Signalbehandling til lydsøgning

Torbjørn Andreas Lisberg

Kongens Lyngby 2007
IMM-THESIS-2007-36

Danmarks Tekniske Universitet
Informatik og Matematisk Modellering
Bygning 321, DK-2800 kongens Lyngby, Danmark
Telefon +45 45253351, Fax +45 45882673
reception@imm.dtu.dk
www.imm.dtu.dk

IMM-THESIS: ISSN 09009-3192

Resume

Denne afhandling omhandler klassifikation af musiksignaler. Der undersøges hvorvidt det er muligt, at benytte en rytmisk feature, baseret på trommesignalet fra et polyfonisk musiksignal, til at klassificere det pågældende musiksignal efter tilhørende musikgenre.
Der undersøges to metoder til, at udføre blind separation af kildesignaler fra et enkeltkanals polyfonisk musiksignal, hvor der er specielt fokus på at kunne separere trommesignalet fra det oprindelige signal. Der undersøges også to metoder til automatisk identifikation af de separerede kildekomponenter der tilhører trommesignalet.
Der udtrækkes korttidslige features fra trommesignalet, hvor der benyttes en Multivariabel autoregressiv model, til at integrere de korttidslige features op på en længere tidsskala. De korttidslige features, der undersøges i dette projekt består av såkaldte MPEG-7 features og Mel Cepstrale koefficienter. Der undersøges også en kernel model til tidslig feature integration.
Der udføres en række forsøg, hvor de introducerede metoder til separation af kildesignaler, testes for at finde ud af hvilken metode giver den bedste kvalitet for de separerede trommesignaler.
Derefter benyttes trommesignalet til at udtrække features fra og der undersøges hvor velegnede disse features er til at klassificere musiksignaler efter musikgenre.
Af de to metoder der benyttes til separation af kildesignaler, fremkommer de bedste resultater, ved at benytte en perceptionelt vægtet ikke-negativ faktorisering (PWNMF). Der findes også frem til en forholdsvis robust metode til at identificere de kildekomponenter, der tilhører et trommesignal.
Til at undersøge klassifikationsperformance benyttes to musikdatabaser hvor den ene danner basis for de udførte tests. De features der giver bedst resultater benyttes derefter på den større database.
Der er også i projekt-forløbet blevet implementeret en Windows applikation, der demonstrerer nogle af de introducerede metoder i praksis. Applikationen er i stand til at anbefale brugeren musik ud fra valgte eksempler.

Forord

Denne afhandling er udarbejdet på Institut for Informatik og Matematisk Modellering ved Danmarks Tekniske Universitet og er sidste led i at opfylde kravene til titlen Cand. Polyt.

Forfatter er Torbjørn Andreas Lisberg (s991648).

Vejleder er Prof. Jan Larsen fra IMM.

Lyngby, april 2007

_____

Torbjørn Andreas Lisberg

Anerkendelse

Jeg vil gerne benytte denne lejlighed til at takke min vejleder Jan Larsen, for at tage sig tid til at svare på mine mange spørgsmål og komme med gode forslag til de forskellige problematikker, som jeg har haft forestående. Også en tak til Mikkel N. Schmidt for deltagelse i nogle af de indledende diskussioner. Jeg vil også gerne takke Anders Meng for lån af musikfiler.

Sidst men ikke mindst vil jeg takke min familie og mine venner for deres støtte og forståelse. En særlig tak til min dejlige kæreste Randi, for hendes uvurderlige støtte og optimisme igennem projektforløbet.

Indholdsfortegnelse

KAPITEL 1 15

INTRODUKTION 15
1.1 Opdeling af musik 2
1.2 Musik klassifikationssystem 3
1.3 Overblik 5

KAPITEL 2 6

MUSIK KLASSIFIKATION 6
2.1 Opfattelse af lyd 7
2.2 Ørets anatomi 8
2.3 Pitch 9
2.4 Loudness 10
2.5 Kritiske frekvensbånd 11
2.6 Opfattelse af musikgenre 12
2.7 Opsætning af forudsætninger 14

KAPITEL 3 15

MUSIK FEATURES 15
3.1 Valg af features 16
3.2 Korttidslige feature 18
3.3.1 Mel-frekvens cepstrale koefficienter (MFCC) 21
3.1.3 Zero Crossing Rate (ZCR) 22
3.1.5 Audio Spectrum Envelope (ASE) 23
3.1.6 Audio Spektrum Centroid 23
3.1.7 Audio Spektrum Spread 24
3.1.8 Spectral Flattness Measure 24
3.2 Tidslig feature integration 25
3.2.1 AR-features 25
3.3.2 Estitmation af parametre 27
3.2.3 MAR features 27
3.2.4 DAR features 28
3.4 Kernel model til feature behandling 28
3.3.1 Kernel Orthonormalized Partial Least Squares (KOPLS) 29
3.4 Diskussion 31

KAPITEL 4 32

SEPARATION AF KILDESIGNALER 32
4.1 Ikke-negativ matrice faktorisering (NMF) 34
4.2.1 Ikke-negativ faktorisering til separation af enkel-kanals polyfoniske lydsignaler 35
4.3 Perceptionelt vægtet NMF til separation af mono polyfoniske lydsignaler 38
4.3.1 Perceptionelt motiverede vægte 38
4.3.2 Vægtet ikke-negative faktorisering (WNMF) 41
4.4 Ikke-negativ matrix 2D faktorisering til separation af lydsignaler 41
4.5 Automatisk identifikation af trommesignaler 44

4.5.1 Trommedetektor 44

KAPITEL 5 51

KLASSIFIERS 51
5.1 Lineær klassifier 52
5.2 Generaliseret lineær classifier 53
5.2 Gaussisk klassifier 54
5.3 Gaussisk miksning classifier 56

KAPITEL 6 57

FORSØG OG RESULTATER 57
6.1 Separation af kildesignaler 58
6.1.2 Separation af et syntetisk signal 60
6.1.3 Perceptionelt vægtet NMF på syntetisk signal 62
6.1.4 NMF2D på et syntetisk signal 64
6.1.5 Automatisk identifikation af kildesignaler 66
6.2 Trommesignaler til klassifikation af musik genre 68
6.3 Diskussion 71

KAPITEL 7 75

WINDOWS APPLIKATION 75
7.1 Funktionalitet 76
7.2 Struktur 78
7.3 Implementering 79
7.3.1 Analysator 80
7.3.2 Front end 81
7.3.3 Externe data 81

KAPITEL 8 82

OPSUMMERING OG KONKLUSION 82
8.1 Opsummering 83
8.2 Konklusion 85
8.3 Fremtidigt arbejde 86

BILAG 1 87

KLASSIFIKATIONSRESULTATER 87
Referencer 91

Kapitel 1

Introduktion

I 2007 rapporterede repræsentanten for den globale musikindustri IFPI, at salget af digital musik på Internettet står i dag for 10 % af det samlede musik marked, med en markedsværdi på 2 milliarder dollars. Samtidig er antallet af tilgængelige musikfiler på Internettet i 2006 fordoblet til 4 millioner [41].
Dermed fortsætter det globale musikmarked med at gennemgå nogle af de store ændringer i forhold til forbrugerens indkøbsmønster, som den har oplevet over de sidste 10 år.

I dag er det muligt at gå på Internettet og hente musik på kun få sekunder. Samtidig forbliver opbevaringsmulighederne hele tiden forbedrede. Det er i dag muligt at opbevare tusindvis af musiknumre på en Pc og de bærbare mediaafspillere

forbedres hele tiden i takt med at de bliver billigere. Her kan nævnes, at på en iPod fra Apple, er det muligt at gemme op til 1200 musiknumre på en gang, også er det blevet mere almindeligt at benytte mobiltelefonen, som et alternativ til mp3-afspilleren.
Med disse nye og spændende muligheder, er der fulgt et øget fokus på automatiske klassifikationssystemer, der er i stand til at klassificere musik efter forbrugerens præferencer, der f.eks. kan være musikgenre eller stemning for en melodi.
Den voksende interesse for systemer af denne art, skyldes derfor i høj grad, at forhandleren af musik på Internettet, har et ønske om, at blive bedre til at vejlede og anbefale kunden ny musik. Et system med denne mulighed, vil have et stort forretningsmæssigt potentiale, men vil også være en attraktiv kundeservice, hvor de mange valgmuligheder ofte kan virke uoverskuelige.
Et automatisk klassifikationssystem vil også kunne være til stor gavn for den private forbruger, da det vil kunne assistere i, at holde styr på den musik, som forbrugeren har liggende på sine private drev.

Et af de nyere områder der benytter musikklassifikation til kommercielt brug, er Internet-tjenester der tilbyder forbrugeren ubegrænset adgang til musik klassificeret efter i forvejen valgte musiknumre. Den mest kendte leverandører af disse tjenester er i øjeblikket www.Pandora.com [42], hvor al musik bliver klassificeret manuelt af eksperter. Her ville et effektivt automatisk klassifikationssystem, uden tvivl være af meget stor værdi, da man ville kunne spare både tid og penge.
Tjenester af denne type, hvor forbrugeren kan streame radio og tv direkte ind på en computer, står i stadig vækst og ny muligheder præsenteres med jævne mellemrum. I dag tilbyder de fleste større radio og tv kanaler streamingstjenester.
Flere af de store computersystem-udbydere, har også vist stor interesse for forskning relateret til søgning efter musik. Her kan nævnes Microsoft Research, Sun Microsystems, Philips og HP-Invent, der sponsorerede sidste års internationale konference om søgning efter musik-information med navnet ” Symposium on Music Information Retrieval (ISMIR)” [11].

1.1 Opdeling af musik

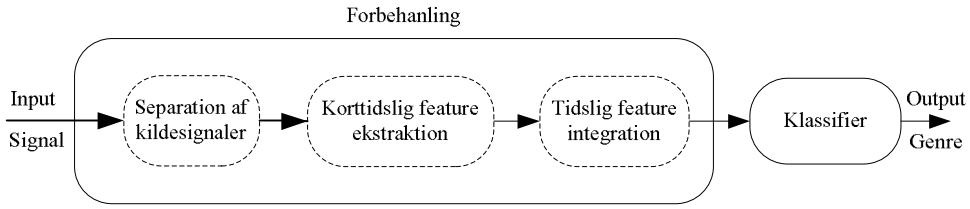
Der findes mange måder at dele musik op på. Der skelnes typisk imellem objektiv og subjektiv opdeling. En objektiv opdeling er typisk baseret på kunstnernavn, sprog, antal instrumenter etc. Dette er information, som ikke kan opfattes forskelligt af forskellige personer. Denne type information er ofte vedhæftet en musikudgivelse i form af metadata.
En subjektiv opdeling kan afhænge meget af den der betragter oplysningerne. Dette kan være stemning, genre eller tema for et musiknummer. Der kan være stor forskel på hvordan mennesker opfatter disse egenskaber. Kunstneren og lytteren kan have to forskellige opfattelser af f.eks. temaet for et musiknummer.

Ved en hurtig undersøgelse af forskellige musikportaler på Internettet, kan der nemt ses, at de mest almindelige kriterier til at dele musikken op efter, er kunster, titel, album og musikgenre. Her er de tre først nævnte objektive kriterier, imens musikgenre er et subjektivt kriterium.

Det er meget almindeligt for mennesker, at dele musikc op efter tilhørende musikgenre. Musikgenre gør det muligt at give en kort, men præcis beskrivelse af en kunstner eller musiknummer og vil for de fleste, give en klar forestilling af hvilken type musik der er talen om. Hvis et musiknummer beskrives som f.eks. at tilhøre Jazz-genren, da vil de fleste mennesker, få en ide om, hvilken type musik der er talen om, selv om de aldrig har hørt det pågældende musiknummer.

1.2 Musik klassifikationssystem

Det klassifikationssystem der benyttes i dette projekt, til klassifikation af musik, består af to hoveddele.
Første del udfører såkaldt *forbehandling*, hvor der udtrækkes nogle repræsentative data fra det musiksignal der ønskes klassificeret. Anden del består af en *klassifier*, der er i stand til at klassificere det pågældende musiksignal ud fra de data der er indkomne fra forbehandlingen.
.
Figur 1.1 nedenfor viser de forskellige dele af det klassifikationssystem, der benyttes i dette projekt.



Figur 1.1. Illustrerer strukturen af klassifikationssystemet, der benyttes i denne afhandling. Input består af lydsignal, der separeres ned i dets kildekomponenter. Kildekomponenterne benyttes derefter til at beregne de korttidslige features. Derefter integreres de korttidslige features op på en større tidsskala. Den fundne featurevektor benyttes som input til klassifieren, der siden hen estimerer de indkomne data. Til sidst efterbehandles de fundne data fra klassifieren og det endelige output dannes i form af estimeret musikgenre label.

De repræsentative data der udtrækkes under forbehandlingen, kaldes for *features*. De kan f.eks. bestå af det frekvensmæssige eller det rytmiske indhold for det

pågældende musiksignal. Features betragtes som en meget vigtig del af de systemer der forsøger at klassificere musik.

Der findes flere måder at repræsentere et stykke musik på. En musiker ville typisk benytte sig af noder. I denne afhandling benyttes det digitale lydsignal, der naturligt findes på computere og Internettet.

Denne afhandling omhandler klassifikation af musik, hvor der undersøges, hvorvidt det er muligt, at benytte trommesignalet fra et komplet musiksignal, til at klassificere det pågældende musiksignal efter. Dermed opnås en feature, der forsøger at udnytte den rytmiske kontekst for et musiksignal, til at genkende eller adskille det pågældende musiksignal ud fra en større mængde musiksignaler. En feature af denne art, kaldes normalt for en rytmisk feature.

En rytmisk features af denne art kan have flere applikations-muligheder. I et system der søger efter musik i en database, kan en velfungerende rytmisk feature, fungere som et stærkt søge-kriterium, hvor der søges efter musiknumre med bestemte rytmiske mønstre eller stemning.

Musikgenre er et ofte benyttet kriterium, til at klassificere musik efter, i et automatiseret klassifikationssystem [11,12]. Dette synes også naturligt, i betragtning af de før nævnte overvejelser.

Denne afhandling omhandler automatisk klassifikation af musik, hvor der benyttes musikgenre til at klassificere musiksignaler efter. Der fokuseres hovedsagligt på udnytte en rytmisk feature til klassifikation.

Separation af kildesignaler er den del af processen, hvor det indkomne digitale lydsignal separeres ned i dets kildesignaler, der består af lydsignalerne fra de musikinstrumenter tilstede i det pågældende polyfoniske musiksignal

I denne afhandling er feature-delen delt op i separation af kildesignaler, hvor trommesignaler sorteres fra de separerede kildesignaler, efterfulgt af ekstraktoin af korttidslige features fra trommesignaler og til sidst tidslig feature integration.

De korttidslige features beregnes ud fra et separet kildesignal over tidsrammer på 10-40 ms. De opfanger dermed kun den information, der lever på en kortere tidsskala.
Derfor benyttes tidslig feature integration, til at kombinere de korttidslige features op på en længere tidsskala. Dermed fås en mere realistisk repræsentation af den information, der lever på en længere tidsskala, f.eks. det rytmiske indhold.

Efter at den pågældende featurevektor er fundet, kan den benyttes som input til klassifieren.

Klassifieren vil forsøge, at klassificere musiksignalet, baseret på de indkomne features. En klassifier kan bestå af en lineær model, der modellerer de indkomne featuredata.
Givet et træningsæt, vil det være muligt at estimere parametrene for klassifieren. Derefter kan de fundne parametre, benyttes til at estimere nye indkomne data.

Denne afhandling fokuserer hovedsagligt på at undersøge hvorvidt det er muligt at benytte trommesignalet fra et musiksignal, til at danne basis for en rytmisk feature, der kan indgå som del af et musikgenre klassifikation system

1.3 Overblik

Kapitel 2 introducerer nogle grundlæggende psykoakustiske begreber, samt diskuterer opfattelse af musik og musigenre

Kapitel 3 beskriver de features der benyttes i dette projekt. Beskrivelsen omfatter korrtidslige features, tidslig feature integration

Kapitel 4 beskriver separation at musiksignaler og hvordan et separeret trommesignal kan udgøre en rytmisk feature i et klassifikationssystem.

Kapitel 5 forklarer klassifiers og hvordan to klassifiers benyttes i dette projekt.

Kapitel 6 beskriver og diskuterer de forsøgsresultater der opnåede ved brug af de valgte metoder.

Kapitel 7 giver en beskrivelse af den Windows applikation, der er implementeret til demonstration af nogle af de introducerede metoder i praksis.

Kapitel 8 diskuterer de opnåede resultater og konklusion.

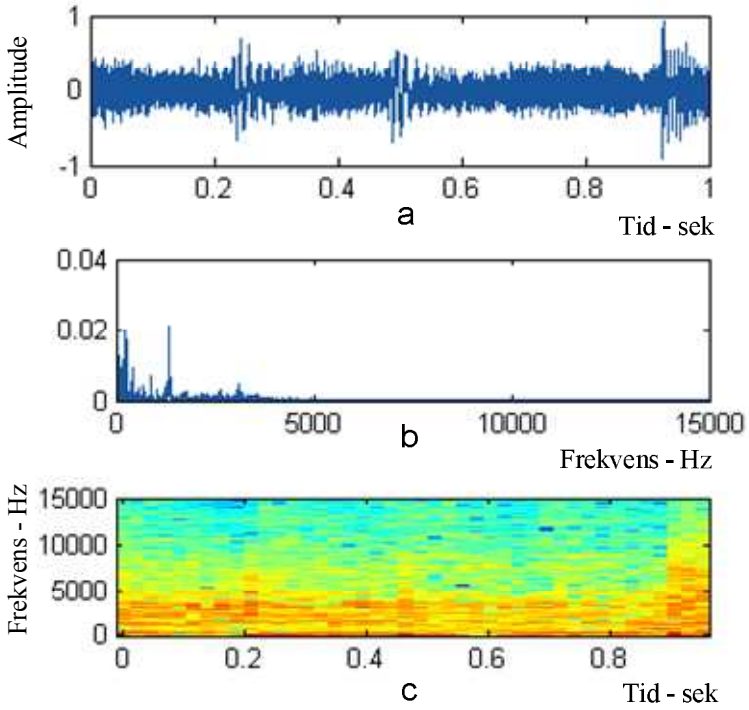
Kapitel 2

Musik klassifikation

Dette kapitel omhandler nogle af de basale egenskaber, der gælder for musik og opfattelse af musik.
Der gives også en kort beskrivelse af hvordan den menneskelige hørelse fungerer, som vil danne grundlag for nogle af de beregningsmodeller der benyttes.

Fra et matematisk synspunkt består digitaliseret musik af en række diskrete talværdier.
I fysiologisk forstand er musik lufttryksændringer, der opfattes af øret til videre behandling af det auditive system.

Musik opstår ved at række toner formes ud fra en eller flere lydkilder, til at danne en lydmæssig struktur. Denne struktur kan beskrives ved hjælp af amplitude eller frekvens som vist i henholdsvis figur 2.1.a og 2.1.b.
I flere tilfælde kan det være en fordel at beskrive lydsignalet ved hjælp af et spektrogram som vist i figur 2.1.c. Spektrogrammet afbilder frekvensen hen over tiden.



Figur 2.1 Det øverste plot viser amplituden for et musiksignal i forhold til tiden. Det midterste plot viser frekvensspektrum. Det nederste plot er et spektrogram – frekvens i forhold til tiden.

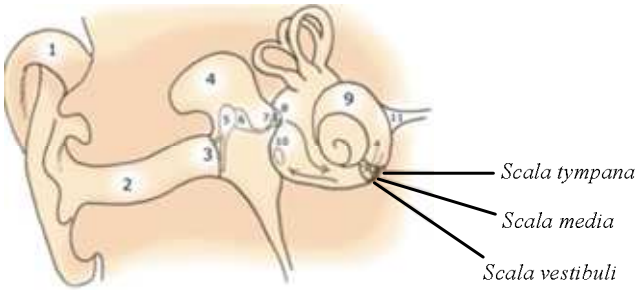
2.1 Opfattelse af lyd

Alle de steder hvor menneskets bevæger sig hen i samfundet, er vi omgivet af lyd i forskellige former. Ofte er det sådan at vi ikke bemærker de lyde vi omgiver os med. Dette kan være fordi vi er vant til at høre den pågældende lyd og derfor ikke lægger mærke til den mere. En anden årsag kan være, at vi ikke er i stand til at høre lyden, fordi vores høresystem ikke er i stand til at opfatte den. Her kan nævnes den velkendte hundefløjte, som ligger udenfor menneskets hørbare rækkevidde.

2.2 Ørets anatomi

Menneskets hørbare rækkevidde afhænger af de fysiologiske begrænsninger som vores høresystem sætter.
Det ydre øre, er den synlige del af øret, som sidder uden på hovedet. Den muslinge-lignende form gør, at det ydre øre fungerer som en tragt, der opfanger lydene og leder dem ind til mellemøret via øregangen. Ørets opbygning er illustreret i fig. 2.2.

1. Ydre øre
2. Øregangen
3. Trommehinden
4. Mellemøret
5. Hammeren (*Mallus*)
6. Ambolten (*Incus*)
7. Stigbøjlen (*Stapes*)
8. Det ovale vindue
9. Sneglen
10. Det runde vindue
11. Nerve



Figur 2.2 Viser ørets anatomi. De forskellige indeks refererer til den pågældende legemsdels benævnelse (er modificeret fra [32]).

Hovedet, torso og det ydre øre kan betragtes som et filter, der filtrerer lyden ved hjælp af maskning og refleksion.

Øregangen ligger i forlængelse af det ydre øre og er et ca. 2,5 cm langt hudbeklædt rør. Længden på øregangen svarer til ca. en fjerdedel af bølgelængden for frekvenser omkring de 4000 Hz. Derfor fremhæves ørets følsomhed i forhold til disse frekvenser [37].

For enden af øregangen sidder trommehinden, som er en membran der adskiller øregangen fra mellemøret. Når der er lyd tilstede, vil trommehinden svinge med lydens frekvens. Disse svingninger vil overføres til mellemøret.

Mellemøret er et luftfyldt hulrum mellem trommehinden og øresneglen. Mellemøret indeholder de tre små knogler kaldet hammeren, ambolten og stigbøjlen. Det er disse knogler der overfører lyden igennem mellemøret [32].

I det indre øre omformes lydbølgen til nervesignaler, der sendes til hjernen via hørenerven.

Det indre øre er et hult, væskefyldt 32 mm langt knogleorgan snoet som et sneglehus. Derfor kaldes det for øresneglen eller *cochlea*. Øresneglen indeholder de sansetråde, som opfatter lydimpulserne, ved at omsætte dem til elektriske nerveimpulser.

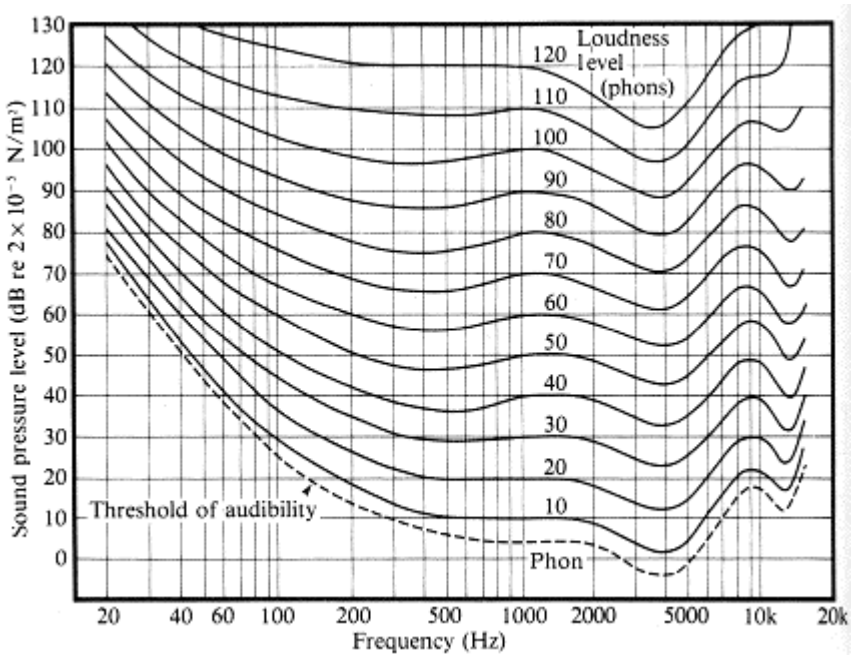
Sansetrådene sidder på en membran kaldet basilarmembranen, der er omgivet af væske. Lydimpulserne forplantes til væsken, der bevæger sig og påvirker sansetrådene. Sanseindtrykket bliver sendt med hørenerven til hjernen hvor det til sidst bliver opfattet som lyd. Et menneskets hørbare rækkevidde ligger i gennemsnit imellem 20 og 20.000 Hz [37].

