

Udtræk af melodien fra et musikalsk nummer

Et datalogisk projektarbejde om identifikationen af melodien i polyfon musik

Morten Wendelboe

Datalogisk Institut
Københavns Universitet
17. juni 2008

Vejleder: Kim Steenstrup Pedersen

Extracting the melody from a musical number

Abstract

This paper treats the extraction of the melody from a polyphone piece of music using an automated method. Only one sound channel is processed. The problem of separating the melody from the rest of the sound signal is still unresolved, but a lot of attempts to get closer to a solution have been made in the last few years.

The proposed method which is based on frequency analysis is a simplified SHS with some minor modifications added. The method conjecture the assumed melody's fundamental frequency, even though there might not be a melody.

At MIREX 2006 one of the given tasks actually was to identify the melody in a polyphone piece of music. This year there was 5 participants and it is the evaluation method at MIREX 2006 which is used to evaluate the proposed method. This means that the results for the proposed method is comparable with the results for the 5 methods that participated at MIREX 2006.

The evaluation shows that the proposed method is given almost as high a score as the two best methods at MIREX 2006.

Udtræk af melodien fra et musikalsk nummer

Resumé

Denne opgave drejer sig om dét at finde melodien i et polyfont stykke musik på automatiseret vis. Der arbejdes kun med én lydkanal. Problemet med at separere melodien ud af et lydsignal er stadig uløst, men der har i de seneste par år været gjort mange forsøg på at komme nærmere en løsning af problemet.

Den foreslåede metode, som baserer sig på frekvensanalyse, er en simplificeret udgave af SHS med små modifikationer. Der gives altid et bud på fundamentalfrekvensen for den formodede melodi, også selv om der måske slet ikke er nogen melodi.

Ved MIREX 2006 gik en af de givne opgaver netop ud på at identificere melodien i et polyfont stykke musik. Der var 5 deltagere dette år, og det er evalueringsmetoden, som blev anvendt ved MIREX 2006, der også anvendes til at evaluere den metode, som foreslås i denne opgave. Det betyder, at resultatet for den foreslåede metode kan sammenlignes med resultaterne for de 5 metoder, der deltog ved MIREX 2006. Evalueringen viser, at den foreslåede metode har fået næsten lige så høj score som de 2 bedste metoder ved MIREX 2006.

Forord

Til denne opgave hører en problemformulering/projektbeskrivelse, som definerer, hvad der er målet med opgaven, samt at det er en opgave på 15 ects-point.

Desuden medfølger en cd, som indeholder dette dokument i pdf-format, hvilket kan være nyttigt, da det er muligt at følge henvisninger og forstørre billeder. Cd'en inkluderer også kildekoden til 3 versioner af en foreslået metode, samt de testprogrammer og testfiler der har været anvendt under evalueringen af den foreslåede metode. Den udviklede kode er skrevet i Matlab og har været kørt i Matlab v. 7.4. Det burde også være muligt at bruge Octave, men i skrivende stund går det meget langsommere i Octave.

Skulle cd'en på en eller anden måde være bortkommet, eller hvis der er problemer med at læse denne, kan opgaven, kildekoden og test-filerne findes via <http://www.morwen.dk/university/melody-project/index.html>.

Indhold

1	Indledning	1
1.1	Søgning efter musik	1
1.1.1	Søgning ved hjælp af lyd	2
1.1.2	Query by Melody	2
1.2	Egenskaberne ved en melodi	3
1.3	Repræsentation af melodien	3
1.4	MIREX	3
1.5	Overblik over følgende kapitler	4
1.6	Målgruppen	4
2	Værktøjskasse	6
2.1	Tonedannelse	6
2.1.1	Harmoniske lyde	7
2.2	Begreber fra den musikalske analyse	8
2.3	Signalbehandling	10
2.3.1	Fouriertransformation	10
2.3.1.1	STFT	11
2.3.1.2	Nyquistfrekvensen	11
2.3.1.3	Vinduer	12
2.3.1.4	Window overlap	13
2.3.1.5	Zero padding	14
2.3.2	Spektrogram	15
2.3.3	Foldning	17
2.3.3.1	Foldningssætningen	17
2.4	Det menneskelige øre	18
2.4.1	Psykoakustik	19
2.4.1.1	Decibel	20
2.4.1.2	Phon	20
2.4.1.3	Lydskalaer	20
2.4.1.4	Manglende fundamentalfrekvens	21
3	Metoder for bestemmelse af melodien	22
3.1	Estimering af fundamentalfrekvensen	22
3.1.1	Metoder i tidsdomænet	23
3.1.1.1	Zero-crossing rate (ZCR)	23
3.1.1.2	Autocorrelation	23
3.1.2	Metoder i frekvensdomænet	24
3.1.2.1	Cepstrum Analysis	24
3.1.2.2	Component Frequency Ratios	24
3.1.2.3	Spektral kompression	25

3.1.2.4	Kam-metoder	25
3.1.2.5	Subharmonic Summation	26
3.1.2.6	En ændret SHS	28
3.1.3	Statistiske metoder	28
3.2	Multiopløsningsspektrum	29
3.2.1	Metoder for generering af et multiopløsningsspektrogram	29
3.2.1.1	Den simple løsning	30
3.2.1.2	Constant Q spectral transform	30
3.3	Støjundertrykkelse	31
3.4	Sporing	31
4	Den foreslåede metode	33
4.1	Version 0a	33
4.2	Version 0b	33
4.2.1	Det anvendte multiopløsningsspektrogram	34
4.3	Version 1	34
4.3.1	Inspiration fra den ændrede SHS	35
5	Generelle Resultater	37
5.1	Testmetode	37
5.1.1	MIREX-2006	37
5.1.2	ISMIR-2004	38
5.1.3	Sammenligning af resultater	38
5.1.4	Raw pitch og raw chroma	39
5.2	Afprøvning med metoden fra MIREX-2006	39
5.3	Afprøvning med metoden fra ISMIR-2004	40
6	Resultatanalyse og forslag til forbedringer	42
6.1	Nærmere analyse af 2 melodier samt forslag til forbedringer . . .	44
6.2	Flere idéer til forbedringer	46
7	Konklusion	51

Kapitel 1

Indledning

Musik, musernes kunst, har i årtusinder været af stor betydning for mennesker. Mange mennesker lytter aktivt til musik hver dag; nogle hører musik i baggrunden mens de arbejder, nogle spiller selv musik. Nogle kan lide jazz, andre klassisk og nogle er mere til hård rock. Men ligegyldig hvad vi er mest til, så kan vi ofte huske melodien i et nummer, som vi har hørt flere gange, eller har hørt i en speciel situation. Mange oplever at de kommer i tanke om en bestemt hændelse eller følelse, når de hører et nummer, som blev spillet, da de oplevede netop dén hændelse eller følelse. Men der sker ofte også det modsatte; nemlig at vi på en eller anden måde bliver inspirerede til at ville høre et bestemt nummer. Selv om man har adgang til netop dét nummer, som man ønsker at høre, er det dog ikke nødvendigvis let at finde nummeret.

1.1 Søgning efter musik

For en række år siden, inden computere blev almindeligt udbredt, var det ganske normalt at man indekserede sin musiksamling alfabetisk efter artist- eller kunstner-navn. Nogen opdelte endda musikken i forskellige genrer. En søgning efter et bestemt nummer (eller musikalsk stykke om man vil) kunne være en langsommelig proces. Dette ville især være tilfældet, hvis det ønskede nummer ikke befandt sig på sin rette plads i den indekserede samling.

Efter indførslen af computere er det blevet meget nemmere at finde frem til et bestemt nummer, såfremt de relevante oplysninger er blevet noteret i en database. Men selv om det er nemmere og hurtigere at finde et nummer via et opslag i en database på en computer, så har måden hvorpå man søger efter musik i bund og grund ikke ændret sig. Man ønsker at finde noget lyd, men søger ved hjælp af nogle tekst-strenger; og disse tekst-strenger har ikke nødvendigvis noget direkte med lyden at gøre. Man kan måske godt huske at komponisten hedder Mozart; men det kan være svært at gætte, hvad nummeret hedder, blot ved at tænke på hvordan violinerne lyder. (I mange popsange kan det dog nogle gange være lettere at gætte titlen på nummeret, da titlen ofte vil blive sunget igen og igen.)

Et problem, ved den måde vi søger musik på, er, at vi først skal komme i tanke om nogle tekst-strenger, som relaterer sig til et nummer, og dernæst kan vi foretage søgningen ved hjælp af disse tekst-strenger. Problemet er, at vi ofte kun kan huske den lyd, vi ønsker at høre, og ikke kan huske nogen form for metadata, som knytter sig til det pågældende nummer. I nogle situationer ville det altså være ønskeligt at kunne søge efter musik ved at søge med en lyd. Det

kunne f.eks. være et brudstykke af det pågældende nummer, eller det kunne være en abstrakt gengivelse af en del af nummeret.

1.1.1 Søgning ved hjælp af lyd

I de seneste par år er blandt andet en række mobiltelefonselskaber begyndt at tilbyde kunderne en service, hvor man ved hjælp af mobiltelefonen kan optage et brudstykke af et nummer, som f.eks. spilles i radioen, og dernæst kan man få oplyst hvilket nummer, der er tale om, eller få mulighed for at downloade nummeret i sin fulde længde. For at det virker, kræves der altså et brudstykke af et nummer, som findes i mobiltelefonselskabets database, og brudstykket kan eventuelt være tilføjet lidt støj.

Den omtalte funktionalitet kan altså ikke anvendes, hvis man ikke har adgang til den originale kilde (evt. tilføjet lidt støj). Hvis et menneske nynner et brudstykke af melodien i et nummer, vil mange mennesker, som har hørt det pågældende nummer, kunne genkende brudstykket. Så på en eller anden måde har vi mennesker en evne til at opfatte musik på en abstrakt måde. Vi opfatter en nynnet gengivelse som noget, der minder om det nummer, der nynnnes. I de fleste tilfælde er det melodien i nummeret, som vi fæstner os ved, men det kan også være bestemte rytmiske figurer.

Hvis man kan lave en algoritme, som kan finde melodien i et nummer, vil det måske være muligt at søge i en musiksamling blot ved at nynne melodien. Der kan også være andre anvendelsesmuligheder. Man kunne f.eks. tænke sig at flere artister har indspillet det samme nummer. Ved hjælp af den omtalte algoritme kan man måske finde frem til flere udgaver af det samme nummer, eller måske finde numre, som indeholder melodiske brudstykker fra andre numre. Det ville måske endda være muligt at lave en repræsentation af et nummer ved hjælp af noder. Så ville en komponist f.eks. kunne spille en komposition på et instrument og automatisk få genereret et nodeark.

1.1.2 Query by Melody

Den primære motivation for at finde melodien i et nummer er, at søgning efter et bestemt nummer muligvis kan gøres lettere. Søgning ved hjælp af en melodistump går under betegnelsen "Query by Melody" eller QbM, og hvis melodistumpen bliver nynet, kan det eventuelt kaldes "Query by Humming" eller QbH.

Der har igennem tiderne været foreslået QbM-systemer, hvor der anvendes 2 algoritmer: først en algoritme, som kan finde melodien i et stykke, og dernæst en algoritme, som ved hjælp af mønstergenkendelse kan undersøge om 2 melodier minder om hinanden.

Der er dog en anden udbredt tilgang til problemet. I stedet for at finde melodien eksplicit og dernæst sammenligne med data fra de numre, som man søger iblandt, kan man lade flere personer nynne melodierne i en samling. Når man søger ved hjælp af en nynnet melodi, er det altså ikke udtrukne melodier der foretages mønstergenkendelse på, men derimod de nynnede repræsentationer. På denne måde undgår man at skulle forholde sig til, hvad der er melodien i et stykke, og man behøver ikke at beskæftige sig med, hvordan vi mennesker opfatter melodier. Problemet med disse systemer er dog, at de netop kræver at nogen har nynet til hvert enkelt nummer i en samling.

Især i systemer, hvor der anvendes kunstig intelligens, berøres spørgsmålet om, hvordan vi mennesker egentlig opfatter melodier, stort set ikke, for i disse systemer er det netop op til systemet selv at finde ud af, hvad det er, der får flere nynnede melodier til at minde om hinanden.

1.2 Egenskaberne ved en melodi

Hvad forstår vi egentlig ved en melodi, og hvordan repræsenterer man en melodi? De fleste vil nok være enige i, at en melodi er en følge af toner (underforstået frekvenser), som har nogle tidsmæssige længder. Tonerne følger ikke nødvendigvis direkte efter hinanden, men kan være adskilt af pauser. Melodiens toner vil for os mennesker ofte træde tydeligt frem i forhold til de øvrige toner, som et polyfont nummer består af. Desuden vil melodien oftest blive spillet af ét instrument (eller sunget af en stemme) gennem en længere passage. Hvis man præsenterer visse numre for 2 forskellige personer, kan man komme ud for, at de har hvert deres bud på, hvad der er melodien i de pågældende numre. Dette kan være tilfældet, hvis man har et nummer, hvor 2 personer synger forskellige toner. Opfattelsen af melodien kan altså være subjektivt betinget.

Det kan være svært at sige præcist, hvad det er, der gør, at vi mennesker opfatter nogle lyde som mere fremtrædende end andre. Dynamikken, en tones overtoner, tonehøjden, tonens ændring gennem tiden (f.eks. vibrato) og formentlig en masse andre parametre har betydning - men hvor meget betyder de enkelte dele? I denne opgave vil vi stort set kun fokusere på dynamikken, da en dynamisk set kraftig melodi, vil være let at høre, mens en svag udgave af den samme melodi kan være næsten umulig at høre i polyfon musik.

1.3 Repræsentation af melodien

Som det blev antydnet i afsnit 1.1.2, findes der systemer, hvor man søger ved hjælp af et kompliceret lydbillede. Det komplicerede lydbillede stammer fra et forsøg på at nynne melodien i et nummer, og dette komplicerede lydbillede bliver ved hjælp af kunstig intelligens sammenlignet med forskellige bud for nynnede melodier. I de fleste tilfælde vil lydbilledet blive anskuet i et frekvensspektrum. Dette er ét bud på en repræsentation af melodien.

En meget mere kompakt repræsentation består blot af den fundamentalfrekvens, som melodien har til en given tid. Ønsker man en endnu mere kompakt repræsentation, kan man beskrive melodien ved hjælp af noder, og man kan endda ignorere forskelle i oktaver. Hvordan melodien bedst repræsenteres afhænger af det enkelte formål, men hvis man repræsenterer melodien som en enkelt frekvens som funktion af tiden, er det let at gøre denne repræsentation mere kompakt. Derfor vil melodien i denne opgave blive repræsenteret som en række frekvenser.

1.4 MIREX

MIREX¹ kan betragtes som et forum med fokus på lyd og forsøg på at udtrække informationer fra lydsekvenser. Hvert år stilles en række opgaver, som man kan

¹http://www.music-ir.org/mirex/2008/index.php/Main_Page

tilmelde sig. I år 2006 gik en af opgaverne ud på at identificere melodien i et polyfont musikstykke. Deltagerne fik mulighed for at afprøve deres metoder ved hjælp af nogle lydfiler samt korrekte angivelser af melodiernes frekvensgange i lydfilerne. Det blev desuden specificeret hvordan et bud på en frekvensgang skulle evalueres i forhold til den korrekte frekvensgang. Deltagernes endelige forslag til metoder blev afprøvet på en hemmelig musiksamling, og metoderne fik tildelt en score ved hjælp af den specificerede evalueringsmetode.

I denne opgave vil den foreslåede metode blive evalueret ved hjælp af den evalueringsmetode, som blev anvendt ved MIREX 2006.

1.5 Overblik over følgende kapitler

I det følgende gives en kort oversigt over de enkelte kapitlers indhold:

- Først starter vi med en værktøjskasse i kapitel 2. I dette kapitel introduceres begreber og metoder, som er relevante eller nyttige, når det handler om analyse af signaler og menneskets opfattelse af lyde.
- I kapitel 3 præsenteres nogle af de metoder, som gennem tiderne har været mest interessante eller betydningsfulde, eller som er værd at nævne for simpelt design. Desuden introduceres nogle specialdesignede værktøjer, så som et multiopløsningsspektrogram.
- Der præsenteres 3 versioner af den foreslåede metode i kapitel 4. De første 2 versioner tager udgangspunkt i den hypotese, som blev stillet i problemformuleringen til denne opgave, mens det endelige forslag er bygget op omkring SHS-metoden, som bliver introduceret i kapitel 3.
- Den foreslåede metode bliver evalueret med den evalueringsmetode, som blev anvendt ved MIREX 2006. Det overordnede resultat af denne evaluering bliver præsenteret i kapitel 5, og her bliver metodens score sammenlignet med den score, som deltagerne ved MIREX 2006 opnåede.
- En nærmere analyse af resultaterne er at finde i kapitel 6, hvor resultater for hver enkelt melodi også kan findes. Her gives der desuden forslag til eventuelle ændringer af metoden, så den forhåbentlig kan give bedre resultater.
- Endelig konkluderes der i kapitel 7, hvor det bliver vurderet at den foreslåede metode har været succesfuld, specielt når man tager med i betragtning, at den er meget simpel.

1.6 Målgruppen

Der er sigtet mod, at en kandidatstuderende ved en datalogiuddannelse eller lignende kan læse denne opgave med udbytte. Dette er også den primære årsag til at kapitel 2 er inkluderet, fordi det f.eks. ikke er alle kandidatstuderende, som tydeligt kan huske hvad fouriertransformation går ud på. Det nævnte kapitel kan også være nyttigt at læse for en person, for hvem begreber, som omhandler beskrivelsen af musik, er lidt fjerne. Hvis man føler sig sikker på emner omkring signalbehandling og kender til musikalske begreber, samt kender til den menneskelige opfattelse af lyde, kan kapitel 2 blot overspringes.

Opgaven er skrevet på dansk², men hvis der er tekster på selve figurerne, er disse på engelsk. Dette skyldes til dels, at nogle af figurerne er taget fra engelsksprogede dokumenter, dels at figurerne ikke behøver at blive lavet om i tilfælde af, at det kunne komme på tale at oversætte dele af opgaven til engelsk. Der optræder også en del engelske termer i løbet af opgaven. Det drejer sig enten om navne på metoder, eller om begreber, hvor det har virket bedre at bruge de engelske termer, eller hvor der ikke har været nogen kendt dansk term.

²Dette burde vist efterhånden være klart for læseren.

Kapitel 2

Værktøjskasse

Dette kapitel har til formål at klæde læseren på til de øvrige kapitler, og de fleste oplysninger i dette kapitel kan findes i diverse opslagsværker og grundlæggende lærebøger. Her berører vi kort nogle centrale emner inden for signalbehandling, samt nævner nogle begreber inden for musikanalyse. Derudover berøres den menneskelige hørelse kort. Vi lægger ud med at beskrive hvad en tone er, samt hvordan man set ud fra et signalbehandlingssynspunkt kan anskue den.

