UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Domen Strupeh

KLASIFIKACIJA VOKALNIH POSNETKOV LJUDSKE GLASBE

DIPLOMSKO DELO
NA UNIVERZITETNEM ŠTUDIJU

Mentor: doc. dr. Matija Marolt

Ljubljana, 2010

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. naloge: 01644/2010
Datum: 15.03.2010

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **DOMEN STRUPEH**

Naslov: **KLASIFIKACIJA VOKALNIH POSNETKOV LJUDSKE GLASBE**
CLASSIFICATION OF VOCAL FOLK MUSIC RECORDINGS

Vrsta naloge: Diplomsko delo univerzitetnega študija

Tematika naloge:

V diplomski nalogi preučite algoritme za klasifikacijo pevskih posnetkov in testirajte delovanje izbranih algoritmov na posnetkih slovenske ljudske glasbe. Pri tem se osredotočite na algoritme, ki ločujejo eno in večglasne posnetke ter inštrumentale in jih preizkusite na terenskih posnetkih iz digitalnega arhiva Etnomuza.

Mentor:

doc. dr. Matija Marolt

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

Dekan:

prof. dr. Franc Solina

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

diplomskega dela

Spodaj podpisani/-a **Domen Strupeh**,

z vpisno številko **63040154**,

sem avtor diplomskega dela z naslovom:

Klasifikacija vokalnih posnetkov ljudske glasbe

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom

  doc. dr. Matije Marolta

- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela

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

V Ljubljani, dne 15.9.2010 Podpis avtorja:

Zahvala

Zahvaljujem se mentorju, doc. dr. Matiji Maroltu, za vsakršno pripravljenost pomagati pri tej diplomski nalogi, za vse nasvete, napotke in potrpežljivost.
V drugi vrsti gre zahvala mojim staršem, ki so me podpirali in spodbujali tekom celotnega šolanja in mi s tem omogočili nemoten študij na fakulteti.
Za podporo bi se rad zahvalil tudi ostalim sorodnikom ter predvsem Hani, ki je z razumevanjem in bodrenjem poskrbela, da sem napisal to nalogo.
Na koncu bi se rad zahvalil še sošolcem na fakulteti za pomoč, kadar sem jo potreboval in prijateljem za izjemno preživljanje prostih trenutkov med študijem.

Kazalo

Seznam uporabljenih kratic in simbolov

MFCC – (ang. Mel-frequency cepstral coefficients) značilnice, ki predstavljajo amplitudni spekter zvoka v kompaktni obliki

ΔMFCC – (ang. Delta mel-frequency cepstral coefficients) značilnice, ki so derivati koeficientov MFCC

GMM – (ang Gaussian mixture model) klasifikator na osnovi statističnega modela kot mešanice Gaussovih porazdelitev

SVM – (ang. Support vector machine) klasifikator na osnovi razmejevanja razredov z uporabo podpornih vektorjev

DFT – (ang. Discrete Fourier transform) diskretna Fouirierjeva transformacija

FFT – (ang. Fast Fourier transform) hitra Fourierjeva transformacija

EM algoritem – (ang. Expectation-maximization) optimizacijska metoda za ocenjevanje parametrov GMM modela

Povzetek

V tej nalogi je predstavljeno področje avtomatskega prepoznavanja sestavov na posnetkih s petjem. Klasifikacija avdio posnetkov ali njihovih delov v posamezne razrede je uporabna predvsem pri velikih zbirkah posnetkov, ki vsebujejo veliko uporabnih informacij za raziskovanje. Ročno označbo posnetkov bi lahko nadomestila avtomatska, toda uspešnost klasifikacije je v veliki meri pogojena s tem, kakšne posnetke želimo klasificirati. V tem primeru sem se omejil na glasbene posnetke s petjem in preučil specifičnost etnomuzikoloških zbirk posnetkov.
Za klasifikacijo sem implementiral dva sistema, ki prepoznavata akustične vzorce v podatkih. Prvi za klasifikacijo uporablja metodo podpornih vektorjev (SVM), drugi pa Gaussove modele (GMM). V obeh sistemih je binarni klasifikator uporabljen za ločevanje posnetkov z enoglasnim in večglasnim petjem ter posnetkov petja in petja z glasbeno spremljavo. Ločevanje je izvedeno na osnovi koeficientov MFCC in ΔMFCC, ki predstavljajo kompakten zapis zvočne barve in se pogosto uporabljajo pri avtomatski klasifikaciji posnetkov.
Testiranja so pokazala, da je natančnost klasifikacije odvisna od izbire števila uporabljenih koeficientov MFCC, medtem ko izbira klasifikacijske metode nima bistvenega vpliva. Pri ločevanju posnetkov z enoglasnim in večglasnim petjem je implementirani sistem pravilno klasificiral 78,6% primerov z uporabo metode GMM, pri ločevanju posnetkov petja in petja z glasbeno spremljavo pa je sistem z metodo SVM dosegel natančnost 89,5%.

Ključne besede

avtomatska klasifikacija, prepoznavanje akustičnih vzorcev, vokalna glasba, etnomuzikološke zbirke

Abstract

In this thesis the automatic recognition of groups in singing recordings is presented. The classification of audio recordings or their parts into defined classes is useful particularly at large record sets that carry a variety of useful research information. The manual record annotation could be replaced with automatic. Nevertheless, the classification accuracy is in great deal conditioned by the type of classification recordings. The scope of my research in this case is singing with accordance to specificity of ethnomusicological record sets.
Two systems that are based on the acoustic pattern recognition are implemented for classification. The first one uses Support vector machines method (SVM) and the second one Gaussian mixture models (GMM). In both systems the binary classifier is used to distinguish solo singing and multi-voice singing or singing and singing with accompaniment. Classification is based on Mel-frequency cepstral coefficients (MFCC) and delta-MFCC, which are a compact representation of tone colour and are frequently used in automatic classification of recordings.
Tests have showed that the classification accuracy depends on the number of used MFCC coefficients while the choice of classifier has no significant impact. With classification of solo singing and multi-voice singing, the implemented system correctly classified 78,6% of instances by using the GMM method and 89,5% of instances by using the SVM method on records of singing and singing with accompaniment.

Key words

automatic classification, acoustic pattern recognition, singing, ethnomusicological record set

1. Uvod

V naših življenjih se glasba pojavi prevečkrat, da bi jo lahko zanemarili. Spremlja nas na poti in v prostem času z vseh mogočih avdio naprav, ponuja nam priložnosti za obisk takšnih in drugačnih koncertov in tudi tistim ljudem, ki jih glasba popolnoma nič ne zanima, je že prav gotovo zaznamovala kak dogodek – če ne drugega, spotikanje pri plesanju tretje slike četvorke na maturantskem plesu. Glasba nam torej odseva neponovljive trenutke, zato nam tudi ostane v spominu. Na ta način se je ohranjala skozi stoletja in se med ljudmi prenašala v obliki petja, igranja in plesov. Včasih je bila to edina glasba, ki je "živela" med ljudmi, zato jo danes označujemo z nazivom ljudska glasba.

Namen ljudske glasbe ni le zabava, kot je to v navadi pri današnji "zahodni" glasbi, ampak ima še svojo globljo funkcijo. Tako lahko pesmi in skladbe ločujemo glede na namen kot napitniške, ljubezenske, vojaške, žalostinke… Prav tako se ta glasba ni razširjala v komercialne namene, kot se razširja pop glasba, temveč se je prav preko svoje funkcije prenašala ustno iz roda v rod. Ustno zato, ker za ljudsko glasbo ne obstaja nobenega ogrodja, ki bi imel določena pravila za zapis, kot jih ima zahodna glasba, ki temelji na dobro definiranem glasbenem sistemu. Če želimo torej preučevati takšno glasbo, jo je najprej potrebno posneti. Posnetki ljudske glasbe so skoraj zmeraj narejeni na terenu in ne v idealiziranem okolju kot je studio. S tem posnetki po svoji naravi niso čisti, vendar vsebujejo boljše informacije o kulturnem ozadju.

Zaradi teh razlik naj bi bil pristop raziskovalcev do ljudske oz. etno glasbe, kot jo ponekod imenujejo, drugačen in bolj specifičen, kar pa se odraža v manjšem obsegu raziskav algoritmov za to področje. Motivov, ki spodbujajo nastanek sistemov za analizo posnetkov, je gotovo več v komercialni glasbi. Ta glasba je bolj razširjena, zanjo obstajajo boljše baze podatkov, znanje pa se lahko uporablja v različnih sistemih, kot je prepoznavanje skladbe z mrmranjem (t.i. *Query by humming*), izločevanje vodilne melodije s posnetka, identifikacija pevcev itd. To področje raziskovanja doživlja razmah predvsem v zadnjem desetletju, ko zbirke posnetkov skokovito naraščajo in se išče načine za brskanje po teh zbirkah in avtomatsko označbo posnetkov z relevantnimi oznakami.

V strokovnih člankih je opazen napredek pri izluščanju informacij z glasbenih posnetkov in rezultati so obetajoči, ampak avtorji člankov delujejo z različnimi glasbenimi zbirkami, zato rezultati med seboj največkrat niso primerljivi. Revolucionarnega univerzalnega sistema, ki bi znal svojo nalogo opraviti stoodstotno in s posnetka pridobiti vse korektne informacije, še ni. Vedno gre za kompromis. Na eni strani se uporabljajo sintetične, za namene raziskav ustvarjene baze, kjer so rezultati večinoma boljši (pravilno delovanje v več kot 90% primerih). Na drugi strani pa so uporabljeni realni posnetki, izseki skladb in pesmi, pri katerih se sicer tudi dosega visok odstotek pravilnega delovanja, srednja vrednost rezultatov pa je večinoma manjša, pa najsi gre za klasifikacijo, segmentacijo ali pa izločevanje določene metrike s posnetka.

Analize zvočnih posnetkov na področju instrumentalne glasbe so usmerjene predvsem v zaznavo in razpoznavo inštrumentov ter v transkripcijo igranih linij v simbolično predstavitev. Na govornem področju se izvajajo analize semantike, kot tudi razpoznava

govora in govorca. Področje petja združuje oboje od zgoraj navedenega; petje ustvarja melodijo, ki je nosilec tona, značilnega tudi za instrument in mu tako lahko določimo osnovno višino, intenziteto, trajanje, poleg osnovne višine pa nastajajo tudi alikvoti, ki dajo zvoku barvo. Hkrati za petje veljajo vsa pravila človeškega vokalnega trakta, ki ga uporabljamo tudi med govorom. Z izgovarjanjem besed nastajajo različni vokali in konsonanti, z oblikovanostjo žrelne, ustne in nosne votline dobi glas značilen zven. Človeškim ušesom ni problem razlikovati med govorom, petjem in igranjem na inštrumente, celo razlikovanje enoglasnega in večglasnega petja ali igranja se ljudem zdi trivialna naloga, toda kako približati razpoznavo zvoka zmožnostim človeškega ušesa ostaja velik problem za računalniške algoritme. V tej nalogi sem hotel preučiti sposobnosti računalniških algoritmov za obvladovanje takšnih problemov.

