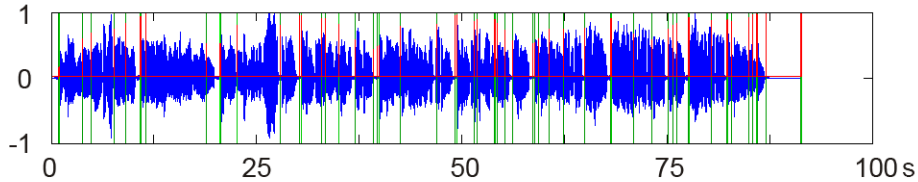

(a) Iskanje delov posnetkov z nizko energijo signala




(b) Iskanje delov posnetkov z nizko glasnostjo

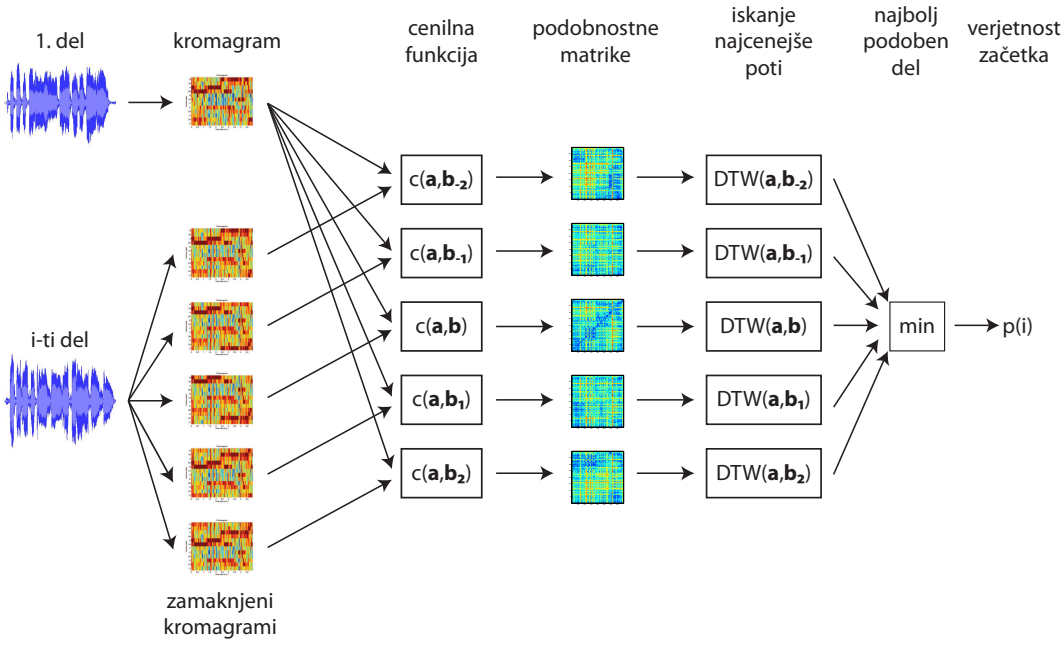



(c) Iskanje delov posnetkov brez zaznavne prevladujoče frekvence




(d) Združeni vsi kandidatni začetkov kitic

Slika 3.2: Iskanje vokalnih pavz




Slika 3.3: Prikaz iskanja kandidatov za začetke kitic.

kjer sta $p_1$ in $p_2$ dela posnetkov, ki ju med seboj primerjamo. $p_1(l)$ in $p_2(l)$ sta ustrezna kromatska vektorja (v predhodnih enačbah **a** in **b**), indeks $l$ pa zavzema vrednosti od začetnega (1) do končnega kromatskega vektorja ($L$) delov posnetka.

3. Metodo DTW uporabimo za izračun cene med deloma posnetka kot:

$$c_{min}(d_j) = DTW(d_0, d_j) = min\{c_p(d_0, d_j)\}, \tag{3.10}$$

kjer je $c_{min}$ minimalna cena (razdalja) med deloma posnetka $d_0$ in $d_j$.

4. Poleg neposredne primerjave delov posnetka, primerjamo še dele posnetka z zamaknjenimi kromatskimi vektorji za polton in cel ton navzdol in navzgor (kar se odraža v ciklični rotaciji 12 dimenzionalnih kromatskih vektorjev oz. kromagramov) ter izberemo tisto ceno izračunano z DTW, ki je najmanjša od izračunanih, kar lahko predstavimo kot:

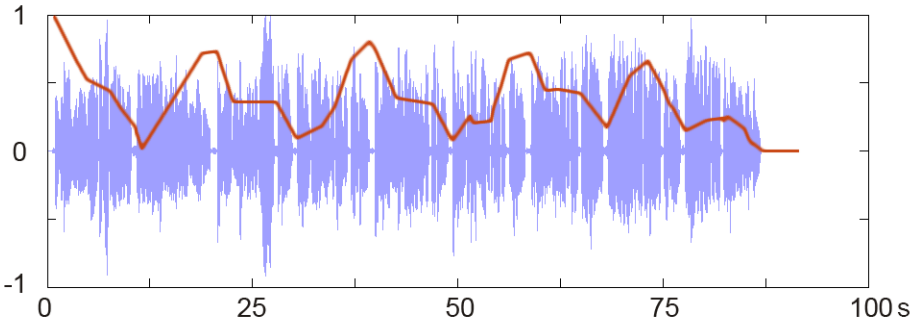
$$\begin{aligned} dist_{min}(d_0) &= 0, \\ dist_{min}(d_j) &= \min_{d_j^f, f \in [-2,2]} \{c_{min}(d_0, d_j^f)\}, \end{aligned} \tag{3.11}$$

kjer, $d_j^f$ predstavlja rotacijo kromatskih vektorjev izbranega dela posnetka od enega tona navzdol do enega tona navzgor, s koraki po polovico tona.

Predpostavimo, da so verjetnosti resničnih začetkov kitic določene s postopkom opisanim v nadaljevanju. Verjetnost, da je nek kandidat za začetek $k_i$ resnični začetek kitice definiramo kot funkcijo:

$$p(i) = \begin{cases} 0, & d_j \notin D \\ 1 - (\frac{dist_{min}(d_j)}{\max_j dist_{min}(d_j)})^2, & d_j \in D \end{cases} . \tag{3.12}$$

Primer takšne funkcije je prikazan v sliki 3.3.3, kjer je z modro barvo prikazan zvočni signal, z rdečo pa funkcija verjetnosti začetkov kitic definirana z enačbo 3.12.




Slika 3.4: Funkcija verjetnosti, da so kandidati za začetke kitic tudi resnični začetki.

### 3.3.4 Izbira dejanskih začetkov

Problem kako izbrati dejanske začetke izmed kandidatov je rešljiv na več načinov. V našem pristopu smo uporabili metodo t.i. *iskanja vrhov* v kombinaciji z uporabo pragu. Vrhove v signalu lahko opišemo kot:

$$v(i) = \begin{cases} p(i), & p(i-1) < p(i) \geq p(i+1) \\ 0, & \text{sicer} \end{cases} , \tag{3.13}$$

kjer je $v(i)$ vrednost vrha verjetnostne funkcije na $i$-tem mestu in $p(i-1), p(i), p(i+1)$ zaporedne vrednosti funkcije verjetnosti.

V tako predstavljeni funkciji verjetnosti, vrhovi predstavljajo mesta z lokalno največjo verjetnostjo začetka kitice. Ker iščemo vrhove glede na lokalno okolico funkcije, so lahko nekateri vrhovi na globalnem nivoju celotne pesmi "nizki". V takšnih vrhovih je verjetnost začetka kitice v primerjavi z ostalimi vrhovi nizka. Za izbiro najboljših vrhov uporabimo prag na nivoju posamezne pesmi. Prag predstavlja srednjo vrednost verjetnostne funkcije $v(i)$:

$$v^*(i) = \begin{cases} v(i), & v(i) > v_m \\ 0, & \text{sicer} \end{cases} , \tag{3.14}$$

kjer $v_m$ predstavlja srednjo vrednost verjetnostne funkcije $v$. Izbrani vrhovi predstavljajo dejanske začetke kitic, izračunane s predstavljeno metodo.a

## 3.4 Izboljšava visokonivojske segmentacije z nizkonivojskim pristopom

Pristopa segmentacije na posameznem nivoju sta med seboj odvisna samo v smeri od segmentacije na višjem nivoju k segmentaciji na nižjem nivoju, saj rezultate segmentacije na višjem nivoju neposredno uporabimo kot vhod za segmentacijo na nižjem nivoju. Rezultate pridobljene s segmentacijo nižjega nivoja lahko posredno ali neposredno uporabimo za izboljšavo segmentacije višjega nivoja. Za povečanje natančnosti visokonivojske segmentacije smo vpeljali tri izboljšave opisane v nadaljevanju. Prva izboljšava izboljša natančnost ocene začetka pesmi, druga izboljšava poišče morebitne zamike začetkov pesmi, tretja pa poišče morebitne zgrešene enote pri višjenivojski segmentaciji v primeru zaporednih enot solo in večglasnega petja.