2.3 Pitch

Pitch benyttes som benævnelse for den opfattede frekvens fra et lydsignal. Enheden for pitch er *mel*. For en tone på 1 kHz, er et pitch pr. definition lig med 1000 mel. Der er udført psykoakustiske forsøg med henblik på at relatere pitch til frekvens.
Forsøgene foregår ved at en række personer hører en 1 kHz tone, med en pitch på 1000 mels. Toenens frekvens ændres og deltagerne forsøger bedst muligt at vurdere pitch for den nye tone. Derefter findes den gennemsnitlige pitch i forhold til den virkelige frekvens. Det er dog stadig uvist hvordan øret opfatter pitch.
Selv om der afspilles den samme pitch node på forskellige musikinstrumenter, så kan fordelingen i frekvens domænet være meget forskellig. F.eks. for det franske horn, er under 0,4% af effekten for bestemte toner samlet ved pitch frekvenserne .
De højere pitch toner, fra forskellige instrumenter, forekommer at lyde mere ens end de ved lavere pitch toner. Dette er fordi at overtonerne fra de højere pitch toner, har en tendens til at falde udenfor menneskets hørbare rækkevidde. Det er strukturen på de hørbare overtoner der får instrumenter til at lyde forskelligt [8].

2.4 Loudness

Begrebet *loudness*, benyttes til at beskrive den opfattede intensitet for et lydsignals. Øret er, som nævnt før, mere følsomt overfor visse frekvenser end andre. Da den menneskelige hørelse kan variere meget fra en person til anden, er der er blevet indsamlede en række psykoakustiske data for et stort antal mennesker. Ud fra disse data er der fundet en gennemsnits frekvens respons for det menneskelige øre i forhold til lydtrykket. Resultatet er afbildet i den såkaldte *equal loudness kurver* (ELC) vist i figur 2.3.



Figur 2.3 Equal loudness kurven for rene toner [38].

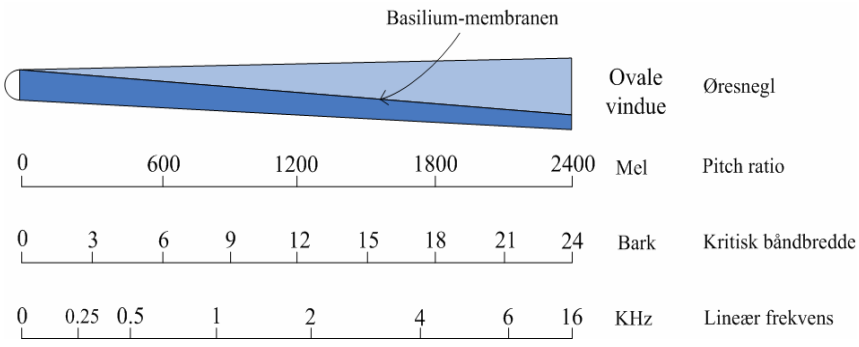
Equal loudness kurverne afbilder lydtrykket som funktion af frekvensen for toner, der opfattes at have den samme loudness. Loudness niveauet for hver enkel kurve er lig med lydtrykket ved 1 kHz. Dette kaldes også for *phon* niveauet for den enkelte kurve.
De kurver der ligger på et lavere niveau, tiltager meget mere, ved lave frekvenser end ved høje frekvenser. Dette er fordi at øret er mindre sensitivt ved lave frekvenser end ved høje frekvenser ved disse lave niveauer. Dette er årsagen til at bas frekvenser i musik lyder mere uklare ved lavere volumen [8].

2.5 Kritiske frekvensbånd

Forskellen imellem frekvenserne på to rene toner, hvor følelsen af ”ruhed” er bort og tonerne lyder fladede ud, kaldes for kritiske bånd. Ved lave frekvenser har man fundet ud af, at disse kritiske bånd viser en næsten konstant brede rundt om 100 Hz, imens for 500 Hz er båndbredden ca. 20 % af center frekvensen [37]..
Til at afbilde de kritiske bånd er der blevet defineret en såkaldt Bark-enhed. Frekvensspektret for et signal kan overføres til bark-skalaen ved hjælp af

$$b(f) = 13\,\mathbf{arctan}(0{,}00076\,f) + 3{,}5\,\mathbf{arctan}((f/7500)^2) \qquad (2{,}1)$$

De mekaniske egenskaber for øresneglen, der gør at disse egenskaber kan modelleres ved hjæp af en filterbank, hvor båndbredden, givet en lineær bevægelse bort fra øresneglens åbning, er omtrent logaritmisk aftagende [37].
Figur 2.3 illustrerer Bark og mel skalaen i forhold til basilarmembranen i udstrakt form.



Figur 2.4 Mel og Bark skalaerne i forhold til den udstrakte ørsnegl og den lineære frekvens skala. Mel og Bark er lineært afbildet imens frekvensen vises på en ulineær skala. Mel er en logaritmisk frekvens skala der er baserede på den menneskelige opfattelse af pitch (figuren er tilpasset fra [37]).

2.6 Opfattelse af musikgenre

I den del af litteraturen, der omhandler musikgenre klassifikation, synes der at være enighed om, at det auditive system har stor betydning for hvordan mennesker sorterer imellem musikgenrer.

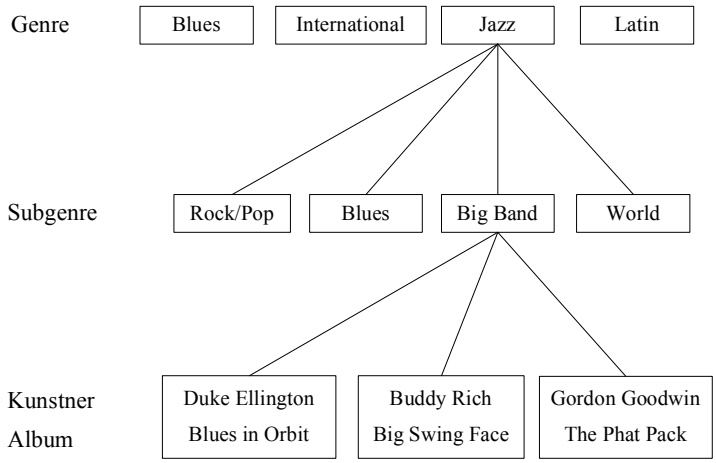
Mennesket opfanger igennem hele sit liv sanseindtryk, som lagres i hukommelsen, dette gælder også for lydindtryk.
Hvordan mennesker opfatter og fortolker et lydsignal, afhænger dels af hjernens hukommelse om det pågældende lydsignal.
Denne måde som hjernen kan huske en lyd på, gælder også for musik. Derfor kan forventes at den kulturelle og sociale baggrund, for en persons opfattelse af musikgenre, kan påvirke udkommet af et klassifikations-scenarium.

For mennesker er det ofte i ungdomsårene, at vi begynder at blive bevidste om musikalske præferencer. Det er meget almindeligt, at en teenager identificerer sig med en bestemt musikgenre.
Da er det ikke afgørende, om det pågældende musiknummer er melodisk fangende, men mere om det tilhører en bestemt genre, der ligger inden for de acceptable livsstil-rammer. Dette fænomen var særdeles tydeligt i 1960-70'erne da Elvis Presley og senere Beatles og Rolling Stones var i sin højde. Da opstod nærmest en helt ny ungdomskultur i forlængelse af musikkulturen.

I [11] fremhæves også at det sociale miljø kan have stor indvirkning på, hvordan en person skelner imellem forskellige musik-genrer. På den anden side påpeges også, at dette ikke nødvendigvis er altafgørende, idet at undersøgelser har vist, at musikere der ikke har været udsat for vestlig musik tidligere, alligevel kan sortere imellem de forskellige genrer. Alligevel viste undersøgelsen, at vestlige musikere og ikke-musikere, var bedre til at sortere imellem genrerne end de ikke-vestlige musikere. Dette indikerer, at man ikke nødvendigvis behøver at kende de forskellige musikgenrer, for at kunne skelne imellem dem, men at det alligevel kan være er fordel, hvis resultatet skal være optimalt.
Menneskers måde at klassificer musik efter genre, er en subjektiv vurdering og kan variere meget fra en person til en anden. Klassifikationen kan også i høj grad afhænge af hvem kunstneren er og hvilken genre denne normalt forbindes med. De fleste kunstnere forbindes med en bestemt musikgenre og derfor vil denne genre, for de fleste mennesker, være den første mulighed, der tages i betragtning, når musikken skal klassificeres (hvis kunstneren er kendt i forvejen). Dette synes også at være en accepteret "sandhed" for mange musikportaler, hvor den vedhæftede

genre-oplysning, minder mere om at være delt op efter kunstneren end den egentlige genre.

Flere af de musikportaler der findes på Internettet, benytter sig af at dele de forskellige musikgenrer op i subgenrer (undergrupper), her kan nævnes Amazon Musik Store og Cd Univers. Subgenrer kan ses som en udvidelse af de øverste genrer i hierarkiet. Andre musikportaler benytter sig af andre måder at vejlede kunden. Itunes.com har f.eks. en undergruppe der oplyser medarbejdernes foretrukne musik. Figur 2.5 nedenfor viser et lille udsnit af genre hierarkiet hos Cd Universe [38]. Her ses bl.a. at en hovedgenre som Blues, også kan fungere som en subgenre.



Figur 2.5 Viser et lille udsnit af genre hierarkiet fra **www.CdUniverse.com**. Yderst til venstre vises hvad de forksllige niveauer i hierarkiet indikerer: genre, subgenre og kunstner/album indikeret.

Selv om musikgenre deles op efter nogle stramme forudsætninger, så vil der altid være overlap genrerne imellem. Dette er hovedsagligt på grund af musikkens natur, da der findes et indbygget overlap. De fleste komponister lader sig inspirere af andres kompositioner, hvoraf en del ofte tilhører forskellige musikgenrer. Derfor vil der fra komponistens side være tilføjet elementer til den pågældende komposition, der oprindeligt kan stamme fra andre genrer. Et meget markant eksempel på dette er heavy metal musikere, der slår sig sammen med symfoniorkestre, for at omdanne deres oprindelige kompositioner, til at være en blanding af klassisk og heavy metal musik.

Selv om begrebet musikgenre kan virke lidt diffust, så forekommer det at være en accepteret og meget brugt opdeling musik og benyttes bl.a. af pladeselskaber, biblioteker og musikbutikker. Ligeledes benytter den private bruger sig også af musikgenre til at dele sin private musiksamling op efter. Derfor vurderes der at eksistere et grundlag for at klassificere musik efter musikgenre i et automatisk klassifikationssystem.

2.7 Opsætning af forudsætninger

Til at benytte et klassifikationssystem til musikgenre, er det nødvendigt, at opstille nogle indledende specifikationer, som benyttes til videre udvikling og test af systemet.

I dette projekt benyttes overvåget indlæring, hvor et klassifikationssystem trænes op til at kunne klassificere fremtidige musiknumre efter genre. Et system af denne type kræver, at hvert musiknummer eller lydklip har et tilhørende genrelabel, som antages at repræsentere den sande genre.

Det antages at et musiksignal kun kan tilhøre én musikgenre. Dermed benyttes en flad opdeling i forhold til genrehierarkiet vist før i figur 2.3. Genrerne antages også at have den samme indbyrdes afstand. Derfor antages en ligeligt fordelt sandsynlighed for at klassificere en genre forkert. Dette kan dog ikke siges at afspejle virkeligheden, da nogle genrer ligger tættere på hinanden end andre.

Der benyttes kun musik fra enten wav-filer eller MP3-filer konverterede til wav-filer med en samplingsfrekvens på 44100 Hz. Stereosignaler omdannes til Monosignaler.
Der efterstræbes at den musik der benyttes afspejler den musik, som den almindelige bruger lytter til.

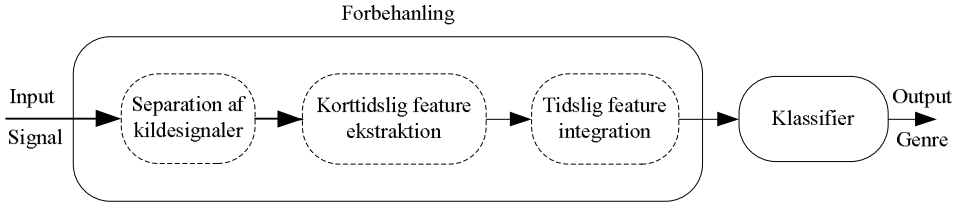
Kapitel 3

Musik features

For at et klassifikationssystem skal kunne klassificere et musiksignal optimalt, er det nødvendigt med nogle inputs, der virker beskrivende for det pågældende signal. Derfor gennemgår signalet en *forbehandlingsproces*, hvor der udtrækkes nogle repræsentative data fra signalet. Disse data kaldes, som nævnt før, for *features*.

Den del af projektet der omhandler features, er delt op i tre dele, der samlet indgår i forbehandlingsprocessen for klassifikationssystemet vist i fig. 3.1.

Dette kapitel giver indledelsesvis en generel beskrivelse af features og valg af features. Her forklares også hvordan *blind separation af et musiksignal* kan indgå som del af en klassifikationsproces med særlig fokus på trommesignalet. Derefter beskrives henholdsvis korttidslig feature ekstraktion og tidslig feature integration.



Figur 3.1 Dette kapitel beskriver korttidslig features integration og tidslig features integration for klassifikationssystemet som vist på figuren.

3.1 Valg af features

Features kan bestå af enkelte eller flere forskellige slags talværdier der repræsenterer et musiksignal der ønskes klassificeret. En feature kan bestå af f.eks. middelværdien for det pågældende datasæt, men den kan også bestå af værdier opnåede ved komplicerede kalkulationer af data-sættet.
En velegnet feature skal virke diskriminere. Derfor er det vigtigt, at to features tilhørende den samme musikgenre, ligger tæt på hinanden i feature-rummet. Omvendt skal to features der ikke tilhører samme genre, gerne ligge langt væk fra hinanden. I praksis vil dette betyde at features tilhørende f.eks. rockmusik og popmusik, er meget forskellige. Dette er dog sjældent tilfældet i praksis, da de typisk overlapper hinanden.
På nuværende tidspunkt eksisterer der ikke nogle features, som fuldstændig er i stand til at diskriminere musiksignaler efter genre. Alligevel betragtes features af meget stor betydning for performance af et klassifikationssystem, da nogle features er bedre velegnede end andre.
Flere af de features, der er udviklede i løbet af de sidste mange år, forsøger at opfange nogle af de perceptionelle kendetegn, der gælder for musiksignalet. Dermed forsøger de at efterligne det menneskelige høresystems opfattelse af musik (eller rettere sagt lyd). Dette har i flere tilfælde vist sig at give gode resultater.
Der findes også en række features, der umiddelbart ikke giver så megen mening for den menneskelige musikopfattelse, men alligevel kan give rimelige resultater, dette kunne f.eks. være variansen at lydsignalet.
Musik opstår ved en sammenblanding af lydsignaler, der udsendes fra de lydkilder til stede i den pågældende sang. Disse lydsignaler er i høj grad afhængige af hinanden, da de følger det samme tempo og er satte sammen til at uddanne en tonermæssig enhed, der målrettet er sat sammen af kunstneren til at opfattes på en bestemt måde af tilhøreren

Musikkomponisten har fuld kontrol over det lydbillede der dannes. Det er dog ikke sikkert, at komponisten har en interesse i, at at det pågældende musiksignal skal opfattes som at tilhøre en bestemt genre. Ligeledes er der ikke nogen garanti for, at komponisten har den samme opfattelse af sin musik, som publikum har.
Mennesket har dog en tendens til inddele deres opfattelse af musiksignaler. Da det er kildesignalerne der udgør det samlede lydbillede, så kan siges at kildesignalerne der struktureret i forhold til hinanden, afgør hvordan et musiksignal opfattes af den menneskelige hørelse. Derfor er det instrumenterne, der afgør hvordan et musikstykke lyder. Musikgenren er da kun en individuel fortolkning af denne lyd.

I musiker-verdenen benyttes musikteori til at beskrive hvordan et instrument spilles. Denne musikteori afspejler sig også i de forskellige musikgenrer. En musiker med en musikteoretisk baggrund, vil typisk kunne identificere akkorder og rytmiske mønstre som tilhørende en bestemt musikgenre elle subgenre. Dette ses blandt i benævnelserne af flere de kendte musikskalaer, hvor navne som jazzskala, bluesskala, sigøjnerskala, kinesisk skala etc er almindelig kendte.
Derfor er der nogle musikteoretiske elementer, som gør sig gældende for hvordan et musiknummer lyder og dermed ubevidst eller bevidst påvirker lytterens opfattelse af bl.a. musikgenre. Dog er det ikke sådan at de genre-baserede musikteoretiske begreber, kun benyttes indenfor den genre de referer til. Det er f.eks. meget almindeligt, at at høre en et rocknummer med musikteoretiske elementer fra andre genrer. At musikerne blander de forskellige stilarter sammen, må også antages at være en af hovedårsagerne til, at genrerne kan være svære at klassificere.
Det element, som måske har mest at sige, for hvordan et musiksignal lyder er slagtøjs instrumenterne. Det er slagtøjerne, der defininerer det rytmiske kontekst af en sang. Samtidig har trommerne ofte en let genkendelig måde at spille på for de enkelte genrer.
At der er forskel på hvordan slagtøjerne spilles i forhold til genren, ses også på akkompaniment-keyboards[1], hvor der f.eks. kan vælges rock som hovedgruppe og så eksisterer der en række rytmer at vælge imellem såsom som hard rock, slow rock og pop-rock, der afspejler subgenrerne.
At rytmiken har påvirken på genre-opfattelsen er også annerkendt i den forskningsorienterede verden, hvor man har benyttet rytmiske features bl.a. i [11,12]

[1] Disse bliver ofte referede til som ”enmandsorkestre” hvor alle eller flere intstrumenter er indbyggede i keyboardet, der bl.a. har en meget udbygget rytmesektion.

Selv om der vurderes at eksistere en forskel på hvordan trommerne spilles for de forskellige genrer, så forventes også et vist overlap genrene imellem. Dette kan bl.a. ses på dance og pop genrerne hvor måden slagtøjs-insturmenterne spilles på, ikke er så dingtistiv i forhold til hinanden.

At klassificere musik efter genre, drejer sig i høj grad om at finde forskelligheder imellem de forskellige musikgenrer. Der findes mange indgangsvinkler til hvordan dette kan gøres. De mest succecfulde metoder benytter lydklip udtaget fra det samlede musiksignal, for derefter at transformere det over i feature rummet.
Baseret på de overvejelser beskrevne før, vil der i dette projekt blive fokuseret på trommesignalet for det pågældende musiksignal.
Dermed vil der blive foretaget for hvert musiksignal en separation af kildesignalerne. Derefter vil der blive udtrukket korttidslige features fra trommesignlet, efterfulgt af tidslig feature integration. Ved brug af denne metode opnås en rytmisk feature, der er fundet ud fra selve fundamentet for det rymiske kontekst af et musiksignal.
Til forfatterens kendskab er separation af kildekomponenter ikke blevet benyttet før i forbindelse med musikgenre. Årsagen kunne være, at der ikke eksisterer nogen færdig løsning på, hvordan separation af enkeltkanals musiksignaler skal udføres. Der er dog sket nogle fremskridt på området indenfor de sidste to-tre år. Samtidig vurderes det ikke at være af afgørende betydning, om de separererede kildekomponenter er af kvalitet der er egnet til menneskelig lytning. Det vigtigste er at de afspejler, et musiksignals rytmiske kontekst, som er relevant for musikgenreklassifikation. Alligevel vurderes det at være vært at tilstræbe en så god kvalitet som muligt for trommesignalerne.
.

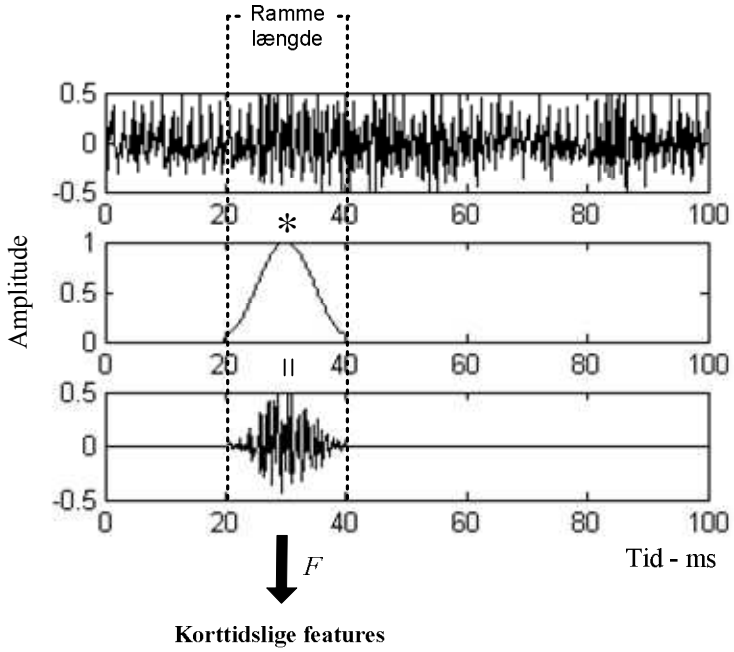
3.2 Korttidslige feature

Kort-tidslig feature ekstraktion, er den proces, hvor der udtrækkes nogle beskrivende data fra et eller flere segmenter af et digitalt lydsignal. Disse segmenter kaldes for *tidsrammer*. Ved brug af tidsrammer, antages at signalet er stationært indenfor rammernes grænser. I denne afhanling benyttes overlappende tidsrammer, hvor der for udtrækkes korttidslige features for hver enkel ramme.
Normalt multipliceres hver enkel tidsramme med en *vindue-funktion* før der udtrækkes features. Dette er hovedsageligt for at undgå spektral lækage så vidt muligt og dermed få den bedst mulige opløsning af frekvensspektret.

Kort-tidslig feature ekstraktion for signalet **s,** kan for en enkel tidsramme matematisk udtrykkes ved hjælp af featurevektoren $\mathbf{x}_n$ til tidsindeks *n* som

$$\mathbf{x}_n = \mathrm{F}\left(w_0\, \mathrm{s}_{n-(N-1)}, \ldots, w_{N-1}\, \mathrm{s}_n\right) \qquad (3,1)$$

hvor $w_0, w_1, \ldots, w_{N-1}$ er vindue-koefficienterne og *N* er længden på den enkelte ramme. Funktionen *F* repræsenterer den metode der benyttes til at finde de ønskede features. Beregning af features kan betragtes som en lineær eller ulineær transformation af $\mathbf{s}_n$ over i $\mathbf{x}_n$ ved hjælp af *F*. Processen er beskrevet for en enkel tidsramme i figur 3.2 nedenfor.

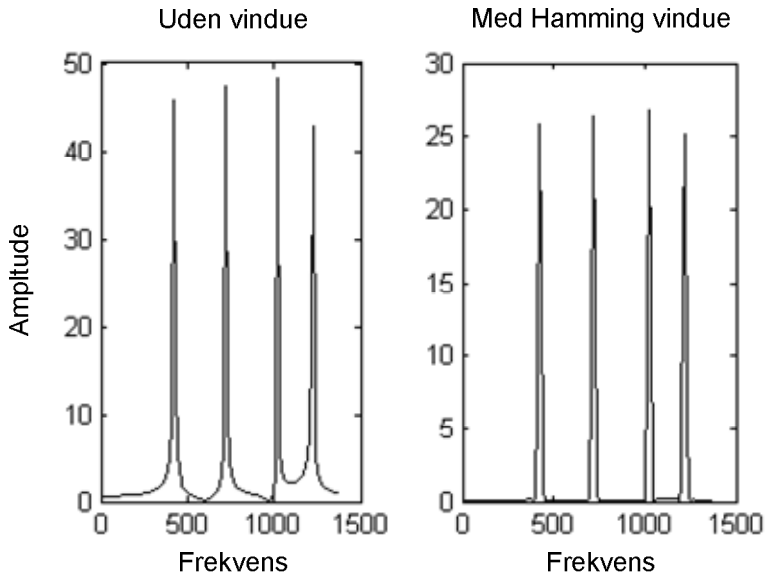


Figur 3.2 Illustration af kort-tidslig feature ekstraktion. Den øverste af de 3 delfigurer afbilder et ubehandlet digitalt musiksignal. Figuren markerer hvordan en enkel ramme på 20 ms udtrækkes fra signalet og multipliceret med et såkaldt Hamming vindue vist i den midterste delfigur. Det endelige indhold af rammen, er afbildet i den nederste delfigur. Det kan nemt ses, at signalet i den nederste delfigur gradvis aftager hen imod ydersiderne. Til sidst bliver indholdet af rammen transformeret af *F*, som derefter returnerer de korttidslige features. *F* kan f.eks. bestå af den diskrete Fourier transformation, hvor de absolutte værdier repræsenterer frekvenskoefficienterne for tidsrammen.