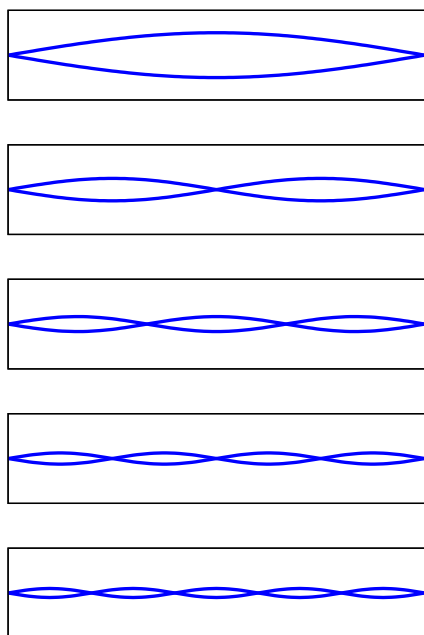
2.1 Tonedannelse

Når vi hører en tone, som bliver sunget eller spillet på et instrument, er det en kombination af mange frekvenser vi hører. Den dybeste frekvens, som dannes på instrumentet, kaldes *fundamentalfrekvensen* (forkortet: f_0 eller F_0), og er den frekvens, som vi opfatter som den vigtigste. De øvrige frekvenser kaldes for *overtoner*.

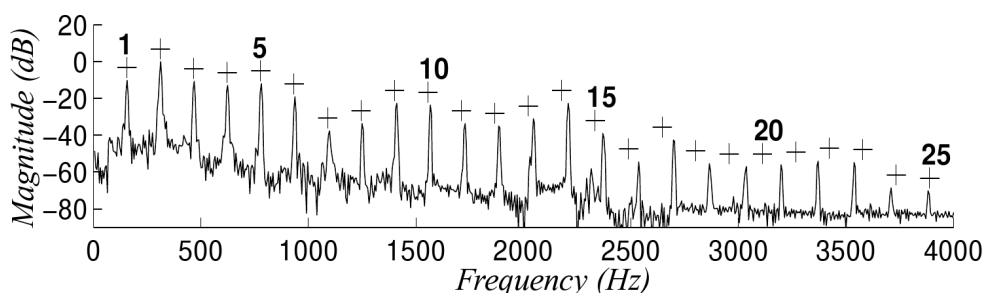
På et strenginstrument vil de 2 ender af en streng være fastlåst, og strengen kan svinge i mellem disse punkter. Ved den største svingning, som kan lade sig gøre, er strengen netop kun fastlåst i de 2 ender. Det er denne svingning, som giver anledning til fundamentalfrekvensen. Derudover er det muligt for strengen at svinge med et helt antal fastlåste punkter fordelt med lige stor afstand. Disse svingninger giver anledning til hver enkelt overtone. Bemærk at begrebet *tone* dækker over fundamentalfrekvensen og eventuelle overtoner, mens en *overtone* svarer til en bestemt frekvens.

Frekvensen afhænger af bølgehastigheden på strengen og bølgelængden. ($F = V/L$, hvor F er frekvensen, V er bølgehastigheden og L er bølgelængden.) Man kan øge en strengs frekvens ved at gøre bølgelængden kortere, eller ved at spænde strengen, så bølgehastigheden stiger. Den første overtone vil have en frekvens, som er dobbelt så høj som fundamentalfrekvensen, da bølgelængden er den halve. På samme måde kan vi konkludere, at den anden overtone har en frekvens, som er 3 gange større end fundamentalfrekvensen. Se figur 2.1.

Det er instrumentets krop og strengtypen, som har betydning for, hvor kraftige de enkelte frekvenser i en tone er. Nogle frekvenser kan svinge med i kroppen, mens kroppen for andre frekvensers vedkommende har en destruktiv effekt.



Figur 2.1. På den øverste figur svinger strengen med den maksimale bølgelængde - den tilsvarende frekvens er fundamentalfrekvensen. På de øvrige figurer er afbildet de svingninger som svarer til de første 4 overtoner.



Figur 2.2. Her vises frekvensspektrummet for en klaverstreng, hvor fundamentalfrekvensen er på 156 Hz. De ideelle harmoniske overtoner er markeret ved "+". De faktiske overtoner får gradvist større interval imellem sig, og den faktiske 24. overtone ligger ved samme frekvens som den ideelle 25. harmoniske overtone. (Figuren er taget fra [18, afsnit 3.1].)

2.1.1 Harmoniske lyde

I foregående afsnit blev det forklaret, hvordan der opstår en række frekvenser, når en streng (eller for den sags skyld en luftsøjle i et blæserinstrument) svinger. Når de opståede frekvenser netop ligger med et fast interval imellem sig, kalder vi det for en *harmonisk* lyd, mens det omvendte tilfælde kaldes for *inharmonisk* lyd. For at en lyd kan kaldes harmonisk, er det dog ikke et strengt krav, at frekvenserne ligger med et fuldstændig fast interval imellem sig. Dette er tilfældet for mange strenginstrumenter, som slås an. På figur er vist frekvensspektrummet for en streng, som er slået an på et klaver. Det inharmoniske aspekt opstår på grund af strengens stivhed, som i sidste ende forårsager, at de forskellige frekvenser udbreder sig med forskellige hastigheder på strengen.[18, afsnit 3.1] Se figur 2.2.

2.2 Begreber fra den musikalske analyse

Her præsenteres et par begreber fra den musikalske analyse. Disse begreber kan være anvendelige til beskrivelsen af forskellige fænomener, som kan høres i et stykke musik. Disse fænomener skal man måske være opmærksom på, når man beskæftiger sig med analyse af et lydsignal, som indeholder musik. Dét, at begreberne findes, siger noget om den menneskelige opfattelse af musik, og et kendskab til begreberne kan måske hjælpe til en mere konstruktiv analyse af et lydsignal.

Tonehøjde og klang Foregående afsnit havde megen fokus på de frekvenser, som en tone består af. Vi opfatter toner som forskellige primært ved 2 parametre. Den ene er *tonehøjden*, og den anden er *klangen*. Tonehøjden beskriver i bund og grund fundamentalfrekvensen og kan tilføjes adverbier så som *høj* (når man taler om høje frekvenser) og *dyb* (når man taler om lave frekvenser). Klangen siger noget om, hvordan det dynamiske forhold er imellem overtonerne, og kan beskrives som mørk eller lys (man kan også bruge adverbier så som *rund*, *hård* eller endda *gennemsigtig*).

Formant Specielt operasangere uddannes til at fremhæve de overtoner, som ligger mellem 2000 og 4000 Hz. En *formant* er et område i frekvensspektrummet, som træder frem som følge af resonans. Det er især operasangernes formant omkring de 3000 Hz, som gør det muligt at høre sangeren tydeligt, selv om der er et helt orkester, som spiller i baggrunden.

Skala, trin og oktav På dansk har ordet *tone* også betydningen af et trin i en skala. Mellem 2 toner, hvor den ene tone har den dobbelte frekvens af den anden, siges der at være et interval på en *oktav*. Dette interval er inddelt i 12 dele, dvs. trin i en skala. De 12 trin er *de* 12 toner, som anvendes i stort set al vestlig musik. Hvis vi vælger en tilfældig tone, og tæller den som den 1. tone, så vil den 13. tone ligge i en oktavs afstand til den 1. tone, og den gives samme navn som den 1. tone. Når en mand og en kvinde synger den samme melodi, vil de som oftest synge de samme toner, men mandens toner ligger en oktav under kvindens toner. Vi opfatter det blot, som om manden har en dybere stemme.

Ambitus Det toneomfang, som et instrument kan spille, kaldes for instrumentets *ambitus*. Melodien i et stykke musik består oftest af toner, som ligger inden for den menneskelige stemmes ambitus - også selv om melodien spilles på et instrument. Faktisk bruges hele ambitus sjældent - toner i randområdet har ofte en dårligere karakter, og specielt i tilfældet med den menneskelige stemme, kan der være problemer med at kontrollere disse toner. Hvis en mand synger melodien, vil den dybeste tone i melodien sjældent have en frekvens under 150 Hz, mens den højeste tone i melodien sjældent har en frekvens over 1000 Hz, når en kvinde synger melodien. Specielt i klassisk musik ses det dog, at melodien godt kan have meget højere toner.

Akkord I et stykke musik vil den samme tone ofte blive spillet af forskellige instrumenter, og det er normalt, at nogle af disse spiller tonen et helt antal oktaver forskudt. Ser man bort fra oktavforskydelser, er der i vestlig musik meget ofte primært 3 forskellige toner, som klinger samtidigt, og disse har mindst 3

trin imellem sig. En sammensætning af sådanne 3 toner, kaldes for en *akkord*. Mens akkorderne skifter forholdsvis langsomt i et stykke musik, kan melodien godt skifte tone hurtigere. Så i en del af tiden, vil melodien have en tone, som befinder sig i akkorden, men det forekommer også at melodien (som oftest i kortere tid) har en tone, som ikke findes i akkorden.

Pitch Når man på engelsk omtaler en tone, vil man anvende ordet *tone*, når man henviser til et trin i skalaen, mens ordet *pitch* henviser til tones frekvens.¹ Hvis et instrument spiller en forkert tone i skalaen, siger vi, at instrumentet spiller forkert. Men hvis instrumentet blot spiller en tone, som har en lidt forkert frekvens, og denne frekvens ligger tættere på den tone, der skal spilles, end på de omliggende toner, siger vi, at der spilles falsk. På engelsk vil man sige at instrumentet eller tonen er "out of pitch".

Vibrato og glissando Når en tones frekvens ændrer sig på en måde, som minder om en sinusbølge, siges det, at tonen vibrerer, eller at der anvendes *vibrato*. Det er ikke ualmindeligt, at en operasanger anvender et vibrato, som får frekvensen til at dække over mere end ét trin i skalaen. Vibrato er altså en opadgående og nedadgående gentagende ændring af frekvensen. Hvis der kun er tale om en opadgående eller nedadgående ændring kaldes det for *glissando*. Så vibrato og glissando bryder med de trin, som findes i skalaen.

Rytme og tempo Ind til videre har vi beskæftiget os med termer, som har fokus på frekvenserne. Vi skal lige nævne 2 begreber, som fokuserer mere på det tidsmæssige aspekt. *Rytme* dækker over en gentaget ændring af en eller flere toners amplitude. I det ekstreme tilfælde er det sammensætningen af de tidsintervaller, hvori der er klingende toner, og de tidsintervaller, hvor der er stilhed. Forskellige instrumenter kan godt have forskellige rytmer samtidigt, men det er helt normalt at en rytme spilles af flere instrumenter. De instrumenter, som oftest foretrækkes til frembringelse eller understregning af en rytme, er slagøjinstrumenter. Der er tale om instrumenter, som giver en kraftig lyd, lige når de slås an; men lyden forsvinder hurtig. Derudover har de fleste slagøjinstrumenter ikke nogen decideret tone. I afsnittet om tonedannelse blev det illustreret, hvordan et strenginstrument kan danne frekvenser med et bestemt interval imellem sig. Slagøjinstrumenter udsender stort set alle frekvenser i et vist frekvensområde.

Den overordnede rytme danner grundlag for en periodisk gentagelse, hvis hastighed opfattes som et *tempo*. Tempoet i et stykke musik kan beskrives med termer så som hurtig og langsom, men det kan også beskrives som et bestemt antal slag pr. minut. Identifikationen af et tempo kan være essentiel, hvis man ønsker at beskrive en melodi ved hjælp af noder.

¹Dette er en meget forenklet forklaring af begrebet *pitch*. Der er i virkeligheden tale om den oplevede fundamentalfrekvens. Afhængig af tonens dynamiske styrke, det indbyrdes styrkeforhold mellem overtonerne, og om lydbølgen minder mest om en sinusbølge eller om en savtaksbølge, kan den oplevede fundamentalfrekvens adskille sig fra den faktiske fundamentalfrekvens. En kort forklaring kan findes i [12].

2.3 Signalbehandling

I det følgende vil vi beskæftige os med signalbehandling. En lydbølge kan beskrives som en følge af positioner, men kan også beskrives som en sammensætning af frekvenser. Det kan være meget nyttigt at anskue et lydsignal i frekvensdomænet; derfor vil stort set hele dette afsnit handle om, hvordan man transformerer et lydsignal til frekvensdomænet, samt om, hvordan man kan repræsentere lydsignalet i frekvensdomænet.

2.3.1 Fouriertransformation

Vi lader $f(x)$ være en kontinuert reel funktion af en reel variabel x , og f er integrabel; da er fouriertransformationen defineret som:

$$F\{f\}(u) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i u x} dx \quad (2.1)$$

, hvor i er den imaginære enhed, og variablen u repræsenterer frekvensen.

Hvis vi har et signal og udtager stikprøver (eller samples), som ligger med lige stor afstand mellem sig, kan vi også beregne den diskrete fouriertransformation, da denne er givet ved:

$$F\{f\}(u) = \frac{1}{N} \sum_{x=0}^{N-1} f(x)e^{-2\pi i u x/N} \quad (2.2)$$

, hvor N er antallet af samples.

Fouriertransformationen skal ses som en anden måde at repræsentere en funktion (eller et signal) på. Hvis vi har et signal i tidsdomænet (det kunne være en lydbølge som funktion af tiden), har vi ved hjælp af fouriertransformationen også mulighed for at anskue signalet i frekvensdomænet. Omvendt, hvis vi har et signal, som er repræsenteret i frekvensdomænet, kan vi anvende den inverse fouriertransformation til at anskue signalet i tidsdomænet. Den inverse fouriertransformation ligner fouriertransformationen bortset fra, at eksponenten har modsat fortegn:

$$f(x) = F^{-1}\{F\}(x) = \int_{-\infty}^{\infty} F(u)e^{2\pi i u x} du \quad (2.3)$$

Den diskrete inverse fouriertransformation kan skrives som:

$$f(x) = F^{-1}\{F\}(x) = \frac{1}{M} \sum_{u=0}^{M-1} F(u)e^{2\pi i u x/M} \quad (2.4)$$

, hvor M er antallet af frekvenser.

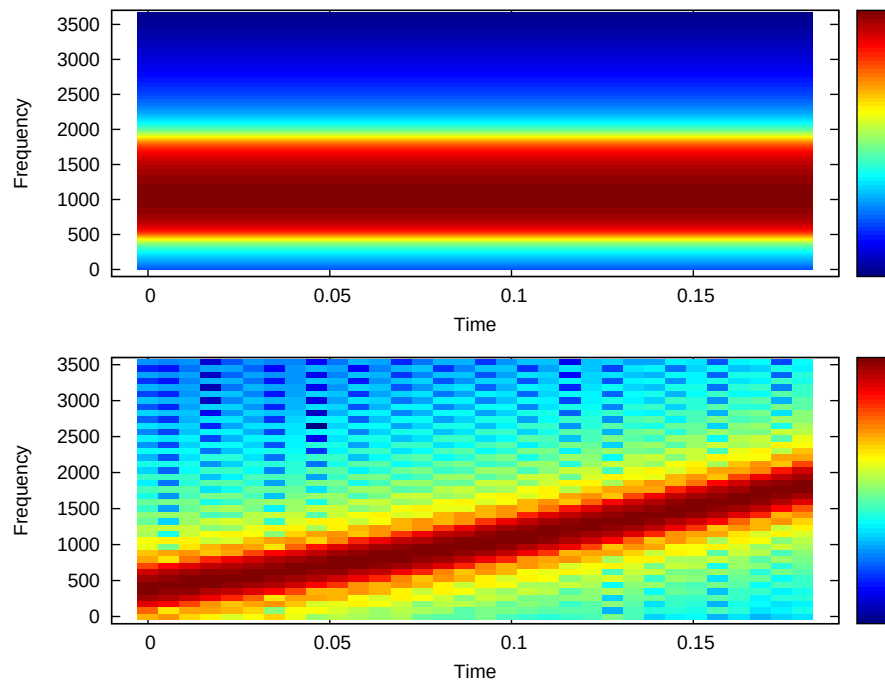
Bemærk at når man fouriertransformerer et reelt signal, får man et komplekst resultat. Hvis man blot ønsker styrken (amplituden) af hver frekvens, kan denne udregnes som den absolutte værdi af det komplekse resultat:

$$A(u) = \|F(u)\| \quad (2.5)$$

Det er også muligt at finde fasen af det komplekse resultat:

$$\phi(u) = \arg(F(u)) \quad (2.6)$$

Hvis man først har foretaget en fouriertransformation af et reelt signal, evt. har ændret resultatet, og så ønsker at foretage en invers fouriertransformation for at gengive det ændrede signal i tidsdomænet, skal man huske, at det er den komplekse fremstilling, som skal anvendes ved den inverse fouriertransformation.



Figur 2.3. På det øverste billede er vist fouriertransformationen af et signal. Det er ikke umiddelbart tydeligt at signalet er dannet ud fra en sinusbølge, hvis frekvens stiger med tiden. På det nederste billede er der foretaget en STFT af signalet, og her er det tydeligt at frekvensen stiger.

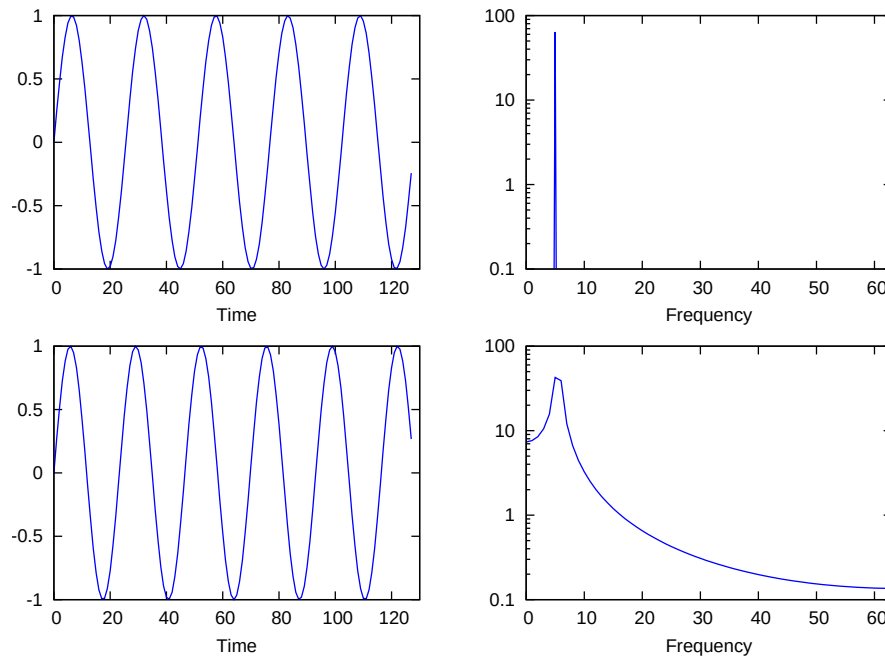
2.3.1.1 STFT

Hvis man tidsmæssigt set opstykker et signal, er det muligt at betragte hvert delstykke som et isoleret signal og foretage en fouriertransformation af hvert delstykke. Dette er nødvendigt, hvis man vil vide, hvordan et signal, som er repræsenteret i frekvensdomænet, ændrer sig i tiden. Netop dette er idéen bag STFT, som står for "short-time fourier transform".