1.1. Motivacija in cilji

Namen naloge je raziskati področje algoritmov za klasifikacijo pevskih posnetkov in preizkusiti delovanje nekaterih na slovenskih ljudskih pesmih. Za raziskovanje na posnetkih ljudskega petja obstaja več razlogov. Eden je ta, da se tudi sam ukvarjam z dejavnostjo petja in me to področje interesira. Drugi je ta, da za tovrstno petje obstaja manj raziskav. Pregledati je potrebno, kakšne pristope uporabljajo raziskovalci in kako se razlikujejo od pridobivanja informacij iz ostalih oblik glasbe.
Nekatere rešitve želim preizkusiti na arhivski zbirki slovenskih ljudskih pesmi. S pridobivanjem informacij iz te zbirke se že ukvarja Laboratorij za računalniško grafiko in multimedije na Fakulteti za računalništvo in informatiko v Ljubljani, posebej s projektom Etnokatalog[1]. Ta naloga bi lahko prispevala k natančnejšemu razvrščanju posnetkov s petjem.
V okviru naloge želim ustvariti svojo bazo posnetkov. Homogeni posnetki, ki bi vsebovali le eno vrsto sestava (solo pevec, več pevcev, petje in inštrumentalna spremljava, glasba brez petja), bi služili za testiranje algoritmov. Ustvariti je potrebno še primerno testno okolje in implementirati izbrane metode za klasifikacijo posnetkov. Z uporabo metod bi bilo s posnetkov smiselno določiti, v katerih se poleg petja pojavlja še inštrumentalna spremljava. Na posnetkih, kjer se pojavlja le petje, pa bi želel ugotoviti, koliko pevcev poje sočasno.

1.2. Sestava naloge

Diplomska naloga je sestavljena sledeče. V drugem poglavju so predstavljene tehnologije klasifikacije posnetkov s podrobnejšim pregledom najbolj uporabnih metod za ločevanje pevskih, govornih in inštrumentalnih posnetkov. Temu poglavju sledi pregled raziskav v zadnjem času in evalvacija pristopov. Sklop se konča z opisom arhivske zbirke slovenskih ljudskih pesmi.
V četrtem poglavju sklopu je naveden potek izdelave baze posnetkov ter uporabljen pristop za klasifikacijo posnetkov. Predstavljeni so rezultati testiranj in vrednotenje rezultatov.
V zadnjem poglavju je predstavljen zaključek s pregledom doseženega in predstavitvijo možnosti za nadaljnje delo.

[1] Več o katalogu na naslovu: http://lgm.fri.uni-lj.si/matic/ethnocatalogue/

2. Klasifikacijske tehnike

2.1. Nadzorovano razpoznavanje vzorcev

V splošnem lahko *razpoznavanje vzorcev* [1] opišemo kot poskus "učenja" nekih funkcionalnih razmerij, ki jih nato prenesemo na poljubne vhodne podatke. Če poteka učenje sistema za razpoznavanje vzorcev na parih podatkov, ki imajo pri želeni funkciji določen vhod in izhod, potem govorimo o *nadzorovanem* sistemu. Vhodna vrednost sistema za nadzorovano učenje ima tipično vektorsko obliko in se imenuje *vektor značilnic (ang. feature vector).*

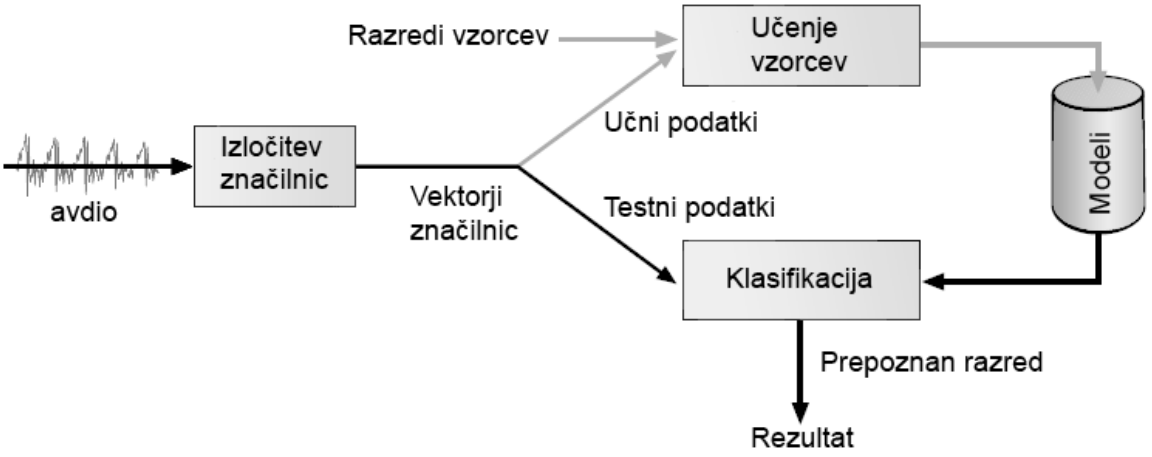
Sisteme za nadzorovano razpoznavanje vzorcev lahko razdelimo glede na obliko izhoda iz funkcije. Če lahko funkcija vrne neskončno število vrednosti, potem sistem izvaja *regresijo.* Primer regresije je prilagajanje linearnega ali polinomskega modela k podatkom z veliko šuma. Obratno lahko funkcija vrne končno število diskretnih vrednosti – v tem primeru sistem izvaja *klasifikacijo* in izhodne vrednosti se imenujejo *razredi*. Klasifikacijski sistem lahko izhodne podatke loči v dva razreda (kot npr. sistem za ločevanje govora in glasbe) ali pa, recimo, v 25 (kot npr. sistem, ki s slike prepozna eno izmed črk abecede). Klasifikacija v dva razreda se lahko prevede tudi kot *detekcija,* kar problem razpoznavanja vzorcev privede že v drugo območje raziskav v teoriji detekcije. V tej nalogi se bom posvečal le sistemom za klasifikacijo.

Cilj pri snovanju sistema za nadzorovano razpoznavanje vzorcev je doseči čim manjši delež napak. Pri klasifikaciji se napaka pojavi ob razvrstitvi v napačni razred. Verjetnost razvrstitve v napačni razred se imenuje *stopnja napake,* obratna verjetnost pa je *stopnja klasifikacije,* ki v odstotkih prikaže delež pravilnih klasifikacij.

Za vrednotenje sistemov za razpoznavanje vzorcev se običajno uporablja *testna množica,* ki je komplementarna *učni množici*. Obe množici dobimo, če množico vseh podatkov (s pripadajočimi oznakami) razdelimo na dve podmnožici in uporabimo eno za učenje sistema ter drugo za testiranje. Delitev lahko opravimo ročno, naključno, ali pa uporabimo t.i. *prečno preverjanje* [22]. Pri *pregibni* metodi prečnega preverjanja (*ang. k-fold cross validation)* je začetna množica razdeljena na več podmnožic oz. "pregibov" – eden se uporabi za testiranje, medtem ko se ostali uporabijo za učenje sistema. Proces prečnega preverjanja se nato ponovi še na ostalih pregibih. Pri metodi *izpuščanja enega* (*ang. leave-one-out cross-validation)* se za testiranje obdrži natanko en primer iz množice, na ostalih se sistem uči, nato pa se postopek ponovi na več drugih primerih. Za naključno delitev podatkov na dve množici običajno potrebujemo več podatkov kot za prečno preverjanje, zato je prečno preverjanje ugodno, kadar želimo kljub manjšemu naboru podatkov preizkusiti, kako se sistem obnese, če bi lahko skoraj vse podatke uporabili za učenje.

2.2. Klasifikacija posnetkov

Na spodnji sliki je predstavljen model klasifikacije avdio posnetkov. Temelji na prepoznavanju vzorcev v avdio signalu, ki ga predstavljajo značilnice. Prikazana blokovna shema je skupna vsem sistemom za prepoznavanje vzorcev, razlika je le pri njenih gradnikih.



Slika 1: Blokovna shema sistema za prepoznavanje vzorcev

Sistem je sestavljen iz treh osnovnih delov: izločevanje značilnic s posnetka, mehanizem za učenje vzorcev in mehanizem za klasifikacijo. Za delovanje sistemov za klasifikacijo sta najpomembnejša:

- mehanizem za učenje – biti mora zmožen posamezne vzorce med seboj čim bolje razmejiti
- zgradba podatkov - pri podatkih moramo poiskati takšen nabor, da bo le-ta čim bolj diskriminativen in bo odseval značilnosti svojega (ter samo svojega) razreda, zato te podatke imenujemo *značilnice*.

Ene najbolj diskriminativnih značilnic zvočne barve so koeficienti MFCC, ena najbolj uporabnih mehanizmov za učenje in razpoznavo glasbenih vzorcev pa sta metodi SVM in GMM.

2.3. Koeficienti MFCC

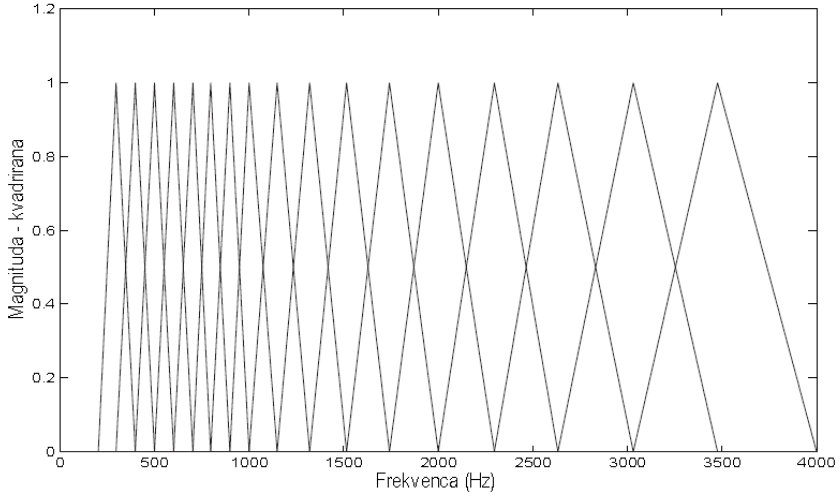
Mel-frekvenčni koeficienti (*ang. mel-frequency cepstral coefficients - MFCC*) so zelo široko uporabne značilnice, ki predstavljajo amplitudni spekter zvoka v kompaktni obliki [8]. Pogosto se jih uporablja pri prepoznavi govora ter prepoznavi govorca, vse pogosteje pa tudi pri nalogah, povezanih s petjem in inštrumentalno glasbo.
Postopek kreiranja koeficientov MFCC je naslednji. V prvem koraku se razdeli vhodni signal na okvire, po navadi z uporabo okenske funkcije. Okenska funkcija, tipično Hammingova funkcija, zmanjša učinke prehoda med okni. Za vsak okvir ustvarimo (kepstralni) vektor značilnic.
V naslednjem koraku naredimo *diskretno Fourierjevo transformacijo* (DFT) za vsak okvir in s tem pretvorimo signal iz časovne domene v frekvenčno domeno. Časovna zahtevnost DFT

je N^2, kar pomeni, da postane ob naraščajočem N ta algoritem zelo potraten. Zato se namesto DFT v večini primerov uporablja *hitra Fourierjeva transformacija* (fast Fourier transformation - FFT), ki zniža časovno zahtevnost na $N\log_2 N$. Pogoj pri tem je, da je dolžina N enaka potenci števila 2, kar ni težko, saj lahko vsako dolžino N dopolnimo z ničlami do izpolnjenega pogoja.