### 3.4.1 Izboljšana natančnost ocene začetka pesmi

Prva izboljšava uporabi čas prve zaznane kitice kot boljšo oceno časa začetka celotne pesmi iz višjega nivoja. To lahko opišemo kot:

$$S_i^1 = \begin{cases} S_i, & pri\ S_i = s_{i,0} \\ s_{i,0}, & sicer \end{cases}, \tag{3.15}$$

kjer je $S_i$ začetek $i$-tega dela segmentacije na višjem nivoju in $s_{i,0}$ začetek prve kitice $i$-tega dela izračunan na nižjem nivoju segmentacije. S tem izboljšamo natančnost ocene začetka pesmi s pomočjo najdenih vokalnih pavz.

### 3.4.2 Odprava morebitnega zamika začetka pesmi

Druga izboljšava pride do izraza kadar segmentacija višjega nivoja zgreši začetek pesmi (meja izračunana s segmentacijo na višjem nivoju je postavljena naprej od dejanskega začetka pesmi). Postopek delovanja metode je sledeč:

1. Predpostavimo, da trenutni izračunan začetek $S_i^1$, pridobljen po prvi izboljšavi, ni pravilen.

2. Kot nov začetek vzamemo začetek $S_i^2$ zamaknjen za 5 sekund pred začetek $S_i^1$ in ponovno izračunamo segmentacijo nižjega nivoja. 5 sekundni zamik smo izbrali, ker to predstavlja približno četrtino povprečne dolžine kitice v eksperimentalni zbirki pesmi.

3. V kolikor je najdenih več ponavljajočih delov kot pri začetku $S_i^1$ smo ugotovili, da je bil začetek napačno postavljen. Ponavljamo postopek, dokler se število

najdenih ponavljajočih delov povečuje. Ko se število delov ne povečuje več smo našli bolje ocenjen začetek $S_i^2$.

Tudi s to izboljšavo povečamo natančnost ocene začetka pesmi. Ta izboljšava odpravlja napake večjega velikostnega razreda kot predhodna izboljšava. Izboljšava ima doseg izven mej segmentov izračunanih z višjenivojsko segmentacijo, česar prejšnja izboljšava nima.

# Poglavje 4

# Rezultati eksperimentov

Poglavje vključuje predstavitev rezultatov eksperimentov z uporabo metode razvite v okviru magistrske naloge in predstavljene v predhodnem poglavju. Metodo smo preizkusili na zbirki terenskih posnetkov slovenske ljudske glasbe iz glasbenega arhiva EtnoMuza. Podana je tudi primerjava z nekaterimi drugimi predstavljenimi metodami.

## 4.1 Priprava posnetkov

Za potrebe testiranja razvite metode smo potrebovali množico ustrezno pripravljenih posnetkov. Podatki v našem primeru predstavljajo digitalne zvočne posnetke vzorčene z 11 kHz. Iz arhiva EtnoMuza smo izbrali 10 terenskih posnetkov dolgih med 45 in 60 minut, ki skupaj predstavljajo 463 minut zvočnega zapisa. Za ovrednotenje rezultatov smo v vseh posnetkih ročno označili segmentacijo na posameznem nivoju. Kot smo predstavili že v predhodnih poglavjih je posamezni terenski posnetek sestavljen iz različnih tipov enot. V Tabeli 4.1 je predstavljena statistika sestave terenskih posnetkov.

Tabela 4.1: Tabela statistike tipov enot v terenskih posnetkih.

| **tip zvoka** | **število enot** | $\overline{\mathbf{u}}$ | $\sum_{\mathbf{i}} \mathbf{u_i}$ |
|---|---|---|---|
| govor | 85 | 1,09 | 92,95 |
| solo petje | 99 | 1,46 | 144,35 |
| večglasno petje | 36 | 1,27 | 45.57 |
| inštrumentalna glasba | 56 | 2,44 | 136,73 |

Stolpec $\overline{u}$ predstavlja povprečno dolžino enote v terenskih posnetkih, stolpec $\sum_i u_i$ v tabeli pa predstavlja skupno trajanje enot v minutah. Pri segmentaciji drugega

nivoja obravnavamo zgolj enote solo in večglasnega petja. V enotah višjega nivoja segmentacije na nižjem nivoju iščemo ponavljajoče dele (kitice). Statistika enot solo in večglasnega petja je predstavljena v Tabeli 4.2.

Tabela 4.2: Tabela statistike ponavljajočih delov v enotah solo in večglasnega petja.

| **tip zvoka** | **število delov** | $\mathbf{c_p}$ | $\overline{\mathbf{p_i}}$ |
|---|---|---|---|
| solo petje | 345 | 4,06 | 16,17 |
| večglasno petje | 393 | 3,97 | 22,07 |

Stolpec $c_p$ predstavlja povprečno število delov na enoto za posamezni tip zvoka, stolpec $\overline{p_i}$ pa predstavlja povprečno trajanje posameznega ponavljajočega dela.

## 4.2 Segmentacija višjega nivoja

Podrobna evalvacija metode višjenivojske segmentacije je predstavljena v delu [30], kjer je metoda tudi opisana. Del zbirke pesmi, na kateri je bila metoda testirana smo uporabili tudi za testiranje in evalvacijo metode nižjenivojske segmentacije. Rezultati višjenivojske segmentacije na tem delu imajo priklic $0,70$, natančnost $0,73$ in mero F 0.71 pri določanju mej med enotami s 5 sekundno natančnostjo zaznave.

## 4.3 Rezultati segmentacije nižjega nivoja

Eksperimente za segmentacijo nižjega nivoja smo izvedli na isti zbirki terenskih posnetkov kot eksperimente višjenivojske segmentacije. Meje pridobljene z višjenivojsko segmentacijo smo uporabili kot vhod v metodo nižjenivojske segmentacije. Za uspešno delovanje metode, smo najprej izbrali manjšo množico pesmi, s pomočjo katere smo izračunali eksperimentalno določene parametre. Eksperimentalno določeni parametri so naslednji:

$\xi_1$ ... določa prag tišine v zvočnem posnetku;

$\xi_2$ ... določa večkratnik povprečne dolžine vokalne pavze;

$\xi_3$ ... določa število poltonov razlike od srednje tonske višine zvočnega posnetka. Dele posnetka z večjo razliko označimo za vokalne pavze;

$\xi_4$ ... določa minimalno dolžino posnetka brez zaznavne tonske višine, ki jo še upoštevamo za vokalno pavzo;

$\xi_5$ ... določa minimalno dolžino zvočnega posnetka, ki še določa kitico;

$\xi_6$ ... določa prag jakosti zvoka, ki ga ne upoštevamo kot vokalno pavzo;

$\xi_7$ ... določa minimalno dolžino zvočnega posnetka, ki še določa kitico;

$\xi_8$ ... določa dolžino delov posnetka z začetki v zaznanih začetkih kitic, ki jih med seboj primerjamo.

Vsi eksperimentalno določeni parametri in njihova uporaba so predstavljeni v poglavju 3.3.2. V nadaljevanju uporabljamo $\delta$ kot dopustno natančnost zaznavanja začetka kitice. Vrednost $\delta$ predstavlja koliko sekund zamika še predstavlja pravilno najden začetek kitice.

### 4.3.1 Določanje optimalnih parametrov

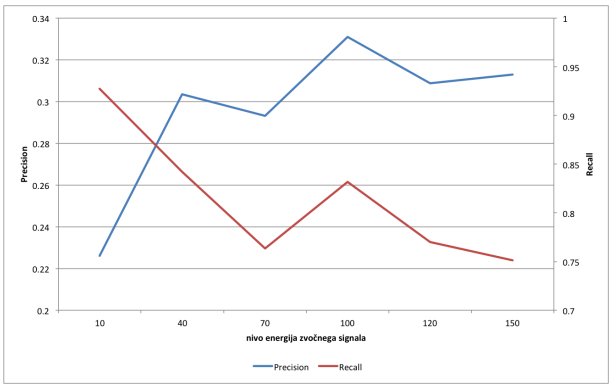
#### Določanje parametrov $\xi_1$ in $\xi_2$

Parametra sta uporabljena pri prvem izmed treh postopkov iskanja vokalnih pavz v zvočnem posnetku. Njuno vrednost smo določili z uporabo izbranega postopka za iskanje vokalnih pavz. Pri tem smo za izračun rezultatov uporabili samo prvi korak predstavljene metode. Za izbrano zbirko posnetkov smo najprej določili parameter $\xi_2$. Za parameter $\xi_2$ smo preizkusili različne vrednosti, ki odstranijo prekratke vokalne pavze. Izbira vrednosti parametra in metode za njegov izračun ne vpliva bistveno na končno natančnost delovanja metode. Razlika pa je v hitrosti izračuna začetkov končne metode, saj z višjim pragom izločimo več kandidatov za začetke kitice in s tem zmanjšamo število računanj podobnosti med deli pesmi. S pomočjo parametra $\xi_8 = 0.7$ odstranimo 15% vokalnih pavz, kar predstavlja skoraj enak odstotek pohitritve izračuna drugega dela metode.