Er et signal genereret ud fra en sinusbølge, hvis frekvens er stigende eller faldende, vil en fouriertransformation ikke direkte vise, at frekvensen ændrer sig, for det fouriertransformerede resultat fortæller kun noget om hver enkelt frekvens amplitude og fase. Det øverste billede i figur 2.3 viser amplituden for et signal og er genereret ved hjælp af en fouriertransformation. Farven rød markerer at amplituden er stor. Hvad man ikke umiddelbart kan se af billedet er, at der er tale om en frekvens, som stiger med tiden. I stedet for ser det ud som om, der er mange forskellige frekvenser i signalet. Den kraftigste frekvens svarer til den gennemsnitlige frekvens i det originale signal. På det nederste billede i figur 2.3 er der anvendt en STFT på signalet, og nu er det tydeligt at der er tale om en stigende frekvens.

2.3.1.2 Nyquistfrekvensen

Nyquistfrekvensen er den halve af samplingsfrekvensen af et diskret signal. Den markerer den laveste frekvens, som ikke kan identificeres korrekt med den gældende samplingsfrekvens. Her gives et eksempel: Hvis vi har en sinusbølge med en frekvens på 20 Hz, samt sampler signalet med en frekvens på 40 Hz, vil vi sample netop dér, hvor bølgen krydser 0 (og dette gør bølgen jo 2 gange pr. svingning). Hvordan skal vi ud fra samplerne kunne afgøre om frekvensen af signalet er 0 Hz, 20 Hz, 40 Hz, eller at andet helt multiplum af 20 Hz? Det



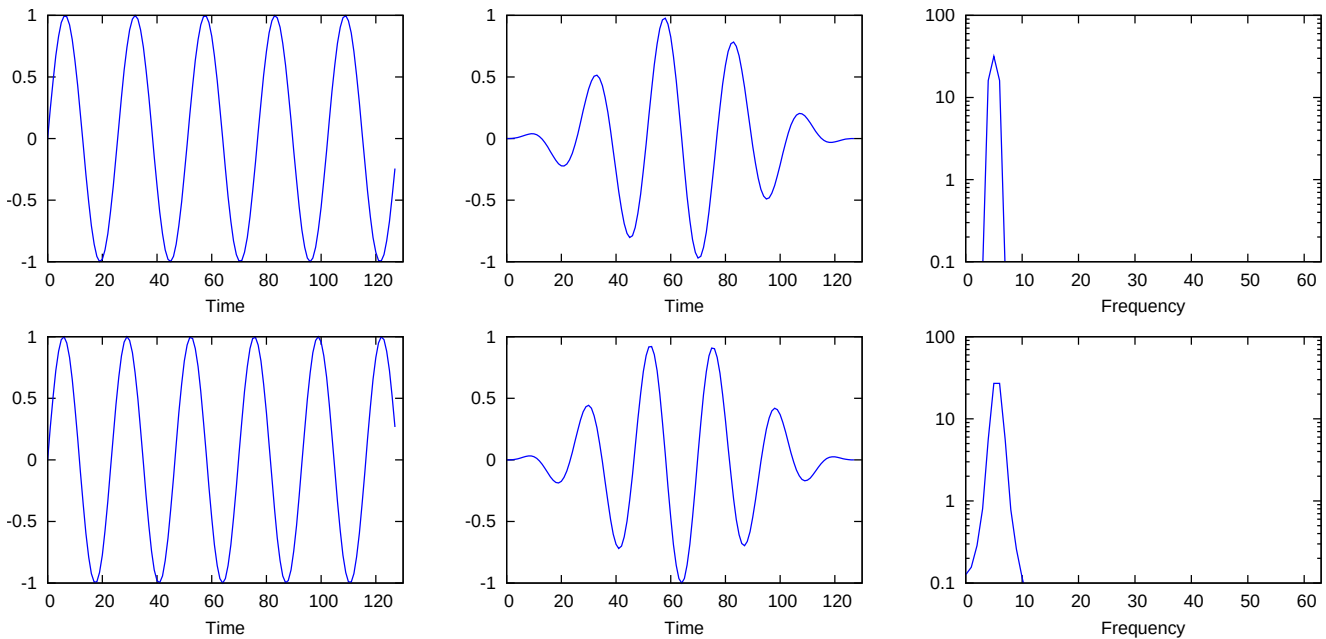
Figur 2.4. På billedet øverst til venstre er vist en sinusbølge, som svinger 5 gange. Der er 128 samples. På billedet øverst til højre er fouriertransformationen vist. Man kan se, at den eneste frekvens, som træder frem, er på 5 Hz. På billedet nederst til venstre svinger sinusbølgen 5,5 gange. Da der er anvendt en diskret fouriertransformation på billedet nederst til højre, er det kun heltallige frekvenser, som kan lade sig gøre. For at danne et signal, som ligner signalet nederst til venstre, er det nødvendigt at anvende alle de frekvenser, der er til rådighed. Frekvenserne på 5 og 6 Hz bliver vægtet højest.

kan vi ikke.

Dét, at vi ikke kan afgøre hvilken frekvens der er tale om, i tilfældet hvor frekvensen er større end eller lig med nyquistfrekvensen, giver anledning til *aliasing*. I denne sammenhæng dækker begrebet over, at én frekvens ligner en anden. Det oplagte eksempel, har vi allerede set, da vi ikke kunne afgøre, om der var tale om en frekvens på 0 Hz eller 20 Hz. For at undgå aliasing, er det nødvendigt at sikre sig, at det signal, som bliver samplet, kun indeholder frekvenser, som er lavere end nyquistfrekvensen. Så hvis man har et signal ved en bestemt samplingsfrekvens og ønsker at repræsentere signalet ved en lavere samplingsfrekvens, er det nødvendigt først at filtrere de høje frekvenser fra og dernæst foretage samplingen.

2.3.1.3 Vinduer

Den diskrete fouriertransformation foretages på et lukket interval, og dette er et problem, hvis frekvensen af signalet ikke passer med intervallets længde. På figur 2.4 kan man se 2 signaler og deres fouriertransformationer. Begge signaler er blevet samplet 128 gange. Da der er anvendt 128 samplinger, kan signalet repræsenteres ved 64 frekvenser - fra 0 til 63 Hz. De mulige frekvenser ligger altså på alle heltallige værdier op til 63 Hz. Signalet i det øverste venstre billede har netop 5 svingninger, og på fouriertransformationen på det øverste højre billede kan man se, at der kun er et udslag ved netop 5 Hz. På det nederste venstre billede er der 5,5 svingninger. Da det kun er muligt at repræsentere signalet med heltallige frekvenser, består det fouriertransformerede signal af alle frekvenserne fra 0 til 63 Hz. De 2 kraftigste frekvenser er 5 og 6 Hz.



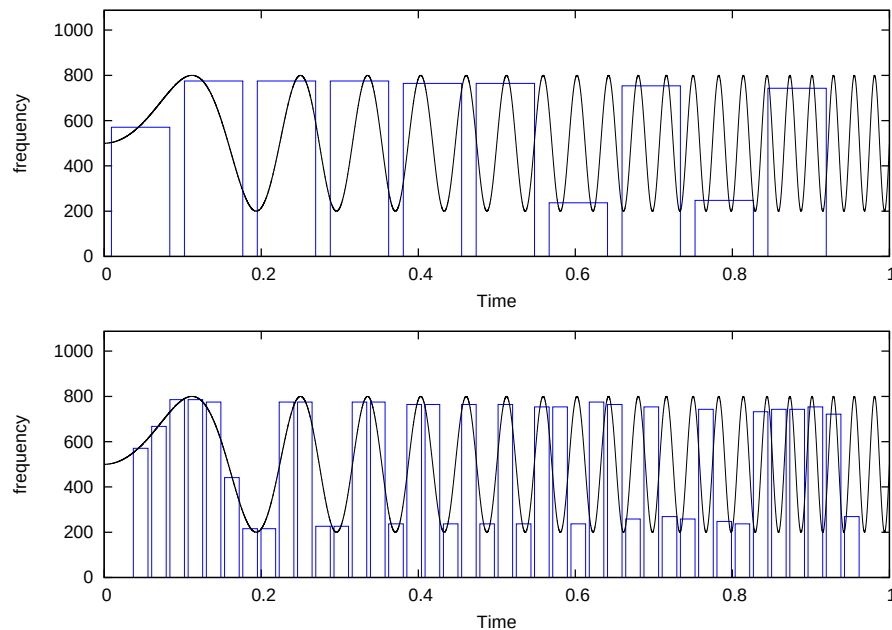
Figur 2.5. Denne figur svarer til figur 2.4, men her er der blevet anvendt et hanning-vindue, hvilket kan ses på billederne i midten. Det kan ses, at fouriertransformationen i de to tilfælde ser meget mere ens ud.

Det store problem ved signalet på det nederste venstre billede er, at det ikke passer sammen i de 2 ender. For at mindske dette problem kan man bruge et andet vindue end det, der indirekte er blevet anvendt. I denne sammenhæng er et vindue en funktion, som ganges på signalet og bestemmer i hvor høj grad de forskellige dele af signalet vægtes. Vi har indirekte ganget med en funktion, som er lig 1 for hele signalet. Men en mulighed er, at man ganger med en klokkeformet funktion, som har værdi 0 i siderne og 1 på midten. Efter den klokkeformede funktion er ganget på signalet, ligner det sig selv inde på midten, men ude i siderne bliver det gradvist svagere for til sidst, at blive lig 0. På denne måde passer signalet fint sammen i de 2 ender. Figur 2.5, viser de samme signaler som på figur 2.4, men der er blevet anvendt et såkaldt hanning-vindue, som netop har den beskrevne klokke-form.

Det kan ses af figuren, at resultatet er blevet meget bedre for tilfældet med en frekvens, som ikke passer til vinduets størrelse. Nu er det kun de frekvenser, som ligger meget nær den rigtige frekvens, der bliver påvirket af, at frekvensen ikke passer til vinduets størrelse. Desværre er der opstået problemer med de omliggende frekvenser i tilfældet, hvor frekvensen passer til vinduets længde. Det er dog helt klart at foretrække brugen af et vindue så som hanning-vinduet, da det jo næsten altid vil gælde, at frekvensen ikke passer til vinduets længde.

2.3.1.4 Window overlap

Ved at lade vinduerne i en STFT dække ind over hinanden (window overlap), således at den samme sample bliver anvendt i flere fouriertransformationer, kan man opnå en kunstigt højere tidsmæssig opløsning. Dette er illustreret på figur 2.6, hvor man kan se en STFT af en bølge, hvis frekvens ændrer sig hurtigere og hurtigere. Den sorte streg markerer den rigtige frekvens, mens de blå søjlerne markerer de frekvenser, som er blevet bestemt ved brug af STFT. På det



Figur 2.6. På det øverste billede er der ikke anvendt overlap. Det ses at STFT er dårlig til at gengive signalet allerede ved 0,2 sekunder markeringen. På det nederste billede er der anvendt et overlap på $3/4$, hvilket giver en kunstig høj opløsning, som er 4 gange højere. Gengivelsen er i dette tilfælde bedre i de første 0,6 sekunder, men herefter ligner gengivelsen ikke det originale signal, fordi afstanden mellem "højene" i signalet er mindre end vinduets størrelse.

øverste billede er der ikke anvendt noget overlap. Det ses at søjlerne allerede passer dårligt til signalet efter ca. 0,2 sekunder. Men afstanden mellem "højene" i signalet, er faktisk så langt fra hinanden, at det godt kunne have ladet sig gøre at repræsentere signalet bedre, hvis blot søjlerne havde været tidsmæssigt forskudt. Dette udnyttes ved brugen af overlap. På det nederste billede er der brugt den samme STFT, men her er anvendt et overlap på $3/4$ af størrelsen af vinduet - hvilket giver en kunstig høj tidsmæssig opløsning, som er 4 gange større. Så længe afstanden mellem højene er større end størrelsen af vinduet, går det godt, men efter ca. 0,6 sekunder går det galt, fordi afstanden mellem højene er for lille.

Herover blev overlappet beskrevet som en andel af vinduets størrelse. En anden måde at beskrive overlappet på, er ved at angive hvor meget vinduerne *ikke* overlapper, og dermed hvor langt der tidsmæssigt er mellem hver STFT. Denne størrelse kaldes for "hop-size" og kan både angives som et bestemt antal samples eller som en andel af vinduets størrelse.

2.3.1.5 Zero padding

Man kan også opnå en kunstigt højere opløsning i frekvensdomænet ved at anvende zero padding. Idéen er, at man forlænger et signal med en række nuller. Da det er længden af signalet (og dermed antallet af samplinger), der afgør hvor mange frekvenser, der anvendes i frekvensdomænet, kan det være en fordel med et langt signal, da der derved opstår kortere afstand mellem frekvenserne. Og hvis man har husket at anvende et godt vindue (så som hanning-vinduet), er det ikke noget problem at forlænge signalet med nuller, for signalet er alligevel lig 0 i kanten af vinduet.

Som nævnt er der tale om en kunstigt højere opløsning. I tilfældet hvor

signalet består af flere frekvenser, kan man altså ikke opnå en bedre adskillelse af frekvenserne. Men, hvis frekvenserne ligger tilstrækkelig langt fra hinanden, kan man bedre bestemme den enkelte frekvens.

2.3.2 Spektrogram

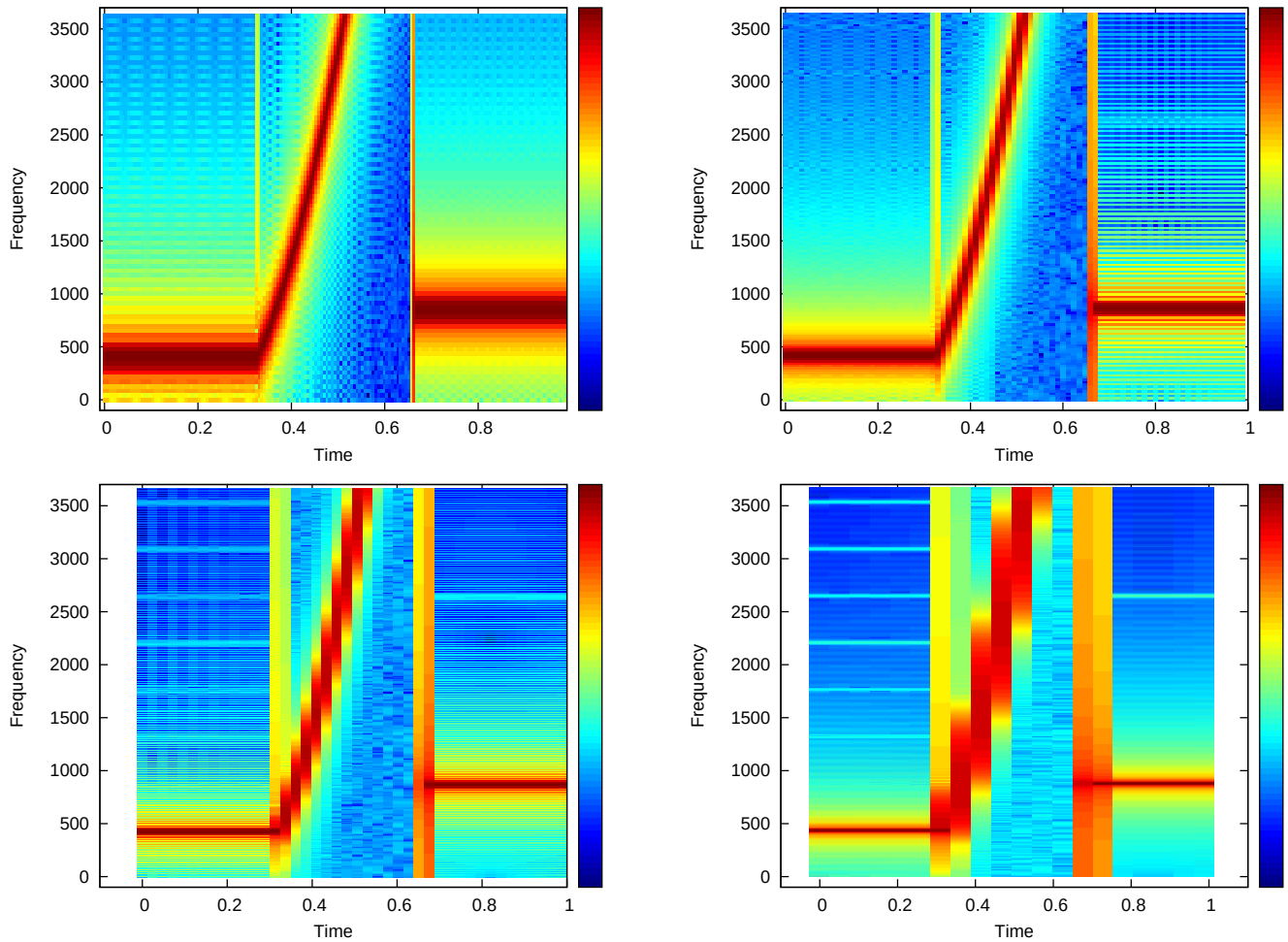
Et spektrogram viser styrken af forskellige frekvenser som funktion af tiden. Man kan f.eks. generere et spektrogram ved hjælp af STFT, eller ved hjælp af wavelettransformation. Faktisk har vi allerede set et eksempel på et spektrogram, da figur 2.3 indeholdt 2 spektrogrammer. Bemærk at når man visualiserer spektrogrammer, anvender man som regel en logaritmisk styrkeskala, således at det er logaritmen til amplituden, som bestemmer farven i spektrogrammet. Som vi skal se i afsnit 2.4.1.1 er dette meget fornuftigt, da den menneskelige opfattelse af lydstyrker er noget nær logaritmisk.

I det følgende vil vi se på nogle eksempler, som er genereret ved hjælp af STFT, og disse er genereret med forskellige vindueslængder, men i alle tilfælde anvendes et overlap på 50 % samt en zero padding, som har samme længde som det anvendte vindue. Derved opnås en kunstig opløsning, som er den dobbelte både på tidsaksen og på frekvensaksen.