Der findes flere slags forskellige vinduefunktioner. I dette projekt benyttes et Hamming der kan udtrykkes ved

$$w_n = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right) \qquad n = 0,\ldots,N-1 \qquad (3{,}2)$$

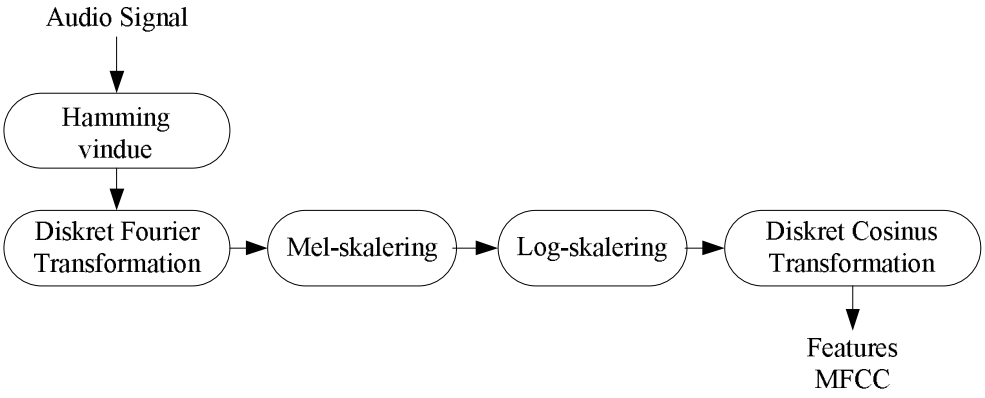
hvor *n* er antal samples. Figur 3.3 nedenfor viser et udklip fra frekvensspektret for et syntetisk sinusformet signal, før og efter det er blevet multipliceret med et Hamming vindue. Det fremgår klart af figuren hvordan den spektrale opløsning forbedres ved brug af vinduet.



Figur 3.3 Illustrerer effekten af et Hamming vindue. Figuren længst til venstre viser et udsnit af frekvensspektret for et syntetisk fremstillet sinusformet signal. Figuren længst til højre viser effekten af et Hamming vindue. Det fremgår klart hvordan et Hamming vindue forbedrer opløsningen af frekvensspektret. Læg også mærke til at vinduet formindsker styrken på signalet.

3.3.1 Mel-frekvens cepstrale koefficienter (MFCC)

Mel-frekvens cepstrale koefficienter (MFCC) er blevet brugt i en række sammenhænge til behandling af lydsignaler. Heraf er de benyttede som features til musik genre klassifikation med gode resultater i bl.a. [44].
Motivationen for at benytte MFCC, er at de til en vis grad repræsenterer de frekvensmæssige karakteristika, som også gælder for den menneskelige hørelse. Dermed forventer man, at de er i stand til at opfange nogle af de kendetegn for musik, som mennesker også opfanger.
I Figur 3.4 nedenfor vises et flowdiagram over hvordan *MFCC* beregnes ud fra den enkelte tidsramme.



Figur 3.4. Overordnet flowdiagram for hvordan Mel-frekvens Cepstral koefficienterne beregnes. Der eksisterer forskellige metoder, for hvordan MFFC implementeres, men langt de fleste følger det ovenforstående flowdiagram.

Efter at have benyttet et Hamming vindue, udføres den Diskrete Fourier transformation, hvor der kun bruges den reelle del af resultatet, dermed opnås en *N*-dimensional spektral repræsentation af tidsrammen. Fasen ignorereres da den ikke vurderes at have den store indvirkning på hvordan mennesker opfatter musik.
Næste skridt er såkaldt Mel-skalering. Formålet med mel-skalering er som nænvt i kapitel 2 at estimere forholdet imellem opfattet pitch og frekvens. Pitch er relevant, da det benyttes af mennesker til at dele lyd op efter på en musikskala.
Til mel-skalering benyttes normalt en filterbank bestående af triangulere filtre i frekvens domænet, hvor center frekvenserne er delt op efter melskalaen.
Logskalering benyttes hovedsageligt til at tilnærme menneskets af Loudness.

Til sidst benyttes den diskrete cosinus transformation, der har til formål at dekorellere de mel-spektrale log-skalerede koefficienter. Der benyttes "voicebox" [27] til at implementere MFCC i Matlab.

3.1.3 Zero Crossing Rate (ZCR)

Zero Crossing Rate (ZCR) er en de oftest brugte kortidslige features. Oprindeligt stammer den fra tale analyse, men er senere blevet benyttet til musik genre klassifikation i bl.a. [12]. ZCR er ganske enkelt, antallet af nullpunkts krydsninger i tidsdomænet. Formelt kan den skrives som:

$$ZCR_n = \sum_{i=n-N+1}^{n} \left| sgn(s_i) - sgn(s_{i-1}) \right| \qquad (3,3)$$

hvor sgn-funktionen returnerer fortegn for det pågældende input. ZCR kan siges at være en kompakt spektral repræsentation for signalet.

3.1.4 MPEG-7 features

Det såkaldte "MPEG-7 framework" [2] er udviklet til at repræsentere multimedia data i en kompakt løsning. I denne afhandling benyttes de såkaldte *spektrale basis feaures* der består af *Audio Spectrum Envelope* (ASE), *Audio Spektrum Spread* (ASS), *Audio Spektrum Centroid* (ASC) og *Audio Spektral Flattness* (ASF). Disse features er blevet undersøgte til klassifikation af musik i bl.a. [11,12]. Der findes flere variantioner af hvordan disse features kan implementeres, følgende fire delafsnit beskriver den generelle struktur.

[2] "Multimedia Content Description Interface"

3.1.5 Audio Spectrum Envelope (ASE)

Formålet med Audio Spectrum Envelope er at beskrive det effektmæssige indhold for et lydsignal, hvor frekvensbåndene er delt op på den logaritmiske skala.
Første skridt er at beregne den diskrete Fourier transformationen over en tidsramme på 30 ms og derefter finde effektspektret.
Derefter benyttes en ¼-oktav opdelt filterbank, af ikke overlappende rektangulære filtre, til at opsummere effekten for de enkelte frekvensbånd. Der benyttes en sekvens af filtre der går fra såkaldt *loEdge* til *hiEdge*, opdelt på den logaritmiske skala. Der benyttes kun et enkelt filter fra 0Hz til *loEdge* og ligeledes fra *hiEdge* til den halve samplingsfrekvens *fs*.
loEdge og hiEdge kan relateres til et 1 kHz reference punkt ved at benytte

$$f_m^e = 2^{rm} 1000\text{Hz} \tag{3,4}$$

hvor f_m^e er edge-frekvensen for den oktave filterbank, *m* er et reelt tal og *r* er oktave opløsningen[3].
Den spektrale repræsentation der opnås ved ASE benyttes som features i et klassifikationssystem.

3.1.6 Audio Spektrum Centroid

Audio Spektrum Centroid (ASC) består af de normaliserede vægtede middelværdier for de logaritmiske frekvens-værdier for et signal og kan udtrykkes ved

$$ASC = \frac{\sum_{i=1}^{N} \log_2 (f_i / 1000) P_i}{\sum_{i=1}^{P} P_i} \tag{3,5}$$

[3] Bemærk at dette gælder ikke for *1/8*-oktav

f_i er frekvensværdierne for den *i*'te frekvenskoefficient med effekten P_i. *N* er det samlede antal samples i tidsrammen, dermed er *N* også antallet af Fourier koefficienter. Denne feature indikerer til hvilken frekvens den dominerende effekt ligger, dette gør sig specielt gældende for signaler med smal båndbredde. Denne menes også være den fysiske correlation af den såkaldte opfattede skarphed [12].

3.1.7 Audio Spektrum Spread

I lighed med ASC, beregner *Audio Spektrum Spread* den vægtede standard afvigelse for den logaritmisk opdelte frekvens. Med samme notation som før, kan dette udtrykkes ved

$$ASS = \sqrt{\frac{\sum_{i=1}^{N} \left(\log_2\left(f_i/1000\right) - ASC\right)^2 P_i}{\sum_{i=1}^{N} P_i}} \qquad (3,6)$$

Dermed måler ASS spedningen af effekten ved middelværdien. Den vurderes at være i stand til at diskriminere imellem tone-agtig og støj-agtig lyd.

3.1.8 Spectral Flattness Measure

Spektral Flattness Measure (SFM) udtrykker afvigelsen fra et fladt effekt spektrum for den pågældende tidsramme. Store afvigelser menes at indikere tonale komponenter. SMF er brugt til audio fingreaftryk i bl.a [34] og musik klassifikation i bl.a. [34]. SFM udtrykkes ikke for hele tidsrammen på en gang, men for hvert enkelt frekvens bånd *k* ved

$$\mathrm{SFM} = \frac{\sqrt{\prod_{i=n(k)}^{n(k+1)} \tilde{P}_i}}{\frac{1}{N_k} \sum_{i=n(k)}^{n(k+1)} \tilde{P}_i} \qquad (3,7)$$

hvor *n(k)* er en indeks funktion for power spektrum koefficienterne $\tilde{P}_i$ imellem kanterne f_k og f_{k+1} hvor N_k er det tilsvarende antal koefficienter.

3.2 Tidslig feature integration

Tidslig feature integration er den proces hvor man forsøger, at sammensætte flere korttidslige features sammen til at danne en enkel feature vektor over et længere tidsinterval. Håbet er dermed at opfange den information, som de korttidslige feature ikke kan opfange. Samtidig fås et mål for de korttidslige features afhængighed i forhold til hinanden.
Processen kan udtrykkes ved

$$x_n = T\left(x_{n-(N-1)}, \ldots, x_n\right) \tag{3,8}$$

Hvor x_n er den nye feature vektor. x_n er dermed en tidsserie af korttidslige features og *N* er længden på tidsrammen. Transformationen *T* udfører den tidslige feature integration.
Selv om korttidslige features er i stand til at opfange vigtig information, så er de ikke i stand til at opfange den information der lever på en længere tidsskala. Dette kunne bl.a. være den rytmiske kontekst for et musiksignal eller den melodiske sammensætning. Derfor betragtes det som en nyttig egenskab, at kunne integrere de korttidslige features op på en længere tidsinterval og dermed opfange den information som de enkeltvis går glip af.
Der findes en række metoder i litteraturen, der udfører korttidslig features integration. I dette projekt benyttes en metode baseret på en Multivariabel Autoregressiv model (MVA), hvor parametrene fra modellen derefter indgår som features i klassifikationssystemet.

3.2.1 AR-features

Ved at benytte AR-features, er det muligt at modellere de dynamiske forhold for de multivariable tidsserier bestående af korttidslige feature vektorer. Derfor er det muligt, at få et mål for de tidsmæssige variationer for et musiksignal over et længere tidsinterval.

Hvis x_n består af tidsserien for de korttidslige features, da kan en AR-model beskrives matematisk som

$$x_n = \sum_{p=1}^{P} \mathbf{A}_p x_{n-p} + \mathbf{v} + \mathbf{u}_n \tag{3,9}$$

hvor matricen $\mathbf{A}_p$ består af de autoregressive koefficienter, $\mathbf{v}$ er den såkaldte intercept vektor og $\mathbf{u}_n$ er en drivende støj proces, der her består af hvid støj. *P* er modellens orden.
Intercept vektoren kan findes ved hjælp af middelværdien μ for x_n og udtrykkes ved

$$\mathbf{v} = \left(\mathbf{I} - \sum_{p=1}^{P} \mathbf{A}_p\right)\mu \tag{3,10}$$

Dermed ses at intercept vektoren gør det muligt, at benytte en fastlagt *middelværdi* for feature vektoren.
I tidsdomænet kan en AR-model ses som en prediktor af fremtidige værdier.
Hvis der eksisterer en featurevektor, der strækker sig fra x_n til x_{n-p}, hvor de pågældende AR-parametre er kendte, kan den næste feature vektor forudsiges ved brug af

$$\hat{\mathrm{x}}_\mathrm{n} = \sum_{\mathrm{p}=1}^{\mathrm{P}} \mathbf{A}_\mathrm{p}\, \mathrm{x}_{\mathrm{n}-\mathrm{p}} + \mathbf{v} \tag{3,11}$$

hvor der benyttes såkaldte *residualer* udtryk ved

$$\mathrm{e}_\mathrm{n} = \mathrm{x}_\mathrm{n} - \hat{\mathrm{x}}_\mathrm{n} = \mathrm{x}_\mathrm{n} - \sum_{\mathrm{p}=1}^{\mathrm{P}} \mathbf{A}_\mathrm{p} \mathrm{x}_{\mathrm{n}-\mathrm{p}} - \mathbf{v} \tag{3,12}$$

til at vurdere hvor godt estimerede de givne data er.

Hvis frekvensspektret estimeres ved hjælp af AR-modellen, er der i observeret [12] at effektspektret for den autoregressive model kan ses som en udglattet version $\hat{P}(\omega)$ af det sande effektspektrum $P(\omega)$. Der observeres også at de globale egenskaber viser, at det modellerede effektspektrum tilnærmer det sande effektspektrum med uniform performance over hele frekvens breden, dette uafhængigt af formen på effektspektret. Dermed tilskrives alle frekvensværdierne den samme betydning, uanset om de har lav eller høj energi. Resonante strukturer (peaks) for det sande effektspektrum modelleres bedre end de støjende dele af signalet

3.3.2 Estitmation af parametre

Der eksisterer flere tilgange til at esitmere parametrene for AR-modellen. Her benyttes ARFIT pakken [39].
Parametrene der estimeres, er AR-matricerne $\hat{\mathbf{A}}_1, \ldots \hat{\mathbf{A}}_p$, intercept termet $\hat{\mathbf{v}}$ og støj covarians matricen $\hat{\mathbf{C}}$.
Efter at parametrene er estimerede benyttes de autoregressive koefficienter $\mathbf{A}_p$, covarians matricen $\hat{\mathbf{C}}$ og de estimerede middelværdier $\mathbf{u}_n$ som features, hvor de benyttes i to forskellige udgaver, den ene kaldet *Multivariate autorregressive* features (MAR) og den anden *Diagonale autoregressive* features (DAR).

3.2.3 MAR features

De førnævnte MAR features kan opstilles i en feature vektor x_n udtrykt ved

$$\mathbf{x}_n = \begin{pmatrix} \mu_n \\ vek\left(\hat{\mathbf{B}}_n\right) \\ vekh\left(\hat{\mathbf{C}}_n\right) \end{pmatrix} \tag{3,13}$$

hvor *"vek"*-operatoren transformerer en matrice til en kolonne matrice, ved at stable de enkelte kolonner i matricen oven over hinanden. *"vekh"*-operatoren gør stort set det samme, men kun for de elementer der tilhører eller ligger ovenfor diagonalen. Dette gøres da $\hat{\mathbf{C}}_n$ er symmetrisk

3.2.4 DAR features

DAR features formes på samme måde, men kun diagonalerne for $\mathbf{A}_{\mathrm{p}}$ og $\hat{\mathbf{C}}_n$ benyttes. Dette kan for feature vektoren x_n udtrykkes som

$$\mathbf{x}_n = \begin{pmatrix} \mu_n \\ diag\left(\hat{\mathbf{A}}_{1n}\right) \\ diag\left(\hat{\mathbf{A}}_{2n}\right) \\ \vdots \\ diag\left(\hat{\mathbf{A}}_{Pn}\right) \\ diag\left(\hat{\mathbf{C}}_n\right) \end{pmatrix} \tag{3,14}$$

hvor "*diag*"-operatoren former en kolonne vektor fra diagonalerne for de pågældende matricer til tiden *n*.

3.4 Kernel model til feature behandling

I dette afsnit præsenteres en kernel model til tidslig feature integration. Kernel modeller benyttes hovedsagligt til at finde latente strukturer i et datasæt, der typisk vurderes at være overrepræsenteret i sin værende størrelse.

I dette projekt benyttes en kernel model kaldet "*Reduced Kernel Orthonormalized Partial Least Squares (rKOPLS)*" i forlængelse af feature transformationerne som musiksignalet gennemgår.
Metoden er første gang introduceret i [41]. Metoden er specielt udviklet til at håndtere store datasæt. I [41] påpeges at metoden har givet lovende resultater til klassifikation af musikgenre.

3.3.1 Kernel Orthonormalized Partial Least Squares (KOPLS)

I dette projekt benyttes KOPLS til at projektere AR – koefficienterne, opstillede i en feature vektor, over i et meget større featurerum hvor de kan håndteres ved hjælp af lineære metoder.

Til at beskrive KOPLS betragtes datasættet $\{\phi(x_i), y_i\}_{i=1}^{l}$ hvor $x_i \in R^N$ består af vektorer indeholdende AR-koefficienter og $y_i \in R^M$ består af tilsvarende labels. Funktionen $\phi(x): R^N \rightarrow F$, benyttes til at mappe de aktuelle inputdata over i et "Reproducing Kernel Hilbert space" (også kaldet featurerum) af meget stor eller uendelig størrelse. Der introduceres også matricerne $\Phi = [\phi(\mathbf{x}_1)\ldots\phi(\mathbf{x}_l)]^T$ og $\mathbf{Y} = [\mathbf{y}_1 \ldots \mathbf{y}_l]^T$ hvorfra følgende matricer opstilles

$$\Phi' = \Phi\mathbf{U} \text{ og } \quad \mathbf{Y}' = \mathbf{YV} \tag{3.15}$$

Disse udgør hver især n_p projektioner af input og output data, hvor **U** og **V** er projektionsmatricerne af henholdvis størrelsen *dim(F) x* n_p og *M x* n_p. Formålet med kernel funktionen er at søge efter projektionsmatricerne **U** og **V** sådan de projekterede input og output data er aligned mest muligt.
Til dette benyttes en kernel model der er en udvidet version af den såkaldte "Orthonormalized Partial Least Square (OPLS)". Modellen udtrykkes ved

$$\begin{aligned} &\text{Maksimer } \mathrm{Tr}\left\{\mathbf{U}^T\tilde{\Phi}^T\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T\tilde{\Phi}\mathbf{U}\right\} \\ &\text{I forhold til } \mathbf{U}^T\tilde{\Phi}^T\tilde{\Phi}\mathbf{U} = \mathbf{I} \end{aligned} \tag{3.16}$$

Hvor $\tilde{\Phi}$ og $\tilde{\mathbf{Y}}$ er centrerede versioner af henholdsvis Φ og $\mathbf{Y}$. **I** er identitets matricen af størrelsen n_p. Bemærk at denne model kun udtrækker projektioner fra input dataene.
KOPLS udfører optimal lineær multi-regression i featurerummet. Dermed minimerer (3.16) også den kvadratiske sum der udgør residualerne for approksimationen af label matricen:

$$\left\| \tilde{\mathbf{Y}} - \tilde{\Phi}' \hat{\mathbf{B}} \right\|_F^2, \quad \hat{\mathbf{B}} = \left(\tilde{\Phi}^T \tilde{\Phi}' \right)^{-1} \tilde{\Phi}'^T \tilde{\mathbf{Y}} \tag{3.17}$$

hvor $\hat{\mathbf{B}}$ er den optimale regressions matrice og $\|\cdot\|_F$ indikerer Forbenius normen af en matrice. Det ses at KOPLS kan være en meget nyttig metode hvor **Y** repræsenterer klasse tilhørsforhold. De optimale betingelser gør features opnåede fra KOPLS vil være bedre til at opfange relevant information i forhold til and MWA metoder, forstået sådan at de vil give lige god eller bedre resultater ved at benytte færre projektioner.

Når data projekteres over i et dimensions rum af uendelig størrelse, er det nødvendigt at benytte *Representer* teorien der siger at hver enkel projekteret vektor i **U** kan udtrykkes som en lineær kombination af træningsdataene. Dog er det hvis der benyttes et stort antal data, som forklaret i 41, mere passende at imposere sparsitet til at repræsentere projektion vektorerne. Derfor benyttes approksimationen $\mathbf{U} = \mathbf{\Phi}_R^T \mathbf{B}$ hvor $\mathbf{\Phi}_R$ er en delmængde af træningsdataene, der kun indeholder *R* mønstre (R<*l*) og $\mathbf{B} = \lfloor \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{n_p} \rfloor$ består af parametrene for den kompakte model. Selv om der findes mere sofistikere metoder til at udvælge træningsdataene til $\mathbf{\Phi}_R$, benyttes der her tilfældig udvælgelse, som gjort i 41.

Ved at erstatte **U** i (2) med den fremførte approksimation, fås et alternativt maksimerings problem der danner basis for KOPLS algoritmen med reduceret kompleksitet (rKOPLS):

$$\begin{aligned} &\text{Maksimer } \mathrm{Tr}\left\{ \mathbf{B}^T \mathbf{K}_R \tilde{\mathbf{Y}} \tilde{\mathbf{Y}}^T \mathbf{K}_R^T \mathbf{B} \right\} \\ &\text{I forhold til } \mathbf{B}^T \mathbf{K}_R \mathbf{K}_R^T \mathbf{B} = \mathbf{I} \end{aligned} \tag{3.18}$$

Hvor $\mathbf{K}_R = \mathbf{\Phi}_R \tilde{\mathbf{\Phi}}^T$ er en reduceret matrice af størrelsen *R x l*.

3.4 Diskussion

Der er i dette kapitel givet en beskrivelse af korttidslig features ekstraktion og tidslig features integration. Korttidslige features som Mel Cepstrale koefficienter og MPEG-7-features er beskrevne, ligeledes er der givet en beskrivelse af hvordan multivariabel autorregressiv regression kan benyttes som tidslig features integration hvor de såkaldte MAR og DAR features indgår. Der er også givet en forklaring på hvordan separation af kildesignaler kan indgå i forbehandlingen i et musikgenre klassifikationssystem.

Kapitel 4

Separation af kildesignaler

Der er i løbet af de sidste 10 år blevet præsenteret flere algoritmer der forsøger at løse problemet med separation af musiksignaler. Alligevel findes der ikke nogen algoritme, der er i stand til at udføre separationen med et perfekt resultat. De fleste algoritmer, der benyttes til separation af musiksignaler, er oftest baseret på velkendte algoritmer som ICA (eng.: *Independent Component Analysis*) [23,24] og NMF (*eng.: Non-negative Matrix Factorisation*) [15,16].
Den del af litteraturen der omhandler separation af kildesignaler, påpeger at ICA ikke er velegnet til separation af enkeltkanals musiksignal, da den forudsætter at kildesignalerne er uafhængige af hinanden. Dette vurderes ikke at være en velegnet antagelse gældende musiksignaler, da de oftest er afhængige af hinanden i form af rytmik og harmoni. En anden begrænsning vedrørende ICA er nødvendigheden for at have minimum det samme antal forskellige udgaver af det miksede signal, som det antal kildesignaler, som man vil separere signalet ned i. Dette besværliggør situationen, da der kun findes en udgave af det pågældende musiksignal, hvis der er

talen om et enkeltkanals signal. Der findes alligevel forsøg hvor man har forsøgt at separere enkeltkanals musiksignaler med algoritmer der bygger på ICA heriblandt *Uafhængig subbånd analyse* (eng: Independent Subband Analysis) og "Blues". Resultaterne er dog ikke særlig overbevisende, selv om de viser at det er muligt at fremhæve af kildikomponenterne.
Der er i løbet af de sidste ti år, blevet præsenterede flere metoder, baserede på NMF, som til en vis grad, er i stand til at udføre separation af enkelkanals musiksignaler. Resultaterne er dog i fleste tilfælde ikke meget bedre end dem man har opnået med andre metoder. Alligevel er der over de sidste to-tre år lykkedes at komme frem til nogle lovende resultater.
Den del af litteraturen, der omhandler separation af musiksignaler, benytter sig ikke af nogen generel metode til at vurdere resultaterne med. Derfor er det svært at sammenligne de opnåede resultater, der opnås af forskellige personer, da de hver har deres egen måde at vurdere resultaterne med. Oftest bliver der lyttet til resultaterne, hvor den person der udfører forsøget, vurderer hvor vellykket den enkelte separation er. Dette er en forholdsvis subjektiv vurdering af resultatet. Samtidig afhænger separationen oftest af de benyttede musiksignaler, f.eks. er der forskel på hvor mange instrumenter de enkelte musiknumre indeholder og hvordan musikinstrumenterne overlapper hinanden i frekvens og tid. Da konklusionerne ikke er baserede på de samme forudsætninger i form af målemetoder og valg af musiksignaler, er det svært at vælge en metode ud fra de konklusioner der fremgår af litteraturen.
Der findes publikationer der afprøver flere separationsmetoder i forhold til hinanden ved brug af de samme succeskriterier. Samtidig har enkelte forskere lagt deres resultater ud på Internettet i form af musikfiler. Disse publikationer og online demonstrationer, vil danne grundlag for hvilke metoder der vælges til separation af musiksignaler i dette projekt.
Den del af litteraturen, der til forfatterens kendskab, sammenligner forskellige metoder, kan i alle tilfælde påvise bedre resultater ved brug af NMF i forhold til andre metoder. Samtidig er der blevet lyttet til demoer der ligger på Internettet og vurderet at de lyder lovende til videre brug i form af musikgenre klassifikation.
Baseret på de forrige observationer, er der i dette projekt valgt at implementere og afprøvet tre metoder til separation af enkelkanals musiksignaler, der alle er baserede på NMF.
Efterfølgende afsnit beskriver tre metoder til at separere et mono musiksignal ned til dets kildesignaler. Derefter gennemgås tre metoder til at identificere de fundne kildekomponenter og sammensætte dem til at danne et lydsignal fra et musikintrument.