V naslednjem koraku smo pri izbrani vrednosti $\xi_2$, izbrali vrednost parametra $\xi_1$. V poglavju 6.1 se v Tabeli 6.1 nahajajo vrednosti parametra in vrednosti posameznih mer za različno dopuščeno natančnost $\delta$. Tako iz tabele kot iz grafa v Sliki 4.1 lahko razberemo, da nižja vrednost parametra $\xi_1$ poveča priklic in zmanjša natančnost, zaradi povečanega števila najdenih vokalnih pavz. Izračun ni zelo občutljiv na izbiro parametra zato smo izbrali vrednost $\xi_1 = 100$, ki je dober kompromis med natančnostjo in priklicem.

#### Določanje parametrov $\xi_3$, $\xi_4$ in $\xi_5$

Za postopek zaznavanja delov posnetka brez zaznavne prevladujoče osnovne frekvence ($F0$) predstavljen v poglavju 3.3.2 smo eksperimentalno določili parametre $\xi_3$, $\xi_4$ in $\xi_5$. Za izračun rezultatov smo uporabili samo prvi korak metode. Najprej smo določili parameter $\xi_3$, ki določa koliko poltonov od srednje frekvence v zvočnem posnetku obravnavamo kot zaznavno frekvenco. Vrednosti parametra s posameznimi merami so predstavljeni v Tabeli 4.3. Parameter smo določili na majhni




Slika 4.1: Primerjava mer precision in recall za pristop iskanja vokalnih pavz na podlagi energije signala pri meri natančnosti $\delta = 2$ sekundi, pri izbiri parametra $\xi_1$.

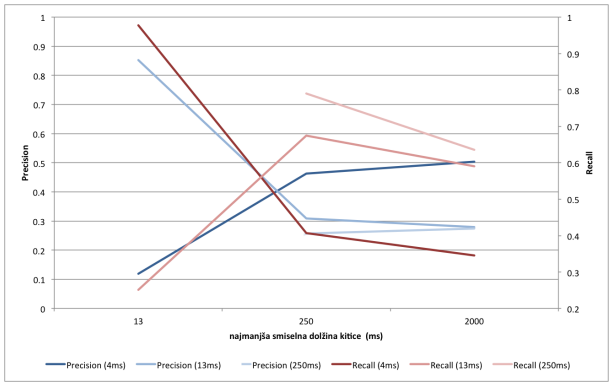
zbirki zvočnih posnetkov, kjer smo ročno označili dele brez zaznavne tišine. Vrednost parametra vpliva na število najdenih vokalnih pavz. V kolikor je vrednost parametra nizka ($\xi_3 = 10$), s čimer je dopusten razpon odstopanja tonske višine manjši, najdemo veliko vokalnih pavz, kar se odraža v višji vrednosti natančnosti in nizki vrednosti priklica. Pri visoki vrednosti parametra ($\xi_3 = 30$), kar pomeni večji dopusten razpon odstopanja tonske višine, najdemo veliko vokalnih pavz in je s tem posledično vrednost priklica visoka. Ker želimo imeti čim bolj uravnovešeni meri priklica in natančnosti izberemo parameter med omenjenima vrednostima $\xi_3 = 20$ kar se odraža z večjo vrednostjo mere F. V naslednjem koraku smo določili še parametra $\xi_4$

Tabela 4.3: Nabor vrednosti parametra $\xi_3$ in vrednosti posameznih mer.

| $\xi_3$ | Precision | Recall | F-Measure |
|---|---|---|---|
| 10 | 0,5426 | **0,6241** | 0,5805 |
| 20 | 0,7261 | 0,6184 | **0,6679** |
| 30 | **0,8122** | 0,4861 | 0,6082 |

in $\xi_5$. Vrednost $\xi_4$ določa prag za minimalno dolžino delov brez zaznavne frekvence. Dele krajše od praga ignoriramo. Parameter $\xi_5$ določa minimalno dolžino kitice. Deli krajši od $\xi_5$ tako predstavljajo vokalne pavze. Za pare parametrov smo izvedli eksperiment na manjši množici zvočnih posnetkov za različno dopuščeno natančnost. Rezultati so predstavljeni v Tabeli 6.2 v poglavju 6.1. V tabeli so označene vrstice z najboljšimi vrednostmi mere F. Vrstice obarvane s sivo barvo so končne izbrane vrednosti parametrov. Z grafom v Sliki 4.2 so prikazane vrednosti mer natančnosti in priklica glede na podan nabor parametrov. V tabeli in na grafu manjkajo vrednosti nesmiselne kombinacije parametrov, kjer bi bila dolžina smiselne kitice krajša od

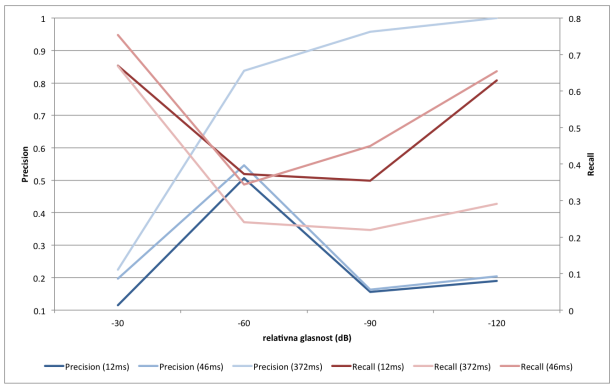
minimalne dolžine vokalne pavze. Iz grafa in slike je razvidno da tako priklic kot natančnost pri nizkih vrednostih parametra $\xi_5$ precej varirajo, pri večjih vrednostih pa se mere ustalijo. Zaradi tega smo izbrali vrednost parametra $\xi_4 = 2$ sekundi in vrednost parametra $\xi_5 = 250$ ms, ki predstavljata najboljše razmerje med priklicem in natančnostjo.




Slika 4.2: Primerjava mer precision/recall za pristop iskanja vokalnih pavz na podlagi zaznavnosti tonske višine pri meri natančnosti $\delta = 2$ sekundi za posamezne vrednosti parametra $\xi_5$.

### Določanje parametrov $\xi_6$ in $\xi_7$

Pri postopku zaznavanja delov posnetka z detekcijo tišine sta eksperimentalno določena dva parametra $\xi_6$ in $\xi_7$. Rezultati so izračunani samo s prvim korakom metode. Parameter $\xi_6$ predstavlja prag tišine v dB. Deli posnetka z glasnostjo pod parametrom $\xi_6$ so deli s tišino in jih povežemo v vokalne pavze. Parameter $\xi_7$ določa minimalno dolžino vokalne pavze. Dele posnetka, ki so krajši od $\xi_7$ in smo v njih zaznali tišino ignoriramo. Nabori parametrov so podani v Tabeli 6.3, ki se nahaja v poglavju 6.1. V tabeli so poudarjene maksimalne vrednosti posameznih mer za določeno stopnjo natančnosti $\delta$. S sivo barvo so obarvane vrstice, ki predstavljajo izbrane parametre, za uporabo v metodi nižjenivojske segmentacije. Izbrane vrednosti parametrov imajo najvišje vrednosti mere F. Z grafom prikazanim v Sliki 4.3 so prikazane vrednosti mer natančnosti in priklica glede na podan nabor parametrov. Iz grafa je razvidno da mere priklica in natančnosti za vrednosti parametra $\xi_6$ nižje od -90 in višje od -60 zelo varirajo. Vrednost parametra $\xi_7$ nima bistvenega vpliva na priklic in natančnost. Zaradi navedenih dejstev smo izbrali vrednosti parametrov $\xi_6 = -60$ in $\xi_7 = 128$, ki predstavljata najboljše razmerje med priklicem in natančnostjo.




Slika 4.3: Primerjava mer precision/recall za pristop iskanja vokalnih pavz na podlagi energije signala pri meri natančnosti $\delta = 2$ sekundi za posamezne vrednosti parametra $\xi_7$.

### Določanje parametra $\xi_8$

Pri postopku izračuna podobnosti med posameznimi deli posnetkov se medsebojno primerjajo deli posnetkov dolgi $\xi_8$. Vrednost parametra smo določili na manjši tesni množici pesmi, kjer smo izračunali mere za posamezne dolžine. V Tabeli 4.4 so predstavljene mere uspešnosti za posamezno vrednost parametra. Iz tabele je razvidno, da najboljše rezultate doseže ravno pri vrednosti 10 sekund. Prikazani rezultati veljajo za natančnost $\delta = 2$ sekundi. Iz tabele lahko razberemo, da vrednost mere F do vrednosti parametra $\xi_8 = 10$ narašča, pri višjih vrednostih pa zopet pada. Zaradi navedene ugotovitve smo izbrali za vrednost parametra $\xi_8 = 10$ sekund, kar predstavlja skoraj točno polovico povprečne dolžine kitice v naši zbirki.