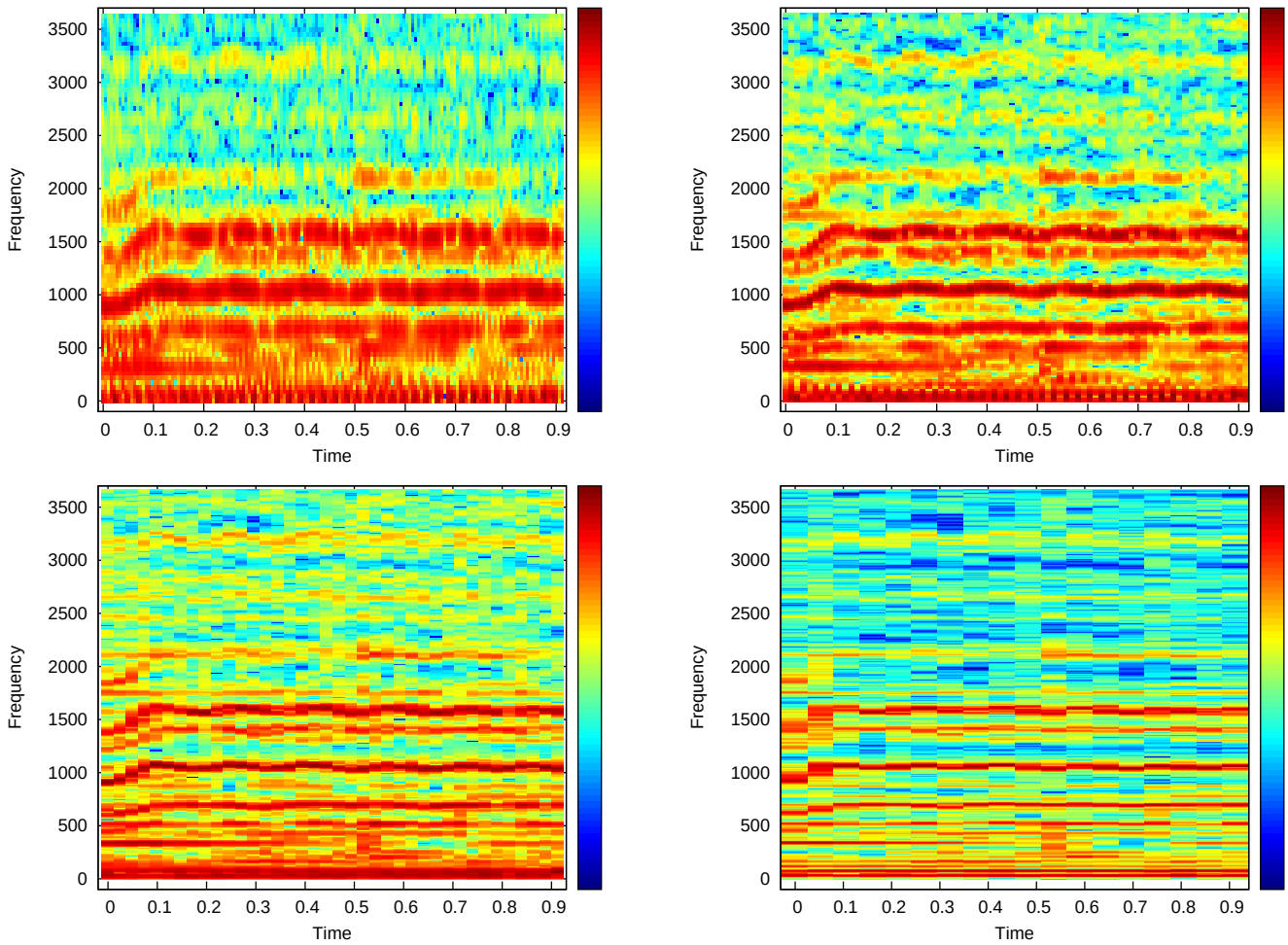
STFT har en fast opløsning, og da størrelsen af frekvensopløsningen er omvendt proportional med tidsopløsningen, gælder det om at finde en opløsning, som er passende. På figur 2.7 er 4 spektrogrammer af et kunstigt signal afbilledet. Signalet består først af en sinusbølge, dernæst øges sinusbølgens frekvens gradvist, og endelig afløses sinusbølgen af en anden sinusbølge. Signalet er samlet ved 44100 Hz og de anvendte vinduesstørrelser er 512, 1024, 2048 og 4096. På figuren kan man se, at en vinduesstørrelse på 4096 er god til at bestemme en frekvensen for en sinusbølge, som ligger fast over et længere interval, men der er problemer med at bestemme, hvornår én frekvens stopper og en anden begynder. Omvendt er dette meget tydeligt, når der anvendes en vinduesstørrelse på 512. I dette tilfælde kan man tydeligt se hvordan frekvensen stiger gradvist. Én af de 2 andre vinduesstørrelser er måske at foretrække som et godt kompromis.

Som nævnt er det et kunstigt signal, som er anvendt på den omtalte figur 2.7. På figur 2.8 er det derimod et virkeligt signal fra en musik-cd. I den korte sekvens, kan man høre 2 sangere, som begge starter med at glide op på hver deres tone. Derudover er der et svagt akkompagnement med en bastone, som har en fundamentalfrekvens ved 40 Hz. Lad os starte med denne bastone.

Man skal helt op i en vinduesstørrelse på 4096, før bastonens fundamentalfrekvens og første overtone kan adskilles. Ved en vinduesstørrelse på 2048 flyder disse frekvenser sammen, og ved de mindre vinduesstørrelser bevirker de små udsving i frekvensen, at det ser ud som om de lave frekvenser dynamisk set svinger hurtigt (og sådan *lyder* det altså ikke). Det ser ud som om en vinduesstørrelse på 2048 er bedre at anvende for frekvenser på omkring 300 Hz og derover. Man kan se, at en af sangerne har en overtone på omkring 1000 Hz, samt at sangeren glider op på tonen inden for de første 0.1 sekunder. Dét, at sangeren glider op på tonen, samt dét, at sangeren ikke synger en helt stabil tone, kan meget tydeligere ses, når der anvendes en vinduesstørrelse på 2048. Måske er det endda bedre at anvende en vinduesstørrelse på 1024. Under alle omstændigheder ser det ud, som om en vinduesstørrelse på 1024 er at foretrække for frekvenser omkring 2000 Hz og derover. At dømme efter det valgte



Figur 2.7. Her er vist 4 spektrogrammer for et kunstigt signal. De anvendte vinduesstørrelser er: 512 (øverst til venstre), 1024 (øverst til højre), 2048 (nederst til venstre) og 4096 (nederst til højre).



Figur 2.8. Her er vist 4 spektrogrammer for en kort sekvens fra en musik-cd. De anvendte vinduesstørrelser er: 512 (øverst til venstre), 1024 (øverst til højre), 2048 (nederst til venstre) og 4096 (nederst til højre).

eksempel, ser det ikke ud til at være nogen fordel at bruge en vinduesstørrelse på 512 for nogen frekvenser.

2.3.3 Foldning

Her præsenteres en matematisk funktion, som kan være nyttig at kende til, når man skal analysere et signal: Den endimensionale foldning er defineret som:

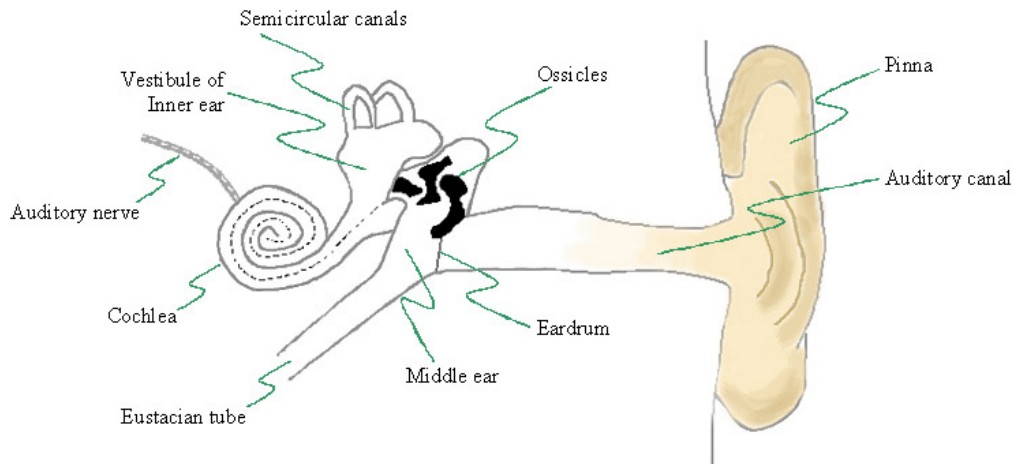
$$(f \star g)(t) = \int_{-\infty}^{\infty} f(\alpha)g(t - \alpha)d\alpha \quad (2.7)$$

Her opskrives den tilsvarende diskrete foldning, hvor n er en indekxværdi som løber fra 0 til N :

$$(f \star g)(n) = \sum_{m=0}^N f[m]g[n - m] \quad (2.8)$$

2.3.3.1 Foldningssætningen

Her skal vi se hvordan foldning og fouriertransformation hører sammen:



Figur 2.9. Her er vist en skematisk repræsentation af det menneskelige øre. Det vigtigste er, at lyde efter at være blevet overført via de 3 små øreknogler, som er markeret med sort, forplanter sig til øresneglen (cochlea). Det er i cochlea, at lydbølgerne omsættes til signaler, som sendes til hjernen. (Figuren er taget fra <http://www.diracdelta.co.uk/science/source/e/a/ear/source.html>.)

Foldningssætningen fortæller, at hvis man folder 2 funktioner, svarer det til at fouriertransformere dem, gange resultatet sammen, og endelig foretage en invers fouriertransformation:

$$f(x) \star g(x) = (f \star g)(x) = F^{-1} \left\{ F\{f(x)\} F\{g(x)\} \right\} \quad (2.9)$$

Ofte vil det gå hurtigere at foretage denne transformation, frem for at udføre foldningen.

Der gælder også den omvendte sammenhæng. Hvis man ganger to funktioner sammen, kan man i stedet for fouriertransformere dem, folde resultaterne og foretage invers fouriertransformation:

$$f(x)g(x) = F^{-1} \left\{ F\{f(x)\} \star F\{g(x)\} \right\} \quad (2.10)$$

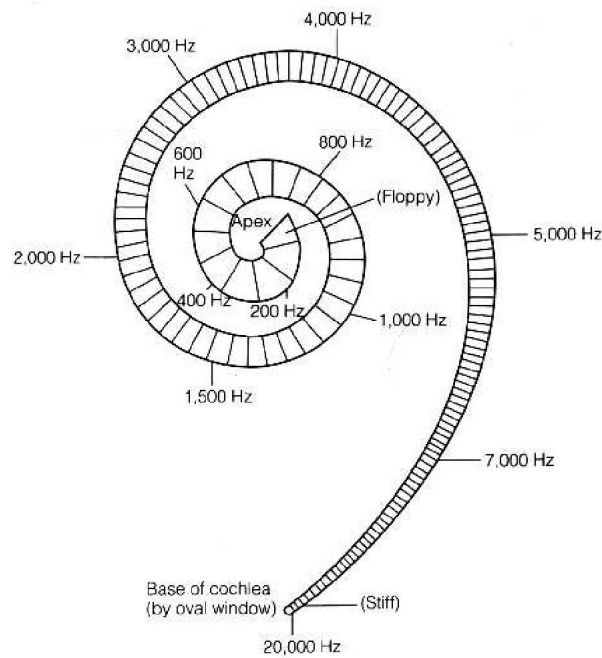
Denne sammenhæng anvendes i praksis kun sjældent, men der kan være eksempler, hvor den fouriertransformerede af den ene funktion er meget simpel. Dette er f.eks. tilfældet, hvis man har et hanning-vindue i tidsdomænet og ønsker at anvende det i flere forskellige opløsninger. Mere herom i afsnit 3.2.

2.4 Det menneskelige øre

På figur 2.9 er vist en skematisk repræsentation af det menneskelige øre. Lyden overføres via de 3 små øreknogler til øresneglen (cochlea), hvis form ligner et sneglehus. Cochlea er væskefyldt, og lydbølgerne, som forplanter sig i væsken, får basilarmembranen og derved også sansehårene på basilarmembranen til at svinge med. Svingningerne i sansehårene giver nerveimpulser, som sendes til hjernen.

De forskellige steder på basilarmembranen, svinger mere eller mindre godt afhængig af frekvensen. Således svinger membranen bedst for høje frekvenser nær indgangen og lave frekvenser nær enden - se figur 2.10.

Der er to teorier om, hvordan signalerne fra sansehårene anvendes til bestemmelse af frekvenser [6, afsnit 3.3.1]. Sted-teorien går ud på, at hjernen kan



Figur 2.10. Her er vist en skematisk repræsentation af cochlea. Ved indgangen (som ses i bunden af tegningen) er basilarmembranen ikke så blød og svinger godt ved høje frekvenser. Mod enden bliver membranen blødere og svinger bedst ved lave frekvenser. (Figuren er taget fra [6, afsnit 3.3.1].)

bestemme hvilke dele af basilarmembranen, der svinger, og derved fastlægge frekvenserne. Tidsteorien går ud på, at sansehårene svinger i takt med de frekvenser, som aktiverer dem, og derfor også udsender signaler til hjernen i takt med frekvenserne. Hjernen kan derfor bestemme frekvenser, ved at undersøge hvor ofte der sendes signaler. Det er muligt, at der faktisk anvendes en kombination af de to teorier, og at tidsteorien måske anvendes mest for frekvenser under 4 - 5 kHz, og sted-teorien anvendes mest for frekvenser herover.

2.4.1 Psykoakustik

Det kan være svært objektivt set at udtale sig om en lyd og dens frekvensielle bestanddele, blot ved at lytte til lyden; for i selve øret er der rent mekaniske begrænsninger for, hvordan forskellige frekvenser kan opfattes. En stærk begrænsning er ørets evne til at reagere på forskellige frekvenser. Normalt kan et øre reagere i frekvensområdet fra 20 Hz op til omkring 20.000 Hz.

Ud over de fysiske begrænsninger, spiller psykologiske effekter også ind på vores opfattelse af en lyd. Således kan man i nogle tilfælde få opfattelsen af, at man hører en bestemt frekvens, selv om denne slet ikke er tilstede i lyden.

Der har været foretaget mange undersøgelser af den menneskelige reaktion på forskellige frekvenser, frekvensernes styrker samt ændringer i tiden. Nogle af resultaterne har med fordel kunnet bruges til at komprimere et signal, uden at mennesker vil opdage, at der er sket en ændring af signalet. Dette bruges f.eks. i MP3-kompression, hvor specielt maskeringseffekten bliver anvendt. (Denne effekt bliver dog ikke berørt nærmere her).

I de følgende underafsnit nævnes nogle af det vigtigste begreber, når det handler om den menneskelige opfattelse af lyde.

2.4.1.1 Decibel

Til at bestemme styrken af et signal, anvendes som oftest decibel-skalaen som måling af lydtrykket. Decibel-skalaen er en relativ skala, dvs. den fortæller noget om lydtrykket, relativt i forholdt til et referencelydtryk, som definerer 0 dB. Referencelydtrykket sættes ofte til at være et lydtryk på $20 \mu\text{Pa}$, hvilket generelt betragtes som den svageste lyd et menneske kan høre. Amplituden af et signal angiver det maksimale lydtryk, så det relative lydtryk kan både udregnes ved hjælp af selve lydtrykket, men også ved hjælp af amplituden. Det relative lydtryk kan beregnes som:

$$L = 20 \log_{10} \left(\frac{A}{A_0} \right) = 20 \log_{10} \left(\frac{P}{P_0} \right) \quad (2.11)$$

, hvor L er det relative lydtryk målt i dB, A er amplituden af signalet og A_0 er amplituden for et signal, som er defineret til at have et lydtryk på 0 dB, P er lydtrykket, og P_0 er referencelydtrykket.

Mennesker har generelt svært ved at høre en forskel i lydtrykket, hvis der sker en ændring på kun 1 dB. Men den menneskelige opfattelse af, hvor kraftig en lyd er, afhænger af frekvensen. Dette er der forsøgt kompenseret for med forskellige vægte af decibel-skalaen. Nogle vægte er specielt beregnet til særlige former for støj, mens andre er lavet ud fra menneskets opfattelse af rene toner. Den mest anvendte vægtning er A-vægtningen. Når denne vægtning tages i brug anvendes enheden dBA eller dB(A). Kurven for A-vægtningen er baseret på 40 phon-kurven, som vi hører om i næste afsnit.

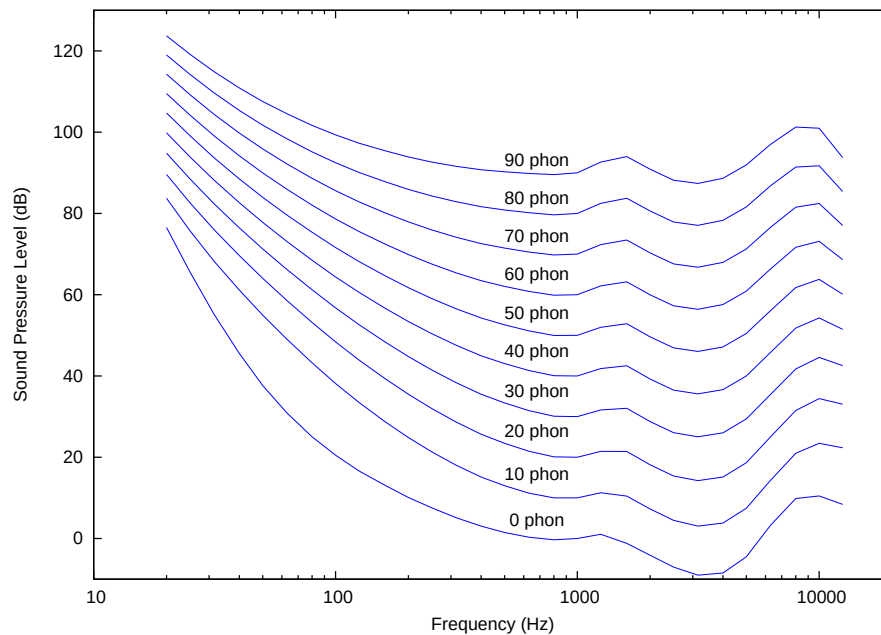
2.4.1.2 Phon

Enheden phon bruges til at angive hvor kraftigt en tone lyder. Idéen med enheden er, at når 2 forskellige toner med forskellige frekvenser lyder lige kraftige, så har de samme phon-værdi. For en tone på 1 kHz svarer phon-skalaen til dB-skalaen. Figur 2.11 viser phon-kurver for den seneste standard, som kan findes i ISO 226 :2003. Det ses af figuren, at dybe lyde skal være meget kraftige, før vi kan opfatte dem; men når først lydstyrken er på omkring 100 dB, har lydens frekvens ikke så stor betydning for, hvor kraftigt vi opfatter lyden. Det ses desuden, at mennesker er mest sensitive i området fra omkring 2000 til 5000 Hz. Desværre er kurverne på figuren ikke definerede for frekvenser helt op til 20.000 Hz, som er grænsen for, hvad de fleste mennesker kan høre; men det antages at kurverne blot stiger kraftigt.