V tretjem koraku s trikotnimi filtri pretvorimo frekvenčno lestvico v t.i. mel-frekvenčno lestvico[7], ki je definirana s formulo (1):

$$m = 1127 \ln\left(1 + \frac{f}{700}\right) \qquad (1)$$

"Mel" je psihoakustična enota frekvence, ki je povezana s človeško zaznavo. Študije so namreč pokazale, da človeško uho ne sledi linearni lestvici, ampak mel-frekvenčni. Zgornja formula določa, da je mel-frekvenčna lestvica linearna do 1000 Hz in se nato nadaljuje logaritemsko. To pomeni, da človek ton z določeno mel vrednostjo sliši 2x višje kot ton, ki ima polovico te mel vrednosti. [21]



Slika 2: Mel-frekvenčna lestvica

V končnem koraku so elementi vektorjev mel spektra visoko korelirani, zato je uporabljena *diskretna kosinusna transformacija*. Ta odpravi korelacijo in mel spekter zopet pretvori v časovnega, s čimer dobimo mel frekvenčne koeficiente.

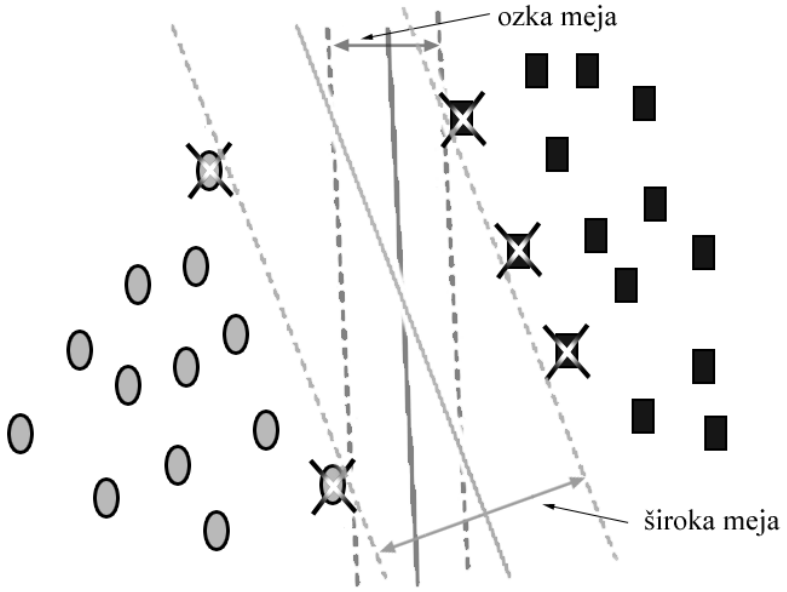
Koeficienti MFCC, pridobljeni z zgornjimi koraki, ne vsebujejo informacij o časovni soslednosti. Da bi v njih vključili tudi podatke o časovnem spreminjanju skozi več okvirjev, jim pogosto dodamo tudi njihove časovne derivate, ki se imenujejo koeficienti Delta (Δ). Koeficienti Delta so izračunani po naslednji formuli za linearno regresijo:

$$\Delta c[m] = \frac{\sum_{i=1}^{k} i(c[m+i] - c[m-i])}{2\sum_{i=1}^{k} i^2} \qquad (2)$$

kjer je *2k+1* velikost regresijskega okna in *c[m]* je *m*-ti koeficient MFCC.
Koeficiente Delta navadno v značilnice vključimo z dodajanjem na konec vektorja koeficientov MFCC; tako je ob uporabi 20 koeficientov MFCC in 20 ΔMFCC vektor značilnic dolg 40 koeficientov.

2.4. SVM – metoda podpornih vektorjev

Metoda podpornih vektorjev (ang. support vector machines - SVM) je klasifikacijska metoda, ki s pomočjo hiperravnine razdeli vhodne podatke na dve množici. Uporablja se pri mnogih problemih klasificiranja, kot npr. identifikacija govorca, kategorizacija besedil, prepoznavanje obrazov, v zadnjem času pa je vedno bolj prisotna tudi pri prepoznavanju inštrumentov[4,14].
Če imamo dva razreda, vsakega s svojimi značilnostmi, poskuša SVM poiskati mejo, ki ločuje značilnosti obeh razredov. To je lahko premica, ravnina ali pa, v N-dimenzionalnem prostoru, hiperravnina. SVM določi mejo tako, da je ta čim širša in s tem najbolje razmejuje podatke. Vektorji, ki ležijo na robu hiperravnine, se imenujejo podporni vektorji in določajo položaj hiperravnine v prostoru.



Slika 3: Iskanje najširše razmejitve podatkov v metodi SVM. Podporni vektorji so označeni s križem.

Formalno SVM poišče hiperravnino $w \cdot x + b = 0$, ki predstavlja učne primere $x_1, \ldots, x_p$ s pripadajočimi oznakami $y_1, \ldots, y_p$ ($y_i \in \{-1,1\}$), tako da velja (3)

$$y_i(x_i \cdot w + b) - 1 \geq 0, \forall i \qquad (3)$$

Vektor *w* predstavlja vektor uteži, *b* pa odmik.

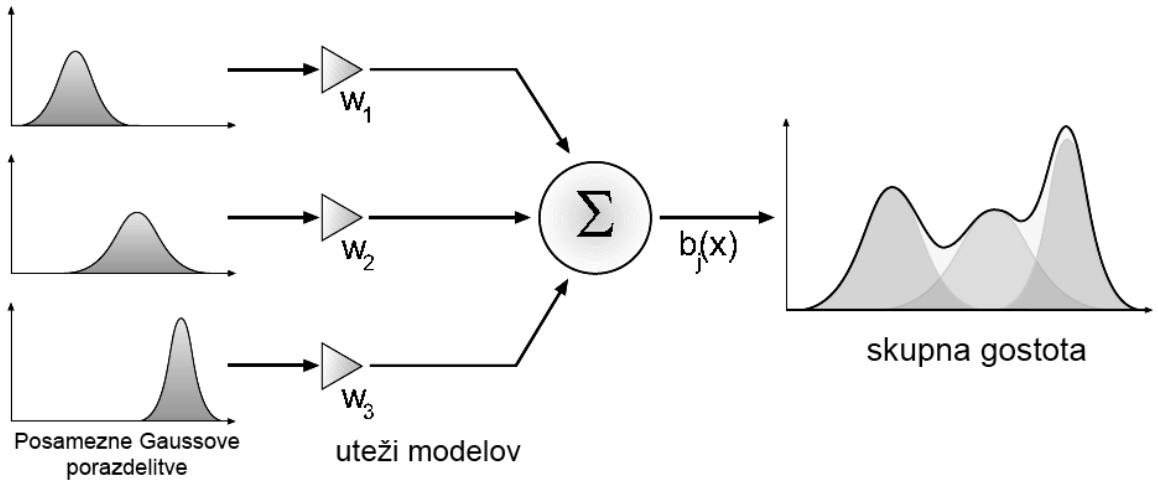
Kadar predmeti v množicah niso linearno ločljivi, kar je pri realnih primerih zelo pogosto, SVM ne skuša poiskati krivulje, ki bi natančno razmejila podatke, ampak uporabi t.i. "jedrno funkcijo". Ta preslika podatke v višjedimenzionalni prostor, kjer je možno za razmejitev zopet uporabiti hiperravnino.
Kot jedrno funkcijo se največkrat uporablja ena od naslednjih treh:

- linearna
- polinomska
- funkcija z radialnimi nevroni (radial basis function)

SVM kot binarni klasifikator v osnovi deli množice v dva razreda. Kljub temu jo lahko uporabimo tudi za delitev v K-razredov z uporabo strategij "ena proti ena" ali "ena proti vsem".

2.5. GMM – Gaussovi modeli

Druga metoda, ki se pogosto uporablja za klasifikacijo inštrumentov, petja in govora, so Gaussovi modeli (ang. Gaussian mixture model - GMM). Ta metoda uporabi statistične podatke učne množice za oceno gostot porazdelitve in izgradnjo modela [14].



Slika 4: Združevanje Gaussovih porazdelitev v mešanico

Za izgradnjo modela metoda potrebuje metoda množico učnih vektorjev značilnic

$$X=\{\bar{x}_1, \dots, \bar{x}_m\}, \text{ kjer } \bar{x}_i \in R^d. \quad (4)$$

Če predvidevamo, da je *m* vektorjev statistično neodvisnih in enako porazdeljenih, potem je verjetnost, da celotna množica pripada razredu (npr. inštrumentu) C_1 enaka

$$p(X = \{\bar{x}_1, \dots, \bar{x}_m\}|C_1) = \prod\nolimits_{i=1,m} p(\bar{x}_i|C_1) \quad (5)$$

Ob predpostavki, da je verjetnost vektorja lahko izražena s skupkom Gaussovih porazdelitev, potem velja

$$p(\bar{x}_i|C_1) = \sum_{l=1}^{K} P(l|C_1)\, p(\bar{x}_i|l, C_1) \quad (6)$$

kjer

$$p(\bar{x}_i|l, C_1) = \frac{\exp\left(-1/2(\bar{x}_i - \mu_{l,1})^t \sum_{l,1}^{-1}(\bar{x}_i - \mu_{l,1})\right)}{\sqrt{(2\pi)^d|\Sigma_{l,1}|}} \quad (7)$$

$P(l|C_1)$ je predhodna verjetnost komponente l za razred C_1 in $p(\bar{x}_i|l, C_1)$ je verjetnost, da vektor $\bar{x}_i$ pripada komponenti l znotraj razreda C_1. Parametri te Gaussove porazdelitve so vektor srednjih vrednosti $\mu_{l,1}$ in diagonalna kovariančna matrika $\Sigma_{l,1}$.

Za učenje sistema zberemo vse vektorje, ki pripadajo določenemu razredu (inštrumentu) in z njimi določimo parametre Gaussovega modela: uteži, vektorje srednjih vrednosti in diagonalne kovariančne matrike. Za ta namen uporabimo EM algoritem (*ang. Expectation-*

Maximization algorithm). To je iterativni algoritem, ki izračuna maksimalne verjetnosti parametrov. Začetni parametri Gaussovega modela (srednje vrednosti, kovariance, in predhodne verjetnosti), ki se uporabijo v EM algoritmu so ustvarjene z metodo gručenja s k-povprečji (ang. k-means clustering).