Tabela 4.4: Primerjava metod z različno dolgimi primerjanimi odseki $\xi_8$ za mero natančnosti $\delta = 2$.

| $\xi_8$ | Precision | Recall | F-Measure |
|---|---|---|---|
| 5 | 0,6049 | 0,5758 | 0,5900 |
| 7 | 0,6472 | 0,5918 | 0,6182 |
| 10 | 0,6773 | 0,6435 | **0,6600** |
| 12 | 0,6797 | 0,6203 | 0,6486 |
| 15 | 0,6667 | 0,5954 | 0,6290 |
| 20 | 0,6834 | 0,5811 | 0,6281 |

### 4.3.2 Ocena prvega koraka metode

V nadaljevanju je podana primerjava pristopov prvega koraka metode. Izbor parametrov vsakega posameznega pristopa je podan v predhodnih poglavjih. Primerjava med pristopi je opravljena za mero natančnosti $\delta = 2$ sekundi. V Tabeli 4.5. so predstavljeni naslednji pristopi za izračun prvega koraka metode:

**max Energy** ... metoda uporablja za zaznavanje začetkov kitic zgolj pristop zaznavanja delov signala z nizko energijo;

**max Amplitude** ... metoda uporablja za zaznavanje začetkov kitic zgolj pristop zaznavanja delov zvočnega posnetka z nizko amplitudo;

**max NoPitch** ... metoda uporablja za zaznavanje začetkov kitic zgolj pristop zaznavanja delov zvočnega posnetka brez zaznavne prevladujoče frekvence ($F0$);

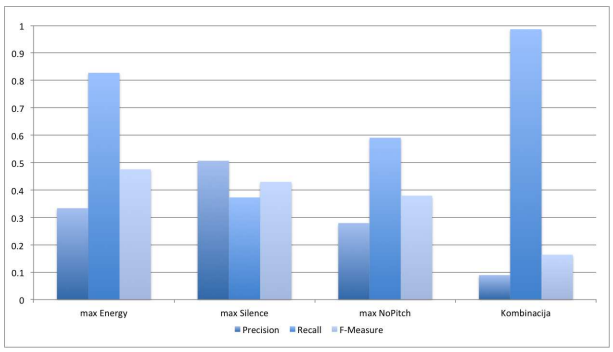
**kombinacija** ... metoda uporablja pristope nizkonivojske segmentacije z maksimalnimi vrednostmi mere F.

Med predstavljenimi pristopi se najbolje obnese *max Energy*, ki ima najvišjo vrednost mere F. Ker pa nas v drugem koraku zanima to, da najdemo čim več vokalnih pavz, smo uporabili kombiniran pristop, ki uporablja parametre posameznih metod za katere je mera F zavzela maksimalne vrednosti. V drugem koraku metode bomo izločili čim več napačno zaznanih začetkov kitic. Iz tabele je razvidno, da se kombiniran pristop za izračun prvega koraka metode obnese bolje, kot zgolj posamezni pristopi za izračun prvega koraka metode. Kombinirani pristop najde veliko več vokalnih pavz, kar se odraža v visokem priklicu, za več kot 0,15 višja vrednost od najvišje vrednosti priklica posameznega pristopa.

Tabela 4.5: Primerjava metod po meri F za mero natančnosti $\delta = 2$.

| Pristop | $\xi_1$ | $\xi_4$ | $\xi_5$ | $\xi_6$ | $\xi_7$ | Precision | Recall | F-Measure |
|---|---|---|---|---|---|---|---|---|
| max Energy | 120 | - | - | - | - | 0,3336 | 0,8276 | 0,4755 |
| max Amplitude | - | - 60 | 128 | - | - | 0,5066 | 0,3729 | 0,4296 |
| max NoPitch | - | - | - | 32 | 256 | 0,2793 | 0,5908 | 0,3793 |
| kombinacija | 120 | - 60 | 128 | 32 | 256 | 0,0894 | **0,9866** | 0,1639 |

Primerjave med merami izračunanimi za posamezne pristope so predstavljene tudi z grafikonom v Sliki 4.4. Iz grafa je razvidno, da imajo vsi pristopi izrazit priklic. To še posebej velja za kombiniran pristop, ki pokrije preko 98% primerov. Iz takšnega pokritja primerov drugi korak metode na podlagi podobnosti delov posnetkov izračuna kateri najbolj ustrezajo pravim začetkom kitic.




Slika 4.4: Primerjava mer precision, recall in mere F med posameznimi metodami, tako enostavnimi, kot sestavljenimi. Mera natančnosti je $\delta = 2$ sekundi za posamezne vrednosti parametra.

### 4.3.3 Ocena delovanja celotne metode

Medsebojno smo testirali tudi delovanje celotne metode. V Tabeli 4.6 so predstavljeni rezultati delovanja metode z uporabo različnih pristopov za izračun prvega koraka. Rezultati so primerjani z delovanjem končne metode. Iz tabele je razvidno, da je delovanje končne metode boljše, če združimo pristope prvega koraka, kar se odraža v višji vrednosti mere F.

Tabela 4.6: Primerjava osnovne metode višjenivojske segmentacije in predstavljene izboljšave.

| Metoda | Precision | Recall | F-Measure |
|---|---|---|---|
| Energija | 0,7592 | 0,4430 | 0,5595 |
| Amplituda | 0,6886 | 0,2574 | 0,3747 |
| NoPitch | 0,7447 | 0,3597 | 0,3597 |
| Sestavljena | 0,6773 | 0,6435 | **0,6600** |

### 4.3.4 Primeri delovanja

#### Primeri dobrega delovanje metode

V nadaljevanju sta predstavljena dva primera dobrega delovanja metode za nižjenivojsko segmentacijo. Izbrana primera imata mero F vrednosti 0.95 in 1. Prvi primer je prikazan v Sliki 4.5(a). Metoda za izbran primer pravilno najde devet od desetih kitic v okviru natančnosti $\delta = 1$ sekunda. V primeru je metodi spodletelo zgolj pri začetku zadnje kitice. Razlog, da metoda ni odkrila tudi te kitice tiči v

statičnem pragu izbire vrhov verjetnostne funkcije začetkov kitic in neupoštevanja dolžine tipične najdene kitice pri iskanju ostalih. Vsi ostali začetki kitic so bili zaznani z visoko natančnostjo.

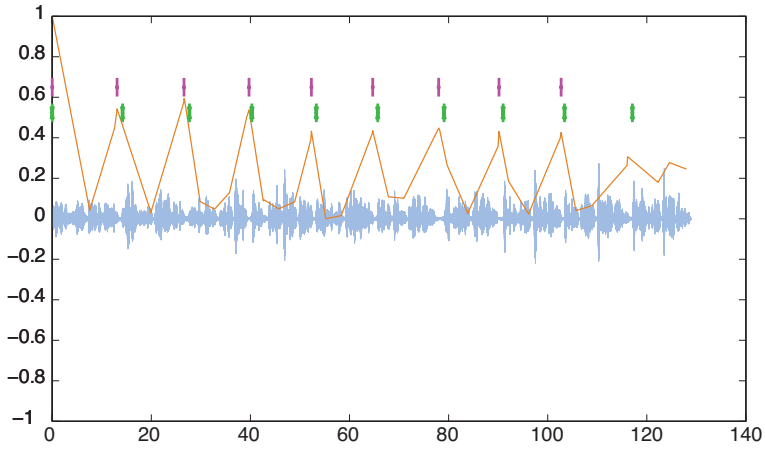
Drugi primer dobrega delovanja metode je prikazan v Sliki 4.5(b). Metoda je bila v tem primeru še uspešnejša, kar nam pove že vrednost mere F, ki je enaka 1. Metoda v pesmi uspešno odkrije začetke vseh šestih kitic znotraj natančnosti $\delta = 1$ sekunda.

### Primeri slabega delovanja metode

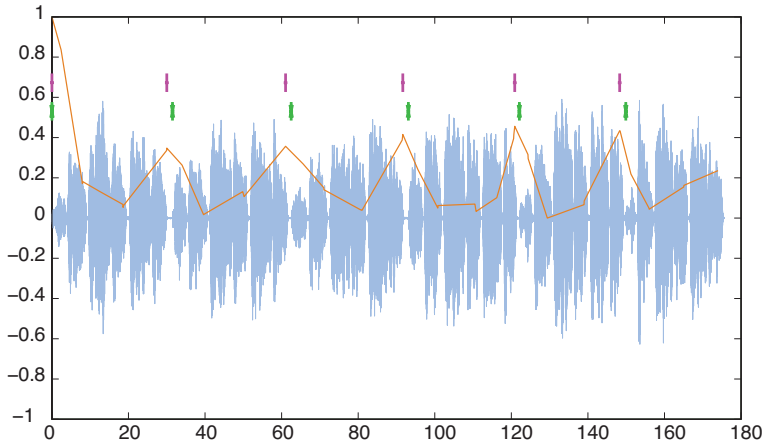
Predstavili bomo dva primera, kjer metoda odpove oz. deluje slabo. Primera sta prikazana v Sliki 4.6. Za vsakega izmed primerov velja, da metoda odpove zaradi različnih vzrokov.