2.4.1.3 Lydskalaer

Det blev nævnt i afsnit 2.2, at den musikalske skala i vestlig musik består af 12 trin eller toner. Det 13. trin opfattes som en gentagelse af det 1. trin - blot er der tale om en lysere udgave af den samme tone. Vi opfatter afstandene mellem 2 trin, som følger efter hinanden i skalaen, som lige store, selv om skalaen har et logaritmisk forhold til frekvens-skalaen. Så ligesom med vores opfattelse af lydtryk, som ser ud til at kunne beskrives med noget nær en logaritmisk skala, kan vores opfattelse af toner også beskrives med noget, som minder om en logaritmisk skala.

Der er flere bud på, hvordan en sådan skala ser ud. Det ene bud er indirekte givet ved 12-tone-skalaen, som er logaritmisk. Ellers er mel-skalaen og bark-



Figur 2.11. Her er vist phon-kurver. To toner, hvis frekvenser er forskellige, men som har den samme phon-værdi, opfattes som lige kraftige, selv om de muligvis har forskellige lydstyrker.

skalaen de to mest kendte skalaer og er baseret på eksperimenter. Begge skalaer har overvejende et logaritmisk udseende.

2.4.1.4 Manglende fundamentalfrekvens

Overtonerne har tilsyneladende stor betydning for, hvordan vi mennesker opfatter en tone. Der er ikke bare tale om, at overtonerne giver en tone klangfarve, således at den samme tone spillet på 2 forskellige instrumenter kommer til at lyde forskelligt.

Overtonerne har indflydelse på vore opfattelse af en tones tonehøjde. For nogle mennesker har overtonerne større betydning end for andre. På mange lydsystemer (så som telefoner), kan dybe toner under 100 Hz slet ikke eller kun i ringe grad gengives, hvilket bevirker, at fundamentalfrekvensen for de fleste bastoner slet ikke er til stede i lydsignalet. Dette er dog ikke nødvendigvis et problem, fordi overtonerne findes i signalet, og ud fra disse får hjernen opfattelsen af, at der høres en tone med en frekvens, som svarer til den manglende fundamentalfrekvens. Når en operasanger kan høres gennem et helt orkester, skyldes det primært, at operasangerens overtoner, som ligger omkring 3000 Hz, er kraftige. Det er overvejende disse overtoner, som vi danner toneopfattelsen på baggrund af.

Kapitel 3

Metoder for bestemmelse af melodien

I dette kapitel beskæftiger vi os med nogle af de metoder, der har været anvendt i forsøget på at bestemme melodien i et lydsignal. Vi vil også se nærmere på repræsentationen af signalet, da en god repræsentation (f.eks. i form af et multiopløsningsspektrogram) har betydning effektiviteten for nogle af metoderne.

Selv om det måske er den oplevede tonehøjde, som er mest interessant, når man ønsker at bestemme melodien, så er fundamentalfrekvensen og den oplevede tonehøjde så tæt knyttet, at det virker fornuftigt at omtale melodiens fundamentalfrekvens som synonym for melodiens tonehøjde. Ofte bliver der ikke skelnet mellem metoder for bestemmelse af fundamentalfrekvensen og metoder for bestemmelse af tonehøjden.[7, med henvisning til W. Hess (1983)] Men specielt når det handler om polyfon musik, hvor der optræder flere fundamentalfrekvenser, er det svært at finde fundamentalfrekvenserne, da de skal findes blandt mange andre frekvenser.

Ind til Masataka Goto i år 1999 præsenterede sin PreFEst-algoritme [15] [14], havde det meste arbejde omhandlende estimering af fundamentalfrekvensen i et lydsignal, drejet sig om lydsignaler, hvor der kun er ét instrument, eller hvor der er én stemme samt dæmpet underliggende støj. De få stykker arbejde, som tidligere havde beskæftiget sig med polyfon musik, krævede som regel enten, at instrumenterne var kendt på forhånd, eller at der kun var tale om ganske få instrumenter.

I det følgende præsenteres nogle betydningsfulde eller interessante metoder for estimering af fundamentalfrekvensen. Det vil vise sig at de fleste interessante metoder anvender frekvensdomænet. Derfor vil vi efterfølgende se på repræsentationen af data i dette domæne. Endelig vil vi se på forslag til måder, hvorpå man kan spore fundamentalfrekvensen blandt flere kandidater.

3.1 Estimering af fundamentalfrekvensen

Med henvisning til W. Hess (1983), så nævner Gómez og Klapuri [11], at den proces at finde fundamentalfrekvensen kan opdeles i 3 dele: Først foretages en pre-processing af lydsignalet (dvs. der udføres eventuelt en form for datareduktion), dernæst kommer den centrale del, hvor fundamentalfrekvenser findes, og endelig kan der udføres en post-processing, for at fjerne eventuelle fejl eller udglatte resultatet. I dette afsnit vil vi beskæftige os med den centrale del.

Der er igennem tiderne foreslået mange forskellige metoder, så her vil vi kun berøre nogle af det mest kendte eller interessante. Først koncentrerer vi os om metoder, som anvender tidsdomænet, og dernæst fortsætter vi med metoder, som anvender frekvensdomænet.

3.1.1 Metoder i tidsdomænet

Disse metoder beskæftiger sig med det "rå" signal. Det er ændringen af lydtrykket som funktion af tiden, der anvendes af metoderne. Den vigtigste metode i tidsdomænet er autocorrelation. Der er også metoder, som anvender faserummet¹; men disse metoder er ikke særlig udbredte.

3.1.1.1 Zero-crossing rate (ZCR)

I denne metode måles det, hvor ofte lydbølgen krydser 0 på amplitudeaksen. Denne metode er meget simpel, men desværre ikke særlig anvendelig, da en tone, som består af nogle kraftige overtoner, vil have en lydbølge, som krydser 0 oftere end en lydbølge, som kun er dannet ud fra fundamentalfrekvensen. Og hvis man har mere end én tone, vil det næsten være umuligt at bestemme fundamentalfrekvensen for den relevante tone.

3.1.1.2 Autocorrelation

Her måles det, hvor meget lydbølgen ligner sig selv et bestemt tidsinterval forskudt. Idéen er, at hvis lydbølgen ligner sig selv (og dermed gentager sig), så svarer tidsforskydelsen til en periode for fundamentalfrekvensen. Så ved at anvende forskellige tidsintervaller og derved finde det tidsinterval, hvor lydbølgen bedst ligner sig selv, kan man finde fundamentalfrekvensen. Selvfølgelig vil lydbølgen ligne sig selv fuldstændig, hvis tidsintervallet er lig 0, så for at undgå dette problem, er man nødt til at kræve, at tidsintervallet har en længde, som overstiger en bestemt størrelse. Den matematiske definition for autocorrelation i diskret tid er:

$$R(n) = \frac{1}{K} \sum_{k=0}^{K-n-1} x[k] \cdot x[k+n] \quad (3.1)$$

, hvor x er en vektor, som beskriver lydsignalet og n er tidsintervallet. Når R er beregnet, kan man lave en afbildning af R som funktion af n , og så kan fundamentalfrekvensen findes ved den første "store høj" i afbildningen. Som det vil blive forklaret i afsnit 3.1.2.4, kan autocorrelation også formuleres ved hjælp af en kam-funktion.

Autocorrelation kan ses som en foldning af signalet med signalet selv - se afsnit 2.3.3. Så en effektiv måde at beregne $R(n)$ på, foregår ved "fouriertransformation, opløst i anden, invers fouriertransformation" [6, side 68]:

$$R = F^{-1} \left\{ (F\{x\})^2 \right\} \quad (3.2)$$

Problemet ved autocorrelation er, at overtonerne også skaber høje i det omtalte billede, og at signalet gentager sig ved den dobbelte bølgelængde (og dermed den halve frekvens) af fundamentalfrekvensen. Så hvis en overtone er kraftigere end fundamentalfrekvensen, er det overtonen, som bliver valgt.

¹For en kort introduktion til dette, se [13].

Cheveigné og Kawahara præsenterede i 2001 en algoritme kaldet YIN [7]. Denne er baseret på autocorrelation, men tilføjer en række modifikationer, som skal mindske mængden af fejlagtige estimater for fundamentalfrekvensen.

3.1.2 Metoder i frekvensdomænet

Frekvensdomænet kan ses som det mest oplagte domæne at anvende, da det jo netop er en frekvens, som ønskes estimeret.

Metoderne i frekvensdomænet anvender en konvertering af signalet fra tidsdomænet til frekvensdomænet. Den mest anvendte konvertering er fouriertransformationen, som udføres på signalet i korte tidsintervaller. Afsnit 3.2 ser vi nærmere på mulighederne for at give metoderne i frekvensdomænet en god signalrepræsentation at arbejde med.

3.1.2.1 Cepstrum Analysis

Denne metode, som egentlig blev introduceret for undersøgelse af seismiske ekkoer, var, ifølge Philippe Martins artikel fra 1982 [19], formegentlig den mest populære metode. I slutningen af 70'erne og i 80'erne blev der dog præsenteret en række bedre metoder.

Frekvensdomænet kaldes også for frekvensspektrummet (eller blot spektrummet), og en analyse af spektrummet kaldes derfor "spectral analysis" på engelsk. Cepstrum Analysis dækker over en analyse af fouriertransformationen af en logaritmisk repræsentation af spektrummet. Da overtonerne til fundamentalfrekvensen oftest er meget svagere end fundamentalfrekvensen, er det en fordel at anvende logaritmen til styrkerne af frekvenserne. Idéen er, at overtonerne ligger med et fast interval imellem sig, så hvis man fouriertransformerer det logaritmiske spektrum, vil den største høj i det resulterende cepstrum angive den mest dominerende periode, og derved kan fundamentalfrekvensen findes.

Problemet med Cepstrum Analysis er, at den er afhængig af, at der for det første er mange overtoner, og for det andet at disse er tilstrækkelig kraftige. Desuden skal alle overtonerne helst være til stede.

3.1.2.2 Component Frequency Ratios

Denne metode, som blev præsenteret af Martin Piszczalski i 1979 [21], er stærkt inspireret af Schoeders idéer fra 1968 [24]. Den baserer sig på overtonerne, og hvis blot nogle af disse er tilstrækkelig tydelige, så har metoden mulighed for at finde fundamentalfrekvensen. Da metoden baserer sig på overtonerne, er den robust over for en svag eller måske endda manglende fundamentalfrekvens.

Metoden finder først "høje" i frekvensdomænet. Disse høje betragtes som kandidater til at være overtoner til den fundamentalfrekvens, der ønskes fundet. Der genereres nu par af høje, således at hver høj er parret med hver af de øvrige høje. For hvert par, bestemmes den højeste fundamentalfrekvens, for hvilken begge høje kan være overtoner. I praksis tillades det, at højenes frekvenser kan flyttes lidt, da der er en vis opløsning på frekvensaksen, og da overtonernes frekvenser ikke altid overholder matematiske formler. (Se afsnit 2.1.1.) Afhængig af hvor godt hvert par passer sammen, samt højenes styrker, tildeles hvert pars fundamentalfrekvenskandidat en vægt. Endelig er det disse vægte, som afgør det endelige bud på et gæt på en fundamentalfrekvens.

3.1.2.3 Spektral kompression

I spektral kompression antages det indirekte at alle frekvenser er fundamentalfrekvenser til toner, som har en række overtoner. Vægtningen af hver frekvens (underforstået fundamentalfrekvens) sker ved en sum eller et produkt af kvadratet på overtonernes amplituder [5]. Noll (1970) beregner således det harmoniske sum-spektrum ved:

$$S(f) = \sum_{n=1}^N \|F_n(f)\|^2 \quad (3.3)$$

, hvor f er frekvensen og F_n er den n 'te kompression af frekvensspektrummet beregnet ved en fouriertransformation. Idéen er, at den fundamentalfrekvens, som man søger, vil fremstå som en høj i det harmoniske sum-spektrum S .

3.1.2.4 Kam-metoder

Der findes flere forslag til metoder, som på en eller anden måde anvender en funktion, som minder om en kam - dvs. en funktion, som er skiftevis høj og lav. Nogle metoder anvender kam-funktioner, hvor tænderne i kammen er meget tynde, mens andre anvender tykke tænder. Som vi skal se, er metoder, som anvender kam-funktioner ikke begrænset til frekvensdomænet; men vi behandler metoderne her, fordi de mest interessante metoder, som anvender kam-funktioner bruger frekvensdomænet.

Man kan egentlig godt opfatte autocorrelation som en metode, der anvender en bevægelig kam med 2 tænder. Hvis vi ser på formel 3.1, så kunne der have været anvendt en kam, hvor afstanden mellem de 2 tænder er angivet ved variablen n .

I 1974 introducerede James A. Moorer en metode kaldet Optimum Comb [20], som i bund og grund blot er en anden betegnelse for autocorrelation. I stedet for at maksimere en multiplikation, søges det at minimere kvadratet på afstanden. I følgende formel for Optimum Comb angiver m afstanden mellem de 2 tænder på kammen:

$$\min_m \left(\sum_{i=0}^{k-1} (X_{n+i} - X_{n+i-m})^2 \right) \quad (3.4)$$

, hvor X beskriver lydbølgen i *tidsdomænet*, og k er længden af den lydbølge der ønskes analyseret. Der minimeres over afstanden mellem tænderne på kammen.

Optimum Comb Filter I [13] henvises der til James A. Moorer (1979) under beskrivelsen af en metode, som arbejder i frekvensdomænet og anvender en kam-funktion med brede tænder. Metoden, som kaldes Optimum Comb Filter, har en række "pass-band" filtre ligeligt fordelt på frekvensaksen, og disse filtre, som tillader et bånd af frekvenser at slippe igennem, har maksimal åbning ved den centrale frekvens.

I Optimum Comb Filter, afgør det første filter de øvrige filters placering. Afstanden fra det første filters centrale frekvens til 0 Hz, er den afstand der er mellem hvert af filtrene. Dette er meget anvendeligt, da filtrene i teorien vil komme til at ligge ved hver overtone, såfremt det første filter ligger ved en fundamentalfrekvens. Da et filter kan tillade et bånd af frekvenser at slippe igennem, gør det ikke så meget, hvis overtonerne ikke ligger helt præcist.

Det første filter lader man løbe over det frekvensområde, som fundamentalfrekvensen søges i. I de højere frekvenser placeres filtre som beskrevet ovenfor. Summen af styrkerne af de filtrerede frekvenser, samt en vægtning af hvert filter, afgør bestemmelsen af fundamentalfrekvensen.

Beskrivelsen i [13] svarer til beskrivelsen af en metode, som Philippe Martin i 1981 selv mener at have introduceret [19]. I Philippe Martins terminologi kaldes metoden for Spectral Comb Analysis eller Spectral Comb Method; og her specificeres det, at vægten af filtrene skal foregå ved en faldende funktion, som f.eks. kan være $n^{-1/k}$, hvor n angiver nummeret på tanden.

For at undgå problemer med, at der vælges en fundamentalfrekvens, som er lavere end den rigtige fundamentalfrekvens, kan man nøjes med kun at tage de første n overtoner med i betragtning, når der skal tildeles vægte.

Bemærk at Optimum Comb Filter (og Spectral Comb Analysis) ikke er afhængig af overtoner, men heller ikke er afhængige af en fundamentalfrekvens, hvis blot en af delene er tilstede og er tilstrækkeligt stærk.

3.1.2.5 Subharmonic Summation

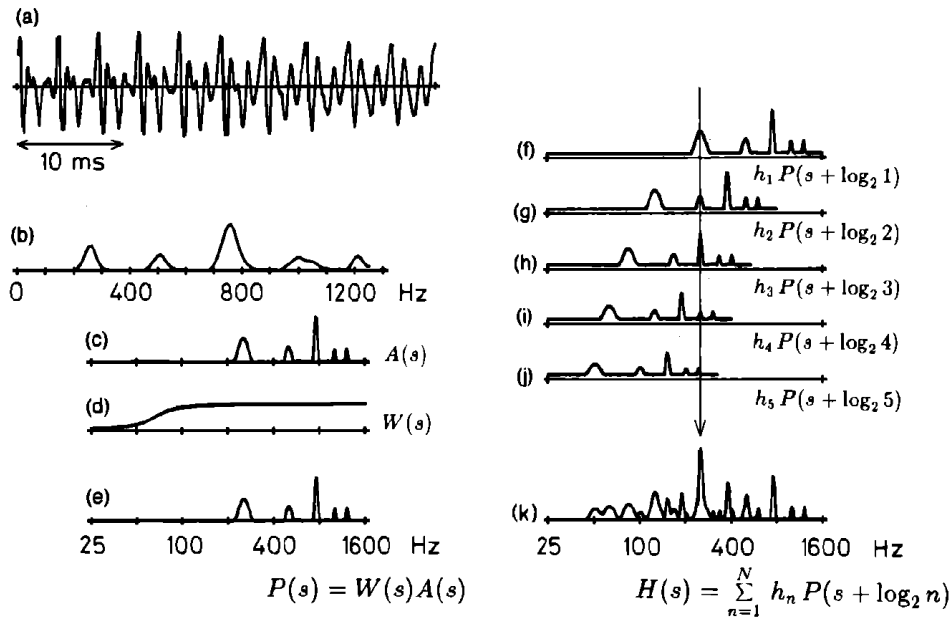
Subharmonic Summation (SHS) blev introduceret af Dik Hermes [16] i en artikel fra 1986, som dog først blev udgivet i 1988. SHS minder i bund og grund meget om Optimum Comb Filter og er baseret på en modifikation af spektral kompression.

Logaritmisk frekvensakse Hermes bemærker at den lineære frekvensakse, som anvendes i spektral kompression er et problem, fordi den begrænsede opløsning sætter en øvre grænse for, hvor mange kompressioner der med fordel kan anvendes (og det lader til at være 5 kompressioner). I SHS beregnes først fouriertransformationen, og ud fra denne bliver punkter på en logaritmisk frekvensakse beregnet ved hjælp af interpolation. Den logaritmiske akse giver mulighed for at have en god frekvensopløsning ved lave frekvenser. Når amplituden af frekvenserne nu er repræsenteret på en logaritmisk frekvensakse, kan summen fra ligning 3.3, beregnes på en nemmere måde.

Logaritmen til en multiplikation er lig summen af logaritmen til hvert af ledene. For at beregne det komprimerede spektrum på en lineær akse, er man nødt til at dele alle frekvenserne med kompressionsfaktoren. Så på en logaritmisk akse, kan vi i stedet for udregne et komprimeret spektrum ved at trække logaritmen til kompressionsfaktoren fra. Kompressionen på den logaritmiske akse svarer altså blot til en flytning. (Problemet er blot, at flytningen netop foregår ved logaritmen til kompressionsfaktoren, hvilket ikke er en heltalsværdi.)