Ko imamo določene vse parametre za vsak razred (inštrument), lahko določimo razred tudi testnemu vektorju. Vektorju s testnimi podatki $\bar{x}$ se določi tisti razred, ki maksimizira $p(C_j|\bar{x})$, kar je enako maksimiziranju $p(\bar{x}|C_j)p(C_j)$ z uporabo Bayesovega pravila. Če ima vsak razred enako predhodno (*a priori*) verjetnost, potem je merilo verjetnosti kar $p(\bar{x}|C_j)$. To pomeni, da je testni vektor $\bar{x}$ klasificiran v tisti razred, ki maksimizira $p(\bar{x}|C_j)$.

3. Pregled področja

Glavni motiv za pridobitev informacij iz glasbe je predvsem olajšati brskanje po vedno večjih zbirkah glasbenih posnetkov. Za učinkovitost takšnega brskalnega sistema so potrebni natančni in relevantni metapodatki, ki posnetku določajo bodisi sestavo, izvajalce, bodisi žanr ali podobnost s kakšnimi drugimi posnetki. V sistemu pa so lahko posnetki z različnih področij: govorni, pevski, inštrumentalni, zato je potrebno najprej razlikovati med tem, v katero področje spada posamezen posnetek ali del posnetka. Pri pridobivanju informacij z zvočnih posnetkov gre torej za interdisciplinarno področje, kjer se s posnetkov avtomatsko odkrivajo relevantne informacije in uporabljajo v namenskih aplikacijah.

3.1. Klasifikacija s posnetkov

Pri razlikovanju področij je osnovno razlikovanje med glasbo in govorom. V [2] so preučili učinek nekaterih značilnic in njihovih derivatov na ločevanje glasbe in govora. Na osnovi velike baze govornih in glasbenih posnetkov popa, jazza, klasične glasbe in nekaj primerov glasbe vzhodnih narodov so se za razlikovanje z 99% uspešnostjo najbolje obnesli MFCC koeficienti in njihovi derivati.
V drugi študiji [16] so MFCC koeficiente z derivati uporabili za bolj zahtevno nalogo, in sicer ločevanje govora ter petja. Zbirka posnetkov je vključevala 75 oseb, ki so brali in peli dele pesmi, sistem pa je posnetke klasificiral s pomočjo dveh 16-komponentnih GMM, za govor in petje. Na dvo-sekundnih signalih so skupaj z merjenjem temeljne frekvence dosegli 90% učinkovitost.
Podobno učinkovitost so dosegli tudi v študiji [17], le da pri razlikovanju petja in glasbe na splošno. Uporabili so orodja za izločitev velikega števila značilnic (389) in potem izbrali najbolj relevantne značilnice. Klasifikacijo so opravili s SVM.

Razlikovanje delov s petjem in delov brez petja je postala del osnovnih metod priprave posnetka za nadaljnjo visokonivojsko analizo. V [20] so razvili algoritem za segmentacijo posnetkov na vokalna in nevokalna področja, z namenom identifikacije pevca v vokalnih področjih. Ugotovili so, da je med petjem in inštrumentalom opazna razlika v spektralni distribuciji, ki se jo z 20 koeficienti MFCC da zajeti. Klasifikacijo so opravili z dvema GMM modeloma z empirično najboljšo konfiguracijo: 64 komponent za vokalni in 80 komponent za nevokalni GMM model. Rezultati pravilne segmentacije algoritma so med 73% in 80%, odvisno od načina segmentacije (po okvirjih, segmentih stalnih dolžin ali segmentih, ki tvorijo homogeno celoto).
Drugačen pristop do segmentacije na vokalna in nevokalna področja so raziskali v [15]. Na zbirki 20 pop pesmi so preizkusili model, zgrajen iz več prikritih modelov Markova (Hidden Markov models – HMM), ki pri odločitvi o prisotnosti vokala upošteva zgradbo pop skladb (uvod, kitica, refren, prehod) ter metrum in osnovno tonaliteto. Z uporabo zgolj MFCC koeficientov in multi-modelnega HMM klasifikatorja so dosegli 78% uspešnost. Za izboljšanje so uporabili koeficiente energije, modelu pa so dodali možnost, da se zanesljivi deli testiranega posnetka še pred klasifikacijo uporabijo za učenje modela. S tem so dosegli 87% uspešnost.

V študiji [3] je uporabljen sistem za ločevanje govora in ostale glasbe. Problem sistema je, da petje velikokrat uvrsti v razred govora. Predstavljena je rešitev, ki razlikuje govor in glasbo v dveh stopnjah – v prvi se na posnetkih detektira samo petje, nato pa se v drugi stopnji preostali signali ločijo še na govor in glasbo. Na prvi stopnji se za detektiranje petja uporabljajo koeficienti MFCC ter harmonični koeficienti z njihovimi 4-Hz modulacijskimi vrednostmi, kot klasifikator pa so uporabljeni modeli GMM. Ugotovili so, da MFCC koeficienti slabo detektirajo petje, medtem ko lahko harmonični koeficienti to nalogo skupaj z njimi opravijo zelo uspešno. Njihova detekcija petja je dosegla 14% stopnjo napake, z uporabo naknadnih pravil pa so napako zmanjšali celo na 0%.

V eni izmed nedavnih študij [5] so primerjali učinkovitost MFCC koeficientov in LPMCC koeficientov (LPC-derived mel cepstral coefficients; MFCC koeficientov z upoštevanjem ovojnice spektra) za modeliranje specifičnega pevskega glasu. Primerjava je pokazala, da LPMCC koeficienti boljše karakterizirajo posameznikov (pevski) glas od MFCC koeficientov, toda MFCC koeficienti dosti bolje predstavijo neko splošno glasovno barvo, ki je bolj skupna skupini kot posamezniku.
V študiji [14] so za namen prepoznave 8 določenih inštrumentov s posnetkov primerjali MFCC koeficiente v mel-lestvici in enake koeficiente v linearni lestvici ter SVM in GMM klasifikatorje. Tudi pri karakteriziranju inštrumentalne barve so se MFCC (z mel-lestvico) izkazali kot značilnice z večjo močjo klasifikacije (za 10%), GMM klasifikator z dvema komponentama pa je bil v tem primeru uspešnejši od SVM za 7%.

Za ločevanje inštrumentalne glasbe od petja z inštrumentalno podlago so v [10] uporabili MFCC koeficiente in dva GMM modela za klasifikacijo med glasbenimi odseki z in brez petja. V študiji so predstavili uporabo avtoregresijske funkcije (ARMA), ki lahko s pomočjo prejšnjih okvirov napove naslednjega in s tem zmanjša napačno klasifikacijo posameznih okvirov. Njihova testiranja na bazi z 10 pevci in 84 1-minutnimi pesmimi so pokazala 82% uspešnost algoritma.

Razlikovanje med posnetki s solistom in posnetki z dvema pevcema so preučili v [19]. Njihov sistem za prepoznavo pevca na posnetkih na prvi stopnji loči tudi solo in duet pesmi. Za tovrstno klasifikacijo so uporabili samo MFCC koeficiente in GMM klasifikacijsko metodo z dvema modeloma za solo ter duet posnetke. Posnetki so bili posneti z desetimi moškimi pevci, kjer je vsak posnel 30 odsekov kitajskih pop pesmi, duet pesmi pa so dobili z združevanjem solo linij. Preizkusi so pokazali, da je med GMM modeli s 16 do 64 komponentami najbolj uspešen model s 16 komponentami in ta model je uspešno klasificiral 95% posnetkov.

3.2. Klasifikacija v ljudski glasbi

Večina raziskav poteka na pop oz. komercialni glasbi, zato so še toliko bolj pomembne raziskave na ljudskih glasbenih posnetkih. Ena izmed študij, ki bi bila primerna kot osnova za reševanje našega problema, je predstavljena v [18]. V njej je kot raziskovalno telo uporabljena etno glasba celega sveta, ki se nahaja v zbirki Lomax. Raziskovalci so uporabili 355 posnetkov več kot 50 različnih kultur iz zbirke. Namen raziskave je bila avtomatska anotacija posnetkov. Z MFCC koeficienti, pridobljenimi s posnetkov in dvorazrednim SVM

klasifikatorjem naj bi sistem ločil med: a)okviri s petjem (ki lahko vsebuje inštrumentalno spremljavo) ter samo inštrumentalnimi okviri, b)okviri s petjem (brez spremljave) ter okviri s petjem in glasbeno spremljavo. Kot značilnice so poleg MFCC koeficientov za doseganje boljšega rezultata bili uporabljeni tudi časovni aspekti, kot so derivati MFCC koeficientov ter uporaba sekvence več vektorjev, hkrati pa so odstranili prvi koeficient v MFCC vektorju, vektorje normalizirali in odstranili okvirje s tišino. Te prilagoditve niso pomembno vplivale na rezultate klasifikacije. Sistem je na 10-150 sekundnih posnetkih pravilno zaznal petje (primer a) v 74% primerih, inštrumente na posnetku (primer b) pa je pravilno zaznal v 77% primerih.

Ena nedavnih raziskav s področja ljudske glasbe [8] se ukvarja z analizo množice različnih avdio deskriptorjev in njihove uporabe za opis posnetkov. Posnetki, ki so uporabljeni v raziskavi, vključujejo zbirko latino glasbe Južne Amerike, zbirko posnetkov etno afriške glasbe in za primerjavo še zbirko komercialne zahodne glasbe. V študiji je prikazana razlika med metapodatki zahodne glasbe, ki po navadi vključuje ime skladbe, ime izvajalca, naslov albuma itd. ter metapodatki etno glasbe. V afriški glasbi so pomembni metapodatki npr. država, od koder prihaja posnetek, katere družine inštrumentov so na posnetku ter kakšno funkcijo ima pesem ali skladba (religiozna, ritualna, pogrebna, jutranja, plesna…). Pri latino posnetkih je pomemben t.i. žanr, ki ima vsak svoje specifično kulturno ozadje in je vsak povezan s svojim okoljem (Tango, Bolero, Bachata, Gaucha…). Analiza na izbranih posnetkih je pokazala, da se klasifikacija glede na značilnice za višino tona na teh posnetkih izkaže za zelo slabo, ter da veliko bolje to glasbo opišejo ritmične značilnice. Slovenske ljudske glasbe takšne ritmične značilnice gotovo ne bi mogle tako dobro opisati, saj je veliko bolj melodična, zato bi za reševanje našega problema bila bolj primerna analiza zvočnega spektra.

Z raziskavami na področju ne čisto ljudske, vendar nekomercialne glasbe so se ukvarjali tudi v [6]. Sistem za avtomatsko prepoznavo pevca so preizkusili na zbirki posnetkov grške rembetiko glasbe. Med posnetki je tudi nekaj takih, ki vsebujejo dosti šuma, saj najstarejši segajo v 30. leta prejšnjega stoletja – na slabše rezultate klasifikacije njihovega sistema naj ne bi vplivali. Ločevanje med okviri z vokalom in tistimi brez vokala je izvedeno na osnovi MFCC koeficientov in GMM klasifikatorja. Z uporabo t.i. maksimum metode, kjer je segment klasificiran v razred, ki ima maksimalno vsoto verjetnosti, je sistem dosegel 83% natančnost. Z uporabo metode prereza, kjer se segmenti s srednjo vrednostjo verjetnosti ne obdržijo, je sistem sicer manj segmentov označil kot vokalne oz. inštrumentalne, toda natančnost označenih se je povečala na 99%.