V prvem primeru metoda odpove zaradi dveh vzrokov. Prvi vzrok je statični prag za zaznavanje začetkov kitic v verjetnosti funkciji, zaradi česar metoda "zgreši" začetek druge kitice. Izračunana verjetnost za začetek druge kitice je prenizka, da bi jo lahko zaznali z uporabljenim pristopom uporabe pragu. Problem zaznavanja začetka tretje kitice je drugačen. Tretja kitica je melodično drugače odpeta kot preostali dve, kar se odraža v kromagramih, ki ponazarjajo melodično predstavitev zvočnega signala. Podobnosti med posameznimi deli lažje razberemo iz podobnostnih matrik prikazanih v Sliki 4.7. Podobnostni matriki prikazujeta podobnost med kromagramom prve kitice in kromagramoma resničnega začetka in zaznanega začetka tretje kitice. Na podobnostnih matrikah so označene diagonale, ki pomenijo podobnost posameznih delov kromagrama. Razvidno je, da sta kromagrama delov iz začetka pesmi in iz zaznanega začetka tretje kitice bolj podobna kot kromagram iz začetka pesmi in pravega začetka tretje kitice. To je tudi vzrok, zakaj je v verjetnostni funkciji začetkov kitic vrh v zaznanem začetku tretje kitice in ne v pravem.

Razlog zakaj metoda odpove pri drugem primeru tiči povsem drugje. Pesem drugega primera je sestavljena iz dveh delov, ki sta melodično povsem različna. Poleg tega je prvi del sestavljen iz krajših melodično ponavljajočih delov kar naša metoda označi za začetke vidne v Sliki —4.6(b). Drugi del pesmi melodično ni podoben prvemu delu, a je prav tako sestavljen iz krajših ponavljajočih delov. Tako problem iskanja začetkov kitic ne tiči v sami metodi, saj je le-ta razvita za iskanje melodično ponavljajočih delov, za kar se tudi v tem primeru dobro izkaže. V tem primeru je problem v samih podatkih, saj za konkreten primer ne velja, da so kitice medsebojno melodično podobne. Metoda podobno kot v prejšnjem primeru začetke najde tudi na koncu pesmi. Razlog za to je enak kot pri prejšnjem primeru.




(a) Prvi primer dobrega delovanja metode. Pesem na sliki ima 10 kitic. Naša metoda je odkrila 9 kitic z visoko natančnostjo $\delta = 1$ sekunda.

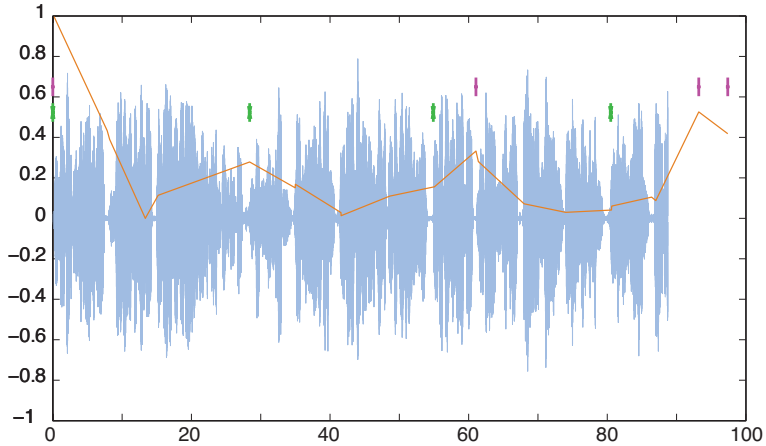



(b) Drugi primer dobrega delovanja metode predstavlja pesem sestavljeno iz 6 kitic. Naša metoda je odkrila vse kitice v visoko natančnostjo $\delta = 1$ sekunda.

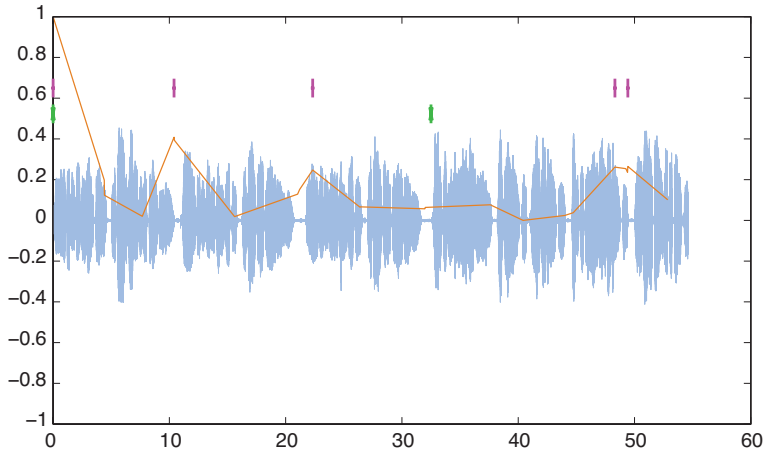
Slika 4.5: Sika prikazuje dva primera pesmi iz arhiva EtnoMuza na katerih deluje metoda nižjenivojske segmentacije zelo dobro. V slikah je z modro izrisan zvočni signal, z oranžno verjetnostna funkcija pojavitve nove kitice, izračunana s pomočjo metode nižjenivojske segmentacije, z zelenimi zvezdicami so označeni dejanski začetki kitic v pesmi določeni ročno, z roza plusi pa so označeni začetki zaznani z metodo opisano v prejšnjem poglavju.

### Povprečni primeri

V nadaljevanju sta predstavljena dva "povprečna" primera nizkonivojske segmentacije. Za oba primera velja, da je njuna mera F blizu povprečne vrednosti v celotni



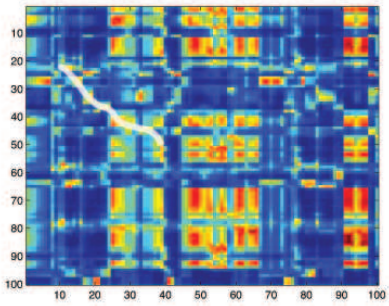

(a) Prvi primer slabega delovanja metode. Pesem ima 4 kitice, naša metoda pravilno zazna zgolj začetek prve kitice.



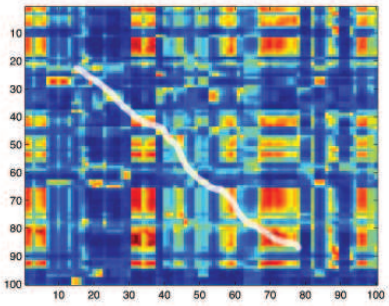

(b) Drugi primer slabega delovanja, kjer metoda ne zazna nobenega začetka kitice.

Slika 4.6: Sika prikazuje dva primera pesmi iz arhiva EtnoMuza na katerih deluje metoda nižjenivojske segmentacije zelo slabo.

testirani zbirki. Vrednosti mere F sta 0,667 za prvi primer in 0.571 za drugi primer. Prvi primer je prikazan v Sliki 4.8(a). Metoda tukaj uspešno zazna tri izmed petih začetkov kitic. Začetek druge kitice ni zaznan zaradi prenizke vrednosti verjetnostne funkcije začetkov. Enako velja tudi za začetek pete kitice, kjer je vrh v verjetnostni funkciji še dosti manj izrazit in s tem tem težje zaznaven. Edini neresnični začetek je zaznan na koncu pesmi. Na tem mestu zaznana vokalna pavza povzroča preglavice delu metode, ki računa podobnost. Pri računanju podobnosti med dvema deloma posnetka, del, ki je krajši od 10 sekund podaljšamo z ničlami, kar v pričujočem pri-




(a) Podobnostna matrika med 10 sekundnima deloma posneta z začetki v začetku pesmi in v pravem začetku tretje kitice.




(b) Podobnostna matrika med 10 sekundnima deloma posnetka z začetki v začetku pesmi in v zaznanem začetku tretje kitice

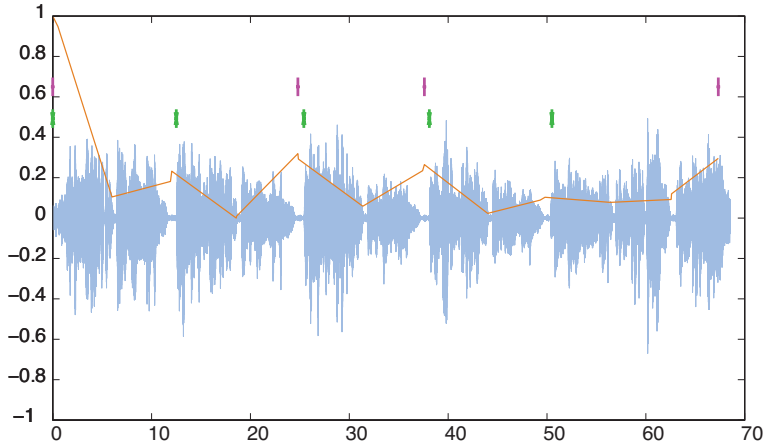
Slika 4.7: Slika prikazuje podobnostni matriki 10 sekundnih delov posnetka z začetki v začetku pesmi, pravem začetku tretje kitice in zaznanem začetku tretje kitice. V matrikah je z belo barvo označena podobnost med deli zvočnih posnetkov.

meru povzroča preglavice. Smiselno bi bilo realizirati drugačen način za računanje podobnosti med deli posnetkov v takšnem primeru.