Algoritmen i korte træk Her oplistes algoritmens overordnede forløb - se også figur 3.1:

- Først foretages en fouriertransformation.
- Dernæst findes høje i frekvensspektrummet, og disse udvælges.
- Da vi har med en numerisk transformation at gøre, vil vi her betragte hver af de mulige frekvenser i frekvensspektrummet som trin. For at fjerne støj bliver alle frekvenser, som befinder sig over x antal trin væk fra en høj, sat til at have en amplitude på 0.



Figur 3.1. Her er de forskellige stadier i SHS algoritmen illustreret. (a) viser bølgeformen af et signal. (b) Her er signalets fouriertransformation vist. (c) Der er foretaget en støjfiltrering og frekvenserne repræsenteres på en logaritmisk akse. (d) Her vises det højpas-filter, som undertrykker de dybe frekvenser. (e) Her vises resultatet af filtreringen. (f-j) Her vises det, hvordan kompressionen foregår ved hjælp af en flytning. (k) Ved at summere over en vægtning af de komprimerede spektre, kan resultatet H bestemmes. Den kraftigste frekvens i resultatet gættes på som fundamentalfrekvensen. (Figuren er taget fra [16].)

- Det filtrerede frekvensspektrum foldes med et hanning-vindue for at udglatte højene.
- Ud fra det filtrerede frekvensspektrum beregnes et nyt interpoleret spektrum, som har en logaritmisk frekvensakse.
- Der foretages en højpas-filtrering, som undertrykker de dybe frekvenser. (Dette foregår med en forhøjet arc-tang funktion, som skal efterligne den dårlige menneskelige evne til at høre dybe frekvenser - se afsnit 2.4.1.2.) Resultatet af denne operation kaldes P .
- For hver kompressionsfaktor genereres nu nye frekvensspektre, som blot er flytninger af det oprindelige spektrum.
- De komprimerede spektre multipliceres med hver deres vægte h_k .
- Endelig bliver spektrene lagt sammen.

Alt dette er summeret i følgende formel for udregningen af det subharmoniske sum-spektrum H på den logaritmiske frekvensakse:

$$H(s) = \sum_{n=1}^N h_n P(s + \log_2 n) \quad (3.5)$$

, hvor n er kompressionsfaktoren, h_n er vægten af den n 'te kompression og $s = \log_2(f)$. Vægten h_n er foreslået sat til 0.84^{n-1} . Estimatet for fundamentalfrekvensen er maksimum af H .

Af hensyn til næste afsnit noteres udregningen af H også som funktion af frekvensen på en lineær akse:

$$H(f) = \sum_{n=1}^N h_n P(nf) \quad (3.6)$$

3.1.2.6 En ændret SHS

Et problem ved SHS er, at der let kan opstå oktavfejl, således at den frekvens, som maksimerer H , er den halve af fundamentalfrekvensen.² Generelt er de lave frekvenser kraftigere end de høje, hvilket bevirker at dybe frekvenser indirekte bliver favoriseret mere end de høje. For at minimere dette problem, er det i [3] foreslået, at der medtages et integral i formlen for udregningen af H . Således kommer formlen til at se således ud:

$$H(f_0) = \sum_{n=1}^N h^{n-1} Q_{f_0}(nf_0) \quad (3.7)$$

, hvor

$$Q_{f_0}(f) = P(f) / \left(\frac{1}{2\rho f_0} \int_{f-\rho f_0}^{f+\rho f_0} P(w) dw \right) \quad (3.8)$$

, ρ bestemmer hvor mange frekvenser, der anvendes til bestemmelsen af den gennemsnitlige frekvensstyrke omkring en given frekvens.³ Denne ændrede udgave af SHS anvender altså information om, hvor kraftige de enkelte frekvenser er set i forhold til de omkringliggende frekvenser. Jo højere frekvens SHS ønskes beregnes for, des bredere er det frekvensområde som gennemsnittet beregnes over. Dette virker fornuftigt, fordi et spektrum ofte har meget kraftige dybe frekvenser, og gradvist flader mere og mere ud som frekvensen stiger.⁴ Forfatterne til [4] opnåede gode resultater ved brug af SHS, men i [3] anvendte de den ændrede udgave, da den efter deres erfaring virker bedre.

3.1.3 Statistiske metoder

Statistiske metoder så som Maximum Likelihood har også været anvendt, men statistiske metoder berøres ikke nærmere, da vi koncentrerer os om de mere simple metoder. Der er dog plads til nogle få bemærkninger.

Goto anvendte en statistisk metode (nemlig Expectation-Maximization) i sin PreFEst-algoritme [15] [14], men som det bliver afsløret i kapitlet, som omhandler de generelle resultater, er der siden foreslået andre simple algoritmer, som ser ud til at virke bedre. (Se kapitel 5.)

Modeller som anvender neurale netværk, kan ses som et forsøg på at efterligne hjernen uden egentlig at vide, hvordan hjernen behandler signaler fra øret. Det primære problem ved neurale netværk er, at de ofte fungerer som en

²I [3] nævnes det blot at der opstår oktavfejl, men egne test peger i retning af, at det er dybe oktaver, som er det største problem.

³Der er desværre anvendt et par forkerte termer i [3] og [4] i forbindelse med forklaringen af ledene i de nævnte formler og ledene i formlen, som anvendes i SHS. Jeg vil antage, at der blot er tale om tastefejl. For det første bliver h_n nævnt som kompressionsfaktoren, selv om det er n , som er kompressionsfaktoren og h_n , som er vægten af de komprimerede spektrogrammer. Derudover bliver P omtalt som kraftspektrummet (dvs. kvadratet på amplituden til hver frekvens), selv om det altså blot er spektrummet (dvs. amplituden til hver frekvens) der anvendes i SHS.

⁴Dette er en påstand, som kun er underbygget af egne erfaringer.

“black box”. Efter optræningen virker modellen muligvis godt på visse mængder data, men det kan være svært at finde ud af, hvorfor modellen virker godt på den aktuelle data, og muligvis dårligt på anden data. Desuden kan det i tilfælde af, at man finder en god model, være svært at gennemskue, hvordan man eventuelt kan optimere modellen.

3.2 Multiopløsningsspektrum

Dette afsnit beskriver ikke nogen metode, men derimod et værktøj, som kan være fordelagtigt at anvende, hvis man bruger en metode, som arbejder i frekvensdomænet.

Et problem ved at arbejde i frekvensdomænet er, at man både ønsker at have en god frekvensbestemmelse, samt en god tidsmæssig bestemmelse. Som beskrevet tidligere er det meget normalt at danne et spektrogram ved hjælp af STFT. STFT giver en bestemt opløsning på tidsaksen og på frekvensaksen. Vi ønsker generelt at have en høj opløsning på begge akser - dette kan dog ikke lade sig gøre, da en bedre frekvensbestemmelse kræver flere målinger i tiden.

Når man ser på frekvensspektrummet for et stykke musik, vil man opdage, at de dybe toner generelt ikke ændrer sig så meget, mens de høje toner ofte har en mere urolig karakter. Desuden er det primært i de høje frekvenser, at percussionlyde (dvs. slagtojslyde) gør sig gældende. Disse lyde kan forvirre lydbilledet, hvis man forsøger at fokusere på de frekvenser, som de øvrige instrumenter spiller. Desuden er lydene kortvarige, og markerer rytmiske figurer. Alt dette taler for, at man ønsker en høj tidsmæssig opløsning for de høje frekvenser.

Hvis man har en høj tidsmæssig opløsning opnår man, at det kun er korte tidsrum, som bliver ødelagt af percussionlydene, samt at man kan bestemme rytmen meget præcist. Til gengæld må man altså leve med, at frekvensopløsningen er dårlig. Dette er dog ikke nødvendigvis et stort problem, for hvis vi er interesseret i at bestemme frekvensen for en relativt dyb tone, vil overtonerne ofte kunne bidrage til en mere nøjagtig bestemmelsen af tones fundamentalfrekvens. Hvis tonens fundamentalfrekvens blot ændrer sig med nogle få Hz, lad os sige x Hz, vil den 10. overtones frekvens ændre sig med 11 gange x Hz. Man kan altså klare sig med en noget dårligere opløsning for de høje frekvenser og stadig bestemme en dyb tone forholdsvis nøjagtigt ved hjælp af de høje overtoner.

Ser vi igen på eksemplet med en relativt dyb tone, som ændrer sig nogle få Hz, hvilket betyder at overtonerne ændrer sig med mange Hz, er det faktisk en ulempe for bestemmelsen af overtonerne, hvis frekvensopløsningen er høj. For den høje frekvensopløsning følger med en dårlig tidsmæssig opløsning, og dermed er det et relativt langt tidsinterval, som de “gennemsnitlige” frekvenser bliver beregnet over, hvilket betyder, at overtonerne bliver bestemt meget unøjagtigt, selv om frekvensopløsningen er høj.

3.2.1 Metoder for generering af et multiopløsningsspektrogram

Med wavelettransformation er det muligt at have en multiopløsning, dvs. en opløsning, som favoriserer tidsbestemmelsen for lave frekvenser og favoriserer

frekvensopløsningen for høje frekvenser. Vi vil dog ikke se nærmere på denne mulighed, men i stedet for se på forslag, som anvender fouriertransformationen.

3.2.1.1 Den simple løsning

Den simple løsning går ud på at foretage en række fouriertransformationer, som anvender forskellige vinduesstørrelser. Man kan f.eks. anvende 3 forskellige vinduesstørrelser og så kombinere dem i et endeligt multiopløsningsspektrogram. Mulige kombinationsformer kommer vi ikke ind på her, men en enkel kombinationsform kan f.eks. gå ud på, at frekvensspektrummet opdeles i zoner på frekvensaksen. Hver zone kan da have sin egen vinduesstørrelse. Denne kombinationsform er anvendt til denne opgaves foreslåede metode for bestemmelse af melodien. Der opstår selvfølgelig nogle bratte overgange mellem hver zone, men det ser ikke ud til at være et stort problem.

Optimering af den simple løsning I 2006 deltog Karin Dressler for anden gang i MIREX (se afsnit 1.4), og denne gang anvendte den foreslåede algoritme et multiopløsningsspektrogram [9]. Her forklares i korte træk hvordan multiopløsningsspektrogrammet kan beregnes hurtigt:

Først antager vi, at vi genererer spektrogrammer ved hjælp af STFT, hvor der er anvendt et rektangulært vindue i tidsdomænet. Det vigtigste er, at fouriertransformationen består af en sum, som kan opstykkes i en række summer. Hver mindre sum svarer til fouriertransformationen med et smallere rektangulært vindue. Så hvis man blot har foretaget fouriertransformationen med den mindste vinduesstørrelse som ønskes, kan de øvrige spektrogrammer med større vinduesstørrelser, hurtig genereres ved en sum af de fouriertransformationer der allerede er beregnet.

Problemet er, at vi ikke ønske at anvende et rektangulært vindue, men derimod f.eks. et hanning-vindue. Heldigvis består den fouriertransformerede af et hanning-vindue kun af 3 led, som er forskellige fra 0. Det betyder, at foldningen i frekvensdomænet kan udføres ved ganske få multiplikationer og summeringer. Dette smarte "trick" er beskrevet i detaljer i [10].

3.2.1.2 Constant Q spectral transform

Denne transformering svarer sådan set til en avanceret udgave af den simple løsning, som lige er blevet præsenteret. Inspireret af den logaritmiske natur i den vestlige 12-tone-skala, præsenterede Judith C. Brown i 1991 [1] Constant Q Spectral Transform algoritmen.⁵ Q står for en kvalitetsfaktor, og angiver forholdet mellem en frekvens og bredden af frekvensopløsningen ved denne frekvens. Dette forhold er altså konstant, hvilket bevirker, at jo højere frekvens, des dårligere frekvensopløsning. Fouriertransformationen bliver så udregnet med forskellige vindueslængder, som svarer til de forskellige frekvenser.

Fordelen ved denne transformation er, at frekvensaksen bliver logaritmisk, hvilket betyder, at det er let at spore ændringer i et instruments tonehøjde. Hvis vi antager, at vi har et instrument, hvis overtonemønster (set ud fra styrken af overtonerne) ikke er forskelligt ved forskellige toner, så vil en ændring af tonehøjden svare til, at overtonemønsteret flytter sig lineært på frekvensaksen. Så hvis fundamentalfrekvensen flytter sig x trin på frekvensaksen, så flytter

⁵ Artiklen stammer faktisk fra 1988, men er først udgivet i 1991.

alle overtonerne sig også x trin. En forslag til en mere effektiv beregning af transformationen er foreslået af Judith Brown og Miller S. Puckette i 1992 [2].

3.3 Støjundertrykkelse

Som en form for pre-processing anvendes der ofte støjundertrykkelse - enten for at mindske mængden af data eller for at forbedre ydelsen af den metode, som finder fundamentalfrekvenser. I nogle metoder er støjundertrykkelsen endda beskrevet som en del af metoden. Dette er f.eks. tilfældet med SHS, hvor frekvenser, som ikke befinder sig i nærheden af en større lokal høj i frekvensbilledet, bliver sat til at have en amplitude på 0.

En anden og mere avanceret form for støjundertrykkelse eller måske snarere signalfremhævelse kan anvende en model af det menneskelige øre. Det kunne være tilfældet, hvis man oversætter frekvensernes amplituder til phon-værdier, og så arbejder på phon-skalaen. Ved at opfatte phon-værdierne som decibel-værdier, kan man oversætte tilbage til amplituder, således at det svarer til en filtrering af frekvenserne - en filtrering, der efterligner den menneskelige opfattelse af lydstyrker.

Støj kan også opfattes som de lyde, der genereres ved at et instruments krop svinger med og på den måde tilføjer nogle frekvenser, som blander sig med fundamentalfrekvensen og overtonerne. En anden form for støj, stammer fra det ekko, der opstår når en lyd reflekteres fra mur. I Paivas metode [6] anvendes en avanceret pre-processing, som skal forsøge at fjerne denne form for støj. Metoden er foreslået af Klapuri [17] og er baseret på Hermanskys RASTA metode fra 1993 [6].

3.4 Sporing

Mennesker sporer lyde i et lydbillede, hvilket gør os i stand til at spore glissando og toneintervaller [13]. Derfor er det oplagt på en eller anden måde at anvende tidsmæssigt adskilte tolkninger af lydbilledet for bestemmelsen af fundamentalfrekvensen til et bestemt tidspunkt.

Sporingen af melodien eller den fundamentalfrekvens, som melodien har, kan foretages på 2 måder. Enten kan man løbende udvælge én kandidat til en fundamentalfrekvens pr. tidsenhed og så generere melodiens forløb eller spor ud fra denne ene kandidat, eller også kan man udtage flere kandidater. Hvis man udtager flere kandidater, har man muligheden for at bestemme den endelige kandidat til hvert tidspunkt, enten alene på baggrund af kandidatudvælgelser på tidligere tidspunkter, eller også kan man inddrage information om kandidatudvælgelser på senere tidspunkter.

I Gotos PreFEst anvendes flere agenter, som arbejder på hver deres spor eller hypoteser for fundamentalfrekvensen. Til enhver tid bestemmes sporenes "pålidelighed" på baggrund af fundamentalfrekvenskandidater og sporenes optjente pålidelighed.[15] Dét spor, som til enhver tid har den største pålidelighed, bliver udvalgt som det melodibærende spor.

Et populært værktøj for sporing af melodien blandt flere kandidater er Hidden Markov Models (HMMs), som er en statistisk tilstandsmodel. Til enhver tid har en række tilstande (hypoteser) en statistisk fordeling, og den tilstand, som

har den største sandsynlighed bestemmer fundamentalfrekvensen. Informationer om fundamentalfrekvenskandidater findes og denne information anvendes i tilstandsmodellen til bestemmelse af nye sandsynligheder, og dermed muligvis et nyt bud på en fundamentalfrekvens.

Som det bliver foreslået af Gerhard [13], så kan en meget simpel sporingsmetode, som blot favoriserer kandidater eller hypoteser nær den sidst-fundne fundamentalfrekvens, forbedre effektiviteten. Dette er baggrunden for valget af den post-processing, som den foreslåede metode anvender (se afsnit 4.3).

Kapitel 4

Den foreslåede metode

I dette kapitel vil vi komme igennem 3 versioner af et forslag til en metode. Der er 2 årsager til, at de første 2 forslag er medtaget: For det første repræsenterer de netop den hypotese, som blev introduceret i projektbeskrivelsen til denne opgave; for det andet er de meget simple, og giver derfor et udgangspunkt for vurderingen af det sidste forslag.

I alle 3 versioner antages det, at det signal der skal analyseres, har en samplingsfrekvens på 44100 Hz. Der foretages en ned-sampling på 4 gange, således at samplingsfrekvensen sænkes til 11025 Hz, hvilket er tilstrækkeligt, da den vigtigste information antages at ligge under ca. 4000 Hz. Ved ned-samplingen er der taget højde for eventuelle problemer med aliasing - se afsnit 2.3.1.2.

4.1 Version 0a

Dette er det mest naive forslag.

Der genereres et spektrogram ved hjælp af en STFT af signalet. Den anvendte vinduesstørrelse er på 512, og hop-size er på $1/8$ af vinduets størrelse. Derudover anvendes zero padding, som er på 3 gange vinduets størrelse. Dette giver en kunstigt høj tidsopløsning på 0,0058 sekunder og en kunstigt høj frekvensopløsning på 5,38 Hz. I afsnit 4.2.1 forklares det indirekte, hvorfor det netop er disse værdier, som er blevet valgt.

I projektbeskrivelsen blev der opstillet en hypotese om, at melodien ofte kan findes som den fundamentalfrekvens, der har størst amplitude. Hvis vi nu blot betragter at alle frekvenser som kandidater, handler det altså blot om at udvælge den frekvens, som har størst amplitude. Så enkel er denne version.

4.2 Version 0b

Denne version adskiller sig stort set ikke fra den forrige.

Inspireret af Karin Dresslers succesfulde metode fra 2006, hvor der anvendes et multiopløsningsspektrogram, og på baggrund af de erfaringer, der blev gjort i afsnit 2.3.2, afsnit 2.4.1.3, samt i afsnit 3.2, er der blevet anvendt et multiopløsningsspektrogram i denne version. Dette er den eneste ændring i forhold til den forrige version. I næste afsnit forklares det, hvordan multiopløsningsspektrogrammet er stykket sammen.

frekvensområde	vægtning af vinduesstørrelse		
	256	512	1024
0 - 200			1
200 - 400		2/3	1/3
400 - 800		1/3	2/3
800 - 1600		1	
1600 - 5000	1/2	1/2	

Tabel 4.1. Her vises vægtene af de 3 vinduesstørrelser, som bruges i multiopløsnings-spektrogrammet.