Nenazadnje se prav s posnetki iz arhiva slovenske ljudske glasbe ukvarjajo v [12]. V raziskavi se osredotočajo na anotiranje posnetkov kot govor, enoglasno petje, večglasno petje, inštrumentalno glasbo ter pritrkavanje. Pri tem gre v prvi vrsti za segmentacijo posnetkov na manjše enote in klasifikacijo le-teh. Zvočni posnetki so predstavljeni z devetimi značilnicami. Med temi so: kvocient RMS energije, srednja vrednost spektralne entropije, varianca derivatov spektralne entropije, varianca prvih treh koeficientov MFCC, ter varianca derivatov prvih treh koeficientov MFCC. Prve tri značilnice so predvsem namenjene razlikovanju med govorom, pritrkavanjem in glasbo. Kot klasifikator je uporabljena logistična regresija, ki je preprosta in daje dobre rezultate, ki so lahko podani kot verjetnostna porazdelitev po vseh razredih. Sistem razvršča segmente v pet skupin s skupno natančnostjo 78%. V moji nalogi se z označevanjem segmentov ne bom ukvarjal, a za nadaljnjo izboljšavo je v članku predlagan podrobnejši pogled v notranjo strukturo segmentov. Tako bi lahko v nalogi preučil

razlikovanje med inštrumentalno glasbo s petjem ter ostalima dvema razredoma (samostojno petje ter inštrumentalna glasba).

3.3. Arhiv terenskih posnetkov

Glasbenonarodopisni inštitut ZRC SAZU zbira posnetke ljudske glasbe že od začetka 20.st. Pri njih se nahaja arhiv terenskih posnetkov, ki so nastajali s snemanjem glasbe in ljudi v njihovem naravnem okolju z namenom ohranjanja ljudske glasbe in dostopnosti tovrstne glasbe za nadaljnje raziskave[13]. V večini primerov gre za intervjuje v živo z ljudmi, ki poznajo glasbo njihovega okolja, vendar niso poklicni glasbeniki. Na posnetkih v avdio obliki so zajeti govor, petje ter inštrumentalno izvajanje. Arhiv vsebuje okoli 30.000 takšnih terenskih zapisov, ki so lahko dolgi od nekaj minut do nekaj ur. Kvaliteta zvoka je različna. Odvisna je od okolice snemanja, saj se lahko pojavljajo različni šumi, kot npr. veter, ter predvsem od starosti posnetka, saj najstarejši digitalni zapisi segajo v leto 1955, ko tehnologija zajemanja zvoka še ni bila tako razvita kot danes.

4. Klasifikacija vokalnih posnetkov

Za klasifikacijo posnetkov potrebujemo najprej bazo testnih primerov. V tem delu je predstavljena zgradba baze testnih primerov, na katerih sem preizkusil klasifikacijski algoritem. Nato sledi opis implementacije izbranih pristopov za klasifikacijo, ob koncu poglavja pa so predstavljeni rezultati testiranja za posamezno vrsto klasifikacije.

4.1. Baza posnetkov

V Laboratoriju za računalniško grafiko in multimedije na Fakulteti za računalništvo in informatiko v Ljubljani imajo za raziskovalne namene shranjen arhiv terenskih posnetkov slovenskih ljudskih pesmi. Iz tega arhiva sem izločil več krajših avdio posnetkov, ki so se mi zdeli relevantni za raziskovanje mojega problema. Izločil sem po več krajših homogenih izsekov s posameznega (tudi večurnega) terenskega avdio zapisa. Skupno sem uporabil 73 različnih terenskih zapisov. Pridobljene posnetke sem ročno anotiral glede na to, ali sta na njem prisotna inštrumentalna glasba in petje ter glede na to, ali se na posnetku pojavi en pevski glas ali več. Med atribute sem dodal še spol pevcev oz. pevk v solo ali večglasni zasedbi, pa tudi govor. Teh dodatnih atributov v okviru svoje naloge sicer ne bom preučeval, je pa s tem baza uporabna za obširnejše raziskave.
Nekatere posnetke sem med poslušanjem označil za nejasne. Ta atribut sem dodal posnetkom, če so imeli v ozadju konstantno motnjo (tiktakanje ure, brnenje v mikrofonu), če sestav s posnetka zaradi specifičnih glasov ni bil določljiv (nejasno večglasje, glas zelo podoben zvoku harmonike), če je bilo petje na posnetku ekstremno tiho ali pa če je bil posnetek preprosto zelo slabe kvalitete in zaradi tega nerazločen. Tovrstnih posnetkov za testiranja nisem uporabil.

POSNETEK	INSTRUMENT	VOKAL	GOVOR	SOLO	VECGLASJE	MOSKI	ZENSKI	MESANO	NEJASEN
354.wav	1	1	0	1	0	1	0	0	0
355.wav	1	1	0	1	0	0	1	0	0
356.wav	0	1	0	0	1	1	0	0	-1
357.wav	0	1	0	0	1	0	0	1	0
358.wav	0	0	1	0	0	1	0	0	0
359.wav	1	1	0	0	1	0	0	1	0
360.wav	1	1	0	0	1	0	0	1	0
361.wav	0	1	0	0	1	1	0	0	0
362.wav	1	1	0	1	0	0	1	0	0

Slika 5: Pogled na strukturo podatkov v bazi

V bazi za učenje in testiranje algoritmov se med 297 posnetki nahaja 229 posnetkov petja, med katerimi je 160 posnetkov večglasja (35 posnetkov moškega, 81 ženskega in 44 mešanega), ter 69 posnetkov enoglasnega petja (v 23 posnetkih nastopa en moški glas, v 46 posnetkih pa en ženski glas). Poleg tega sestavlja bazo tudi 36 posnetkov petja in inštrumentala skupaj, ter 24 posnetkov inštrumentala brez petja. Med inštrumenti se pojavljajo godala, harmonika, ustna harmonika, brenkala, flavta in dude.

Vsi posnetki so shranjeni v formatu wav, njihova frekvenca je 48 kHz in bitna globina 16 bit. Zaradi lažje obdelave in zaradi tega, ker je bila večina posnetkov na voljo le v mono tehniki,

so vsi posnetki spremenjeni v mono tehniko. Dolžina originalnih posnetkov v bazi je 20s, a učenje in testiranje algoritmov sem opravil na skrajšanih posnetkih dolžine 4s.

4.2. Implementacija

Implementiral sem algoritma za izvrševanje dveh klasifikacijskih nalog na avdio posnetkih:

- ločevanje enoglasnega in večglasnega petja
- ugotavljanje prisotnosti inštrumentov na posnetkih petja

Sistem sem zasnoval na osnovi modela za prepoznavanje vzorcev. To pomeni, da potrebuje sistem za svoje delovanje učne podatke, ki jih klasifikacijska metoda uporabi za nastavitve parametrov svojega modela. Metodi klasifikacije, ki sem ju implementiral, sta SVN in GMM. Ti dve metodi sta se že izkazali kot sposobni ločiti različne izvore zvoka, kar pomeni, da dobro delujejo na avdio signalu, če je le-ta dobro predstavljen. Značilnice, ki opisujejo avdio signal v našem sistemu, so koeficienti MFCC in njihovi derivati, ΔMFCC. Te značilnice se v nekaterih sistemih uporabljajo kot ene izmed množice različnih značilnic, v drugih sistemih pa so glavni in edini nosilec podatkov za klasifikator. V slednjem primeru morajo biti zelo univerzalne, da lahko na njihovi podlagi sistemi dosegajo visoko stopnjo natančnosti pri klasifikaciji tako petja in inštrumentala, kot tudi enoglasja in večglasja. Z uporabo koeficientov MFCC v različnih klasifikacijskih nalogah sem preveril univerzalnost teh značilnic.

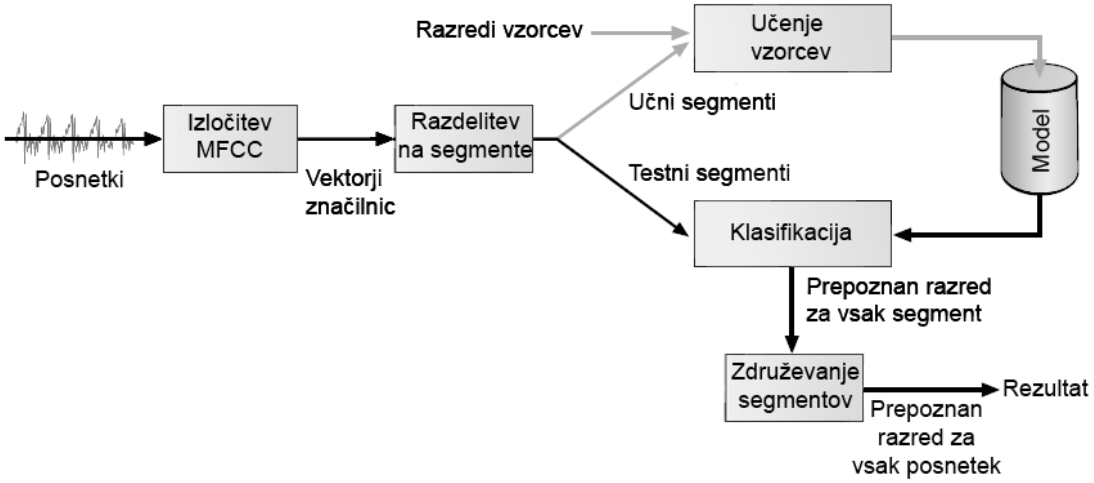
Vse algoritme sem implementiral v testnem okolju s programom *Matlab*, ki je učinkovito orodje za testiranje in vrednotenje algoritmov za obdelavo signalov. Za implementacijo izločevanja koeficientov MFCC sem uporabil *Auditory toolbox* za Matlab [23]. Hitra Fourierjeva transformacija deluje znotraj okna velikosti 1024 točk (21,3ms v 48kHz signalu) s pomikanjem okna za 160 točk. Tako z vsakega posnetka izločimo 1243 okvirjev s koeficienti MFCC.
Te koeficiente nato posredujemo mehanizmu za razpoznavo, ki iz učnih posnetkov zgradi klasifikacijski model. Izgradnja modela je odvisna od izbrane metode.

4.2.1. Klasifikacija z metodo SVM

SVM je v osnovi dvorazredni klasifikator in je kot tak idealen za reševanje zastavljenega problema. Radi bi ločili enoglasne in večglasne posnetke ter paroma inštrumental in petje s spremljavo ter petje s spremljavo in samostojno petje. Metoda SVM med pari razredov ustvari hiperravnino, ki razvršča testne primere glede na to, na katero stran hiperravnine spadajo značilnice testnega primera. Hiperravnina je v našem sistemu določena z modelom, ki je zgrajen ob učenju vzorcev na učnih segmentih. Segmenti so oblikovani s povprečenjem vrednosti koeficientov MFCC (in derivatov) na vnaprej določenem številu okvirjev.