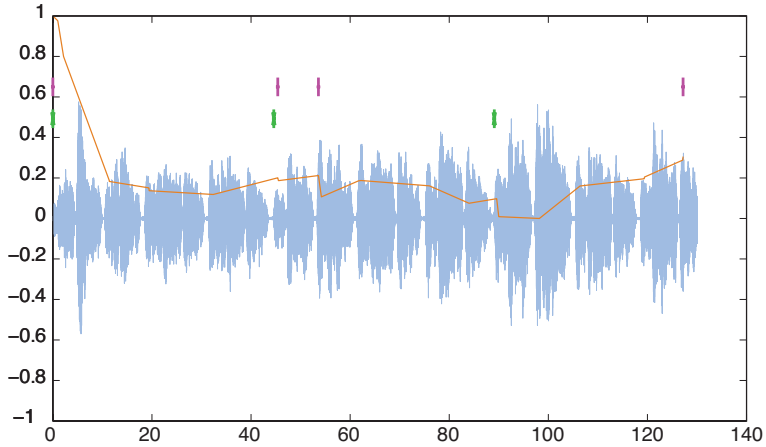
Drugi primer, prikazan v Sliki 4.8(b), je dosti neprijetnejši. V tem primeru je podobnost med poljubnimi deli zvočnega posnetka izredno nizka. Naša metoda uspešno zazna dva začetka kitic, pri enem odpove, napačno pa zazna prav tako dva napačno zaznana začetka. Sama funkcija verjetnosti ima zelo neizrazite vrhove, zato naša metoda zaradi naivnega iskanja vrhov odpove. Vrednosti vseh vrhov so zelo blizu povprečne vrednosti, nekateri pa padejo precej pod njo. Napačna detekcija zadnjega začetka je ekvivalentna napačni detekciji v prejšnjem primeru. Prva napačna detekcija, pa je zopet problem melodične podobnosti znotraj posamezne kitice.

## 4.4 Primerjava z obstoječimi metodami

Metodo nižjenivojske segmentacije smo primerjali z že obstoječimi metodami segmentacije zvočnih posnetkov posameznih pesmi v smiselne dele. Pristopa segmentacije sta predstavljena v poglavjih 2.7.1 in 2.7.2 in se precej razlikujeta od razvite metode. Prvi pristop išče dele posnetka zgolj na podlagi dihalnih pavz, s čimer zanemari pomen posameznega dobljenega dela. Kljub temu gre za segmentacijo na enakem nivoju in je zaradi tega smiselno primerjati rezultate metod. Drugi pristop se razlikuje v tem, da potrebuje za svoje delovanje predznanje eksperta. Metoda potrebuje simbolično predstavitev prototipne kitice, katere ponovitve nato v zvočnem posnetku išče. Tudi ta pristop ima omejeno uporabo, saj pri segmentaciji zvočnih posnetkov takšno predznanje ni na voljo. Vseeno gre za metodo segmentacije na istem nivoji, zaradi česar je prav tako smiselno medsebojno primerjati rezultate.




(a) Prvi primer "povprečnega" delovanja metode. Pesem ima 5 kitic, naša metoda pravilno zazna 3 začetke. Metoda napačno zazna en začetek.




(b) Drugi primer "povprečnega" delovanja metode. Metoda v pesmi zazna 3 pravilne začetke in 2 napačna začetka kitice.

Slika 4.8: Sika prikazuje dva primera pesmi iz arhiva EtnoMuza na katerih deluje metoda nižjenivojske segmentacije "povprečno" dobro.

Metode smo testirali na ločeni zbirki 21 zvočnih posnetkov solo petja. Posnetki obsegajo 62 minut petja, ki skupaj predstavljajo 229 kitic. Povprečna kitica v posnetkih je dolga 15,2 sekundi, posnetki pa imajo povprečno 11 kitic. Zvočni posnetki so del arhiva EtnoMuza.

V Tabeli 4.7 so predstavljeni rezultati naše metode, Kranenburgove metode [47, 46] predstavljene v poglavju 2.7.1 in Müellerjeve metode [32, 33] predstavljene v poglavju 2.7.2. Iz tabele je razvidno, da predstavljena metoda močno zaostaja za rezultati Müellerjeve metode, pri čemer glavni razlog tiči ravno v dejstvu, da naša

Tabela 4.7: Primerjava rezultatov delovanja naše, Müellerjeve in Kranenburgove metode segmentacije za mero natančnosti $\delta = 2$.

| metoda | Precision | Recall | F-Measure |
|---|---|---|---|
| Kranenburg | 0,149 | **0,930** | 0,257 |
| Mueller ($\Delta^{fluc}$) | 0,865 | 0,748 | **0,802** |
| naša metoda | 0,442 | 0,646 | **0,525** |

metoda ne uporablja nikakršnega predznanja za izvedbo segmentacije, Müellerjev pristop pa kot vhod potrebuje predlogo ene ponavljajoče kitice v simbolični obliki. Razvidno je tudi, da predstavljena in Müellerjeva metoda delujeta veliko bolje pri iskanju smiselnih delov, kot Kranenburgova metoda. Potrebno je seveda povedati, da Kranenburgova metoda v osnovi ni namenjena iskanju smiselnih delov ampak zgolj iskanju vokalnih pavz. Skoraj identični pristop je v naši metodi uporabljen kot eden izmed pristopov prvega koraka za iskanje vokalnih pavz, predstavljen v poglavju 3.3.2.

## 4.5 Rezultati izboljšav višjenivojske segmentacije

### 4.5.1 Izboljšana detekcija začetka pesmi

V nadaljevanju so predstavljeni rezultati izboljšane detekcije začetkov pesmi glede na višjenivojsko segmentacijo. Izboljšava uporabi konec vokalne pavze na začetku posnetka, v kolikor tam je vokalna pavza, za izboljšanje določitve začetka enote solo ali večglasnega petja. Rezultati izboljšane detekcije začetkov pesmi so predstavljeni v Tabeli 4.8. Iz tabele je razvidno, da v kolikor pri pesmih v zbirki popravimo začetke s pomočjo nižjenivojske segmentacije, izboljšamo rezultate višjenivojske segmentacije za skoraj 1%. Tukaj velja omeniti, da smo začetke popravljali zgolj na 135 izmed 276 enot. Toliko je v zbirki enot, ki predstavljajo solo ali večglasno petje. Pri enotah ostalega tipa nismo uporabili nižjenivojske metode segmentacije in tako nismo izboljševali višjenivojske zaznave začetkov kitic.

Tabela 4.8: Primerjava osnovne metode višjenivojske segmentacije in predstavljene izboljšave.

| Metoda | Precision | Recall | F-Measure |
|---|---|---|---|
| Marolt | 0,7023 | 0,7302 | 0,7160 |
| Izboljšava | 0,7132 | 0,7340 | 0,7235 |

# Poglavje 5

# Zaključek

Zanimanje za raziskave na področju ljudske glasbe se zadnja leta povečuje zaradi čedalje večjega števila digitaliziranih zbirk ljudske glasbe. S tem privlači tako raziskovalce s področja etnomuzikologije kot tudi s področja MIR in vzpodbuja sodelovanje med raziskovalnima področjema, kot tudi razvoj računalniško podprtih metod za proučevanje glasbe. Nenazadnje se to odraža s pojavitvijo več javno dostopnih zbirk ljudske glasbe, ki omogočajo vsakemu posamezniku vpogled v kulturno dediščino glasbe iz določenega območja. Čeprav so metode razvite za proučevanje ljudske glasbe večinoma specializirane zgolj za ljudsko glasbo, so še vedno zanimive tudi za uporabo pri delu s komercialno glasbo, kjer industrija prav tako ogromno vlaga v razvoj novih metod.

V okviru magistrske naloge smo razvili, preskusili in ovrednotili metodo za večnivojsko segmentacijo terenskih posnetkov ljudske glasbe. Metoda nek terenski posnetek na posameznem nivoju razdeli na vsebinsko smiselne dele pri čemer se zgleduje po že razvitih metodah predstavljenih v pregledu področja, uvaja pa tudi lastne ideje in pristope k reševanju problema. Metoda deluje v treh korakih:

- prvi korak zajema segmentacijo terenskih posnetkov na višjem nivoju z obstoječo metodo [30] za iskanje vsebinsko smiselnih enot (*govor*, *solo petje*, *zborovsko petje*, *inštrumentalno glasbo*, *pritrkavanje zvonov*) v zvočnem posnetku in določanje mej posameznih posnetkov;

- drugi korak zajema nadaljnjo segmentacijo petih enot (*solo petje*, *zborovsko petje*) na krajše dele s ponavljajočo melodijo, pri čemer uporablja značilnice pridobljene na kratkih zvočnih izsekih;

- tretji del zajema izboljšanje segmentacije višjega nivoja s pomočjo informacij pridobljenih s segmentacijo na nižjem nivoju, kar se odraža z natančnejšo določitvijo mej teh enot.