4.2.1 Det anvendte multiopløsningsspektrogram

Ifølge analysen af figurerne i afsnit 2.3.2, skulle en vinduesstørrelse på 4096 være at foretrække op til omkring 300 Hz. Her arbejder vi med en samplingsfrekvens, som er 4 gange lavere, så vinduesstørrelsen skal være på 1024. Analysen viser også at den halve vinduesstørrelse, dvs. 512, er at foretrække omkring 1000 Hz. Endelig ser det ud som om endnu en halvering af vinduesstørrelsen, dvs. 256, er at foretrække over 2000 Hz. Der er foretaget et par yderligere analyser af et par korte sekvenser, og ud fra disse er multiopløsningsspektrogrammet lavet med de vægte, som kan ses i tabel 4.1.

For at de 3 spektrogrammer let kan kombineres, er det en fordel at de anvender den samme opløsning. I afsnit 5.1 forklares det, at den tidsmæssige opløsning skal være på 0,0058 sekunder. Så for at de 3 spektrogrammer opnår en kunstigt høj tidsmæssig opløsning, som svarer til kravene, er det nødvendigt at bruge overlap. De 0,0058 sekunder svarer til en hop-size på 64 samples. Så i tilfældet med en vindueslængde på 256 anvendes en hop-size på 1/4 af vinduets længde. Hop-size er 1/8 ved en vindueslængde på 512 og 1/16 ved en vindueslængde på 1024.

Den bedste frekvensopløsning fås ved vindueslængden på 1024. Men det er altid en god ide at anvende en zero padding på mindst vinduets længde, for at opnå en kunstigt højere frekvensopløsning. Derfor sættes zero padding til at være lig vinduets længde i tilfældet med vindueslængden på 1024. Dermed er kravet til frekvensopløsning 5,38 Hz. For at der opnås denne frekvensopløsning med en vinduesstørrelse på 512, skal der anvendes en zero padding på 3 gange vinduets størrelse; og i tilfældet med en vindueslængde på 256 skal der anvendes en zero padding på 7 gange vinduets størrelse.

4.3 Version 1

Dette er den endelige version. Den anvender multiopløsningsspektrogrammet fra version 0b (afsnit 4.2.1). I afsnit 4.3.1 vil vi se, hvordan kandidater til fundamentalfrekvenser bliver fundet ved hjælp af dette spektrogram. Her vil vi derfor først se på pre-processing og post-processing.

Pre-processing Der er ikke blevet foretaget pre-processing af signalet (se afsnit 3.1), da det er blevet vurderet, at den metode, som tager sig af kandidat-udvælgelsen, indirekte inddrager en form for pre-processing (se afsnit 4.3.1).

Post-processing I afsnit 3.4 nævntes nogle muligheder for sporing af melodien blandt kandidaterne til fundamentalfrekvenser. Dette kan betragtes som en post-processing (se afsnit 3.1). post-processing ved hjælp af HMMs eller mere avancerede statistiske metoder kan måske af nogle ses som noget af det vigtigste, fordi der gives mulighed for at betragte hver lille lydsekvens som en del i den større sammenhæng¹. I denne opgave er det primære fokusområde ikke at finde en god post-processing, så vi nøjes med en meget simpel post-processing.

Det er et problem, at den melodibærende frekvens ofte er noget urolig hvad angår størrelsen af amplituden. Dette er specielt et problem, hvis man har med en sanger at gøre. Problemet bliver værre, når der anvendes glissando eller vibrato i melodien, fordi den glidende frekvensændring medfører at spektrogrammet bliver utydeligt, og at energien i den frekvens, som melodien har, bliver spredt til omliggende frekvenser (se afsnit 2.3.2 og figur 2.7). I håb om at mindske dette problem er der indført en meget simpel post-processing, som i hvert tidsskridt foretager en vægtning af kandidaterne på baggrund af hvilken frekvens, der senest blev gættet på. Hvis der i tidsskridt x blev gættet på frekvens nr. f , så vil kandidaterne i tidsskridt $x + 1$ blive vægtet med en forhøjet gausskurve, som er centreret omkring frekvens nr. f .

Det endelige gæt på en melodibærende frekvens baserer sig på den kandidat, som efter multiplikationen med den forhøjede gausskurve, har den største kredit. For at få et bedre estimat af den melodibærende frekvens, foretages et vægtet gennemsnit af de omkringliggende frekvenser, og dette gennemsnit fastsætter det endelige gæt.

4.3.1 Inspiration fra den ændrede SHS

Til dét at finde fundamentalfrekvenskandidater er der blevet udviklet en metode, som er stærkt inspireret af den ændrede SHS, som blev præsenteret i afsnit 3.1.2.6. Årsagen til valget af SHS er dels, at den er meget enkel, dels er der en personlig årsag².

¹For at give en forklaring på, hvorfor det kan være en fordel at anskue den større sammenhæng fremstilles her et visuelt scenarium: Vi har et billede, som er taget fra en høj i et støvet tørt landskab. En smal utydelig grusvej slynger sig gennem landskabet. Hvis vi skal finde vejen på billedet, kan vi opdele billedet i tynde strimler og så for hver pixel i en strimmel bestemme kandidat-pixler til vejføringen. Hver enkelt strimmel ligner blot en række tilfældige farveværdier, eller hvis man er heldig, kan man måske identificere nogle småsten i hver strimmel. Derfor vil det være overordentligt vanskeligt blot ud fra en strimmel at bestemme vejens placering på strimlen. Men med post-processing kan man sammenholde de udtrukne informationer fra flere strimler, og muligvis finde et mønster, således at vejen kan identificeres.

²Allerede inden jeg stiftede bekendtskab med SHS og kun overfladisk havde læst om blandt andet Dresslers metode[8], Paivas metode[6], Gotos metode[15] og Poliners metode[22][23], tænkte jeg, at der da måtte findes en enklere metode. Da jeg vidste, at overtonerne i hvert fald i opera har stor betydning, forsøgte jeg mig med en fouriertransformation af et spektrogram. Jeg havde taget logaritmen til amplituden, så der var egentlig tale om cepstrum analysis, hvilket jeg først fandt ud af senere. Men allerede inden jeg læste om cepstrum analysis, var jeg i gang med at forsøge mig med andre metoder, fordi metoden var for afhængig af nogle klare overtoner - hvilket ikke var til stede i det Alanis Morissette nummer, som jeg primært havde testet på. Jeg konkluderede, at dét, som jeg havde brug for, var en metode, som virkede, hvis melodien kun havde en enkelt eller få meget kraftige frekvenser, og som også virkede, hvis melodien ikke havde nogen særlig kraftig frekvens, men derimod havde mange tydelige overtoner. Så metoden skulle på en eller anden måde inkludere en vægtet sum (eller multiplikation) af overtonernes amplituder. Da var det, at jeg stødte på SHS og tænkte, at her havde jeg en meget simpel metode, som netop havde de ønskede egenskaber.

For at simplificere SHS-metoden, anvendes hverken støjundertrykkelsen eller konverteringen til den logaritmiske frekvensskala. Til gengæld anvendes integral-tilføjelsen fra [3] - dvs. formel 3.8. Muligvis kunne der have været opnået bedre resultater, hvis den fulde SHS havde været anvendt, men integraltilføjelsen viste sig at være ganske nyttig. Integralet er dog også blevet simplificeret. Således er der ikke en Q -funktion for hver f_0 , men blot en enkelt Q -funktion. Q -funktionen kommer til at se således ud:

$$Q(f) = P(f) / \left(\frac{1}{2\rho(f)} \int_{f-\rho(f)}^{f+\rho(f)} P(w)dw \right) \quad (4.1)$$

, hvor

$$\rho(f) = b^{\log(f)} f \quad (4.2)$$

, hvor b er en konstant, som sættes til 0.80. På denne måde udregnes integralet over et stigende interval, når f stiger; men intervallet stiger langsommere og langsommere. På denne måde kommer den simplificerede Q -funktion til at minde om en form for gennemsnitlig Q_{f_0} -funktion.

Effekten af at anvende integralet er, at de generelt kraftige lave frekvenser ikke dominerer de høje frekvenser. Problemet med integralet er dog, at det nærmest kommer til at undertrykke informationen i de lave frekvenser. For at dæmme op for dette problem multipliceres med P^a . Således kan den endelige formel for udregningen af H skrives som:

$$H(f_0) = \sum_{n=1}^N h_n Q(nf_0) \quad (4.3)$$

, hvor

$$Q(f) = P(f)^a / \left(\frac{1}{2\rho(f)} \int_{f-\rho(f)}^{f+\rho(f)} P(w)dw \right) \quad (4.4)$$

I implementeringen er $h_n = 0,80$, $a = 1,2$ og $N = 12$. Desuden beregnes H kun for frekvenser mellem 100 Hz og 1300 Hz, da dette frekvensområde dækker de frekvenser, som en mandlig og kvindelig sanger normalt synger indenfor, og det antages at melodien ligger inden for et sangbart område.

Kapitel 5

Generelle Resultater

I dette afsnit præsenteres resultaterne af en afprøvning, som overholder de testkrav, der blev anvendt ved MIREX-2006. Der præsenteres også resultater ved brug af den evalueringsmetode, som blev anvendt ved ISMIR-2004, da de foreslåede metoder ved ISMIR-2004 blev evalueret på baggrund af de samme testfiler, som blev anvendt ved MIREX-2006.

Køretiden har ikke være i fokus i denne opgave, men en kørsel i Matlab version 7.4 på en Pentium 4 3GHz tog 6 minutter og 22 sekunder for alle testfilerne; og da test-filerne samlet set varer 6 minutter, kører implementeringen i næsten realtid på det anvendte system.¹

5.1 Testmetode

Her beskrives de testmetoder, som blev anvendt ved ISMIR-2004, MIREX-2005 og MIREX-2006, og som også anvendes her ved afprøvningen af implementeringen.

Deltagerne til MIREX-2005 og MIREX-2006 fik stillet nogle træningssæt (dvs. nogle samlinger af lydfiler, med tilhørende angivelser af melodiens frekvens) til rådighed. Af disse træningssæt, er træningssættet, som hedder "FullSet.zip"², blevet anvendt. Dette træningssæt er blevet valgt, fordi resultaterne for deltagerne i MIREX-2006 blandt andet er baseret på dette træningssæt. Derudover blev træningssættet også anvendt ved ISMIR-2004, hvor deltagerne anvendte halvdelen af sættet under udarbejdelse af metoder, og den anden halvdel blev anvendt ved afprøvningen.³ Træningssættet, som består af 20 korte melodistumper, har en tidsmæssig opløsning på 0,0058 sekunder, så det er denne opløsning, der er blevet valgt i implementeringen af den foreslåede metode (se afsnit 4.2.1).

5.1.1 MIREX-2006

En mere nøjagtig angivelse af, hvordan der udregnes en score, er ikke umiddelbart at finde på hjemmesiden for MIREX-2006, men de anvendte beregninger er at finde under hjemmesiden for MIREX-2005⁴. MIREX-konkurrencerne handlede ikke blot om at finde melodien i et stykke, men også om at bestemme, hvornår der er tale om en melodi. Dette er dog ikke noget problem, da de 2

¹Den anslåede tid inkluderer tiden for opstarten og afslutningen af Matlab.

²<http://www.iaa.upf.es/mtg/ismir2004/contest/melodyContest/FullSet.zip>

³http://ismir2004.ismir.net/melody_contest/results.html

⁴<http://www.music-ir.org/evaluation/mirex-results/audio-melody/index.html>

delopgaver kan betragtes som adskilte. Referenceresultatet indeholder angivelsen af melodiens frekvens til hvert tidsskridt. Hvis der ikke er nogen melodi, angives dette med frekvensen 0 Hz. Hvis deltagerne ønskede at give et bud på, hvornår der er en melodi, kunne de angive dette med positive og negative frekvenser.

Da denne opgave ikke tager sig af, om der overhovedet er nogen melodi, koncentrerer vi os om, hvordan der bestemmes en score for budet på en frekvens. Hvis der i referenceresultatet ikke er nogen melodi, ignoreres budet på en frekvens. Hvis der i referenceresultatet er angivet en frekvens for melodien, godtages bud, som ligger op til $1/2$ trin⁵ fra referenceresultatet. Det betyder, at bud, som ligger inden for den accepterede zone, har en fejl på 0, mens bud, som ligger uden for den accepterede zone, har en fejl på 1.

5.1.2 ISMIR-2004

Ved ISMIR-2004 var der ikke den samme klare adskillelse af de 2 delopgaver: bestemmelse af hvornår der er melodi, og bestemmelse af melodiens frekvens. Hvis man ønskede at angive, at der ikke er nogen melodi, skulle man angive dette med en frekvens på 0 Hz. Det betyder, at deltagerne ikke altid har angivet et bud på en frekvens, selv om der i referenceresultatet er angivet en frekvens. Budene fra ISMIR-2004 kan altså ikke helt sammenlignes med budene fra MIREX-2006, fordi der ikke er den samme mulighed for, at komme med et gæt på frekvensen til ethvert tidsskridt.

Udregningen af en score var også anderledes ved ISMIR-2004. I stedet for at have en klar adskillelse mellem et korrekt gæt på en tone og et ukorrekt gæt, anvendtes der ved ISMIR-2004 en score, som også er afhængig af, hvor god frekvensbestemmelsen er inden for en accepteret zone. En frekvens, som ligger mere end 1 trin fra referencefrekvensen, har en fejl på 1, mens en frekvens, som ligger under 1 trin fra referencefrekvensen, har en fejl, som svarer til afstanden målt i trin.

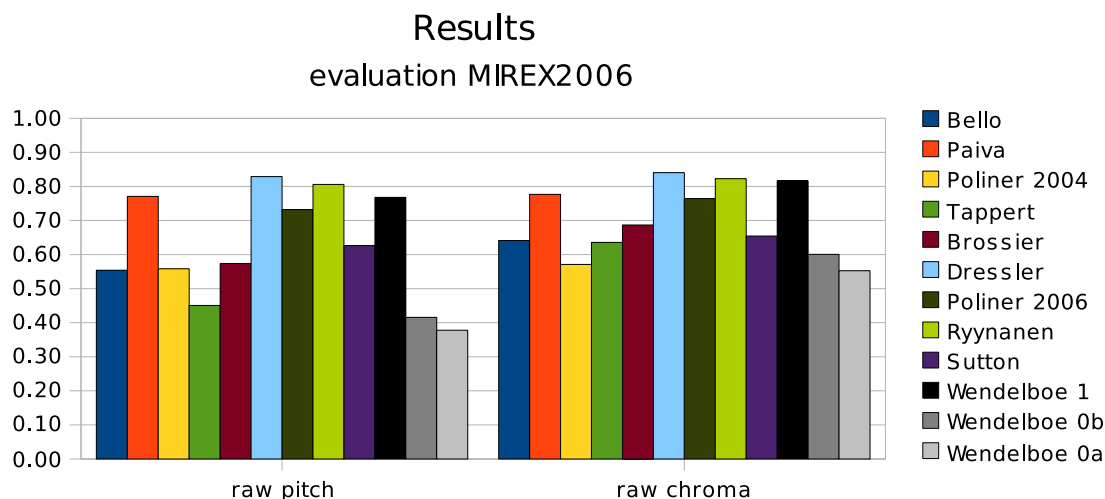
5.1.3 Sammenligning af resultater

Som nævnt tidligere var resultatet ved MIREX-2006 blandt andet baseret på det anvendte træningssæt. En nøjagtig resultatopgørelse over hver enkelt deltagende metode kan findes på hjemmesiden for MIREX-2006⁶, hvor det er muligt at downloade resultaterne som dataark. Resultaterne for de deltagende metoder ved ISMIR-2004 kan findes på hjemmesiden for ISMIR-2004⁷, hvor det også er muligt at downloade filer, som beskriver hver enkelt metodes bud på frekvenser. Desuden kan man finde kildekoden til de funktioner, der anvendes til evalueringen af de enkelte metoder. Bemærk at der på den nævnte hjemmeside angives resultater for "Option 1" og "Option 2". I "Option 1" og "Option 2" opfattes det som en fejl, hvis der er blevet angivet en melodi, men denne ikke er til stede i referenceresultatet. I resultaterne som kan findes i afsnit 5.3 tages de tidsskridt, som ikke indeholder en melodi i referenceresultatet ikke med i betragtning.

⁵Et trin (se afsnit 2.2) er dét, som man kalder for en halv tone, og der er 12 halve toner i en oktav.

⁶http://www.music-ir.org/mirex/2006/index.php/Audio_Melody_Extraction_Results

⁷http://ismir2004.ismir.net/melody_contest/results.html



Figur 5.1. Her vises det samlede resultat målt med metoden fra MIREX-2006. Version 1 ligger i top 3 både hvad angår raw pitch og raw chroma. Version 0b klarer sig bedre end version 0a, og er næsten på højde med Suttons metode når det gælder raw chroma.

5.1.4 Raw pitch og raw chroma

Der anvendes 2 mål ved evalueringen af metoderne. Det ene mål, "raw pitch", angiver, hvor godt gættet på en melodi svarer til referenceresultatet. Med metoden fra MIREX-2006 angiver dette mål altså, hvor stor en andel af tidsskridtne der indeholder et korrekt gæt. Det andet mål, "raw chroma", svarer til "raw pitch", men der accepteres oktavfejl. Mennesker kan finde på at synge i forskellige oktaver, uden at det generer oplevelsen (dette er stort set altid tilfældet, hvis en mand og en kvinde synger sammen). Man kan derfor sige, at "raw chroma" måske bedre svarer til dét, som vi mennesker oplever. Et argument for at "raw chroma" kan være et godt mål er, at hvis man har en referencemelodi, og skal sammenligne med en melodi, som synges af et menneske, så kan mennesket godt finde på at skifte oktav undervejs i melodien; og dette skal helst ikke have indflydelse på resultatet.

5.2 Afprøvning med metoden fra MIREX-2006

Her skal vi se på resultaterne, når testmetoden fra MIREX-2006 anvendes. Vi begynder med at se på figur 5.1.

De første 4 deltagere stammer fra ISMIR-2004, og kan derfor ikke helt sammenlignes med de øvrige deltagere. Blandt disse 4 skiller Paiva sig ud. Paivas metode [6] indeholde en kompliceret model af den menneskelige opfattelse af lyde, hvor filtreringen foregår ved filtre, som skal efterligne dén filtrering, der foregår i cochlea⁸.

De sidste 3 deltagere repræsenterer den foreslåede metode i version 1, version 0b og version 0a. Som man nok kunne forvente af version 0a og 0b, ligger de i bunden. Man kan også se, at anvendelsen af multiopløsningsspektrogrammet giver en klar forbedring; og forbedringen er faktisk tydeligere, end hvad der kan ses på figuren. Resultaterne for de enkelte melodier viser nemlig, at

⁸Se afsnit 2.4.

version 0b er bedre end version 0a i alle tilfælde med undtagelse af 2, hvor de er lige gode⁹. Bemærk at raw chroma for version 0b næsten er på højde med raw chroma for Suttons metode.