Za izgradnjo modela bi lahko npr. uporabili en vektor značilnic za vsak posnetek. Če imamo na voljo veliko množico anotiranih posnetkov, potem tak način predstavitve zelo dobro opiše

celoten posnetek z vidika ohranjanja ključnih značilnosti, toda problem je v pogoju, da potrebujemo zelo veliko množico. V primeru, da imamo malo množico, lahko opravimo klasifikacijo na vsakem posameznem okvirju posnetka. S tem dobimo veliko zalogo podatkov za izgradnjo klasifikatorja, vendar tudi več šuma, ki ga posamezni okvirji vsebujejo. Poleg tega postane preslikovanje v višjedimenzionalni prostor z uporabo velikega števila vektorjev računsko zelo zahtevno. V tej nalogi je implementirana vmesna pot, kjer s povprečji segmentov dobimo večjo zalogo podatkov za izgradnjo modela, hkrati pa lahko z dolžino segmentov prilagodimo število učnih segmentov za izgradnjo modela.



Slika 6: Shema sistema z metodo SVM

.Metoda SVM vsak segment uvrsti v enega izmed dveh razredov: $c_s=0$ ali $c_s=1$. Ker nas zanima klasifikacija celotnega posnetka in ne posameznih segmentov, algoritem po klasifikaciji segmentov naredi vsoto razredov segmentov za vsak posnetek, nato pa celotnemu posnetku določi razred c_p po formuli:

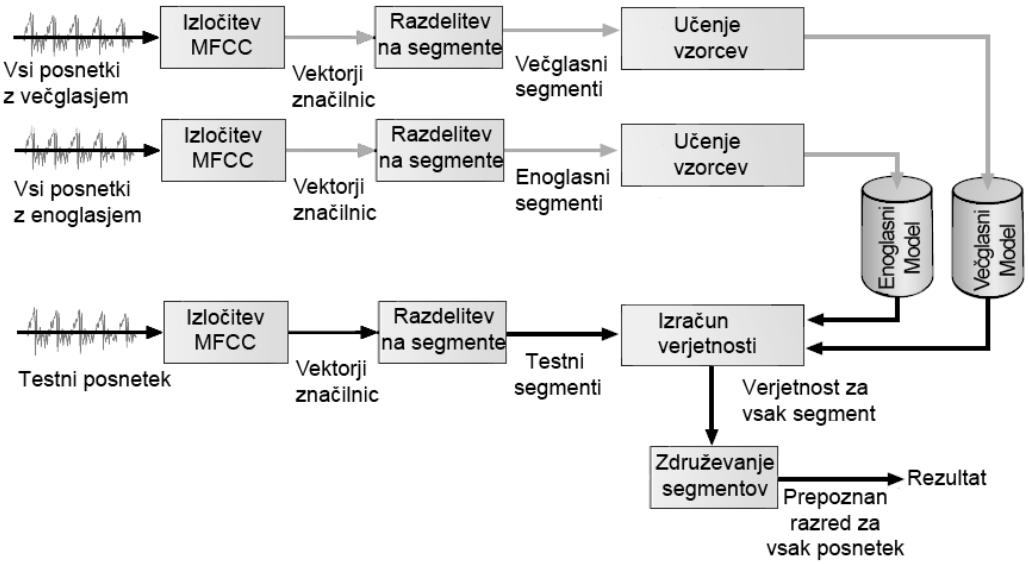
$$c_p = \begin{cases} 1; \sum_{s=1}^{N_{seg}} c_s \geq \frac{N_{seg}}{2} \\ 0; \sum_{s=1}^{N_{seg}} c_s < \frac{N_{seg}}{2} \end{cases} \quad (8)$$

Metodo za klasifikacijo SVM sem implementiral s pomočjo vgrajenih funkcij *svmtrain* in *svmclassify*, ki sta del Matlab-ove bioinformacijske zbirke orodij. Kot jedrno funkcijo za preslikavo v višjedimenzionalni prostor sem uporabil skalarni produkt, za iskanje razmejujoče hiperravnine pa kvadratično programiranje, ki deluje s pomočjo metod Matlabove zbirke orodij za optimizacijo.

4.2.2. Klasifikacija z uporabo modelov GMM

Pri klasifikaciji z uporabo modelov GMM se najprej iz učnih posnetkov izločijo značilnice, ki jih združimo v segmente, tako kot je to opisano pri klasifikaciji z metodo SVM. Zatem algoritem vse segmente, ki pripadajo istemu razredu, uporabi za določitev parametrov modela

GMM. Na sliki je prikazan primer za razreda "enoglasno petje" in "večglasno petje". Enoglasni model na sliki predstavlja akustični vzorec enoglasnih posnetkov, večglasni model pa akustični vzorec večglasnih posnetkov. Takšen model je bil za namene ločevanja enoglasnih in večglasnih posnetkov uporabljen v [19]. Enak pristop je uporabljen tudi pri klasifikaciji petja z in brez spremljave, le da je v tem primeru eden izmed modelov zgrajen na vseh posnetkih enoglasnega petja, drugi pa na vseh posnetkih petja z inštrumentalom.
Za klasifikacijo testnega posnetka se najprej iz značilnic posnetka oblikujejo segmenti, nato pa se za vsak segment izračuna verjetnost, da pripada posameznemu modelu. Za določitev, v kateri razred spada celoten posnetek, se uporabi vsota verjetnosti vseh segmentov posnetka, nato pa posnetek sistem klasificira kot enoglasnega, če je $\log P(X|\lambda_e)-\log P(X|\lambda_v)$ večje od določene meje, ki je v našem primeru 0. $P(X|\lambda_e)$ predstavlja verjetnost, da celoten posnetek pripada enoglasnemu modelu, enak postopek pa je uporabljen za vse razrede.



Slika 7: Shema sistema z metodo GMM

GMM model sem zgradil s pomočjo metode *gaussmix*, ki je del zbirke orodij za Matlab *Voicebox* [24], verjetnosti pripadnosti določenemu modelu pa sem dobil z metodo *gmmprob*, ki je del zbirke orodij za Matlab *Netlab* [25].

4.3. Testiranje in rezultati

Za pridobitev rezultatov natančnosti algoritmov za klasifikacijo sem uporabil prečno preverjanje z 10 pregibi. V vsaki izmed 10 iteracij je sistem 90% naključno izbranih posnetkov uporabil za učenje klasifikatorja, ostalih 10% posnetkov pa je bilo uporabljenih za testiranje. Natančnost klasifikacije sem izračunal po formuli, uporabljeni tudi v [ident in duet record 08]:

$$\frac{\check{S}t.\,pravilno\ razvr\check{s}\check{c}enih\ posnetkov}{\check{S}t.\,vseh\ testnih\ posnetkov} \times 100\%$$

Pri vseh vektorjih koeficientov MFCC sem odstranil prvi koeficient, ki po [10] predstavlja splošno energijo in naj se ga ne bi uporabljalo za klasifikacijo, da se s tem izognemo razlikam pri ojačitvah posnetkov. Za izračun koeficientov ΔMFCC sem uporabil okno širine 9 okvirov, ki se je empirično izkazalo za najboljše.

4.3.1. Ločevanje enoglasnega in večglasnega petja

4.3.1.1. Metoda GMM

Za modeliranje akustičnih vzorcev sem uporabil model GMM s 4 komponentami. S tem modelom sem dobil najboljše rezultate, ki pa se od modela s 16 komponentami, ki sem ga zaradi pogoste rabe v literaturi tudi preizkusil, niso bistveno razlikovali. Za širino segmenta sem uporabil 80 okvirjev (16 segmentov na posnetek). Če se za učenje parametrov modela GMM uporabi 90% podatkov, potem smo enoglasni model GMM naučili na podlagi vzorca 992 enoglasnih segmentov, večglasnega pa na podlagi vzorca 2304 večglasnih segmentov.

Kadar sem uporabil poleg koeficientov MFCC še koeficiente ΔMFCC, sem jih vedno uporabil enako število kot koeficientov MFCC. V preglednici 4-1 so navedene klasifikacijske točnosti ob spreminjanju števila koeficientov. Število koeficientov 4 v preglednici pomeni, da sem uporabil prve štiri koeficiente MFCC. Ker sem prvega izpustil, sem dejansko uporabil le koeficiente na mestih 2,3,4. Z uporabo koeficientov ΔMFCC enakih dolžin je tako dolžina celotnega vektorja značilnic znašala 6.

Število koeficientov MFCC(+delta)	Natančnost
4	**78,6%**
5	77,3%
6	77,3%
7	76,9%
12	68,6%
20	64,4%

Tabela 4-1: Natančnost GMM klasifikatorja pri različnem številu koeficientov MFCC

Ob uporabi dodatnih koeficientov ΔMFCC pri najboljšem primeru, se natančnost klasifikacije poveča predvsem na račun pravilne klasifikacije razreda z večglasjem (skoraj 90% točno), a pri tem natančnost razvrščanja razreda z enim glasom znaša komaj 56%, kar prikazuje matrika razvrščanj v tabeli 4-2.

	Dejansko	
Razvrščeno	En glas	Večglasno
En glas	**39**	19
Večglasno	30	**141**

Tabela 4-2: Matrika razvrščanj za primer s 4 koeficienti MFCC in 4 koeficienti ΔMFCC

Če uporabimo samo štiri koeficiente MFCC brez delta-koeficientov, dobimo sicer manjšo skupno natančnost klasifikacije (74%), toda s tem dosežemo veliko bolj natančno klasifikacijo razreda z enim glasom, kar prikazuje tabela 4-3.

	Dejansko	
Razvrščeno	En glas	Večglasno
En glas	**53**	43
Večglasno	16	**117**

Tabela 4-3: Matrika razvrščanj za primer s 4 koef MFCC in brez koeficientov ΔMFCC

Do napak pri razvrščanju večglasnih posnetkov z metodo GMM je prihajalo pogosteje ob počasnejših pesmih, oz. kadar je nek glas držal en ton dlje časa ali pa če je en glas bil tišji od drugega. Napačna razvrstitev enoglasnih posnetkov se je zgodila, kadar je pevcu na posnetku glas nihal.

4.3.1.2. Metoda SVM

Za širino segmenta pri metodi SVM sem preizkusil segmente dolge od 120 do 300 okvirjev (11 do 5 segmentov na posnetek). Empirično krajši segmenti zanemarljivo malo vplivajo na natančnost klasifikacije, se pa s krajšimi segmenti močno podaljša čas računanja. Pri dolžini segmenta 80 je tako algoritem za 10-pregibno prečno preverjanje potreboval 40min. Odločil sem se za računsko manj potratno možnost in testiral z dolžino segmenta 300 okvirjev.