Predstavljena metoda je bila preizkušena na zbirki terenskih posnetkov slovenske ljudske glasbe iz arhiva EtnoMuza. V nalogi so predstavljeni in ovrednoteni rezultati

razvite metode. Prav tako smo rezultate metode na posameznem nivoju primerjali z obstoječimi pristopi, kjer se je razvita metoda odrezala bolje od nekaterih predstavljenih.

## 5.1 Sklepne ugotovitve

Magistrska naloga in z njo povezano delo nas je privedlo do sledečih sklepnih ugotovitev:

- predstavljena metoda je prva, ki združuje segmentacijo terenskih posnetkov ljudske glasbe na dveh nivojih;
- metoda v pesmih zazna začetke kitic z uspešnostjo 68% brez dodatnega predznanja ekspertov;
- metoda uporabi rezultate nizkonivojske segmentacije za izboljšanje segmentacije na višjem nivoju;
- v primerih, ko kitice med seboj melodično niso podobne ali podobnost med njimi ne presega določene stopnje, predstavljena metoda odpove;
- delovanje metode ni povsem zanesljivo in natančno, a je dovolj dobro za podporo dela etnomuzikologov.

## 5.2 Nadaljnje delo

Na podlagi sklepnih ugotovitev lahko podamo naslednje smernice za nadaljnje delo:

- izboljšava oz. prilagoditev metode za boljše delovanje segmentacije nižjega nivoja na delih z zborovsko in inštrumentalno glasbo;
- razvito metodo preizkusiti na veliko večji zbirki terenskih posnetkov iz arhiva EtnoMuza;
- analiza in uporaba značilnic za iskanje ritmičnih in barvnih podobnostih med kiticami;
- razvito metodo preizkusiti na terenskih posnetkih ljudske glasbe drugih arhivov, ki imajo podoben način hranjenja in evalvacija rezultatov metode na takšni zbirki;
- metodo prilagoditi za uporabo z različnim predznanjem (na primer z znanjem o dolžini posameznih segmentov na nižjem nivoju);
- vključitev razvite metode v uporabniško opremo, ki jo pri svojem delu uporabljajo etnomuzikologi;

- pohitriti delovanje metode za uporabo v realnem času.

## 5.3 Zaključna misel

Magistrsko delo predstavlja izziv prav vsakemu inženirju ter ga sooči z raziskovalnim delom v dosti globljem pomenu kot v okviru diplomskega dela. Soočenje z izzivi lahko predstavlja zanimivo delo iz interesnega področja kot tudi sestavljanje in vrednotenje rezultatov, ki jim ni videti konca. Kljub vsemu je vložen trud in vztrajnost poplačana z občutkom, o doprinosu novega znanja skupnosti, povezani z delom iz okvira magistrske naloge.

# Poglavje 6

# Dodatki

## 6.1 Dodatek A

Tabela 6.1: Nabor vrednosti parametra $\xi_1$ in vrednosti posameznih mer.

| $\xi_1$ | $\delta$ | Precision | Recall | F-Measure |
|---|---|---|---|---|
| 10 | 1 | 0,1964 | **0,8258** | 0,3173 |
| | 2 | 0,2261 | **0,9273** | 0,3635 |
| | 5 | 0,2992 | **0,9712** | 0,4575 |
| 40 | 1 | 0,2730 | 0,7641 | 0,4023 |
| | 2 | 0,3036 | 0,8423 | 0,4463 |
| | 5 | 0,3690 | 0,8929 | 0,5222 |
| 70 | 1 | 0,2932 | 0,7634 | 0,4237 |
| | 2 | 0,3209 | 0,8310 | 0,4630 |
| | 5 | 0,3842 | 0,8740 | 0,5338 |
| 100 | 1 | 0,3054 | 0,7722 | 0,4377 |
| | 2 | 0,3310 | 0,8320 | 0,4736 |
| | 5 | 0,3955 | 0,8837 | **0,5465** |
| 120 | 1 | 0,3088 | 0,7701 | 0,4408 |
| | 2 | 0,3336 | 0,8276 | **0,4755** |
| | 5 | 0,3962 | 0,8801 | 0,5464 |
| 150 | 1 | **0,3129** | 0,7513 | **0,4418** |
| | 2 | **0,3361** | 0,8013 | 0,4736 |
| | 5 | **0,3992** | 0,8631 | 0,5459 |

Tabela 6.2: Nabor vrednosti parametrov $\xi_4$ in $\xi_5$ in vrednosti posameznih mer.

| $\xi_4$ | $\xi_5$ | $\delta$ | Precision | Recall | F-Measure |
|---|---|---|---|---|---|
| 4 | 13 | 1 | 0.1049 | **0.9350** | 0.1886 |
| | | 2 | 0.1191 | **0.9770** | 0.2124 |
| | | 5 | 0.1661 | **0.9905** | 0.2845 |
| | 250 | 1 | 0.4277 | 0.3767 | **0.4006** |
| | | 2 | 0.4622 | 0.4065 | **0.4326** |
| | | 5 | 0.5056 | 0.4295 | 0.4645 |
| | 2000 | 1 | 0.4714 | 0.3238 | 0.3839 |
| | | 2 | 0.5040 | 0.3455 | 0.4100 |
| | | 5 | 0.5711 | 0.3753 | 0.453 |
| 13 | 13 | 1 | **0.8288** | 0.2493 | 0.3833 |
| | | 2 | **0.8525** | 0.2507 | 0.3874 |
| | | 5 | **0.9014** | 0.2602 | 0.4038 |
| | 250 | 1 | 0.2612 | 0.5759 | 0.3594 |
| | | 2 | 0.3089 | 0.6748 | 0.4238 |
| | | 5 | 0.3982 | 0.7710 | **0.5251** |
| | 2000 | 1 | 0.2354 | 0.5014 | 0.3203 |
| | | 2 | 0.2793 | 0.5908 | 0.3793 |
| | | 5 | 0.3843 | 0.7019 | 0.4966 |
| 250 | 250 | 1 | 0.2115 | 0.6558 | 0.3199 |
| | | 2 | 0.2569 | 0.7900 | 0.3878 |
| | | 5 | 0.3536 | 0.9309 | 0.5125 |
| | 2000 | 1 | 0.2258 | 0.5271 | 0.3161 |
| | | 2 | 0.2743 | 0.6355 | 0.3832 |
| | | 5 | 0.3757 | 0.7493 | 0.5005 |

Tabela 6.3: Nabor vrednosti parametrov $\xi_6$ in $\xi_7$ in vrednosti posameznih mer.

| $\xi_6$ | $\xi_7$ | $\delta$ | Precision | Recall | F-Measure |
|---|---|---|---|---|---|
| -30 | 128 | 1 | 0,0823 | 0,4946 | 0,1412 |
| | | 2 | 0,1147 | 0,6694 | 0,1958 |
| | | 5 | 0,1874 | **0,8713** | 0,3084 |
| | 512 | 1 | 0,1588 | **0,6198** | 0,2528 |
| | | 2 | 0,1975 | **0,7532** | 0,3130 |
| | | 5 | 0,2763 | 0,8580 | 0,4180 |
| | 4096 | 1 | 0,1853 | 0,5568 | 0,2780 |
| | | 2 | 0,2246 | 0,6685 | 0,3362 |
| | | 5 | 0,3162 | 0,7982 | **0,4530** |
| -60 | 128 | 1 | 0,4693 | 0,3511 | **0,4017** |
| | | 2 | 0,5066 | 0,3729 | **0,4296** |
| | | 5 | 0,5782 | 0,3850 | 0,4622 |
| | 512 | 1 | 0,5042 | 0,3200 | 0,3915 |
| | | 2 | 0,5466 | 0,3440 | 0,4223 |
| | | 5 | 0,5956 | 0,3573 | 0,4467 |
| | 4096 | 1 | **0,8378** | 0,2412 | 0,3746 |
| | | 2 | **0,8378** | 0,2412 | 0,3746 |
| | | 5 | **0,8533** | 0,2490 | 0,3855 |
| -90 | 128 | 1 | 0,1319 | 0,3045 | 0,1841 |
| | | 2 | 0,1566 | 0,3545 | 0,2173 |
| | | 5 | 0,2169 | 0,4318 | 0,2888 |
| | 512 | 1 | 0,1269 | 0,3517 | 0,1865 |
| | | 2 | 0,1631 | 0,4492 | 0,2393 |
| | | 5 | 0,2711 | 0,6525 | 0,3831 |
| -120 | 128 | 1 | 0,1406 | 0,4781 | 0,2173 |
| | | 2 | 0,1897 | 0,6296 | 0,2915 |
| | | 5 | 0,2806 | 0,7737 | 0,4119 |
| | 512 | 1 | 0,1419 | 0,4646 | 0,2174 |
| | | 2 | 0,2035 | 0,6544 | 0,3105 |
| | | 5 | 0,3087 | 0,8272 | 0,4496 |

# Literatura


[1] Australian Folk Songs, 1994. URL: http://folkstream.com/.