.

4.1 Ikke-negativ matrice faktorisering (NMF)

Ikke-negativ matrix faktorisering (NMF) blev første gang introduceret som et koncept af *Paatero*i [12]. To år senere foreslog Lee og Seung [13] nogle effektive algoritmer til beregning af NMF. Siden hen er NMF blevet benyttet i en række sammenhæng med gode resultater. Det har også vist sig at ikke-negativitet er en nyttefuld egenskab gældende NMF til faktorisering af matricer.
Til uddybning af NMF, defineres en ikke-negativ *MxN* matrice **V,** der udtrykkes ved

$$\mathbf{V} \approx \mathbf{WH} \tag{4,1}$$

hvor **W** er en *MxR* basis matrice og **H** er en *RxN* koefficients matrice, både **W** og **H** er ikke-negative. Formålet med NMF er at rekonstruere **V** som et produkt af de to ikke-negative matricer **W** og **H** hvor $R \leq M$ således at **V** kan tilnærmes som vist i (4.1).
Matricen **V** kan forestilles at bestå af *M* datavektorer, hvor hver enkel datavektor består af *N* datasæt. Hvis *R* vælges mindre end *M* eller *N* bliver matricerne **W** og **H** mindre end den originale matrice **V**. Resultatet af 4.1 bliver da en komprimeret version af den originale matrice **V**.
Til at optimere **W** og **H** benyttes to kostfunktioner introduceret af Lee og Seung [13]. Den ene er den Euklidiske afstand imellem **V** og **WH**:

$$\mathbf{C} = \left\|\mathbf{V} - \mathbf{WH}\right\| = \sum_{ij} \left(\mathbf{V}_{ij} - (WH)_{ij}\right)^2 \tag{4,2}$$

den anden er divergensen imellem **V** og **WH:**

$$\mathbf{D}\left(\mathbf{V} \| \mathbf{WH}\right) = \sum_{ij} \left(\mathbf{V}_{ij}\, log \frac{\mathbf{V}_{ij}}{(\mathbf{WH})_{ij}} - \mathbf{V}_{ij} + (\mathbf{WH})_{ij} \right) \tag{4,3}$$

I dette projekt benyttes to rekursive opdaterings algoritmer, der begge konvergerer hen imod et lokalt minimum [13]. For den Euklidiske distance benyttes:

$$\mathbf{W} \leftarrow \mathbf{W} \bullet \frac{\mathbf{V}\mathbf{H}^T}{\mathbf{W}\mathbf{H}\mathbf{H}^T}, \qquad \mathbf{H} \leftarrow \mathbf{H} \bullet \frac{\mathbf{W}^T\mathbf{V}}{\mathbf{W}^T\mathbf{W}\mathbf{H}} \tag{4,4}$$

og for KL-divergensen benyttes:

$$\mathbf{W} \leftarrow \mathbf{W} \bullet \frac{\frac{\mathbf{V}}{\mathbf{W}\mathbf{H}}\mathbf{H}^{\mathrm{T}}}{\mathbf{l} \cdot \mathbf{H}^{\mathrm{T}}}, \qquad \mathbf{H} \leftarrow \mathbf{H} \bullet \frac{\mathbf{W}^T \frac{\mathbf{V}}{\mathbf{W}\mathbf{H}}}{\mathbf{W}^T \cdot \mathbf{l}} \tag{4,5}$$

hvor $\mathbf{A} \bullet \mathbf{B}$ og $\frac{\mathbf{A}}{\mathbf{B}}$ indikerer henholdsvis elementvis multiplikation og division.

4.2.1 Ikke-negativ faktorisering til separation af enkelkanals polyfoniske lydsignaler

Til separation af lydsignaler benyttes en lineær signal model, hvor den enkelte basis vektor $\mathbf{x}_t$ antages at være en lineær miksning af basisvektorerne $\mathbf{s}_n$. Dette kan udtrykkes ved

$$\mathbf{x}_t \approx \sum_{n=1}^{N} a_{t,n}\,\mathbf{s}_n \qquad t = 1 \ldots K \tag{4,6}$$

Hvor $a_{t,n}$ er miksnings vægten for den *n*'te komponent til observation *t*. *N* er antal komponenter og *K* er antal observationer. En lydkilde kan dermed repræsenteres, som en sum af en eller flere komponenter. Modellen i (4.6) tager for enkelhedens skyld ikke hensyn til den støj der normalt eksisterer i praksis.
Udtrykket i (4.6) kan også skrives på matrice form som

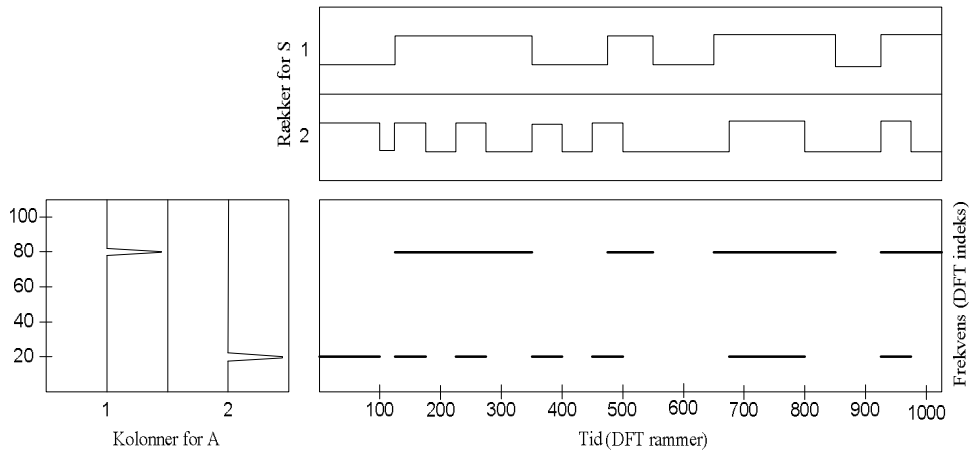
$$\mathbf{X} \approx \mathbf{AS} \tag{4,7}$$

Ved at sammenligne ovenforstående udtryk i (4.7) og NMF approksimationen i (4.1), kan man nemt se, at de viser det samme udtryk. Miksningsmatricen **A** i (4.7) svarer til koefficientmatricen **W** i (4.1) og komponent matricen **S** svarer til basisvektorerne i **H**. Dette indebærer at NMF gør det muligt opløse den observerede matrice **X** op i dens komponenter.
Da det er talen om musiksignaler, er det mest nærliggende at lade **X** repræsentere spektrogrammet for et signal, der findes ud fra den velkendte diskrete korttidslige Fourier Transformation (STFT) udtrykt ved

$$F = DFT\begin{bmatrix} x(t_1) & x(t_2) & & x(t_N) \\ \vdots & \vdots & \cdots & \vdots \\ x(t_1+M-1) & x(t_2+M-1) & & x(t_N+M-1) \end{bmatrix} \tag{4.8}$$

M er størrelsen på DFT og *N* er antal rammer (der benyttes også et vindue på rammerne i praksis). Spektrogrammet **X** er da den absolutte størrelse $|F|$. I praksis er spektrogrammet derfor en frekvensmæssig repræsentation af signalet i forhold til tiden.

Ved at benytte NMF på **X** opnås **A** og **S**. Proceduren er vist i figur 1.



Figur 4.1. NMF på et spektrogram. De tre diagrammer ovenfor viser hvordan et spektrogram ved hjælp af NMF kan opløses i matricerne **A** og **S**. Det nedre plot til højre, viser et spektrogram for to sinusformede signaler med tilfældige amplituder. Nederst til venstre vises to kolonner for matricen **A**, der fortolkes som basis for spektrum af signalerne. Det øverste plot viser to rækker af **S**, der repræsenterer tidsvægtene i forhold til de spektrale baser (figuren er modificeret fra [13]).

Det midterste plot viser et spektrogram for et sinusformet signal, der står og ”tænder og slukker”. Ved at benytte NMF på signalet, opnås faktorerne **A** og **S** som vist i figuren. Diagrammet nederst til venstre, viser at de to kolonner for **A,** kun indeholder energier ved de frekvenser der forekommer i spektrogrammet. Derfor fortolkes disse kolonner som basis funktioner for de spektrale data i spektrogrammet. På samme måde ses at rækkerne i **S**, der vises i det øverste diagram, kun indeholder energier på de samme tidspunkter som spektrogrammet gør. Derfor kan rækkerne i **S** fortolkes som vægte for de spektrale baser for hvert tidspunkt. Vægtene og baserne har 1:1 korrespondance. Den første basis beskriver spektrum for den øverste sinusoid og
den første vægt beskriver ”envelopen”. På samme måde beskrives den anden sinusoid ved hjælp af basis nr. 2. og vægt vektor nr. 2.
Proceduren i figur 1 er et opstillet eksempel for to sinusoider. Alligevel har den i [13] vist sig at kunne adskille selv meget komplicerede klaverstykker op i vægte og spektrale basisfunktioner hvor den er i stand til beskrive hver enkel tone der spilles og dens tidsmæssige placering.

4.3 Perceptionelt vægtet NMF til separation af mono polyfoniske lydsignaler

Den metode der præsenteres her, er baseret på en metode introduceret i [40]. Metoden benytter en vægtet NMF på spektrogrammet for et input signalet. Der benyttes perceptionelt motiverede vægte for de kritiske frekvensbånd, til at modellere det menneskelige høresystems opfattelse af loudness. De enkelte kildesignaler repræsenteres ved hjælp af en eller flere kildekomponenter. Metoden forklares detaljeret i følgende afsnit.

4.3.1 Perceptionelt motiverede vægte

Den menneskelige opfattelse af et lydsignal, kan modelleres ved en komprimere signalet for hvert enkelt kritisk frekvensbånd. Dette gøres ved at vægte hver enkel DFT-komponent separat for hver enkel ramme i spektrogrammet, hvor frekvenskomponenter vægtes ens hvis de tilhører det samme kritiske frekvensbånd. For at beregne loudness for hvert enkelt kritisk bånd, overføres frekvensværdierne til Bark-skalaen, hvor en bark-enhed repræsenterer det kritiske frekvens-bånd. Ved at beregne den estimerede loudness for hver enkel bark-enhed kan vægtene findes for hvert enkelt kritisk bånd.

Energien for hvert enkelt kritisk frekvensbånd kan udtrykkes ved

$$e_{b,t} = \sum_{f \in F_b} x_{f,t} \tag{4,8}$$

hvor $x_{f,t}$ er det pågældende (*f*,*t*)-element i spektrogrammet **X** og $F_b, b = 1,...,24$ indikerer de kritiske frekvensbånd. Der benyttes 24 frekvensbånd da det menneskelige hørelse ikke kan strække sig over en større frekvensskala.

Effekt-response for det ydre og indre øre tages i betragtning ved at multiplicere energien for hvert enkelt kritisk frekvensbånd med den tilsvarende effekt response h_b. Energien efter ydre og indre øre filtrering udtrykkes

$$g_{b,t} = h_b e_{b,t} \tag{4,9}$$

Til at repræsentere h_b benyttes effekt response for den før omtalte equal loudness kurven.
Der benyttes her et *loudness indeks* til at referere til den estimerede loudness for hvert enkelt kritisk frekvensbånd og pågældende ramme i spektrogrammet. Loudness-indeks $L_{b,t}$ for det kritiske bånd *b* og ramme *t* kan udtrykkes ved

$$L_{b,t} = \left[y_{b,t} + \varepsilon_b \right]^v - \varepsilon_b^v \tag{4,10}$$

hvor *v* er en fastlangt kompremeringsfaktor og ε_b er en øvre grænse for hørelsen ved bånd *b*.
Den øvre grænse for den menneskelige hørelse er ikke nødvendigvis kendt i praksis. Derfor kan den beregnes fra input signalet. For enkelhedens skyld sættes ε_b til at være den samme for alle kritiske bånd. Efter at ydre og indre øre filtrering er udført, estimeres niveauet på signalet ud fra variansen σ^2 givet ved

$$\sigma^2 = \frac{1}{K} \sum_{b=1}^{B} g_{b,t} \tag{4,11}$$

hvor *B* er antal bark-enheder. Et godt valg for ε_b findes emperisk til at være $10^{-5}\sigma^2$.

Vægtene $c_{b,t}$ findes for hvert kritisk bånd og tilhørende ramme, ved at minimere den perceptionelle loudness. Vægtene vælges således, at den vægtede sum for DFT-komponenterne, er lig med den estimerede loudness. Dermed vil den kvantitative signifikans for en *t-f*-komponent svare omtrent til den "*perceptionelle signifikans*". Dermed vælges vægtene således at

$$c_{b,t} e_{b,t} = L_{b,t} \tag{4,12}$$

hvor $c_{b,t}$ findes ved

$$c_{b,t} = \frac{L_{b,t}}{e_{b,t}} \tag{4,13}$$

I praksis vil energien indenfor et kritisk bånd aldrig være præcis lig med nul, derfor vil ligning altid være gyldig. Dette kan dog forekomme for et syntetisk fremkaldet signal. Da vil $c_{b,t}$ være $vg_{b,t}^{v-1}$ idet at

$$\lim_{e_{b,t}\to 0} \frac{L_{b,t}}{e_{b,t}} = vg_{b,t}^{v-1} \tag{11}$$

Hvis vægtene $c_{b,t}$ samles i en *F* x *t* matrice $\mathbf{\Omega}_{f,t}$, hvor *f,t* repræsenterer det enkelte element i spektrogrammet for $f \in F_b$.

Til at implementere loudness modellen i matlab findes først spektrogrammet for det pågældende signal. Der benyttes kun de absolutte værdier hvori fasen ignoreres. Fasen for det originale signal gemmes dog til senere syntesering af komponenterne. De kritiske frekvensbånd identificeres, ved at overføre den lineære frekvensskala til barkskalaen ved hjælp af udtryk 2.2 i kap. 2. Effektrespons for equal loudness kurven importeres fra mat-aim-toolboxen [2], der også oveføres til barkenskalaen. De kritisek frekvensbånd for hver enkel ramme i spektrogrammet multpliceres derefter med de tilsvarende frekvensbånd for equal loudness kurven. Dernæs benyttes udtrykkene i (7) og (10) til at beregne vægtene for de enkelte kritiske frekvensbånd. Til sidst overføres vægtene til den lineære frekvensskala, sådan at de kan benyttes som input til den vægtede NMF.

4.3.2 Vægtet ikke-negative faktorisering (WNMF)

Metoden benytter sig af en vægtet udgave af NMF og en vægtet divergens benyttes til at opdatere **A** og **S.** Divergensen benyttes i stedet for den euklidiske afstand, da den er mindre følsom overfor signaler med en høj energi [40]. Opdateringsreglerne for den vægtede NMF er udledt i [40] og udtrykkes ved

$$\mathbf{A} \leftarrow \mathbf{A} \bullet \frac{\frac{\mathbf{\Omega X}}{\mathbf{AS}}\mathbf{S}^T}{\mathbf{\Omega} \cdot \mathbf{S}^T}, \qquad \mathbf{S} \leftarrow \mathbf{S} \bullet \frac{\mathbf{A}^T \frac{\mathbf{\Omega X}}{\mathbf{AS}}}{\mathbf{A}^T \cdot \mathbf{\Omega}} \tag{4,14}$$

hvor $\mathbf{\Omega}$ er en matrice indeholdende de fundne vægte.

Til at opdatere **A** og **S** benyttes normalt en iterativ process, hvor processen standser når **A** og **S** ikke ændrer sig efter et vist antal iterationer.
Til at generere de enkelte komponenter benyttes

$$\mathbf{X} = \mathbf{AS} \tag{4,15}$$

hvor hver enkel kolonne i **X** vil bestå af de fundne komponenter. Dermed er **X** nu en kompimeret version af den originale input matrice **X.**

4.4 Ikke-negativ matrix 2D faktorisering til separation af lydsignaler

Dette afsnit beskriver den såkaldte Ikke-negative matrix 2D faktorisering (NMF2D). Metoden er en forlængelse af NMF og blev første gang introduceret i 2004 i [3].
NMF er i flere henseender mangelfuld, da den ikke tager hensyn til relative positioner i spektrogrammet. Dermed ignoreres ændringer i pitch og tid. Disse frekvensmæssige og tidsmæssige informationer er NMF2D bedre rustet til at håndtere.

Ved brug af NMF2D, antages at et instrument kan modelleres ved en specifik tids-frekvensmæssig signatur, der modellerer lyden fra det pågældende intrument over tiden *t*. Når så et intrument spiller en tone til et bestemt tidspunkt *t*, da vil denne signatur bevæge sig hen over tids-aksen. På samme måde, når en tone spilles med en bestem pitch ϕ, svarer det til, at den det pågældende instruments signatur bevæger sig henad tids-frekvens-aksen.

Til at udtrykke NMF2D udvides udtrykket for NMF i 2.1.1 således at $\mathbf{A}^t$ er afhængig af tiden *t* og $\mathbf{S}^\phi$ afhænger af pitch ϕ. Dermed fås følgende udtryk for X

$$\mathbf{X} \approx \mathbf{\Lambda} = \sum_t \sum_\phi \overset{\downarrow\phi \ \rightarrow t}{\mathbf{A}^t \mathbf{S}^\phi} \tag{4,16}$$

Hvor $\downarrow\phi$ indikerer et nedadrettet skift, der forskyder hvert element ϕ pladser nedad i **A**,. $\rightarrow t$ indikerer at hvert element i pågældende matrice flyttes t pladser til højre f.eks.

$$\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}, \quad \overset{\downarrow 2}{\mathbf{A}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 2 & 3 \end{pmatrix}, \quad \overset{\rightarrow 1}{\mathbf{A}} = \begin{pmatrix} 0 & 1 & 2 \\ 0 & 4 & 5 \\ 0 & 7 & 8 \end{pmatrix}$$

Hvert element i $\mathbf{\Lambda}$ er givet ved

$$\mathbf{\Lambda}_{i,j} = \sum_t \sum_\phi \sum_d \mathbf{A}^t_{i-\phi,d} \mathbf{S}^\phi_{d,j-t} \tag{4,17}$$

Opdateringsalgoritmerne der benyttes til at opdatere $\mathbf{A}^t$ og $\mathbf{S}^\phi$ er baseret på den euklidiske distance og KL divergens [3]. For den Euklidiske distance gælder følgende udtryk for opdatering af $\mathbf{A}^t$ og $\mathbf{S}^\phi$:

$$\mathbf{A}^t \leftarrow \mathbf{A}^t \bullet \frac{\sum_\phi \overset{\uparrow\phi}{\mathbf{X}} \overset{\rightarrow t\ T}{\mathbf{S}^\phi}}{\sum_\phi \overset{\uparrow\phi}{\mathbf{\Lambda}} \overset{\rightarrow t\ T}{\mathbf{S}^\phi}}, \qquad \mathbf{S}^\phi \leftarrow \mathbf{S}^\phi \bullet \frac{\sum_\tau \overset{\downarrow\phi\ T}{\mathbf{A}^t} \overset{\leftarrow t}{\mathbf{X}}}{\sum_\tau \overset{\downarrow\phi\ T}{\mathbf{A}^t} \overset{\leftarrow t}{\mathbf{\Lambda}}} \tag{4,18}$$

For KL divergens gælder:

$$\mathbf{A}^t \leftarrow \mathbf{A}^t \bullet \frac{\sum_\phi \left(\frac{\overset{\uparrow\phi}{\mathbf{X}}}{\mathbf{\Lambda}}\right) \overset{\rightarrow t\ T}{\mathbf{S}^\phi}}{\sum_\phi \mathbf{l} \cdot \overset{\rightarrow t\ T}{\mathbf{S}^\phi}}, \qquad \mathbf{S}^\phi \leftarrow \mathbf{S}^\phi \bullet \frac{\sum_t \overset{\downarrow\phi\ T}{\mathbf{A}^t} \left(\frac{\overset{\leftarrow t}{\mathbf{X}}}{\mathbf{\Lambda}}\right)}{\sum_t \overset{\downarrow\phi\ T}{\mathbf{A}^t} \cdot \mathbf{l}} \tag{4,19}$$

Til implementering af NMF2D benyttes et logaritmisk spektrogram til at repræsentere **X**. Det logaritmiske spektrogram findes på samme måde som det lineære spektrogram i (2), men i stedet for at benytte den lineære frekvens skala benyttes en logaritmisk repræsentation. Motivationen for at benytte det logaritmiske spektrogram, er den tolv tone ligeligt fordelte musikskala, der danner basis for moderne vestlig musik. Ved brug af denne skala, deles en oktav op i tolv halvtoner, hvor forholdet imellem hver halvtone er det samme. Hvis F_1 betegner den fundamentale frekvens for en tone, da vil den fundamentale frekvens for en tone p halvtoner over denne tone udtrykkes ved

$$F_2 = F_1 \cdot 2^{p/12} \tag{4,20}$$

Ved at brug af logaritmen fås

$$\log F_2 = \log F_1 \cdot \tfrac{p}{12} \log 2 \tag{4,21}$$

Hvorfra det ses at den logaritmiske frekvens repræsentation er lineær.

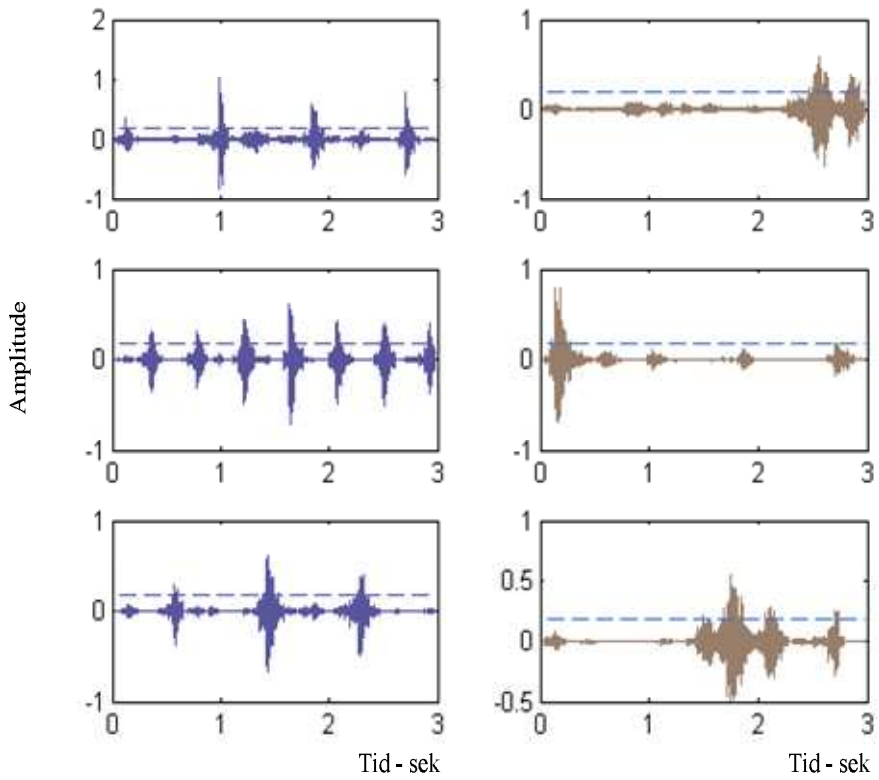
4.5 Automatisk identifikation af trommesignaler

Selv om de ovenfor beskrevne metoder til enkeltkanals separation udfører en vellykket separation, så står man stadigvæk tilbage med det store problem, med at skule identificere de komponenter der tilhører det samme instrument. I største del af litteraturen hvor NMF er brugt til separtion af lydsignaler, benytter forfatteren sig af manuel identifikation af komponenterne, se bl.a. [1,2,3]. Derfor er automatisk identifikation af komponenterne en forholdvis uudforsekt disciplin. I dette projekt er det dog nødvendigt med en automatisk idenfiktion process for at kunne benytte separation af mono musiksignaler i et fuldautomatisk klassifiktionssystem.
I dette projekt er der benyttet to forskellige metoder der udfører fuldautomatisk identifikation af de fundne komponenter. Metoderne er til forfatterens kendskab, ikke blevet brugt før i denne sammenhæng.