Število koeficientov MFCC	Natančnost
3	**75,7%**
4	74,3%
7	74,3%
12	72,8%
20	64,4%

Tabela 4-4: Natančnost SVM klasifikatorja pri različnem številu koeficientov MFCC

Za najbolj učinkovito značilnico se je pri SVM metodi izkazal vektor štirih koeficientov MFCC brez koeficientov ΔMFCC (tabela 4-4). Z njim je sistem uspel klasificirati enoglasne

in večglasne posnetke z natančnostjo 73,5%. Toda, če z njim poskušamo pri SVM metodi klasificirati vse enoglasne in večglasne posnetke naenkrat, dobimo res skupno natančnost 73,5%, ampak v tem primeru matrika razvrščanj izgleda takole (tabela 4-5):

	Dejansko	
Razvrščeno	En glas	Večglasno
En glas	**11**	3
Večglasno	58	**157**

Tabela 4-5: Matrika razvrščanj pri delovanju SVM nad neuravnoteženima množicama učnih posnetkov

Problem SVM metode je, da ne deluje dobro, kadar množici razredov nista uravnoteženi. V naši bazi imamo 69 enoglasnih posnetkov in 160 večglasnih, kar je očitna neuravnotežena razdelitev. V tem primeru SVN ne razmejuje več razredov enakomerno, ampak se neuravnoteženost kaže v favoriziranju večjega izmed razredov.
Pomanjkljivost sem rešil s tem, da sem namesto celotne množice večglasnih posnetkov, za testiranje uporabil le 71 naključno izbranih posnetkov iz te množice, da se je približno ujemala po velikosti z velikostjo množice enoglasnih posnetkov. Natančnost klasifikacije je ostala enaka, le napačno klasificirani primerki so se porazdelili med oba razreda:

	Dejansko	
Razvrščeno	En glas	Večglasno
En glas	**49**	17
Večglasno	20	**54**

Tabela 4-6: Matrika razvrščanj ob prilagoditvi števila učnih posnetkov v večji množici

Do napak pri razvrščanju solo petja je prihajalo pri posnetkih z bolj odmevajočim glasom ali pa pri posnetkih, kjer pevcu glas niha. V obratni smeri je prihajalo do napak, kadar je bil eden glas močnejši od drugega.

4.3.2. Ločevanje petja in petja s spremljavo

4.3.2.1. Metoda GMM

Pri ločevanju petja in petja s spremljavo se je najbolje obnesel GMM model s 16 komponentami, zato sem ga uporabil za testiranje v vseh iteracijah. Poskuse sem opravil tudi s spreminjanjem dolžine segmenta med 20 in 160 okviri, vendar to ni bistveno vplivalo na rezultate klasifikacije.
Obratno kot pri ločevanju enoglasnega in večglasnega petja je tukaj GMM model deloval bolje pri daljših vektorjih značilnic. Pri dolžini segmenta 40 okvirjev je sistem dosegel svojo najvišjo natančnost že z uporabo 12 koeficientov MFCC in 12 koeficientov ΔMFCC (tabela

4-7). Za učenje pevskega GMM modela sem uporabil 6592 posameznih segmentov, za učenje modela z glasbeno spremljavo pa 992 posameznih segmentov.

Število koeficientov MFCC (+Delte)	Natančnost
20	**89%**
12	89%
7	79,2%
4	59,5%

Tabela 4-7: Delovanje GMM metode ob različnem številu koeficientov MFCC

Matrika razvrščanj za najboljši primer z 20 koeficienti MFCC in 20 koeficienti ΔMFCC kaže na enakovredno natančnost klasifikacije nad obema razredoma.

	Dejansko	
Razvrščeno	Petje s spremljavo	Samo petje
Petje s spremljavo	**28**	22
Samo petje	7	**207**

Tabela 4-8: Matrika razvrščanj za primer z 20 MFCC in 20 ΔMFCC

Napake pri klasifikaciji se pojavijo večinoma pri zelo rezkih ženskih glasovih, ki zvenijo skoraj kakor instrument.

4.3.2.2. Metoda SVM

Metoda SVM pri ločevanju petja in petja s spremljavo zopet trči ob oviro neenakomerne razporejenosti učnih posnetkov. V bazi imamo 229 posnetkov samostojnega petja in 35 posnetkov petja z inštrumentalno spremljavo. Da bi dobil relevantne rezultate, sem iz množice vokalnih posnetkov naključno izbral 38 posnetkov, kar pomeni, da se je SVM klasifikator učil na množici 170 segmentov. Rezultati natančnosti klasifikacije pri dolžini segmentov 300 okvirjev so naslednji (tabela 4-9):

Število koeficientov MFCC (+Delte)	Natančnost
20	89,3%
12	81,7%
7	**89,5%**
4	71,6%
3	68,9%

Tabela 4-9: Natančnost ločevanja petja in petja s spremljavo z uporabo SVM

Za vektor značilnic s 7 koeficienti MFCC sem izdelal še analizo, tako da sem 5-krat pognal 10-pregibno prečno preverjanje, vsakič z na novo izbranimi 38 vokalnimi posnetki. V povprečju znaša natančnost klasifikacije celotnih posnetkov s 7 koeficienti MFCC 87,1%.

Do napak pri klasifikaciji je velikokrat prihajalo zaradi ritma. Pri posnetkih, kjer inštrument igra enak ritem kot ga pevec poje, je klasifikator posnetkom v nekaj primerih suvereno določil napačni razred petje. Po drugi strani je do napak pri klasifikaciji v razred petje s spremljavo prihajalo pri posnetkih, kjer je v petju bolj ritmično poudarjena dinamika pa tudi pri rezijanskem petju, ki je po načinu podobno inštrumentalnemu izvajanju. Poleg tega je do napak zopet prihajalo ob rezkih ženskih glasovih.

4.3.3. Klasifikacija po segmentih

Cilj našega sistema je bil, sicer, klasifikacija celotnih posnetkov, a skozi testiranje sem opazoval tudi rezultate klasifikacije posameznih segmentov. Analiza segmentov je pomagala pri ugotavljanju, na katerih delih pesmi pride do napačne klasifikacije glede na razred celotnega posnetka. Tabela 4-10 prikazuje rezultate testa po segmentih z uporabo najboljših konfiguracij.

	Ločevanje eno- in večglasja	Ločevanje petja in petja+spremljava
Metoda GMM	62% (16 segmentov/posnetek)	83,3% (32 segmentov/posnetek)
Metoda SVM	66,8% (5 segmentov/posnetek)	80,8% (5 segmentov/posnetek)

4-10: Klasifikacija po segmentih z uporabo najboljših konfiguracij

4.3.4. Uporaba koeficientov MFCC za ločevanje med inštrumentalno glasbo in petjem s spremljavo

Oba sistema sem preizkusil tudi pri klasifikaciji inštrumentalne glasbe in petja z inštrumentalno spremljavo.
Preizkus je pokazal, da klasifikacija z metodo GMM doseže najvišjo klasifikacijsko točnost z uporabo 16-komponentnega modela GMM, segmentov dolžine 40 in 20 vektorjev MFCC z dodanimi 20 vektorji ΔMFCC. Preizkus je tudi pokazal, da natančnost takšne klasifikacije niha med 56% in 66%.
Še slabše se je odrezala metoda SVM, ki na podatkih iz obstoječe baze ne zmore zadovoljivo klasificirati inštrumentalnih posnetkov ter posnetkov z inštrumentalom in petjem. Najboljša natančnost 56% je dosežena pri velikosti segmenta 300 okvirjev in uporabi vektorja značilnic, dolgega 20 koeficientov MFCC ter 20 koeficientov ΔMFCC.

4.4. Razprava

Poglejmo si od blizu najprej rezultate zadnjih dveh testov. Pri ločevanju med inštrumentalno glasbo in petjem s spremljavo gre pravzaprav za detekcijo petja – v inštrumentalnih signalih skušamo poiskati tiste, ki vsebujejo tudi petje. Pri tej nalogi se MFCC koeficienti slabo odrežejo in skupaj z derivati niso zmožni zastopati distinktivnih značilnosti posameznega razreda. To je v skladu z dognanji v študiji [3], kjer sistem na osnovi zgolj koeficientov MFCC in njihovih derivatov pravilno prepozna petje v manj kot 32% primerov.
Zanimivo je, kako koeficienti MFCC niso zmožni poiskati pevskih signalov znotraj instrumentala, ko pa vendarle v osnovi izhajajo s področja govora in govorno povezanih problemov. Kot kaže, je s koeficienti MFCC res predstavljena neka zvočna barva, ki je pri nekaterih problemih bolj razločno razmejena, pri drugih pa manj. Kako pa torej, da so v raziskavi [10] uspeli prav pri ločevanju delov z inštrumentalnim igranjem ter delov s petjem in inštrumentalno spremljavo doseči natančnost 81,3% (merjeno z rezultatom *F score*) samo z uporabo značilnic MFCC? Pri tej raziskavi so uporabili posebno avtoregresijsko funkcijo v postopku po-obdelave, ki je upoštevala tudi predhodne okvire. Poleg tega so testiranja izvedli na posnetkih pop skladb 10 priznanih pevcev – v našem sistemu takšnih pevcev gotovo ni, včasih se na posnetku zvok inštrumentalnega ozadja celo zlije skupaj s pevskim glasom. Zato bi potreboval naš sistem drugačne značilnice, kot so koeficienti MFCC, če bi želeli učinkovito detektirati petje na inštrumentalnih posnetkih.

Rezultati testa po segmentih so manj natančni od tistih po celotnih posnetkih in na to lahko gledamo z dveh vidikov. Prvi vidik nam kaže, združevanje segmentov pri končni klasifikaciji posnetka je pomembno, saj lahko dosežemo zgladitev odločitve. Drugi vidik pa kaže na to, da rezultat testa po segmentih ni preveč pomemben, saj so bili posnetki anotirani na celotni dolžini in kot takšni tudi veljajo. Namreč, ni nujno, da bi ob ročni anotaciji vsakega posameznega segmenta s poslušanjem tudi sam tem segmentom pripisal isti razred, kot ga imajo sedaj v okviru celotnega posnetka.

Sistem za klasifikacijo sicer na celotnih posnetkih deluje dokaj dobro. Oba klasifikatorja (SVM in GMM) delujeta usklajeno, pri čemer pa z metodo GMM dobimo boljše rezultate pri ločevanju enoglasja in večglasja. To je lahko posledica tega, da je vhodnih podatkov v GMM model več, saj lahko pri prečnem preverjanju zgradimo model z 90% vseh podatkov, ki so na voljo in nam ni treba toliko skrbeti za neuravnoteženost množic. Težave s tem so pri metodi SVM. Zaradi neuravnoteženosti vhodnih razredov za učenje modelov ne moremo izrabiti vseh podatkov, ki so na razpolago. Posledica je manjša zaloga učnih podatkov, ki je tako manjša kot pri GMM, saj uporabljamo manj segmentov na posnetek. Rešitev tega problema bi bila dodaja uteži pri izgradnji SVM modela in uporaba krajših segmentov.