[2] The American Folk Song Collection, 2004. URL: kodaly.hnu.edu/home.cfm.

[3] L. Barrington, A. Chan, in G. Lanckriet. Modeling music as a dynamic texture. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):602–12, 2010.

[4] Mark A. Bartsch in Gregory H. Wakefield. To catch a chorus: Using chroma-based representations for audio thumbnailing. Objavljeno v *Applications of Signal Processing to Audio and Acoustics*, strani 15–19, New Platz, NY , USA, October 2001.

[5] J. C. Brown. Calculation of a constant Q spectral transform. *Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

[6] Matthew Cooper in Jonathan Foote. Summarizing popular music via structural similarity analysis. Objavljeno v *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, strani 127–30, New Paltz, NY, United States, 2003.

[7] Matthew L. Cooper in Jonathan Foote. Automatic music summarization via similarity analysis. Objavljeno v *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, strani 81–85, Paris, France, October 2002.

[8] Alain de Cheveigné in Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[9] Iréne Deliége, Marc Mèlen, in Ian Cross. Musical Schemata in real-time listening to a piece of music. *Music Perception*, 14(2):117–160, 1996.

[10] G. Doretto, A. Chiuso, Y. N. Wu, in S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003.

[11] J. P. Eckmann, S. O. Kamphorst, in D. Ruelle. Recurrence plots of dynamical systems. *EPL (Europhys. Lett.)*, 4:973–977, 1987.

[12] Tuomas Eerola in Petri Toiviainen and. Suomen Kansan eSävelmät. Digital archive of Finnish Folk Tunes. Finnish Literary Society, 2004. URL: esavelmat.jyu.fi.

[13] Antti Eronen. Chorus Detection with combined use of MFCC and Chroma features and image processing filters. Objavljeno v *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, strani 229–36, Bordeaux, France, 2007.

[14] Jonathan T. Foote. Visualizing Music and Audio using Self-Similarity. Objavljeno v *Proceedings of the 7th ACM international conference on Multimedia*, strani 77–80, 1999.

[15] Jonathan T. Foote in Matthew L. Cooper. Media Segmentation using Self-Similarity Decomposition. Objavljeno v *Proceedings of SPIE Storage and Retrieval for Multimedia Databases*, strani 167–175, 2003.

[16] Emilia Gómez. *Tonal Description of Music Audio Signals.* Doktorska disertacija, University Pompeu Fabra, Barcelona, Spain, July 2006.

[17] Masataka Goto. A Chorus-Section Detecting Method for Musical Audio Signals. Objavljeno v *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, stran 50, October 2003.

[18] Masataka Goto. A Chorus Section Detection Method for Musical Audio Signals and its Application to a Music Listening Station. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1783–94, 2006.

[19] Ning Hu, Roger Dannenberg, in George Tzanetakis. Polyphonic Audio Matching and Alignment for Music Retrieval. Objavljeno v *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, strani 185–8, New Paltz, NY, United States, 2003.

[20] KNAW/Meertens Instituut. Dutch Song Database. KNAW/Meertens Instituut, 2008.

[21] J. S. Downie J. Tague-Sutcliffe in S. Dunne. Name that tune: An introduction to musical information retrieval. Objavljeno v *Information as a Global Commodity: Communication, Processing and Use. Proceedings of 21st Annual Conference of the Canadian Association for Information Science*, strani 204–216, Antigonish, Nova Scotia, July 1993. Canadian Association for Information Science.

[22] Tristan Jehan. *Creating Music by Listening.* Doktorska disertacija, Massachusetts Institute of Technology, MA, USA, September 2005.

[23] Kristoffer Jensen. A Causal Rhythm Grouping. Objavljeno v *Computer Music Modeling and Retrieval*, Lecture Notes in Computer Science, strani 83–95. Springer Berlin / Heidelberg, 2005.

[24] Kristoffer Jensen. Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Applied Signal Processing*, 2007(1):159–159, 2007.

[25] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, in Lian-Hong Cai. Music type classification by spectral contrast feature. Objavljeno v *Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International Conference on*, del 1, strani 113–116, 2002.

[26] Florian Kaiser in Thomas Sikora. Music Structure Discovery in Popular Music using Non-negative Matrix Factorization. Objavljeno v *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*, Utrecht, Netherlands, Avgust 2010. Society for Music Information Retrieval. ISBN 978-90-393-53813.

[27] Beth Logan in Stephen Chu. Music Summarization Using Key Phrases. Objavljeno v *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, del 2, strani 749–752, June 2000.

[28] Lie Lu, Muyuan Wang, in Hong-Jiang Zhang. Repeating pattern discovery and structure analysis from acoustic music data. Objavljeno v *Proceedings of the ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR)*, strani 275–82, New York, NY, United States, 2004.

[29] Namunu C. Maddage, Changsheng Xu, Mohan S. Kankanhalli, in Xi Shao. Content-based music structure analysis with applications to music semantics understanding. Objavljeno v *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, strani 112–119, New York, NY, USA, 2004. ACM.

[30] Matija Marolt. Probabilistic Segmentation and Labeling of Ethnomusicological Field Recordings. Objavljeno v *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, strani 75–80, Kobe, Japan, October 2009.

[31] Meinard Müller. *Information Retrieval for Music and Motion.* Springer Verlag, 2007.

[32] Meinard Müller, Peter Grosche, in Frans Wiering. Robust Segmentation and Annotation of Folk Song Recordings. Objavljeno v *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, strani 735–740, Kobe, Japan, Oktober 2009.

[33] Meinard Müller, Peter Grosche, in Frans Wiering. Automated Analysis of Performance Variations in Folk Song Recordings. Objavljeno v *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)*, strani 247–256, Philadelphia, Pennsylvania, USA, 2010.

[34] Jouni Paulus in Anssi Klapuri. Music structure analysis by finding repeated parts. Objavljeno v *In Proceedings of the 1st ACM Workshop on Audio and Music Computing for Multimedia (AMCMM 2006*, strani 59–68. ACM Press, 2006.

[35] Jouni Paulus in Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech & Language Processing*, 17(6):1159–70, 2009.

[36] Jouni Paulus, Meinard Müller, in Anssi Klapuri. Audio-based Music Structure Analysis. Objavljeno v *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*, strani 625–636, Utrecht, Netherlands, August 2010.

[37] Geoffroy Peeters. Toward automatic music audio summary generation from signal analysis. Objavljeno v *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, strani 94–100, Paris, France, October 2002.

[38] Geoffroy Peeters. Sequence Representation of Music Structure Using Higher-Order Similarity Matrix and Maximum-Likelihood Approach. Objavljeno v *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, strani 35–40, Vienna, Austria, 2007.

[39] Aggelos Pikrakis, Theodoros Giannakopoulos, in Sergios Theodoridis. A computationally efficient speech/music discriminator for radio recordings. Objavljeno v *in Proc. of of the 7th International Conference on Music Information Retrieval (ISMIR'06*, strani 107–110, 2006.

[40] Lawrence Rabiner. A tutorial on HMM and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Februar 1989.

[41] Lawrence Rabiner in Biing H. Juang. *Fundamentals of Speech Recognition.* Prentice Hall, united states ed edition, April 1993.

[42] Helmut Schaffrath. The Essen Folksong Collection. D. Huron (ed.), Stanford, CA, 1995. URL: http://essen.themefinder.org/.

[43] S. S. Stevens, J. Volkmann, in E. B. Newman. A Scale for the Measurement of the Psychological Magnitude Pitch. 8(3):185–190, 1937.

[44] Graga Strle in Matija Marolt. Conceptualizing the Ethnomuse: Application of CIDOC CRM and FRBR. Objavljeno v *Proceedings of CIDOC2007*, 2007.

[45] David Temperley. *The Cognition of Basic Musical Structures*. The MIT Press, 2001.

[46] Peter van Kranenburg. *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*. Doktorska disertacija, Utrecht University, 2010.

[47] Peter van Kranenburg in George Tzanetakis. A Computational Approach to the Modeling and Employment of Cognitive Units of Folk Song Melodies using Audio Recordings. Objavljeno v *Proceedings of the 11th International Conference on Music Perception and Cognition*, strani 794–797, 2010.

[48] Ron J. Weiss in Juan Pablo Bello. Identifying repeated patterns in music using sparse convolutive non-negative matrix factorisation. Objavljeno v *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR 2010)*, Utrecht, Netherlands, August 2010.