4.5.1 Trommedetektor

I denne metode instroduceres en metode, der er i stand til at identificere tromme komponenterne fra resten af de andre komponenter, der tilhører de peak-baserede instrumenter. Metoden gør det muligt i en klassifikationsproces, at klassificere et musiknummer efter trommerne og eventuelt efter de harmoniske instrumenter for sig selv.
Det antages at tempoet for moderne vestlig musik ligger imellem 70 og 210 bpm. Derudover antages, at for komponenter der tilhører trommerne i en sang, vil der for hvert trommeslag observeres et tilhørende peak i tidsdomænet. Derfor hvis trommerne spiller en forholdsvis konstant sekvens igennem et musiknummer, vil der være ét slag pr. sekund for hver enkel tromme, hvis tempoet er lavere end 120 bpm. Omvendt vil der eksistere mindst to peaks pr. sekund hvis trommerne spiller i et tempo højere eller lig med 120 bpm. Da trommerne antages at spille nærmest de samme toner igennem en hel sang, antages at der vil være *mindst et peak pr. sekund, hvis en komponent tilhører trommerne.*
De komponenter hvor der ikke observeres et peak for hvert sekund, vil da antages at tilhøre de pitch baserede instrumenter.
Metoden er fremkommet ved at undersøge 100 musiknumre for ligheder imellem trommekomponenter i forhold til de andre fundne komponenter. Figur 4 nedenfor viser 6 komponenter udaf 10 fundne, ved brug af den vægtede NMF beskrevet før. De tre figurer længst til venstre viser tre komponenter, der ikke tilhører trommesignalet, imens de tre figurer længst til højre viser de tre fundne

komponenter, der tilhører trommesignalet. Den punkterede linje viser tærskelværdien for den pågældende sang

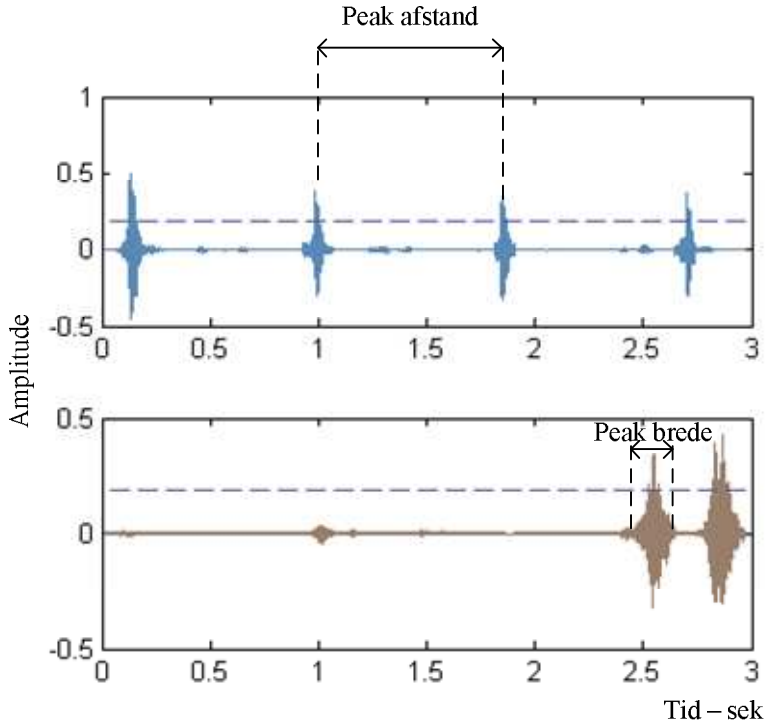


Figur 4.2. De tre delfigurer til venstre afbilder tre komponenter tilhørende et trommesignal fundet ved en separationsproces. De tre delfigurer til højre afbilder tre komponenter tilhørende et *ikke*-trommesignal.

Ved at se på de tre delfigurer til venstre i figur 4, ses der ret tydeligt, at de har en periodisk tendens Derfor vurderes de at repræsentere et rytmisk signal.
De tre delfigure til højre repræsenterer de komponenter, der *ikke* tilhører et trommesignal. Der observeres, at de ikke har den samme periodiske tendens som komponenterne til venstre har. Derfor vil de ikke blive identificerede som trommekomponenter.
Figur 4 underbygger dermed også antagelse, at der for hvert sekund, skal vær mindst et peak tilstede, der samtidig ligger over en bestemt tærskelværdi, hvis den pågældende komponent skal identificeres som et trommesignal.

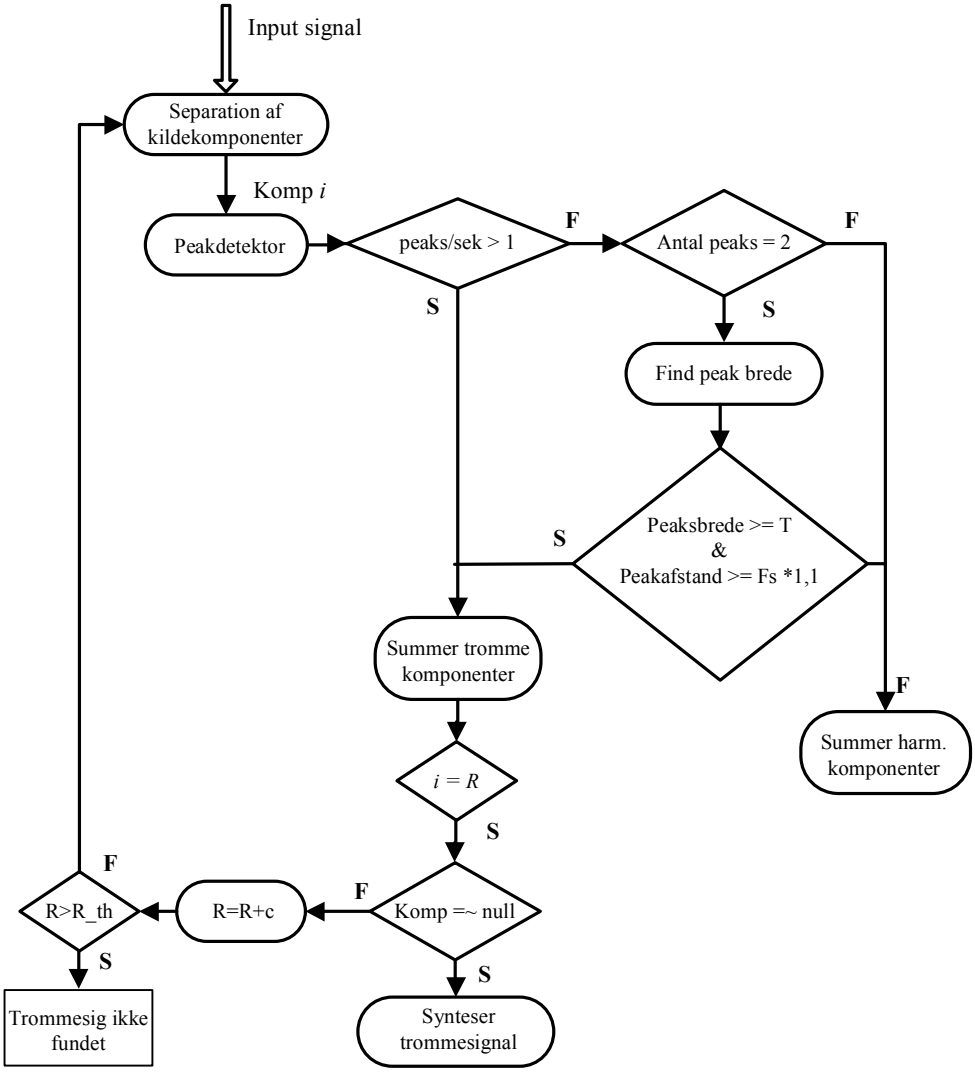
Der er udført lyttetest på 100 musiknumre, hvor den automatiske identifikation af trommekomponenter er sammelignet med tilsvarende manuel identifikation. I alle tilfælde kunne trommedetoktoren påvise et godt resultat, hvor der fremkommer en tydelig rytmesekvens. I flere tilfælde kunne den også påvise et resultat af bedre kvalitet for den samlede trommesekvens end den fremkommet ved manuel sortering.
Resultatet af komponentent-identifikationen afhænger dog af, hvor godt systemet har været til at finde brugbare komponenter.

Til at finde de pågældende peaks for en sang, benyttes en simpel peakdetektor.
Peakdetektoren består af en iterativ process, der søger igennem hele signalsekvensen og finder de værdier der ligger over tærskelværdien. Derefter tælles op hvor mange af disse peakværdier ligger indenfor et sekund. Hvis der eksisterer ét eller flere peaks pr. sekund vil den pågældende komponent tilskrives trommesignalet ellers vil den tilskrives signalet for et peakinstrument.
Til at repræsentere et peak benyttes det lokale maksimum. Dermed er det kun den maksimale værdi der tages i betragtning for hver peak. Dette gøres ved at nulstille de komponenter der ligger i en bestemt afstand til højre og venstre fra det pågældende peak. I dette projekt stilles denne afstand til 10^4 punkter.
Der tages også hensyn til afstanden imellem to sidestående peaks. Dette gøres for at undgå, at to peaks der ligger meget tæt på hinanden, men alligvel tilhører to forskellige sekund-intervaler, ikke identificeres som en rytmisk sekvens. I dette projekt er denne mindste afstand, imellem to peaks, sat til *fs*1,1* punkter.
Figur 5 nedenfor viser de forskellige peaks markeret for henholdsvis en trommekomponent og ikke-trommekomponent. Hvor afstanden imellem to peaks er indikeret i det øverste plot.



Figur 4.3. De to delfigurer afbilder to komponenter fundne ved en separationsproces , hvor den øverste komponent tilhører et trommesignal og den nederste tilhører et *ikke*-trommesignal. De røde komponenter indikerer peakværdier fundet ved brug af en peakdetektor. Den øverste figur viser også afstanden imellem to efterfølgende peaks.

På det nederste plot i figur 11 ovenfor observeres, at at selv om der findes peaks højere end tærskel-værdien, så vil de ikke kunne blive identificeret som en trommekomponent, da der ikke eksisterer peaks for alle tre sekunder.



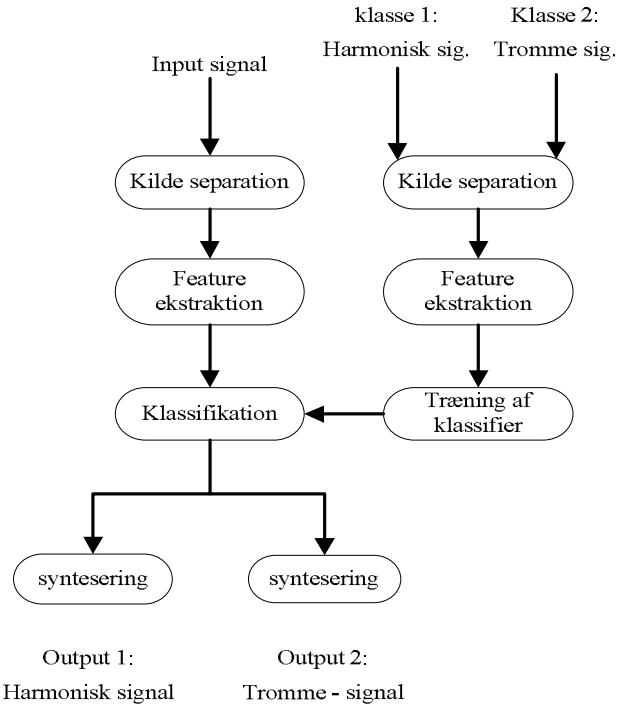
Figur 4.4 viser flowdiagram for trommedetektor.

Figur 11 ovenfor viser flowdiagrammet for den peakbaserede trommedetektor. Et input signal vil blive separet i *R* kildekomponenter. Derefter identificeres for hver enkel komponent de værdier der ligger over en fastlagt tærskelværdi i tidsdomænet. Dette gøres ved hjælp af en iterativ peakdetektor hvor en peakværdi består af værdier der ligger over tærskelværdien. Derefter hvis der er mindst ét peak pr.

sekund så til komponenten blive sorteres som tilhørende et trommesignal og summeret med andre fundne trommekomponenter. Hvis der ikke eksisterer ét peak pr. sekund, da undersøges om der eksisterer to peaks over tre sekund, hvis dette er sandt så undersøges breden for de eksisterende peaks, hvis breden er større end en fastlagt værde, da sorteres komponenten som trommesignal, hvis ikke - så sorteres den at tilhøre de harmoniske kildesignaler.
Hvis der ikke er fundet et trommesignal efter en separationsproces, da sættes antal komponenter op og processen gentages.

4.5.2 Lineær klassifier til detektion af tromme komponenter

Den metode der præsenteres her er baseret på en metode introduceret i [43] til identifikation af trommekomponenter.
Metoden benytter standard tekniker til mønster genkendelse i form af en lineær klassifier (se kapitel 5.1). Først trænes algoritmen op af to klasser der består af trommekomponenter og *ikke*-trommekomponenter (harmoniske signaler). Proceduren er afbildet i figur 12 nedefor.



Figur 4.5. viser flowdiagram for hvordan komponenter tilhørende trommesignler eller harmonikse signaler kan identificeres ved hjælp af en lineær klassfier.

Systemet trænes op ved at benytte en stort antal musiksignaler, der er delt op i tromme signaler og harmoniske signaler. Hvert enkelt signal gennemgår en separationsproces i form af NMF. Derefter udtrækkes fra de separerede komponenter. Disse features benyttes derefter til at træne en lineær klassifier. Efter at systemet er trænet, da vil det være muligt at lade fremtidige signaler gennemgå den samme proces, men hvor parametrene fundet ved træningen af den lineære klassifier, benyttes til at identificere hvilke komponenter der tilhører trommesignaler eller harmoniske signaler.
Årsagen til at trænings-signalerne ikke benyttes direkte, men gennemgår en separationsproces, er fordi at de ønskes at være i samme format, som de fundne kildekomponenter.
Der benyttes en database bestående af 600 små lydklip taget fra forskellige instrumenter (såkaldte loops) til at træne systemet op. Lydklippene er hentet over Internettet fra [1,2,3]. Disse lydklip deles videre op i to klasser, hvor den ene klasse

består af lydklip tilhørende trommesignaler og den anden består af lydklip tilhørende ikke-trommesignaler.
Til at separere de indkomne signaler benyttes NMF, hvor lydklippene deles op i 3 komponenter. Derefter udtrækkes features i form af Mel Cepstrale koefficienter, fra de opnåede komponenter. Der findes 10 koefficienter ud fra tidsrammer på 30 ms med 15 ms overlap. De fundne MFCC features benyttes derefter til at træne den lineære klassifier op, for derefter at kunne skelne i mellem komponenter tilhørende trommesignal og alle andre mulige lydsignaler.
Efter at klassifieren er trænet op, vil det være muligt at lade komponenterne fra alle fremtidige separationsprocesser, blive klassificerede som tilhørende en af de to klasser. Derefter bliver komponenterne syntetiserede til at danne et kildesignal og om nødvendigt summeret med andre komponenter tilhørende trommesignalet.

Kapitel 5

Klassifiers

I denne afhandling benyttes to klassifiers. Den første kaldes for den generaliserede klassfier og er en udvidelse af den såkaldte linære klassifier. Denne type klassifiers hører til de såkaldte diskriminerende klassifiers. Derefter benyttes en generative probalistisk model der kaldes for den Gaussiske miksningsmodel der er en udvidelse af den Gauissiske klassifiere.

5.1 Lineær klassifier

Den lineære klassifier er iblandt de enkleste og mest brugte klassifiers. Den er bl.a. blevet brugt til musik genre klassifikation i [1]. Til at beskrive den benyttes funktionen y(**x**) som funktion af inputvektoren $\mathbf{x} = (x_1,...,x_n)^T$.
En lineær model for y(**x**), kan udtrykkes matematisk ved

$$y(\mathbf{x}) = w_0 + \sum_{i=1}^{N} w_i x_i = w_0 + \mathbf{w}^T \mathbf{x} \tag{5,1}$$

Hvor vektoren **w** består af vægtene for **x**.
Hvis **x** udvides med en ekstra dimension, således at $\mathbf{x} = (1,x_1,...,x_n)^T$ kan (1) reduceres til

$$y(\mathbf{x}) = \sum_{i=0}^{N} w_i x_i = \mathbf{w}^T \mathbf{x} \tag{5,2}$$

Vægtene i **w** kan estimeres ved at minimere den velkandte "sum.of-squares"-error funktion givet ved

$$\begin{aligned} E(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N} \left(y_k(\mathbf{x}^n;\mathbf{w}) - t^n \right)^2 \\ &= \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} \left(y_k(\mathbf{x}^n;\mathbf{w}) - t_k^n \right)^2 \end{aligned} \tag{5,3}$$

hvor $y(\mathbf{x}^n;\mathbf{w})$ repræsenterer output for klasse *k*, som funktion af input vektor $\mathbf{x}^n$ og vægt-vektor ***w***. *N* er antal trænings samples og *K* er antal genrer. Størrelsen t_k^n repræsenterer target værdien for output enheden ***k*** når input vektoren er $\boldsymbol{x}^k$.
Til at finde de optimale vægte og dermed minimum for (3) benyttes udtrykket

$$\mathbf{W}^T = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{T} \tag{5,4}$$

hvor

$$\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2 \ldots\ \mathbf{x}_M] \quad \text{og} \quad \mathbf{W} = [\mathbf{w}_1\ \mathbf{w}_2 \ldots\ \mathbf{w}_M].$$

I dette projekt, repræsenterer ***x***, input features for de enkelte genrer og ***t*** består af labels, der repræsenterer den tilhørende genre. Der benyttes "1-out-c-codning". Dermed udnyttes den indeksmæssige placering for den pågældende label. F.eks. vil en feature tilhørerende genre 2, ud af ialt 4 genrer, skrives som

$$\mathbf{t} = [0, 1, 0, 0]^T \tag{5,5}$$

De input data, der benyttes til at finde vægtene, kaldes for *træningsdata* og indgår i *træningsfasen* af systemet. Da systemet er færdigt trænet, kan man benytte de fundne vægte til at modellere andre data med, disse data kaldes *testdata* og benyttes i *testfasen*.
Hvis der eksisterer en feature vektor $\boldsymbol{x_n}$, kan de estimerede target data $\tilde{t}_n$ findes ved $\tilde{\mathbf{t}}_n = \mathbf{W}^* \mathbf{x}_n$ hvor man ved brug af "1-out-c-codning" finder det estimerede label som indeks for det største element tilhørende $\boldsymbol{x_n}$.

5.2 Generaliseret lineær classifier

Single-layer Neuralt Netværk er en forlængelse af det før omtalte Lineære Neurale netværk. Det kan forklares ved at udtrykke posterier sandsynlighenden $P\langle C \,|\, x_n \rangle$ ved

$$P\langle K = k \,|\, \mathbf{x} \rangle = \frac{exp(\mathbf{w}_k \mathbf{x})}{\sum_{j=1}^{N_c} exp(\mathbf{w}_j \mathbf{x})} \tag{6,6}$$

hvor $\mathbf{x}$ er den udvidede featurevektor benyttet i (2). En af fordelene ved at bruge SNN er at den tvinger $P\langle C \,|\, \mathbf{x}_n \rangle$ til at ligge imellem 0 og 1 og giver dermed et mere realistisk estimat.

Den logaritmiske error funktion kan skrives som

$$\begin{aligned} E &= \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbf{t}_k^n \, log \frac{exp(\mathbf{w}_k \mathbf{x}_n)}{\sum_{j=1}^{K} exp(\mathbf{w}_k \mathbf{x}_n)} \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} -\mathbf{t}_k^n \, log\left(1 + \sum_{k\neq 1} exp(\mathbf{w}_k \mathbf{x}_n)\right) \end{aligned} \tag{5,7}$$

hvor $\mathbf{t}_k$ angiver label for sample *n* ved brug af "1-of-c-coding".
Givet en ny feature vektor $\boldsymbol{x_n}$ kan udtrykket i (6) benyttes til at finde posterier sandsynligheden, der senere i efterbehandlingen kan benyttes til at beregne genre label for den pågældende sang.

5.2 Gaussisk klassifier

Til at klassificere et datasæt, kan det være meget en nyttig antagelse, at det pågældende datasæt har en bestemt form, der kan beskrives ved sansynligheds fordelingen *p*(x), hvor en række justerbare parametre indgår. Disse parametre kan derefter optimeres sådan *p*(x) bedst tilpasser det pågældende datasæt.
En simplel og meget brugt model, er den såkaldte Gaussike fordeling, der har flere fordelagtige analytiske og statistiske egenskaber.
For et feature rum bestående af *d* dimensioner, kan en feature vektor $\boldsymbol{x_n}$ med genre indeks *k* beskrives ved

$$p\langle \mathbf{x}_n \,|\, K = k \rangle = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} exp\left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \tag{6,8}$$

hvor vektoren $\boldsymbol{\mu}_k$ består af middelværdierne for den pågældende genre og Σ_k er en *dxd* kovarians matrice med determinanten $|\Sigma_k|$.
Ved brug at den Gaussiske model antages, at de observerede data for datasæt $(k_{s(j)}, \mathbf{x}_j)$ er uafhængige af hinanden, hvor indekset *j* bevæger sig over alle feature vektorer i datasættet og $c_{s(j)}$ er det tilsvarende genre label. Funktionen *s(j)* repræsenterer indekset for det pågældende musiknummer til feature vektor *j*. Denne antagelse om uafhængighed er diskuteret før, da valget stod imellem at vælge *ICA* og *NMF* til kildesepartion. Alligevel er der valgt at benytte den Gauissiske klassifier, da den i flere tilfælde har vist sig at give gode resultater[4]. Ved hjælp at den nævnte antagelse om uafhængighed, kan den logaritmiske sandsynlighed udtrykkes ved

$$\begin{aligned} L &= log\left(k_{s(1)},..,k_{s(M)},x_1,..,x_M\right) = log\prod_{j=1}^{M} p\left(k_{s(j)},x_j\right) \\ &= \sum_{j=1}^{M} log\, P(K=k_{s(j)}) + \sum_{j=1}^{M} log\, p\left\langle x_j \,\middle|\, K=k_{s(j)}\right\rangle \end{aligned} \tag{5,9}$$

Hvor *M* er det samlede antal feature vektorer i datasættet.
Ved at maksimere udtrykket for *L*, kan man estimere parametrene for modellen [Bishop s.41] og dermed finde $\boldsymbol{\mu}_c$ og Σ_c. Estimatet for *P(C)* bliver da ganske enkelt en normaliseret optællling af forekomster for hver enkel klasse. Efter at have fundet nødvendige parametre er det muligt at forudsige $P\left\langle K \,\middle|\, x_n\right\rangle$ for nye feature vektorer der tilhører et test-datasæt. Ved brug af *Bayes'* regel kan den forudsagde sandsynligehed for hver enkel genre *k* udtrykkes ved

$$P\left\langle K=k \,\middle|\, x_n\right\rangle = \frac{P(K=k)p\left\langle x_n \,\middle|\, K=k\right\rangle}{\sum_{j=1}^{N_c} P(K=j)p\left\langle x_n \,\middle|\, K=j\right\rangle} \tag{5,10}$$

hvor N_k er antal genrer. Udtrykket i (10) er således output for hver feature vektor $\boldsymbol{x_n}$.

[4] *ICA* der ikke har påvist specielt gode resultater til enkeltkanals separtion af kildesignaler.

Den Gaussiske klassifier er bedst egnet til datasæt hvor dimensionen er forholdvis lille. Er datasættet for stort bliver covarians matricen upålidelig.