Ločevanje petja in petja s spremljavo lahko dokaj natančno realiziramo z obema metodama. Pri obeh se najbolje obnese vektor značilnic z 20 koeficienti MFCC in ΔMFCC, kar potrjuje tezo iz [20], da je med vokalom in inštrumentalom opazna razlika v spektralni distribuciji, ki se jo z 20 koeficienti MFCC da zajeti. Pri ločevanju enoglasnega in večglasnega petja pa smo z obema metodama dobili najboljše rezultate pri uporabi majhnega števila koeficientov MFCC, kar pomeni, da v višjih MFCC-jih razlike med enoglasnim in večglasnim petjem niso več tako dobro vidne.

4.4.1. Primerjava s sistemom Etnomuza

Napisal sem, da sistem deluje dokaj dobro. Za ugotovitev, kako dobro, ga lahko morda primerjamo s klasifikatorjem, objavljenim v članku *Etnomuza*[12]. Ker v tem članku ne ločujejo pevskih posnetkov in pevskih s spremljavo, lahko primerjamo le delovanje pri ločevanju enoglasnega in večglasnega petja. V članku je predstavljena matrika razvrščanj in razvidno je, da klasifikator enoglasne posnetke razvrsti med enoglasne v 62% primerov ter med večglasne v 24% primerov. Večglasne pravilno razvrsti v 82% primerov, med enoglasne pa jih uvrsti v 10% primerov. Seveda velja klasifikacija za razvrstitev v 5 razredov, tako da moramo natančnosti prevesti tako, kot če bi se sistem odločal le med dvema razredoma.
S pretvorbo dobimo za klasifikator iz članka natančnosti, ki so prikazane v tabeli 4-10.

	Dejansko	
Razvrščeno	En glas	Večglasno
En glas	**71,8%**	11%
Večglasno	28,2%	**89%**

Tabela 4-11: Matrika razvrščanj pri klasifikatorju iz *Etnomuza*

V odstotke pretvorimo še rezultate naše GMM metode pri uporabi 4 koeficientov MFCC in ΔMFCC. Ob neuporabi derivatov bi dobili bolj enakomeren rezultat, ampak ker so v *Etnomuza* tudi uporabljeni derivati koeficientov MFCC, lahko metodi s tem lažje primerjamo.

	Dejansko	
Razvrščeno	En glas	Večglasno
En glas	**56,5%**	11,9%
Večglasno	43,5%	**88,1%**

Tabela 4-12: Matrika razvrščanj pri našem klasifikatorju GMM

Vidimo, da sta pri razvrščanju večglasnih posnetkov klasifikatorja skoraj poravnana, razlika pa je pri razvrščanju enoglasnih, kjer se je naš odrezal slabše. V *Etnomuza* so za klasifikacijo uporabili varianco koeficientov MFCC, jaz pa sem uporabil kar vrednosti same. Iz tega bi lahko sklepal, da razlike v varianci koeficientov MFCC bolje ločijo enoglasne posnetke od večglasnih. Poleg tega je v *Etnomuza* bilo uporabljenih še nekaj dodatnih značilnic, kot so RMS energija in spektralna entropija, ki so, poleg ločevanja od ostalih razredov, najbrž pozitivno vplivali tudi na klasifikacijo med enoglasnimi in večglasnimi posnetki.

5. Zaključek

V okviru naloge sem preučil pristope za klasifikacijo predvsem pevskih posnetkov ljudske glasbe. Za klasifikacijo se v večini primerov kot značilnice uporabljajo koeficienti MFCC. Te značilnice sem uporabil za implementacijo sistema za klasifikacijo eno- in večglasnih posnetkov ter pevskih in posnetkov z glasbeno spremljavo. Na podlagi klasifikatorjev SVM in GMM sem dobil rezultate, ki so skladni z rezultati v obstoječi literaturi. Natančnost klasifikacije je tudi primerljiva z obstoječimi sistemi, ni pa najboljša. S tem sem pokazal, da so koeficienti MFCC zelo močne značilnice, vendar bi za izboljšavo algoritmov bilo potrebno obstoječe značilnice združiti še s katerimi.

Kot ideje za izboljšave predlagam algoritme za zaznavo osnovnih frekvenc v večglasnih posnetkih, ki bi preko analize spektrograma zvoka ugotavljala, koliko melodij je na posnetku in s tem pomagala razvrstiti enoglasne in polifonične posnetke. Morda bi lahko enako vlogo opravljali algoritmi za določanje melodije v večglasnih posnetkih.

Tudi sicer ponuja diplomska naloga še več možnosti za nadgradnjo. Za začetek nadaljnjega dela bi bilo dobro bazo posnetkov še dopolniti z inštrumentalnimi posnetki, da bi se lahko algoritmi razširili na več klasifikacijskih razredov. Z večjo množico inštrumentalnih posnetkov bi lahko bolj natančno preučili ločevanje inštrumentalne glasbe in petja z inštrumentalno spremljavo. Za to nalogo bi morda lahko uporabili harmonične koeficiente in njihove 4-Hz modulacijske vrednosti, preizkusiti pa bi veljalo tudi avtoregresijsko funkcijo, ki se je za tovrstne probleme že izkazala.

Baza posnetkov, ki sem jo v okviru naloge izgradil, je anotirana tudi z atributi spola pevcev na posnetkih, zato bi se v nadaljnjem delu lahko preučili tudi algoritmi za določevanje spola in preizkusili na obstoječi bazi. Bržkone bi bilo treba pri tem problemu izhajati iz algoritmov za analizo govora in ne glasbe, kar pa ni nujno slabo, saj so tudi eni najbolj uporabnih značilnic za analizo glasbe, koeficienti MFCC, prišli v glasbene vode z govornega področja. Morda jih nasledniki že čakajo tam.

6. Seznam slik

7. Seznam tabel

8. Literatura

[1] M. Bartsch, *"Automatic Singer Identification in Polyphonic Music"*, doktorska disertacija, Univerza Michigan, oddelek za elektrotehniko, ZDA, 2004

[2] M. Carey, E. Parris, H. Thomas, *"A Comparison of Features for Speech, Music Discrimination"*, v zborniku *IEEE International Conference on Acoustics, Speech, Signal Processing*, Phoenix, ZDA, 1999

[3] W. Chou, L. Gu,*" Robust Singing Detection in Speech/Music Discriminator Design"*, v zborniku *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001)*, Utah, ZDA, 2001, str. 865-868

[4] S. Essid, G. Richard, *"Musical Instrument Recognition by Pairwise Classification Strategies"*, *IEEE transactions on audio, speech and language processing*, zv. 14, št. 4, 2006

[5] H. Fujihara, M. Goto, T. Kitahara, HG. Okuno, *"A Modeling of Singing Voice Robust to Accompaniment Sounds and Its Application to Singer Identification and Vocal-Timbre-Similarity-Based Music Information Retrieval"*, *IEEE transactions on Audio, Speech and Language Processing*, zv. 18, št. 3, 2010

[6] A. Holzapfel, Y. Stylianou, *"Singer Identification in Rembetiko Music"*, v zborniku *Sound and Music Computing (SMC '07)*, Lefkada, Grčija, 2007

[7] G. Kumar, K. Raju, *"Speaker recognition using GMM",International Journal of Engineering Science and Technology*, zv. 2, 2010

[8] T. Lidy, C. Silla, O. Cornelis, F. Gouyon, A. Rauber, C. Kaestner, A. Koerich, *"On the Suitability of State-of-the-art Music Information Retrieval Methods for Analyzing, Categorizing and Accessing Non-Western and Ethnic Music Collections"*, Signal Processing, 2009, str. 1032-1048

[9] B. Logan, *"Mel Frequency Cepstral Coefficients for Music Modeling"*, v zborniku *International Symposium on Music Information Retrieval (ISMIR)*, Massachusetts, ZDA, 2000

[10] H. Lukashevich, M. Gruhne, C. Dittmar, *"Effective Singing Voice Detection in Popular Music using Arma Filtering"*, v zborniku *10th International Conference on Digital Audio Effects (DAFx-07)*, Bordeaux, Francija, 2007

[11] M. Mandel, D. Ellis, *"Song-level features and support vector machines for music classification"*, v zborniku *6th ISMIR*, Anglija, 2005

[12] M. Marolt, *"Probalistic Segmentation and Labeling of Ethnomusicological Field Recordings"*, v zborniku *10th International Society for Music Infortmation Retrieval Conference (ISMIR09)*, Kobe, Japonska, 2009, str. 75-80

[13] M. Marolt, J.F. Vratanar, G. Strle, *"Ethnomuse: Archiving Folk Music and Dance Culture"*, v zborniku *Eurocon 2009*, St. Petersburg, Rusija, 2009

[14] J. Marques, P. Moreno, *"A study of Musical Instrument Classification using Gaussian Mixture Models and Support Vector Machines"*, tehnično poročilo, *Cambridge Research Laboratory*, Masshachusetts, ZDA, 1999

[15] T. Nwe, A. Shenoy, Y. Wang, *" Singing Voice Detection in Popular Music"*, v zborniku *12th annual ACM international conference on Multimedia*, New York, ZDA, 2004

[16] Y. Ohishi, M. Goto, K. Itou, K. Takeda, *"Discrimination between Singing and Speaking Voices"*, v zborniku *Interspeech05*, Lizbona, Portugalska, 2005, str. 1141-1144

[17] V. Ourania, *"Singing Phoneme Class Detection in Polyphonic Music Recordings"*, magistrska naloga, Univerza Pompeu Fabra, oddelek za informacijsko in komunikacijsko tehnologijo, Barcelona, Španija, 2008

[18] P. Proutskova, M. Casey, *"You Call That Singing? Ensemble Classification for Multi-Cultural Collections of Music Recordings"*, v zborniku *10th International Society for Music Infortmation Retrieval Conference (ISMIR09),* Kobe, Japonska, 2009, str. 759-764

[19] W. Tsai, S. Liao, C. Lai, *"Automatic Identification of Simultaneous Singers in Duet Recordings"*, v zborniku *9th International Conference on Music Information Retrieval (ISMIR '08)*, 2008

[20] W. Tsai, H. Wang, *"Automatic Singer Recognition of Popular Music Recordings via Estimation and Modeling of Solo Vocal Signals"*, *IEEE Transactions on Speech and Audio Processing*, zv. 14, št. 1, 2006

[21] G. Tzanetakis. P. Cook, *"Musical Genre Classification of Audio Signals"*, *IEEE Transactions on Speech and Audio Processing*, zv. 10, št. 5, 2002, str. 293-302

[22] *Cross-validation, dostopno na:* http://en.wikipedia.org/wiki/Cross-validation_(statistics)

[23] *Auditory toolbox: A Matlab toolbox for Auditory Modeling Work*, dostopno na: http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/

[24] *Voicebox:Speech Processing Toolbox for Matlab*, dostopno na: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[25] *Netlab: Algorithms for Pattern Recognition*, dostopno na: http://www.mathworks.com/matlabcentral/fileexchange/2654-netlab