5.3 Gaussisk miksning classifier

Den gaussiske miksning classifier er tæt knyttet til til den før beskrevne Gaussiske classifier. I stedet for at modellere de pågældende data med en enkel gaussisk fordeling, mikser den flere fordelinger sammen.

$$\mathrm{p}\langle \mathrm{x_n} \,|\, \mathrm{K}=\mathrm{k} \rangle = \frac{1}{(2\pi)^{\mathrm{d}} |\Sigma_{\mathrm{k}}|^{1/2}} exp\left\{ -\frac{1}{2} (\mathbf{x}_{\mathrm{n}} - \boldsymbol{\mu}_{\mathrm{k}})^T \Sigma_{\mathrm{k}}^{-1} (\mathbf{x}_{\mathrm{n}} - \boldsymbol{\mu}_{\mathrm{k}}) \right\} \quad (5,11)$$

hvor *k* er indekset for den pågældende miksning og

$$p\langle \mathbf{x}_n \,|\, K=k \rangle = \sum_{l=1}^{L} P\langle \mathbf{x}_n \,|\, L=l \rangle P\langle L=l \,|\, K=k \rangle \quad (5,12)$$

hvor L er det totale antal miksnings komponenter. På linie med den gaussiske classifier findes miksningins parametrene $\boldsymbol{\mu}_{\mathrm{k}}$ og Σ_{k}, såvel som *P(C)* og $\mathrm{P}\langle \mathrm{K} | \mathrm{C} \rangle$ ved brug af maksimum likelihood metoden og antagelsen om uafhængighed imellem observationerne $(\mathrm{c}_{\mathrm{s(j)}}, \mathbf{x}_{\mathrm{j}})$. Maksimum for den logaritmiske sandsynlighed kan dessværre ikki findes analytisk. Derfor benyttes EM-algoritmen [2], [3] til iterativt at søge efter maksimum. Efter at systemet er trænet, benyttes udtrykket i (10) til at finde sandsynligheden for at en ny featurevektor tilhører en bestem klasse.
Der benyttes Netlab Matlab i forsøgsopstillingerne.

Kapitel 6

Forsøg og resultater

I de foregående kapitler er en række algoritmer introducerede, der benyttes til at udtrække features fra et musiksignal, her iblandt metoder til separation af kildesignaler. Der er også præsenterede to klassificere.
I dette kapitel vil de beskrevne modeller blive undersøgte og testede.

Første afsnit undersøger de beskrevne metoder til separation af kildekomponenter. Der benyttes hovedsagligt lyttetest til at vurdere kvaliteten for separationsprocesserne. Der opstilles et mindre forsøg hvor der udføres tests på kunstigt sammensatte signaler, denne metode giver mulighed for direkte at sammenligne de fundne kildesignaler med de originale kildesignaler.
Der undersøges hvilken af de to beskrevne metoder til identifikation at trommekomponenter giver det bedste resultat. Dette foregår ved at benytte lyttetest fra autentiske signaler.

Efterfølgende afsnit omhandler tests af selve klassifikationssystemet, her benyttes de separerede trommesignaler til klassifikation af musiksignaler efter musikgenre.

Der benyttes to musikdatabaser. Den ene stammer fra den private samling anskaffet til dette projekt. Den anden består af musik lånt fra IMM. Alle tests er implementerede og udførte i Matlab.

6.1 Separation af kildesignaler

Der findes som nævnt før, ikke nogen generel metode til at vurdere separation af kildesignaler. Derfor benyttes der i litteraturen hovedsagligt lyttetest, hvor der lyttes til de fundne kildesignaler, for derefter at vurdere kvaliteten.

Da der ikke eksisterer nogle anerkendte metoder til at vurdere kvaliteten af separationsprocesserne, benyttes der også i dette projekt lyttetest, til at vurdere kvaliteten for de fundne kildesignaler. Kvaliteten vurderes i forhold til hvor godt signalet lyder i forhold til det originale kildesignal.

Der er udført lyttetest på 100 forskellige musiknumre, ligeligt fordelt over musikgenrerne, hvor der meget intensivt er eksperimenteret med de forskellige valgfrie parametre for at finde den optimale opstilling. Samtidig er der taget stikprøver under hele projektforløbet og specielt under selve klassifikationsprocessen og lyttet til de fundne komponenter, for at høre om de har en rimelig kvalitet.

Det er vigtigt at finde en så robust metode som muligt, der er i stand til at identificere kildekomponenterne uafhængig af genre og lydmæssig struktur.

Baseret på de forskellige lyttetest, findes der ud af at et instruments rolle for det samlede lydbillede varierer meget genrerne imellem. I f.eks. Country-genren benyttes forholdsvis sjældent elektroniske musikinstrumenter, imens disse udgør en stor del (måske hoveddelen) af instrumenterne i Dance-genren. Derfor kan det være svært at opstille nogle i forvejen definerede forventninger til et instruments rolle, da den kan variere meget genrerne imellem..

Derfor opleves der også at være en sammenhæng imellem kvaliteten af separationsprocessen og den tilhørende musikgenre. F.eks. er slagtøjsinstrumenterne typisk af meget rimelig god kvalitet for Dance-musik, imens for Country er det andre kildesignaler der er af bedre kvalitet.

En separationsprocedure, der virker godt på en bestemt musikgenre, kan derfor give et relativt dårligere resultat for en anden musikgenre. Derfor er det nogle gange nødvendigt, at gå på kompromis med kvaliteten af kildesignalerne, for at sikre at separationsproceduren er i stand til at finde de ønskede kildesignaler, uafhængig af genre. Dette gælder især for en klassifikationsproces, hvor der benyttes flere hundrede musiknumre til at træne på. Da er det nødvendig at have sikkerhed for, at systemet er i stand til at finde det kildesignal, som features skal udtrækkes fra.
Den varierende kvalitet af kildesignalerne i forhold til genrerne, er en problemstilling, som er svær at komme udenom. I sidste ende kan det medføre, at features fundne fra kildesignaler af dårlig kvalitet, kan give et misvisende billede af musiksignalet og dermed ikke være egnet som basis for sammenligning i featurerummet.

Til at teste WNF2D benyttes Matlab-kode fra [11]. Spektrogrammerne overføres fra og til det oktav opdelte logaritmiske frekvensdomæne ved at benytte Matlab kode fra [3], hvor frekvens komponenterne vægtes tilnærmelsesvis fra det logaritmiske domæne over igen til det lineære domæne

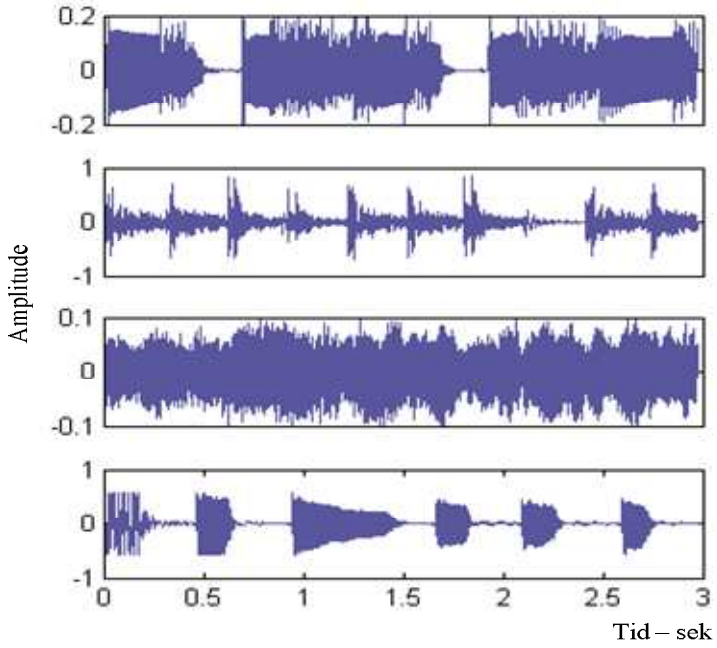
Af de i kapitel 3 beskrevne separations-procedurer, giver metoden baseret på en PWNMF de bedste resultater baseret på lyttetest. Dette gælder både når der benyttes enkle musiksignaler bestående af kun to lydkilder, men også de mere komplicerede lydbilleder, hvor der eksisterer mere en fem kildesignaler. Forskellen imellem WNF2D og PVNMF er dog mindre, når musiksignalet lyder mere enkelt, med ikke for mange toner tilstede.
Der undersøges også hvor mange komponenter er nødvendige for at opnå en god lydklvalitet for de fundne kildesignaler. Der observeras at det kun er nødvendigt med 25-28 komponenter for PWNMF for at opnå en god lydkvalitet der er relativt stabil genrerne imellem. For NMF2D er det nødvendigt at benytte mindst 50 komponenter for at opnå en rimelig kvalitet, hvor tau og phi sætte til henholdvis 30 og 20. Alle de opnåede resultater er fremkomne ved manuelt at identificere de fundne komponenter.
Der er dog indikationer om at kvaliteten forbedres ved at øge om antallet af komponenter, men når antallet er større end 50, bliver det meget svært at skelne imellem komponenterne ud fra hørelsen Denne egenskab er dog meget mere markant for NMF2D

Der undersøges hvorvidt kvaliteten af PWNMF er afhængig af signalets format. Her observeres at den bedste kvalitet fås ved en samplingsfrekvens på 44,1 kHz. Samtidig observeres at down-sampling forværrer kvaliteten af separationsprocessen, der oberserveres også en forværring af kvaliteten ved konvertere filerne fra WAV (PCM) til MP-3 –format.

6.1.2 Separation af et syntetisk signal

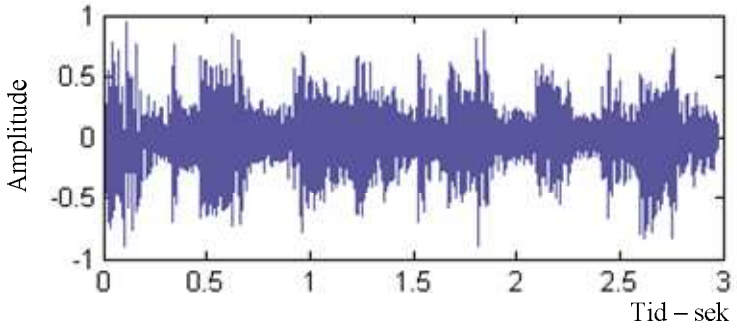
Foruden de før beskrevne lyttetest, vil der også blive benyttet et antal kunstigt sammensatte signaler, til at undersøge kvaliteten af kildeseparations-procedurerne. De sammensatte musiksignaler består af lydsignaler, fra lydfiler hentet fra Internettet, for siden hen at blive satte sammen ved hjælp af summation, til at udgøre det endelige musiksignal.
Kildesignalerne er derfor ikke afhængige af hinanden, i forhold til harmoni og rytmik, da de egentlig intet har med hinanden at gøre, før de sættes sammen. Alligvel vurderes resultaterne at kunne give en god indikation om hvilket separationsprocedure der performer bedst. Fordelen ved at benytte et kunstigt sammensat signal, er at de originale kildesignaler er til rådighed og kan derfor benyttes i en direkte sammenligning i forhold til de funden kildesignaler.
De sammensatte signaler består af fire signaler, der hver især stammer fra fire forskellige musikinstrumenter. Der benyttes lydsignaler fra akustisk guitar, basguitar, trommer og synthesizer, der alle har alle en samplingsfrekvens på 44100 Hz. De sættes sammen sådan, at de lyder (nogenlunde) realistiske, i forhold til hvor fremtrædende de er i lydbilledet.

Til at måle kvaliteten af de separerede kildesignaler benyttes ”signal-to-noise ratio” *SNR*. Udfaldet fra denne målemetode besår kun af et enkelt talværdi og er ikke nødvendigvis sigende om kvaliteten for hele signalet. Der findes avancerede metoder til at måle de lokale ændringer i signalet, de bl.a. [3]. Til illustrere de separerede signaler i forhold til det originale signal over hele signallængden, afbildes et af de kunstigt sammensatte signaler, i forhold til de fundne kildesignaler.
Figur 3.1 afbilder de fire benyttede kildesignaler der benyttes til illustrationen.



Figur 6.1 Afbilder amplituderne i forhold til tiden for de fire benyttede kildekomponenter. Fra øverst er afbildet en akustisk guitar, trommer, synthesizer og basguitar.

Efter at kildekomponenterne er sat sammen ved, fås det miksede signal afbildet i figur 3.2 nedenfor.

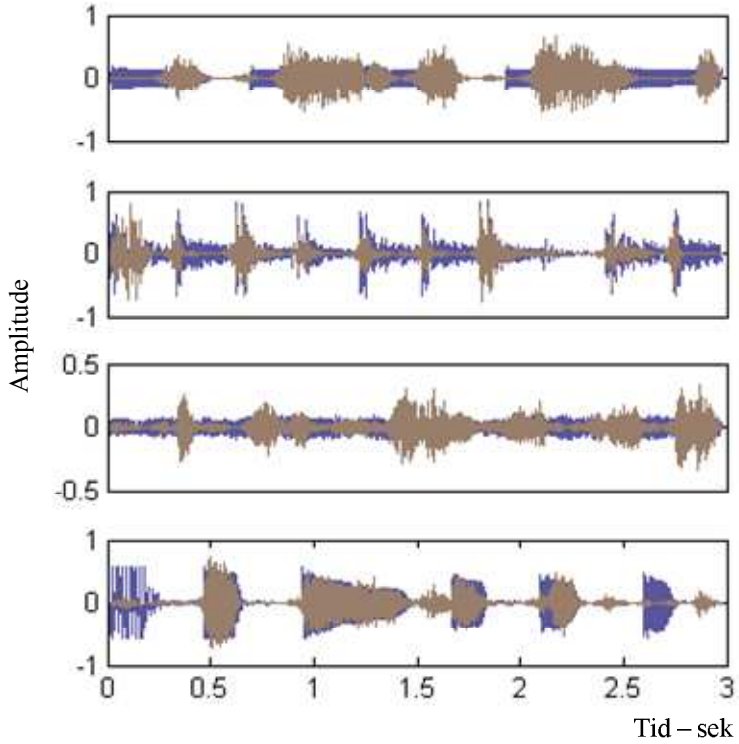


Figur 6.2 Afbilder det sammensatte signal, der benyttes til at teste kvaliteten af kilde-separationen.

Undersøgelsen af separationsmetoderne er delt op i to dele hvor den første del undersøger selve kvaliteten af de separerede komponenter i forhold til de autentiske komponenter. Der benyttes manuel identifikation af de fundne komponenter. Derefter sættes komponenterne sammen ved summation til at danne de endelige kildesignaler. Anden del undersøger hvilken proces er bedst til automatisk at genkende trommekomponenter fra de andre.

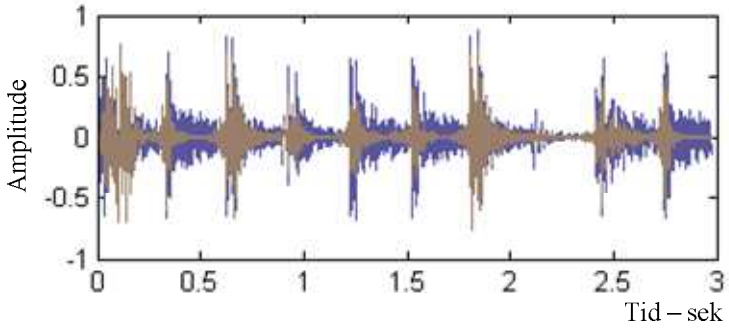
6.1.3 Perceptionelt vægtet NMF på syntetisk signal

Til den direkte perceptionelt WNMF bliver der prøvet en række separationsprocesser hvor antallet komponenter varierer imellem 10 og 30 da dette er det højeste antal komponenter hvor man kan identificere komponenterne manuelt. De bedste resultater fås ved at benytte 25 komponenter hvor længden på NMF sættes til 4096 punkter. De separerede komponenter er vist i figur 3.3 sammen med de originale komponenter.



Figur 6.3. Viser de separerede kildesignaler sammen med de originale kildesignaler afbildet med blåt. Fra øverst er afbildet en akustisk guitar, trommer, synthesizer og basguitar**.**

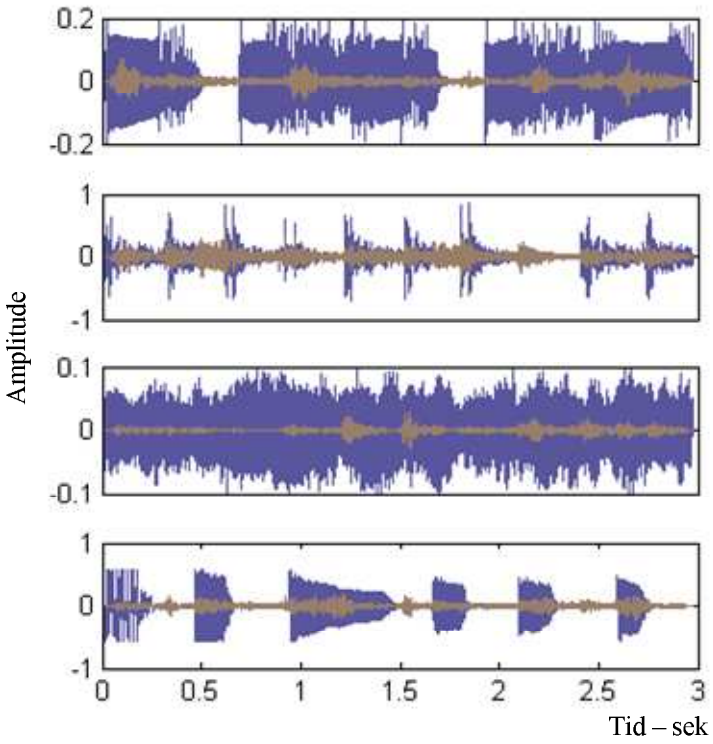
Der ses på figur3.3 at alle de separerede kildesignaler kommer rimelig tæt på de originale signaler. Det bemærkes at for trommesignalet afbildet nr. to fra oven, er processen i stand til at opfange det rytmiske mønster meget godt i forhold til det originale signal. Der observeres også at for basguitar-signalet er der flere steder næsten fuldkomment overlap, dog er det visser steder som ikke passer sammen. De dårligste resultater fås for den akustiske guitar og synthesizer der delvis er overdøvede af støj. Lydmæssigt lyder trommesignalet en hel del bedre end de andre og lyder meget ens med det originale signal. Det er primært lydkvaliteten i form af efterklang effekter der gør at det originale signal lyder bedre.
De tre andre signaler bærer alle præg at, at være påvirkede af hinanden, hvor man kan høre små dele af signalerne være hæftet på hos hinanden. Dette er dog ikke lige mærkbart for trommesignalerne.



Figur 6.4 Ovenfor afbilder kildesignalet for trommerne sammen med det originale trommesignal. Her er der i stedet for at benytte den originale fase for hele signalet, benyttet den originale fase for trommesignalet. Der observeres at forskellen er meget lille i forhold til afbildningen af trommesignalerne i figur 3.3.

6.1.4 NMF2D på et syntetisk signal

Til at test separationsmetoden baseret på NMF2D benyttes 20 til 100 komponenter. Hvis antallet af komponenter bliver større end dette er det meget svært at identificere komponenterne ud fra hørelsen. De bedste resultater fås ved at benytte 70 komponenter, men dette er dog svært at sige da det for mange komponenter var umuligt at høre hvilket kildesignal de hører til. De eneste signaler, som kan identificeres nogenlunde, er basguitaren og trommerne som er afbildet i figur 3.6.



Figur 6.5. Viser de separerede kildesignaler sammen med de originale kildesignaler afbildet med blåt. Fra øverst er afbildet en akustisk guitar, trommer, synthesizer og basguitar**.**

Det observeres fra figuren ovenfor, at det fundne trommesignal passer flere steder godt sammen med det originale signal, dog kan man se og høre at signalet bærer præg af en hel del støj. Basguitar-signalet har ikke mange ligheder med det originale signal, men når man lytter til signalet, kan man sagtens høre at der er tale om basguitaren. Signalerne for guitar og keyboard kunne nærmest ikke identificeres, dette observeres også på figuren ovenfor, hvor de er nærmest helt fladede ud. Man kan dog høre dele af guitarsignalet i baggrunden for tromme - og basguitar - signalet.

SNR Måleresultater

Der benyttes det såkaldte "Signal-Noise-Ratio" (SNR) til at målde kvaliteten på det separerede trommesignal i forhold til det originale trommesignal. Dette gøres ved at dele signalerne op i segmenter på 30 ms for siden at beregne SNR for hvert enkelt segment ved at benytte udtrykket

$$SNR_i = 10\,log\left(\frac{\left[\sum \mathbf{s}_i(n)^2\right]}{\left[\sum \mathbf{e}_i(n)^2\right]}\right) \qquad n = 0,\ldots N-1$$

hvor $e_i(n) = s_i(n) - y_i(n)$ og $y_i(n)$ er det estimerede trommesignal til segment indeks *i og* s(n) er det originale trommesignal. Det samlede SNR findes derefter som middelværdien for de fundne SNR værdier. Beregningerne udføres på de fem kunstigt sammensatte signaler. Tabellen nedenfor viser gennemsnits SNR for de 5 signaler.

Tabel 1 Viser gennemsnits SNR for 5 kunstigt fremstillede signaler.

	PWNMF	**NMF2D**
SNR (dB)	6.24	2,31

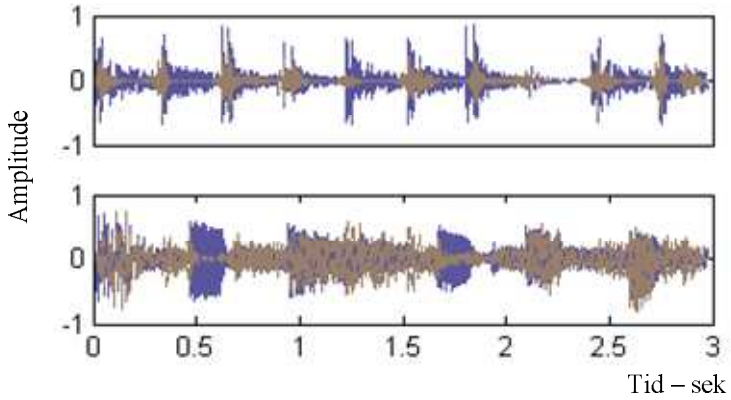
Ud fra de observationer, der er lavede, hvor der er lyttet til en række autentiske musiksignaler der har gennemgået en af de to omtalte separationsprocesser samt undersøgt kvaliteten i forhold til syntetiske signaler, er det klart, at kvaliteten på trommesignalerne, der opnås ved brug af den perceptionelt vægtede NMF er at foretrække i forhold til den kvalitet af trommesignalerne opnået ved NMF2D. Derfor benyttes den perceptionelt vægtede NMF til videre brug i dette projeket.

6.1.5 Automatisk identifikation af kildesignaler

Dette afsnit tester automatisk identifikation af kildekomponenterne ved hjælp af peak-detektor og linæer klassifikation af kildekomponenter, som før beskrevet i kapitel 3. Der benyttes PWNMF og 25 komponenter til at separere 50 udvalgte autentiske signaler.

Det blev hurtigt klart at den metode, der benyttede en peakdetektor, var klart den bedste til at identificere trommekomponenterne. Den metode der benyttede en lineær klassifier, blev trænet op af 600 kildesignler, der var separerede i tre komponenter, hvorfra der blev udtrukket 10 MFCC fra komponenterne og benyttet til at træne den lineære klassifier med.
Årsagen til at denne metode virkede så meget værre end den anden, vurderes hovedsagligt at være på grund af kildesignalerne, der blev benyttede til at træne den lineære klassifier. Det vidste sig at være meget svært at anskaffe musikfiler, bestående kun af enkelte musikinstrumenter. Selv om der blev benyttede 600 kildesignaler at træne på, så vurderes variationen af de forskellige kildesignaler at være for lille, til at kunne dække hele spektret af mulige kildesignaler, hvor der benyttes flere hundrede musiksignaler, fordelt over forskellige musikgenrer.

Den metode der benyttede en peakdetektor vidste sig derimod af være ret effektiv til at identificere trommekomponenterne. Metden blev afprøvet på det syntetikse signal benyttet i forrige afsnit. Resultaterne for det fundne trommesignal vises i figur 3.5 nedenfor.



Figur 6.6 Afbilder resultaterne fra peakdetektoren hvor kildekomponenterne findes ved brug af den vægtede NMF. Den øverste delfigur afbilder det identificerede trommesignal i sammen med det originale trommesignal. Den nedre delfigur viser resten af kildesignalerne i forhold til de originale.

I figur 3.5 ses at det fundne trommesignal har det samme mønster som det originale trommesignal. Samtidig observeres at det ligner meget det trommesignal, fundet ved manuel identifikation af komponenterne. Når der lyttes til det fundne

trommesignal synes der også at være at rimelig god lydkvalitet, dog kan man høre at dele af basguitaren er også kommet på. Dette ville i praksis ikke være et stort problem, da basguitaren typisk følger det rytmiske mønster. For de kunstige signaler, der benyttes her, hører basguitaren og trommerne ikke sammen og derfor følges de selvfølgelig ikke ad.
Trommedetektoren viste sig dog at være betydelig mere effektiv på kildekomponenter der stammer fra autentiske signaler.

6.2 Trommesignaler til klassifikation af musik genre

Dette afsnit beskriver de forsøg, der i dette projekt er udførte, til at undersøge hvordan features udtrukket fra et trommesignal, fungerer som input til et musikgenre klassifikationssystem.

Der er benyttet to musikdatabaser til forsøgene. Musikdatabase A består af 312 musiknumre fordelt over 5 genrer. Musiknumrene stammer fra opsamlingsalbums for de forskellige musikgenrer. Genrerne er Heavy Metal, Country (Oldies), Blues, Pop og Dance. Genrerne er relativt snævert opdelte. Samplingsfrekvensen er 44,1 kHz og WAV-format.

Årsagen til den snævre opdeling, er at en rytmisk feature vurderes at kunne virke mere diskriminerende, hvis genrerne er mere præcist opdelte. En snæver opdeling burde også kunne formindske overlappet genrerne imellem. Dog forventes tradition blues og tradition country at have ligheder, ligeledes som pop og dance også har ligheder.

De 5 genrer og antal musikfiler pr. genre vises i tabellen nedenfor

Tabel 2 **viser den indbyrdes fordeling af klasserne.**

Genre	Antal
Heavy Metal	61
Pop	63
Dance	64
Blues (Tradition)	59
County (Tradition)	65

Musikdatabase B består af 800 musinumre fordelt over 8 genrer. Denne database er benyttet til musikgenre klassifikation i [11,12], hvor der eksisterede tre ekstra klasser. Musiknumrene er opsamplede fra 22.5 kHz til 44,1 kHz og konverteret fra MP-3 til WAV format.

Til at måle performance benyttes den såkaldte generaliserede klassifikationskorrekthed, der i procent beskriver hvor stor andelen af korrekt klassificerede eksempler er, i forhold til det fulde antal sande klasser. Til at afbilde hvilke genrer, systemet er bedst til at klassificere, benyttes den såkaldte konfusion matrice, der afbilder hvilke klasser er forkert klassificere i forhold til den sande klasse.

Der er i projektforløbet udførte en række forsøg, hvor de før beskrevne features er udtrukket fra trommesignalet. I dette kapitel diskuteres hovedsagligt den opstilling, der gav de bedste resultater. Resultaterne for de opstillinger der gav mindre gode resultater er henviste til bilag 1.

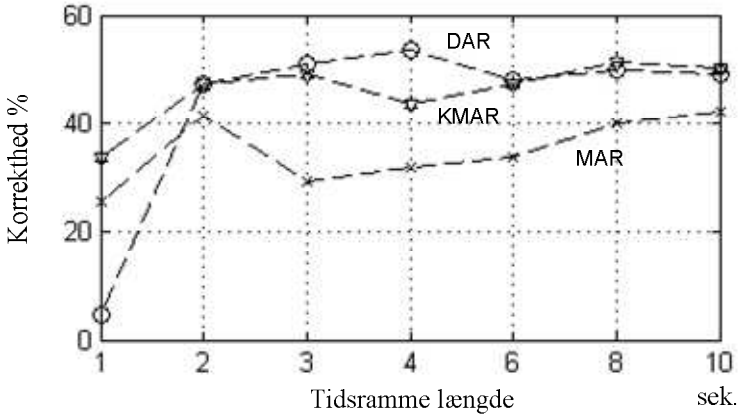
De udførte forsøg benytter først den mindre database A, til at finde den opstilling, der giver den bedste performance i form af klassifikations korrekthed. Derefter benyttes denne opstilling på den større database A, for at se om disse features også fungerer på en database, hvor klasserne i højere grad overlapper hinanden.

Korttidslige features benyttes ikke som enkelt stående input, i nogen af de opstillede forsøg. Dette er fordi at den rytmiske information, som er hovedinteressen her, vurderes at leve på en længere skala, som de korttidslige features ikke opfanger enkeltvis.

Til at integrere korttidslige features op på en længere skala findes DAR, MAR og *KMAR* ud fra de korttidslige features. KMAR er den feature, der er opnås ved projektion af MAR features hvor kernelmodellen rKOPLS (R=100) benyttes.

De korttidslige features der giver de bedste resultater består af 5 MFCC koefficienter (inkl. nulte-koef.) med 40 filtre og tidsrammer på 30 ms med 15 ms overlap. Derfor benyttes denne opstilling for korttidslige features sammen med den generaliserede lineære klassifier.i dette kapitel. Resultaterne for de andre kortidslige features og GMM er henviste til bilag 1.

For at finde det optimale tidsingerval som korttidslige features skal udtrækkes fra, for derefter at omdannes til AR-features. Undersøges tidsintervaller af forskellig længde, der varierer fra 1. til 10. sek. Figuren nedenfor viser hvordan klassifikationskorrektheden for de tre nævnte tidslige features varier i forhold til længden på tidsintervallerne for et trommesignal.



Figur 6.7. Det ses på figuren overfor, at det fås den laveste klassifikations korrekthed, ved at udtrække features fra ét sekunds længde af signalet. Det observeres samtidig, at det bedste resultat fås ved at benytte DAR features over 4. sekunder hvor klassifikationskorrektheden er på 54%. Ved tidsrammer på 8. sek. fås næsten et lige så godt resultat på 52% ved at benytte KMAR, hvor den performer en smule bedre end DAR. Der ses også at MAR giver dårligst resultater for alle tidsinterval længder.
Der er benyttede 1l musiksignaler for hver genre at teste klassifikationskorrektheden. Den performance afbildet i figuren ovenfor, er middelværdierne for klassifikations korrektheden, fundet ved at træne klassifikationssystemet 10 gange, hvor der for hver gang benyttes 55 nye. Dermed vil performance for klassifieren ikke være afhængig af det ene sæt træningsdata.

Figuren nedenfor viser konfusion matricen for database A

Confusion matrice Dataset A – MFFC_4(5) DAR

	H.M	**Pop**	**Country**	**Blues**	**Dance**
Heayv Metal	**36.6**	21.9	4,6	0	36.9
Pop	4,6	**59,1**	9,1	0	27,3
Country	0	4,6	**81,8**	13,6	0
Blues	4,6	9,1	68,2	**18,2**	0
Dance	4,6	18,2	4,55	0	**72,7**

Det ses på konfusion matricen at Country, Blues og Heavy Metal giver de bedste resultater. Dance derimod giver klart de dårligste resultater hvor den overlapper med Heavy Metal, men det er særdeles overlap med Pop der påvirker resultaterne.

Den samme opstilling benyttes også for en større database, hvor der er betydeligt mere overlap genrerne imellem. Da fås en klassifikationskorrekthed på 42 %. Konfusion matricen er vidst nedenfor.

Confusion matrice Dataset B - MFFC_4(5) DAR features

	Cou	**Alt**	**Eas**	**Rock**	**Reg**	**R&H**	**Rb&S**	**Rb&S**
Country	**15,0**	5,0	25,0	15,0	0,0	5,0	10,0	25,0
Alternative	9,5	**28,6**	4,76	23,8	9,52	0,0	0,0	23,8
Easylistning	10,0	15,0	**45,0**	5,0	5,0	5,0	0,0	15,0
Rock	0,0	5,0	10,0	**40,0**	10,0	20,0	15,0	0,0
Reggae	0,0	0,0	0,0	15,0	**15,0**	15,0	55,0	0,0
RapHipHop	9,1	9,1	0,0	0,0	9,1	**36,4**	9,1	27,3
Rb&SOul	0,0	0,0	0,0	15,0	15,0	15,0	**55,0**	0,0
Country	10,0	45,0	0,0	0,0	5,0	15,0	0,0	**25,0**

Der ses ud fra konfusion matricen at bl.a. country og RapHipHop performer forholdsvis dårligt imens Rock giver udmærkede resultater.

6.3 Diskussion

Der i dette kapitel beskrevet to separate undersøgelser af de metoder introducerede i denne afhandling. Afsnit 6.1 omhandler de i kapitel 3 introducerede metoder til separation af kildesignaler. Imens afsnit 6.2 undersøger hvorvidt features udtrukket fra trommesignalet for et musiksignal er egnede til musikgenre klassfikation. Diskussionen her vil først behandle undersøgelserne til separation af kildesignaler.

Da der som nævnt før ikke eksisterer nogen generel metode til at estimere separation af kildesignaler, så er det hovedsagligt lyttetest der benyttes til at vurdere de afprøvede metoder. De indledende undersøgelser gik ud på at finde hvilken af de to metoder PWNMF eller NMF2D var bedst egnet til videre brug i dette projekt. Der var særlig interesse for at finde en metode der var i stand til at separere trommesignalet fra et musiksignal. Til denne undersøgelse blev der udført lyttetest på 100 musiksignaler hvor de fundne kildekomponenter blev manuelt sorterede i forhold til tilhørende kildesignal. Metoderne blev også afprøvede på 5 forskellige kunstigt sammensatte musiksignaler. Her var der mulighed for at sammenligne metoderne ved at benytte støjforholdet SNR som indeks for hvor gode metoderne var til at separere trommesignalerne.

Både lyttetest og SNR gav en god indikation om at PWNMF var den metode der bedst egnet til at separere trommesignalet fra et musiksignal. Specielt var det lyttetestene, hvor der blev brugt autentiske musiksignaler, der klart viste at PWNMF gav en bedre kvalitet af kildesignalerne. Denne forskel var dog mindre åbenlys når der blev benyttet kunstigt sammensatte signaler. Selv om der var en forskel på SNR, var den hørbare kvalitet næsten lige så god og i nogle tilfælde bedre ved brug af NMF2D. Forskellen bestod dog i at PWNMF var bedre til at opfange det fulde signal, hvorimod ved at benytte NMF2D manglede der nogle dele af det fulde trommesignal. Dog syntes kvaliteten at kildekomponenterne fundet ved at benytte NMD2D at være mere klare, hvorimod kildekomponenterne fundet ved hjælp af PWNMF kunne lyde lidt forvrængede eller overstyrede. Hvorfor NMF2D virkede så meget bedre på de syntetiske signaler, kan der være svært at give en præcis forklaring på, men en af årsagerne kunne tænkes at være, at de syntetiske signaler havde en noget enklere lydmæssig struktur end det typiske autentiske signal, derfor vil det være nemmere for algoritmerne at opfange tonerne, også ved brug af forholdsvis få komponenter. Dette kunne være en indikation om, at NMF2D ved brug af flere komponenter, ville kunne give en meget bedre kvalitet af de separerede kildesignaler. Dette forventes dog også være tilfældet for PWNMF. Problemet heri består af, at når der benyttes over de 50 komponenter, bliver det meget svært, ud fra hørelsen, at afgøre hvilke komponenter, der tilhører et bestemt kildesignal. Samtidig eksisterer der ikke nogen algoritme, der effektivt er i stand til at sortere kildekomponenterne ud fra tilhørende kildesignaler.
Til at vælge hvilken af de to metoder der skulle benyttes videre i projektforløbet til at separere trommesignalet fra et musiksignal, var det af interesse, at den valgte metode var i stand til at separere trommesignalet ved brug af så få kildekomponenter som muligt, da dette vile gøre det nemmer for trommedetektoren at identificere komponenterne. Det vigtigste var dog, at metoden var i stand til at separere trommesignalet fra et autentisk musiksignal, af en så god kvalitet som muligt. Her var det klart at PWNMF var bedst på begge områder, til brug på autenstiske signaler. Derfor var dette det foretrukne valg til vide brug i projektforløbet.
Næste skridt bestod i at teste hvilke af de to introducerede metoder til automatiks identifikation af trommekomponenter var bedst egnet til videre brug. Her blev der også benyttede lyttetest til at finde hvilken ad de to metoder var bedst egnet. Her var det meget klart, at de bedste resultater fås ved at benytte den metode der var baseret på en iterativ peakdetektor. Den metode der benyttede en lineær klassfier, viste sig at give en hel del dårligere resultater. Hovedårsagen forventes at være at de lydsignaler som blev benyttet til at træne den lineære klassifier, at de ikke har været særlig velegnede. Muligvis burde der også være benyttet en mere avanceret klassifier.
Til videre brug i musikgenre klassifikationssystemet vil der derfor blive brugt en PWNMF til separation af kildekomponenter og peakdetektor til at identificere trommekomponenterne.

Næste del af dette kapitel omhandler som nævnt musikgenre klassifikation, hvor features ekstraheres fra trommesignalet. Der benyttes primært en database bestående af 5 forskellige musikgenrer til at teste en lang række features. Den opstilling der gave de bedste resultater benyttede 5 Mel Cepstrale koefficienter inklusivt den nulte koefficient, som korttidslige features. Tidslig feature integration blev udført ved at benytte såkaldte DAR featuers. Der blev her opnået er klassifikationskorrekthed på 54 %. Det ser ud til at det specielt er pop, country og dance der opnår de bedste resultater. Dermed har systemet relativt nemt ved at klassficere pop og dance der umiddelbart var forventet i højere grad at overlappe hinanden. Dog ses der at Dance delvis klassificeres som Pop og omvendt. De fejlklassificerede eksempler for Country har en lille overvægt hos Blues-genren. Dette er også forventet da disse to genrer i højere grad minder om hinanden i forhold til de andre genrer. Det ses for Blues, som giver de dårligeste resultater, at hoveddelen af de forkert klassificerede eksempler tilhørerer Country. Dette kunne være en indikation om, at de benyttede features er relativt gode til at opfange de benyttede musiksignaler som de opfattes af mennesker. Dette er fordi at Pop og Dance kan være svære at skelne imellem, her kan nævnes at i f.eks. [11,12] benyttes de som én genre.
Da der benyttes en rytmisk feature at klassificere efter, så er det også af interesse, at undersøge hvordan den fungerer ud fra et rytmisk synspunkt. Største delen af de eksempler, der bliver forkert klassificerede, ser ud til at ophobe sig ved en anden genre f.eks. blues klassficeres som country. Årsagen forventes at være at den rytmiske kontekst for disse signaler ligner meget det rytmiske kontekst for de signaler der tilhører den genre, som er mål for de forkert klassificerede eksempler. Derfor er den rytmiske feature forholdsvis effektiv til at sammenligne den rytmiske kontekst for signalerne, selv om dette indebærer, at de klassificeres at tilhøre en forkert musikgenre. Denne egenskab kan være et effektivt redskab for bl.a. musik rekomendationssystemer hvor fokus er på den rytmiske kontekst og ikke specielt musikgenren.
Til sammenligning med andre rytmiske features, så er der i [11] udført tests hvor to rytmiske features kaldet ”Beat Histogram” og ”Beat Spectrum” benyttes til musikgenre klassifikation af en musikdatabase, specielt opstilet sådan at der ikke eksister noget overlap genrerne imellem. Her opnås en klassifikations korrekthed på ca. 45 %. I forhold til disse resultater vurderes den rytmiske feature, introduceret i dette projekt, at performe relativt godt, selv om resultaterne er baserede på forskellige forudsætninger og derfor ikke er direkte sammenlignlige.

Der undersøges også hvorvidt den benyttede forsøgs opstilling kan gøre sig gældende når der benyttes musikgenrer, der er mere bredt definerede i forhold til den første database. Her opnås en klassifikationskorrekthed på 38 % som er en del mindre end for den første database. Årsagen vurderes hovedsagligt at være overlap genrerne imellem. Dog er dette resultat en indikation om at den rytmiske feature er mindre velegnet til musikgenre klassifikation hvis musikgenrerne er forholdsvis bredt definerede. Dog kan der ikke ses bort fra at disse musikfiler var oprindeligt downsamplede og konverterede til MP-3 format. Undersøgelser har vist at dette i høj grad påvirker kvaliteten af de separerede signaler, selv om de opsamples igen til det originale format. Derfor er der god mulighed for at trommesignalerne ikke har været af en kvalitet der er velegnet til at udtrække features fra.

Der er også undersøgt en række opstillinger af features der benytter MPEG-7 features også er de samme undersøgte udførte med den gaussiske miksningsmodel. Dog giver disse opstillinger ikke lige så gode resultater som den opstilling der giver de bedste resultater. Den MPEG-7 feature der giver de bedste resultater er baseret på Spektral Envelope hvor der benyttes DAR features til tidslig feature integration. Resultaterne for den Gaussiske miksnings model er ikke helt så gode som for den Generaliserde lineære klassfier. Den bedste klassifikationskorrekthed der fås her er 38% og der benyttes den samme opstillinger som for de bedste resultater ved brug af GLM for database A.

Kapitel 7

Windows Applikation

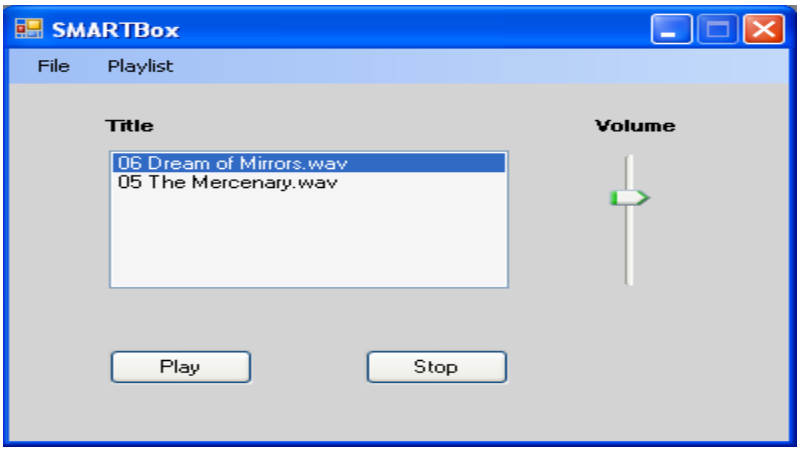
Der findes en række områder hvor de benyttede metoder kan indgå. Nogle af disse områder er omtalte i kapitel 2.
I dette projekt er det ønskeligt at implementere en applikation hvor nogle af de benyttede metoder til musikgenre klassifikation indgår og dermed vise dem i praksis. Derfor er der valgt at implementere en Microsoft Windows applikation, der er i stand til at klassificere et musiknummer valgt af brugeren og derefter generere en afspilningsliste, der stilmæssigt passer sammen med det afspillede musiknummer.

Det følgende kapitel giver en overordnet beskrivelse af hvordan applikationen fungerer og er implementeret.

Programmet er udviklet i objekt-orienteret C#-kode og .Net.

7.1 Funktionalitet

Applikationen bygger hovedsagligt på strukturen af det klassifikationssystem, der er benyttet i dette projekt. Brugerfladen er vist i figur 7.1 nedenfor



Figur 7.1 viser brugerfladen for den Windows applikation der er implemtneret til dette projekt. Det er muligt at gå under File og finde et musiknummer der ønskes afspillet. Under playlist er det muligt at sætte den default mappe som der ønskes søgt igennem for at genrere en afspilningsliste. Det er også her at en forespørgsel sendes om at at generere en playliste ud fra det markerede musiknummer. Det er muligt at afspille et musiknummer ved at enten trykke på play eller markere det i listboksen.

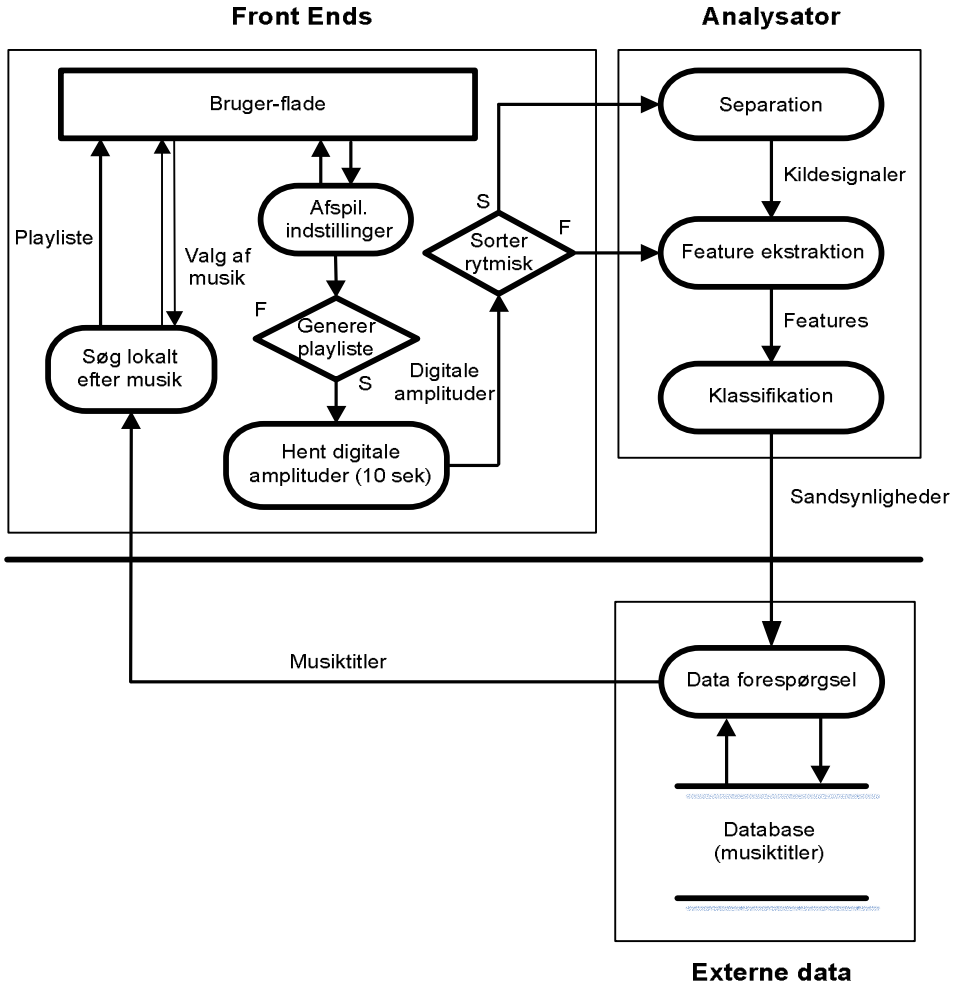
Applikationen er designet hovedsagligt til at håndtere to scenerier fra brugerens side. Det ene vil være, at brugeren kun er interesseret i at afspille musik, som han/hun selv har valgt og dermed danner sin egen afspilningsliste. Det andet scenario består i at systemet automatisk generer en afspilningsliste, baseret på de musiknumre brugeren har valgt.

Det er selvfølgelig det sidste scenario, som er hovedformålet med applikationen. Scenariet udspilles ved at brugeren, efter at have startet applikationen, vælger en mappe indeholdende de musikfiler, hvorfra der ønskes genereret en afspilningsliste. Systemmappen *musik* er defaultmappe.

Når brugeren har bestemt sig for et musiknummer, kan der genereres en afspilningsliste ud fra det valgte nummer. Da vil systemet begynde at analysere musiksignalet. Da analysen er færdig vil systemet søge igennem en database efter musiktitler, der passer sammen med det oprindelige musiksignal. Efter at have fundet det ønskede antal musiktitler, vil systemet gå ind i den mappe, som brugeren har valgt og søge efter de titler fundne fra databasen. Her vil der blive søgt igennem selve mappen og alle undermapper. De titler, der passer sammen med titlerne, der er fundet i databasen, vil derefter blive opstillede i det grafiske brugerinterface på listeform, hvor brugeren frit kan vælge imellem de fremkomne musiknumre.
I en kommerciel applikation, ville det bl.a. være oplagt, at henvise brugeren til en internetbutik, med mulighed for at købe de musiktitler, der ikke kunne findes på den lokale computer

7.2 Struktur

Systemet består af tre hoveddele. Det Front End, Analysator og ekstern database. Sammenhængen er illusteret i figur 6.



Figur 7.2 illusterer strukturen for Windows applikationen.

Via brugerfladen har brugeren mulighed for at søge lokalt på sin computer efter musik og vælge basale afspilnings indstillinger i form af f.eks. volumen, næste sang og stop.
Hvis brugeren indstiller systemet til at generere en afspilningsliste, så vil de digitale amplituder blive udtrukket fra det markerede musiknummer og sendt videre til behandling i analysatoren. Der er mulighed for at vælge at generere en afspilningsliste ud fra trommesignalet eller hele signalet. Hvis der benyttet trommesignalet, vil der blive udført en separation af kildesignalerne, hvor trommesignalet sendes videre i systemet og der udtrækkes features. Hvis trommesignalet ikke benyttes, udtrækkes de samme features direkte fra signalet.
Efter at de pågældende features er udtrukket bliver signalet klassificeret efter musikgenre, til dette benyttes vægtene fundet fra træningsfasen af klassifikationssystemet. I træningsfasen blev der for hvert musiksignal gemt en vektor, der indeholder sandsynlighederne for hvilken musikgenre signalet kan tilhøre. Denne vektor er blevet gemt i en database tabel, sammen med tilhørende sangtitel. Figur 3 viser et udsnit fra database opstillingen.

Da analysatoren har estimeret musikgenre label og sandsynlighederne for hver musikgenre som det valgte musiksignal kan tilhøre, da vil systemet søge i databasen, under den estimerede genre, med sandsynlighederne som søgekriterium. Titlerne hvor sandsynlighederne ligger tættest på sandsynligheder fundet af analysatoren, vil blive udtrukket fra databasen. Derefter søges efter de fundne titler i udvalgte mapper på den lokale computer. De titler der findes og dermed eksisterer på den lokale computer, sættes sammen til at danne en afspilningsliste.

7.3 Implementering

Til at implementere systemet benyttes den procedure, der har givet de bedste resultater, i de opstillede forsøg fra kapitlel 6.

Til at separere et musiksignal benyttes den PWNMF beskrevet i kapitel 5, hvor den separerer signalet ned i 25 kildekomponenter. Trommekomponenterne identificeres ved hjælp af peakdetektor metoden, også beskrevet i kapitel 5.

Da trommesignalet er genereret udtrækkes korttidslige features fra det i form af MFCC hvor de 5 første koefficienter benyttes, da disse gav de bedste resultater. De korttidslige features integreres derefter op på en længere tidsskala ved at

generere tilsvarende DAR features. Vægtene for klassifieren er allederede fundet fra træningsfasen i Matlab og genbruges derfor her.
Det kan variere lidt hvordan nogle af de før benyttede algoritmer kan implementeres. Til implementering af analysatoren, har det været ønskeligt, at resultaterne er identiske med dem opnået i Matlab, for at kunne sammenligne de forskellige beregninger i analysatoren, med tilsvarende beregninger i Matlab. Dermed kan man undgå fejl, som kan være svære at få i øje på i koden.
Derfor følger flere af algoritmerne kodet i C#, strukturen på tilsvarende Matlab-kode.
De funktioner som er indbyggede i Matlab, er ikke alle offentligt tilgængelige, derfor benyttes i disse tilfælde alternative implementeringsmetoder.

I C# er der ikke indbyggede funktioner på samme måde som i Matlab. Derfor er det nødvendigt i C# at implementere også de mere enkle beregningsmetoder..
Der er undersøgt hvorvidt "open source" biblioteker kunne hjælpe til at forkorte kode-processen. Disse er dog forholdsvis sparsomme i en pålidelig facon i C#-kode og ved nærmere undersøgelse viste flere af disse sig at give upålidelige resultater i forhold til tilsvarende Matlab implementeringer. Derfor er der valgt som udgangspunkt, at implementere størstedelen fra grunden af.

7.3.1 Analysator

Udgangspunktet for analysatoren er, at den skal kunne håndtere et input-array bestående af digitale amplitudeværdier.
De algoritmer der driver analysatoren, er matrice baserede og benytter sig også i flere tilfælde af komplekstals repræsentation. Derfor er der implementerede to klasser der benyttes til resten af koden til implementering af selve signalbehandlingen. Den ene klasse kaldes *Matop*, der indeholder de nødvendige metoder til at udføre basale matrice operationer (linear algebra), hvor input-matricerne er opstillede i arrays. Imens den anden kaldes for Complex til at udføre kompleks-tals operationer.
Fast Fourier transformation ingår også flere steder i algoritmerne. Derfor implementeres en klasse kaldet FFT, der indeholder to metoder, en til FFT og en til den inverse FFT. Denne klasse oversat fra java kode [19]. Beregning af spektrogram benyttes også flere steder. Derfor implementeres en klasse kaldet *Spektrogram* der indeholder metoder til at beregne spektrogram, inverst spektrogram og frekvensen for et signal.

Separation af kildekomponenterne og identificering af trommekomponenterne er implementerede i fire klasser. Klassen *Weigths* finder de perceptionelle vægte ved hjælp af loudness metoden beskrevet i kapitel 5. Derefter udføres WNMF implementeret i klassen af samme navn, hvor loudness modellen beskrevet i kapitel 5 benyttes. Kildekomponenterne synteseres i *GenerateComp*-klassen hvor den originale fase fra signalet benyttes. Efter at komponenterne er generede, kan de identificeres som tilhørende et trommesignal eller et harmonisk signal ved at benytte metoden beskrevet i 5.6 implementeret i klassen *DetectDrum*.

Til at udtrække features er der implementerede 2 klasser. Først beregnes MFCC ved at benytte klassen *MelCep*. Derefter Beregnes AR-features hvor *DAR-klassen* benyttes. Her benyttes fremgangsmåderne beskrevet i henholdsvis kapitel 6.1 og 6.4. Implementeringen af MelCep følger den i Matlab implementerede VoiceBox.

7.3.2 Front end

Til at implementere brugerfalden benyttes 6 klasser der delvis har til opgave at styre afspilningsfunktionerne i form af volumen, stop/play, men også tælle op hvor mange musiknumre der vises på afspilningslisten og hvilke er markerede til afspilning eller generering af en ny liste. Det er også denne del der der går ind og søger i forskellige foldere på den lokale PC for at finde musiktitler der er identificerede fra databasen til at passe sammen med det valgte musiknummer.

7.3.3 Externe data

For alle de musiknumre, der blev benyttede i forsøgsopstillingen, er vægtene fra den generaliserede lineære klassifier blevet gemt til senere brug i Windows applikationen. De er derefterv benyttede til at beregne genre sandsynligherne for en række musiknumre hvor titlerne optilles i sql database tabeller sammen med tilhørende sandsynligher. Dermed er der implementeret en klasse der ved at benytte ADO.Net går ind i database tabellerne og sammenligner de vægte fundet uf fra et musiknummer valgt af brugeren og benytter KL-divergens til at sammenligne med vægtene i databasen. De titler hvor KL-divergensen er lavest i forhold til det aktuelle musiknummer vil derefter blive udtrukket og danne basis for søgningen på den lokale computer, for derefter at generere en afspilningsliste.

Kapitel 8

Opsummering og konklusion

Dette kapitel vil opsummer de metoder der er benyttede i denne afhandling. Derefter gives en konklusion ud fra de opnåede resultater. Der gives også til sidst forslag til fremtidigt arbejde.

8.1 Opsummering

Denne afhandling har undersøgt hvorvidt trommesignalet fra et musiksignal kan indgå som en rytmisk feature i et musikgenre klassifikationssystem. Til at separere trommesignalet fra et musiksignal er der afprøvet to metoder, der bygger på ikke-negativ matrix faktorisering (NMF). Metoderne er testede i forhold til hinanden, med fokus på, i hvor høj grad de er i stand til at separere trommesignalet fra et musiksignal. For at få det bedst mulige resultat, blev der til denne del af forsøget, benyttet manuel identifikation af de separerede kildekomponenter. Ligeledes blev kvaliteten af de separerede kildesignaler vurderet ved hjælp af lyttetest. Dette er hovedsagligt fordi, der ikke findes en overordnet metode i litteraturen, der er i stand til at vurdere kvaliteten af en separationsproces. Af denne årsag er der også generede 5 syntetisk sammensatte musiksignaler, bestående af trommer, guitar, keyboard og basguitar. Disse blev siden hen separerede ved hjælp af de to metoder. Derefter blev de separerede kildesignaler direkte sammenlignede med de oprindelige kildesignaler. Her blev der brugt signal-to-noise ratio (SNR) til at vurdere kvaliteten af separationsprocessen.
Baseret på de udførte tests, blev der fundet ud af, at den metode, der gav den bedste kvalitet, var den såkaldte *Perceptionelt vægtede NMF*, der benytter en introduceret loudness model til at beregne vægtene. Separationen kunne udføres ved at benytte i gennemsnit 25 komponenter, som er et relativt lille antal komponenter.
Den sidst benyttede metode var den såkaldte NMF2D. Her var resultaterne en smule værre i forhold til den anden metode. Resultaterne vurderes dog at kunne være forbedrede, hvis der var brugt et højere antal komponenter. Disse ville dog være svære at identificere ved at hjælp af lyttetest. Et meget stort antal komponenter vil dog gøre processen meget tidskrævende og ville derfor også have besværliggjort træningen af klassifikationssystemet, samtidig vil et stort antal komponenter være meget hukommelseskrævende. Derfor blev der valgt at gå videre med den metode, baseret på den perceptionelt vægtede NMF.

For at kunne benytte separation af kildesignaler i et fuldautomatisk system, er det nødvendigt at finde frem til en metode, der automatisk er i stand til at identificere de separerede komponenter, der hører til trommesignalet. Til dette er der afprøvede to metoder hvor den ene benytter sig af lineær klassifier, imens den anden er mere enkel, idet den søger efter peaks i et signal. Metoderne blev testede på ægte

musiksignaler hvor kvaliteten blev vurderet i forhold til hinanden og i forhold til de manuelt identificerede kildekomponenter.
Det viste sig, at den mere enkle peak-søgende metode, var bedre til at identificere kildekomponenterne. Derfor blev denne metode valgt til videre brug i klassifikationssystemet. Årsagen til at metoden med klassifieren ikke virkede så godt kan have været mangel på egnede trænginsdata, samtidig som træningssignalerne havde en længde på et sekund og derfor muligvis ikke være i besiddelse af rytmisk kontekst der gerne varierer over mindst et sekund.

Efter at have fastlagt hvilken procedure, hvilken procedure der skal benyttes til kildeseparation af trommerne, var næste skridt at undersøge hvilke features der kunne udtrækkes fra trommesignalet til videre brug, som input til en klassifier.

Der blev der undersøgt korttidslige features fra MPEG framework og Mel Cepstral koefficienter. Derudover blev der benyttet multivariabel autoregressiv regression til at integrere de korttidslige features op på en længere tidsskala. Derudover er der benyttet en kernel model.

Der er blevet benyttede to velkendte klassifiers i form af den generelle lineære klassifier og den Gauissiske miksningsmodel.
Der blev benyttet to forskellige musikdatabaser til at udføre klassifikationstests med. Den ene bestod af 5 genrer hvor, genrerne var relativt snævert opdelte. Den anden database bestod af 8 musikgenrer med 100 musiknumre til hver genre med den del overlap.

Det viste at den opstilling der gav bedst resultater, var at udtrække MFCC fra trommesignalet, for derefter at benytte DAR-features. Siden hen benyttes en generaliserede lineære klassifier benyttet til at klassificere de forskellig features i forhold til tilhørende musikgenre.

For den mindre musikdatabase blev der opnået en klassifiaktions korrekthed på 54% hvor der blev klassificeret efter trommesignalet.
For den store musikdatabase, opnås en overordnet klassifikations korrekthed på 42% hvor der benyttes 8 musikgenrer. Til sammenligning er der opnået en korrekthed på 50% i [1] på den samme database, hvor features udtrækkes direkte fra musiksignalet.

8.2 Konklusion

Hovedelementet i dette projekt har været det udfordrende område der vedrører blind separation af kilde signaler fra enkeltkanals polyfoniske musiksignaler og hvordan dette kan indgå som del af et musikgenre klassifikationssystem.

Den største udfordring har været at finde en passende metode til separation af kildesignalerne og finde ud af hvordan separationsproceduren kan forløbe fuldautomatisk. Samtidig har det været af stor vigtighed at den benyttede separationsprocedure er meget robust, hvor den ideelt kan håndtere alle former for musiksignaler med lige kvalitet.

Til separation af kildesignaler blev der foretrukket den metode baseret på en perceptionelt vægtet NMF. Dette er en metode der i flere henseender synes at udkonkurrere flere af værende algoritmer på området. Samtidig er der fundet en relativt effektiv og enkel metode til at identificere kildekomponenterne med.

Til klassifikation af musikgenre fra en musikdatabase hvor genrerne er snævert opdelte, fås der en klassifikationskorrekthed på 54% hvor der benyttes tidslig feature integreatoin af korttidslige MFFC udtrukket fra trommesignalet.
For en større musikdatabase hvor musikgenrerne er mere bredt opdelte, opnås en overordnet klassifikationskorrekthed på 42%.

.

Dette indikerer at trommesignalet som rytmisk feature indeholder vigtig information for et musikgenre klassifikationssystem. Der observeres også ud fra hvilke genrer, de forkert klassificerede eksempler, klassificeres at tilhøre, at systemet er ret godt til at opfange rytmik, da det typisk er genrerne med den samme rytmiske stil der er svære at skelne ad.

Der påpeges også at den rytmisk feature benyttet i dette projekt performer den hel der bedre en f.eks. Beat Histogram.

I den software applikation udviklet til dette projekt, er det muligt at lave en database søgning ud fra et valgt musiknummer. Applikationen gør det muligt at søge i forhold til hele musiksignalet eller kun i forhold til trommerne. Her er det muligt at observere hvordan dette fortolkes af systemet i forhold til hvilke musiknumre der vælges.

8.3 Fremtidigt arbejde

I dette projekt er der blevet undersøgte en række aspekter der er interessante i forhold til musikgenre klassifikation. Vedrørende separation af kildesignaler fra mono-kanals polyfoniske musiksignaler, så findes der ikke på nuværende tidspunkt nogen færdig robust løsning. Derfor vil forbedringer på dette område uden tvivl også kunne forbedre mulighederne for effektive systemer til søgning efter musik, hvor brugeren har mulighed for at benytte mere specifikke søgningskriterier.
Derudover er det ønskeligt at finde metoder, der er bedre til automatisk at genkende kildekomponenterne fra en separation.
Det ville også være interessant at undersøge andre features som kunne benyttes på trommesignalerne.

Bilag 1

Klassifikationsresultater

Tabel 3. AR-features – MPEG-7 Spektral Envelope

	Heayv Metal	**Pop**	**Country**	**Blues**	**Dance**
Heayv Metal	36,4	36,4	0,0	18,2	9,09
Pop	45,5	45,5	0,0	0,0	9,09
Country	9,1	0,0	72,7	9,1	9,09
Blues	9,1	0,0	54,6	36,4	0,0
Dance	27,3	9,1	9,1	0,0	54,55

Tabel 4. DAR-features – MPEG-7 Spektral Flatness

	Heayv Metal	**Pop**	**Country**	**Blues**	**Dance**
Heayv Metal	18,2	18,2	18,2	18,2	27,3
Pop	18,2	36,4	9,1	18,2	18,2
Country	9,1	18,2	36,4	27,3	9,1
Blues	18,2	0,0	45,5	36,4	0,0
Dance	27,3	18,2	9,1	0,0	45,5

Gaussisk miksning model

Tabel 5. DAR-features – 5 MFCC

	Heayv Metal	Pop	Country	Blues	Dance
Heayv Metal	36.4	36.4	0,0	27.3	0,0
Pop	27.3	36.4	27.3	0,0	9,1
Country	9.1	18.2	45.5	9.1	18,2
Blues	18.2	18.2	9.1	45.5	9,1
Dance	0,0	18,2	9,1	54,6	18,2

Tabel 6. DAR-features – MPEG-7 Spektral Envelope

	Heayv Metal	Pop	Country	Blues	Dance
Heayv Metal	54,6	36,4	0,0	0,0	9,1
Pop	45,5	36.4	18,2	0,0	0,0
Country	0,0	0,0	72,7	0,0	27,3
Blues	0,0	36,4	27,3	0,0	36,4
Dance	18,3	9,1	9,1	63,6	0

Tabel 7. DAR-features – MPEG-7 Spektral Flatness

	Heayv Metal	Pop	Country	Blues	Dance
Heayv Metal	0,0	0,0	100	0,0	0,0
Pop	0,0	18,2	81,8	0,0	0,0
Country	0,0	45,5	54,6	0,0	0,0
Blues	0,0	36,4	63,7	0,0	0,0
Dance	0,0	0,0	100	0,0	0,0

Tabel 8. DAR-features – MPEG-7 Spektral Envelope

	Heayv Metal	Pop	Country	Blues	Dance
Heayv Metal	54,6	36,4	0,0	0,0	9,1
Pop	45,5	36.4	18,2	0,0	0,0
Country	0,0	0,0	72,7	0,0	27,3
Blues	0,0	36,4	27,3	0,0	36,4
Dance	18,3	9,1	9,1	63,6	0

Tabel 9. MPEG-7 Spektral Centroid

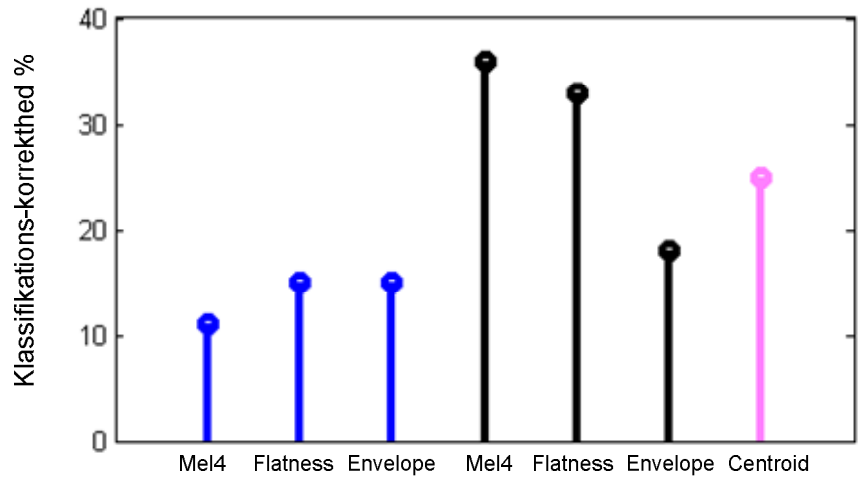
	Heayv Metal	Pop	Country	Blues	Dance
Heayv Metal	18,2	18,2	27,3	27,3	9,1
Pop	0,0	54,6	18,2	0,0	27,3
Country	27,3	9,1	36,4	18,2	9,1
Blues	18,2	18,2	0,0	18,2	45,5
Dance	45,5	9,1	27,3	18,2	0,0

Tabel 10. MPEG-7 Spektral Spread

	Heayv Metal	Pop	Country	Blues	Dance
Heayv Metal	0,0	27,3	45,5	0,0	27,3
Pop	0,0	9,1	54,6	27,3	9,1
Country	27,3	9,1	18,2	18,2	27,3
Blues	0,0	27,3	63,6	9,1	0,0
Dance	0,0	18,2	45,4	18,2	18,2

Tabel 11. MPEG-7 Spektral Spread

	Cou	Alt	Eas	Rock	Reg	R&H	Rb&S
Heayv Metal	0,0	27,3	45,5	0,0	27,3	27,3	27,3
Pop	0,0	9,1	54,6	27,3	9,1	9,1	9,1
Country	27,3	9,1	18,2	18,2	27,3	27,3	27,3
Blues	0,0	27,3	63,6	9,1	0,0	0,0	0,0
Dance	0,0	18,2	45,4	18,2	18,2	18,2	18,2



Figur 3 Viser performance for de forskellige features ved brug af den Gaussiske klassifier. De blå komponenter viser Mel4_KAR, de sorte viser Mel4_DAR og den grønne viser kun brug af Splectral Centroid.

Referencer

[1] Sherrick, John D. (2001) Concepts in Systems and Signals. Prentice Hall, Upper Saddle River, New Jersey.

[2] Van Loan, Charles F. Second Edition (1999). Introduction to Scientific Computing. Prentice Hall, Upper Saddle River, New Jersey.

[3] Råde, Lennart og Westergren, Bertil. Fourth Edition. Mathematics Handbook for Science and Engineering BETA. Studentlitteratur, Lund, Sweden.

[4] Larson, Roland E. og Edwards, Bruce H. Third Edition, (1996) Elementary Linear Algebra, D. C. Heath and Company. USA

[5] Proakis, John G. og Manolakis, Dimitris G. Third Edition (1996) Digital Signal Processing, Prentice Hall, Upper Saddle River, New Jersey.

[6] Walther, Stephen. Second edition (2004) ASP.NET. Sams, Indianapolis, Indiana.

[7] Chassaing, Rulph. (2005) Digital Signal Processing and Applications with the C6713 and C6416 DSK. Wiley- Interscience, USA.

[8] W. Marshall Leach, Jr. Third Edition (2003) Introduction to Electroacoustics and Audio Amplifier Design. Kendall/ Hunt Publishing Company, Iowa, USA.

[9] Bishop, Christopher M. (2005) Neural Networks for Pattern Recognition. Oxford University, Great Britan.

[10] Larman, Craig. Second Edition (2002) Applying UML and Patterns. Prentice Hall, Upper Saddle River, New Jersey.

[11] Ahrendt, Peter. (2006) Music Genre Classification Systems.

[12] Meng, Anders. (2006) Temporal Feature Integration for Music Organisation.

[13] Lee, Daniel D. og Seung, H. Sebastian. Algorithms for Non-negative Matrix Factorization.

[14] Wang, Beiming og Plumbley, Mark D. Musical Audio Stream Separation by Non-Negative Matrix Factorization. Queen Mary University of London

[15] Smaragdis, Paris og Brown, Judith C. (2003) Non-Negative Matrix Factorization for Polyphonic Music Transcription, USA

[16] Smaragdis, Paris. Non-Negative Matrix Factor Deconvolution.

[17] Mørup, Morten og Schmidt, Mikkel N. (2006) Sparse Non-Negative Matrix Factor 2- D Deconvolution. DTU, Kgs. Lyngby

[18] Mørup, Morten og Schmidt, Mikkel N. Nonnegative Matrix Factor 2- D Deconvolution for Blind Single Channel Source Separation. DTU, Kgs. Lyngby.

[19] Zhang, Sheng- Wang, Weihong- Ford, James og Makedon Fillia. Learning from Incomplete Ratings Using Non-Negative Factirization. Hanover, USA

[20] D. Guillamenr et al. Pattern Recognition letters 24 (2003)

[21] Pedersen, Michael Syskind,- Lehn- Schiøler, Tue og Larsen, Jan. Blues from Music: Blind Underdetermined Extraction of Sources from Music. Intelligent Signal Processing IMM, Technical University of Denmark.

[22] Jang, Gil-Jin og Lee, Te-Won. A Maximum Likelihood Approach to Single-channel Source Separation. Journal of Machine Learning Research 4 (2003).

[23] Hyvärinen, Aapo og Oja, Erkki. Independent Component Analysis: Algorithms and Applications. Neural Networks Research Centre Helsinki University of Technology.

[24] Hyvärinen, Aapo. Survey on Independent Component Analysis. Helsinki University of Technology.

[25] Virtanen, Tuomas. Separation of Sound Sources by Convolutive Sparse Coding. Workshop on Statistical and Perceptual Audio Processing SAPA-2004, 3 Oct 2004, Jeju, Korea.

[26] Paul Kienzle's Octave links. http://users.powernet.co.uk/kienzle/octave/

[27] VOICEBOX: Speech Processing Toolbox for MATLAB. http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[28] Complex.java http://www.cs.princeton.edu/introcs/32class/Complex.java.html

[29] Netlab Downloads http://www.ncrg.aston.ac.uk/netlab/down.php

[30] PLP and RASTA http://labrosa.ee.columbia.edu/matlab/rastamat/

[31] Jazz Music at CD universe http://www.cduniverse.com/browsecat.asp?cat=59&BAB=U

[32] Øre- Wikipedia, den frie encyklopædi http://da.wikipedia.org/wiki/%C3%98re

[33] Free Rock Drum loops with drum fills by Jim Dooley http://www.dooleydrums.com/drum_loops_000052.php

[34] MIS http://theremin.music.uiowa.edu/MIS.html

[35] Phatso's Place: Free Funky Drum Loops & Samples
http://www.phatdrumloops.com/old_site/

[36] The Instrument Encyclopedia Database Search
http://www.si.umich.edu/chico/instrument/search.phtml?field=continent&search=Europe&andor=and&type=gbrowse

[37] Jehan, Tristan. Creating Music by Listening
http://web.media.mit.edu/~tristan/phd/dissertation/index.html

[38] http://www.cduniverse.com

[39] ARfit: A Matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models
http://www.gps.caltech.edu/~tapio/arfit/

[40] Virtanen, Tuomas O. Monoural Sound Source Separation by Perceptually Weighted Non-Negative Matrix Factorization
Technical report, Tampere University of Technoligy, Institute of Signal Processing, 2007

[41] Arenas Garcia, Jerónimo. Brandt Petersen, Kaare. Hansen, Lars Kai.
Sparse Kernel Orthonormalized PLS for feature extraction in large data sets.

[42] aim-mat
http://www. Mrc-cbu.cam.ac.uk/cnbh/aimmanual

[43] Virtanen, Tuomas O.Separation of Drums From Poyphnic Music Using Non-Negative Matiric Factorization and Support vector machine