Version 1 er markant bedre end version 0b, specielt når det handler om raw pitch. Faktisk ligger version 1 i top 3 eller 4. I modsætning til de 3 andre gode metoder (Paiva, Dressler og Ryynanen) er der en tydelig forskel på raw pitch og raw chroma.

Sammenligning med andre afprøvninger Som nævnt kan man ikke helt sammenligne med metoderne fra ISMIR-2004, men kan man muligvis forvente at metoderne fra ISMIR-2004, kunne have scoret lidt højere, hvis de på samme måde som ved MIREX-2006, havde kunnet bestemme, hvornår der er en melodi til stede.

Ved MIREX-2006 blev der ud over det anvendte træningssæt også anvendt et testsæt, som er hemmeligt - også for deltagerne. Dette testsæt blev også anvendt ved MIREX-2005, hvor nogle af deltagerne fra MIREX-2006 også deltog. Hvis man ser på resultaterne fra MIREX-2006¹⁰, vil man opdage, at resultaterne fra det hemmelige testsæt og det træningssæt, som anvendes her, ligger på stort set samme niveau. Undlader vi at tage Brossier's metode med i betragtning, så ligger resultaterne for det hemmelige testsæt omkring 5 til 10 %-point lavere set i forhold til træningssættet. (Dresslers metode svinger dog ikke helt så meget som de øvrige metoder.)

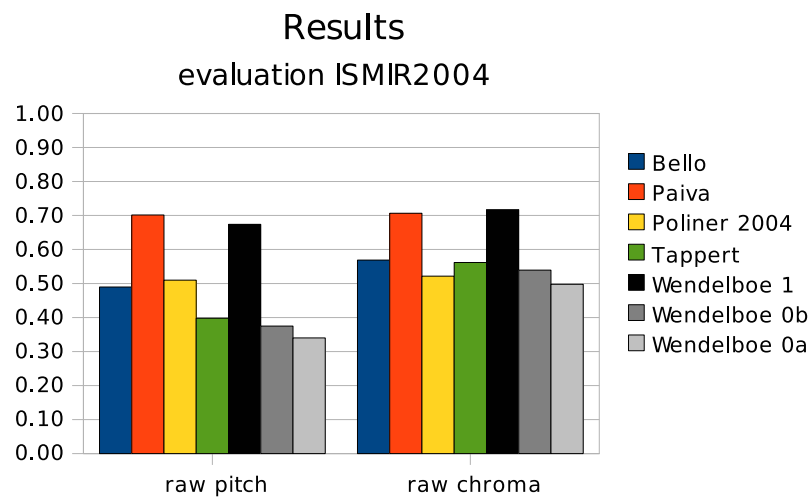
Ved MIREX-2005 er der kun offentliggjort resultater for det hemmelige testsæt. En sammenligning med resultatet fra MIREX-2006 viser, at Dresslers metode er forbedret med omkring 10 %-point, Ryynanens metode er stort set ikke forbedret, og Poliners metode er faktisk blevet lidt dårligere. Desuden kan man se, at Gotos kendte PreFest metode var næsten lige så god som Dresslers, Ryynanens og Poliners. Hvad der også er interessant er, at Paiva også deltog i MIREX-2005, og her havde en score på omkring 65 %, hvilket altså er ca. 5 %-point lavere end de andre omtalte deltagere. Hvis vi antager, at Paivas metode lige som metoderne fra MIREX-2006 kunne score omkring 5-10 %-point højere på træningssættet, giver det en forventelig score på omkring 75 %. Når vi så ser på resultatet af Paivas metode, hvor evalueringen fra MIREX-2006 anvendes (se figur 5.1), kan det ses, at Paivas metode scorer lidt over de 75 %. Så den angivne score for Paivas metode er måske i virkeligheden meget sammenlignelig med resultaterne for de metoder, som deltog i MIREX-2006.

5.3 Afprøvning med metoden fra ISMIR-2004

På figur 5.2 er vist resultaterne ved brug af afprøvningsmetoden fra ISMIR-2004. Som nævnt tidligere kan disse resultater ikke helt sammenlignes, da det også skulle bestemmes, om der er en melodi til stede. Men med udgangspunkt i de sidste konkluderende bemærkninger i forrige afsnit, er det måske alligevel muligt at foretage en svag sammenligning med Paivas metode. Det ses at Paivas metode og version 1 ligger på samme niveau.

⁹Disse resultater er dog ikke vist her i opgaven.

¹⁰Se http://www.music-ir.org/mirex/2006/index.php/Audio_Melody_Extraction_Results.



Figur 5.2. Her vises det samlede resultat målt med metoden fra ISMIR-2004. Version 1 ligger i top sammen med Paivas metode.

Kapitel 6

Resultatanalyse og forslag til forbedringer

Her vil vi se lidt nærmere på de resultater, som fremkom ved evalueringen med metoden fra MIREX-2006. I tabel 6.1 vises resultatet for raw chroma opgjort for hver enkelt melodi.

Version 0b Lad os først se på den sidste søjle, som viser resultater for version 0b. Ved de første 4 melodier er der opnået et godt resultat. Resultatet for “daisy3” er markeret med grøn, da det er det ene af de 3 tilfælde, hvor version 0b har opnået højest score eller lige så høj score som de andre, der har opnået den højeste score. Daisy3 er altså et eksempel på en melodi hvor fundamentalfrekvensen er den kraftigste frekvens, og daisy3 er derfor et pragteksempel, som underbygger hypotesen fra problemformuleringen.

For de 4 jazzmelodier er der opnået svingende resultater, og i 2 af tilfældene er version 0b den dårligste metode, hvilket er markeret med lyserødt.

Resultaterne for de 4 midi-melodier viser både styrkerne og svaghederne ved den simple version 0b. Så længe melodiens fundamentalfrekvens er den kraftigste, sejrer metoden, men når fundamentalfrekvensen er svag scorer metoden meget lavt. Alligevel er det kun i det ene tilfælde, at metoden har scoret lavest af alle.

I de 4 operamelodier scorer metoden, set i forhold til de andre metoder, ganske godt.

Popmelodierne er åbenbart ikke egnede for denne metode, for her har den scoret lavest af alle metoder.

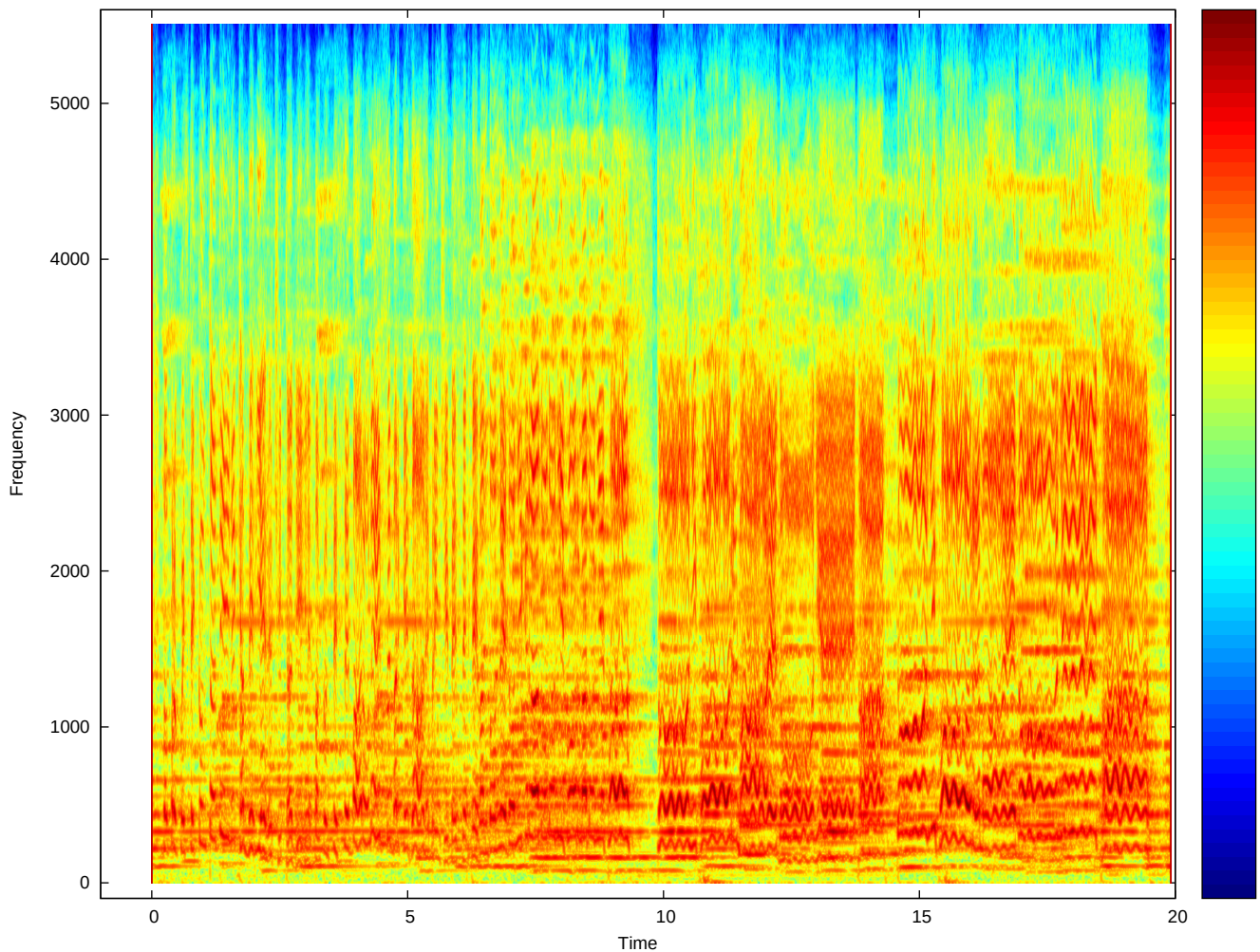
Version 1 Lad os nu fokusere på den næstsidste søjle og dermed på resultaterne for version 1. Lad os først sammenligne med resultaterne for version 0b. Der er 2 tilfælde, hvor version 1 har fået lavere score end version 0b - det drejer sig om “midi2” og “midi3”. Dette er helt klart et problem ved metoden. Specielt ved midi3 er der et problem, for i dette tilfælde har version 1 fået den laveste score af alle.

De 4 popmelodier udgør også et problem for metoden, for her har metoden generelt scoret lavt - i “pop3” og “pop4” har metoden faktisk scoret lavest.

I daisy- og jazz-melodierne har metoden generelt scoret forholdsvis højt. Opera-melodierne ser ud til at have været de mest vanskelige. Specielt “opera_fem2” og “opera_male3” har givet store problemer. Disse melodier ser dog

	B	D	P	R	S	W1	W0b
daisy1	0.92	0.97	0.94	0.96	0.91	0.94	0.94
daisy2	0.92	0.96	0.93	0.96	0.86	0.91	0.90
daisy3	0.83	0.97	0.96	0.96	0.71	0.97	0.97
daisy4	0.76	0.97	0.98	0.97	0.94	0.97	0.81
jazz1	0.79	0.86	0.79	0.84	0.56	0.88	0.76
jazz2	0.77	0.82	0.74	0.83	0.72	0.83	0.67
jazz3	0.82	0.91	0.88	0.89	0.54	0.94	0.72
jazz4	0.88	0.90	0.86	0.92	0.58	0.92	0.44
mid1	0.46	0.84	0.83	0.72	0.13	0.86	0.24
mid2	0.85	0.95	0.93	0.91	0.95	0.93	0.95
mid3	0.73	0.85	0.78	0.80	0.84	0.72	0.87
mid4	0.51	0.93	0.64	0.69	0.33	0.92	0.24
opera_fem2	0.36	0.57	0.46	0.53	0.38	0.62	0.56
opera_fem4	0.45	0.69	0.42	0.68	0.50	0.71	0.68
opera_male3	0.37	0.53	0.47	0.59	0.41	0.55	0.43
opera_male5	0.56	0.76	0.64	0.82	0.55	0.78	0.71
pop1	0.50	0.78	0.71	0.78	0.75	0.70	0.31
pop2	0.61	0.84	0.72	0.84	0.80	0.69	0.51
pop3	0.81	0.81	0.81	0.86	0.82	0.73	0.09
pop4	0.86	0.87	0.80	0.92	0.81	0.75	0.22

Tabel 6.1. Her vises raw chroma resultaterne for hver melodi. B = Brossier, D = Dressler, P = Poliner 2006, R = Ryyanen, S = Sutton, W1 = Wendelboe 1, W0b = Wendelboe 0b. Med farven rød er markeret den metode, som virkede dårligst for hver melodi - her er Wendelboe 0b ikke taget med i betragtning. I kolonnen med Wendelboe 0b markerer farven lyserød, at metoden var den dårligste, mens farven grøn markerer, at det var den bedste metode.



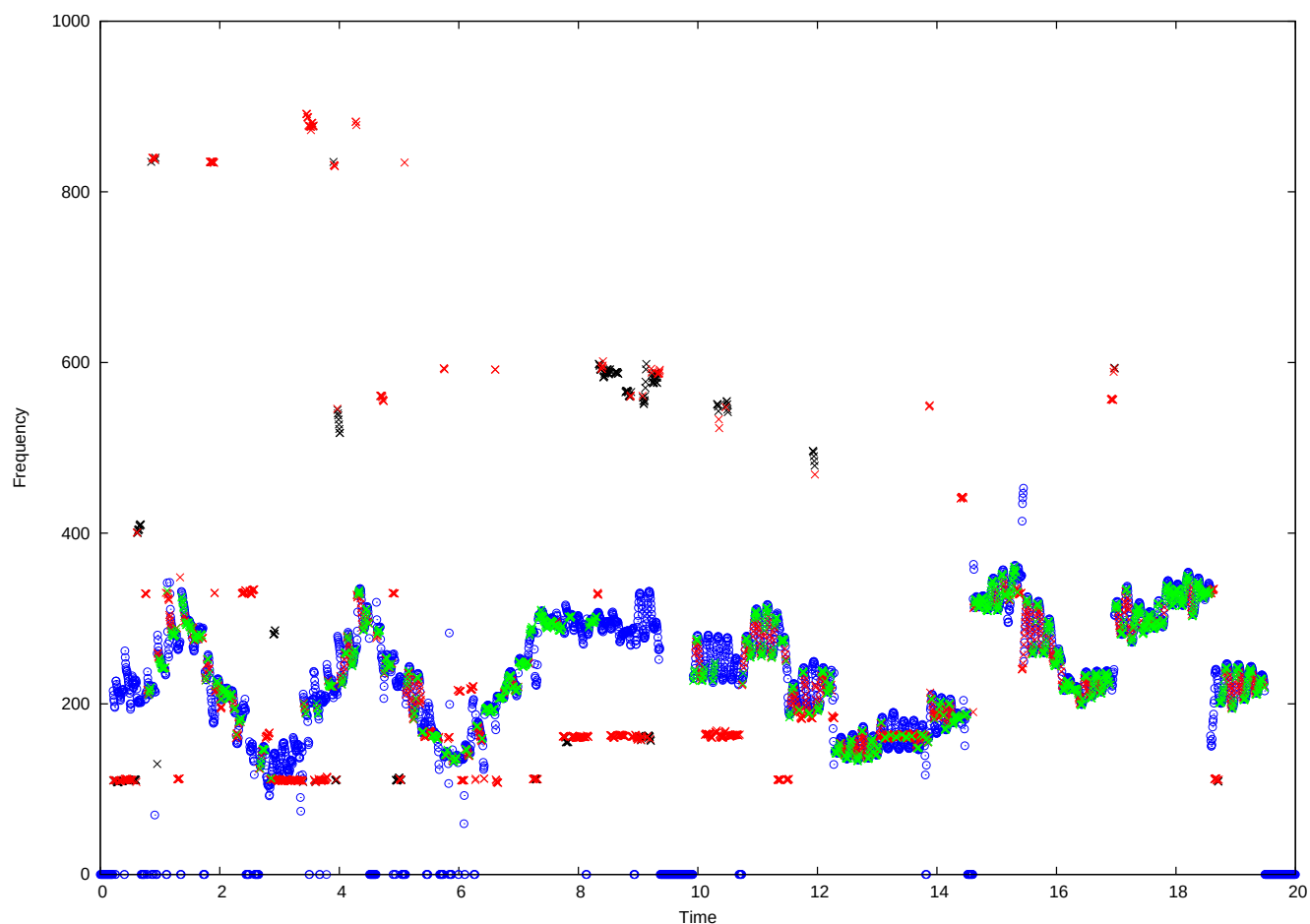
Figur 6.1. Her vises spektrogrammet for opera_male3. I de første 10 sekunder er melodien ikke helt tydelig. Dog kan man visuelt godt identificere fundamentalfrekvensen og den første overtone, da de bevæger sig op og ned omkring henholdsvis 200 og 400 Hz. Der er dog også nogle frekvenser, som konstant er tydelige. Disse frekvenser må stamme fra akkompagnerende instrumenter. I de sidste 10 sekunder kan man også se akkompagnerende instrumenter, men her er melodien mere tydelig. Man kan også se, at operasangeren har en formant omkring 3000 Hz samt anvender et stort vibrato.

ud til at have været problematiske for alle metoder. Faktisk har version 1 scoret højest på opera_fem2 og næsthøjest på opera_male3.

6.1 Nærmere analyse af 2 melodier samt forslag til forbedringer

I forrige afsnit blev et par melodier omtalt som problematiske. Her vil vi se nærmere på 2 af de problematiske melodier, for at undersøge hvad der er problemet.

opera_male3 Først ser vi på opera_male3, som er den melodi version 1 har haft sværest ved. På figur 6.1 er vist spektrogrammet for opera_male3, og på figur 6.2 er vist melodien forløb, samt version 1's gæt på en melodi.



Figur 6.2. Her vises melodien for opera_male3 samt hvilke frekvenser, der er blevet gættet på. Melodien er vist med de blå markeringer. Hvis gættet er korrekt, er dette markeret med grøn, mens et ukorrekt gæt er markeret med rød. Hvis et gæt er korrekt, når der anvendes chroma, er det markeret med sort.

Som det kan ses på figur 6.2, er det største problem, at det ind i mellem er akkompagnerende instrumenter, der bliver identificeret. Dette er specielt et problem i de første 10 sekunder. Muligvis kan melodien i de første 10 sekunder bedre identificeres, hvis overtonerne bliver vægtet højere. Da melodien i forhold til akkompagnementet ser ud til at have en meget mere urolig karakter, kunne man måske også favorisere bevægelser i frekvensmønsteret. Dette strider med den post-processing der anvendes, da denne netop favoriserer frekvenser, som ikke bevæger sig.

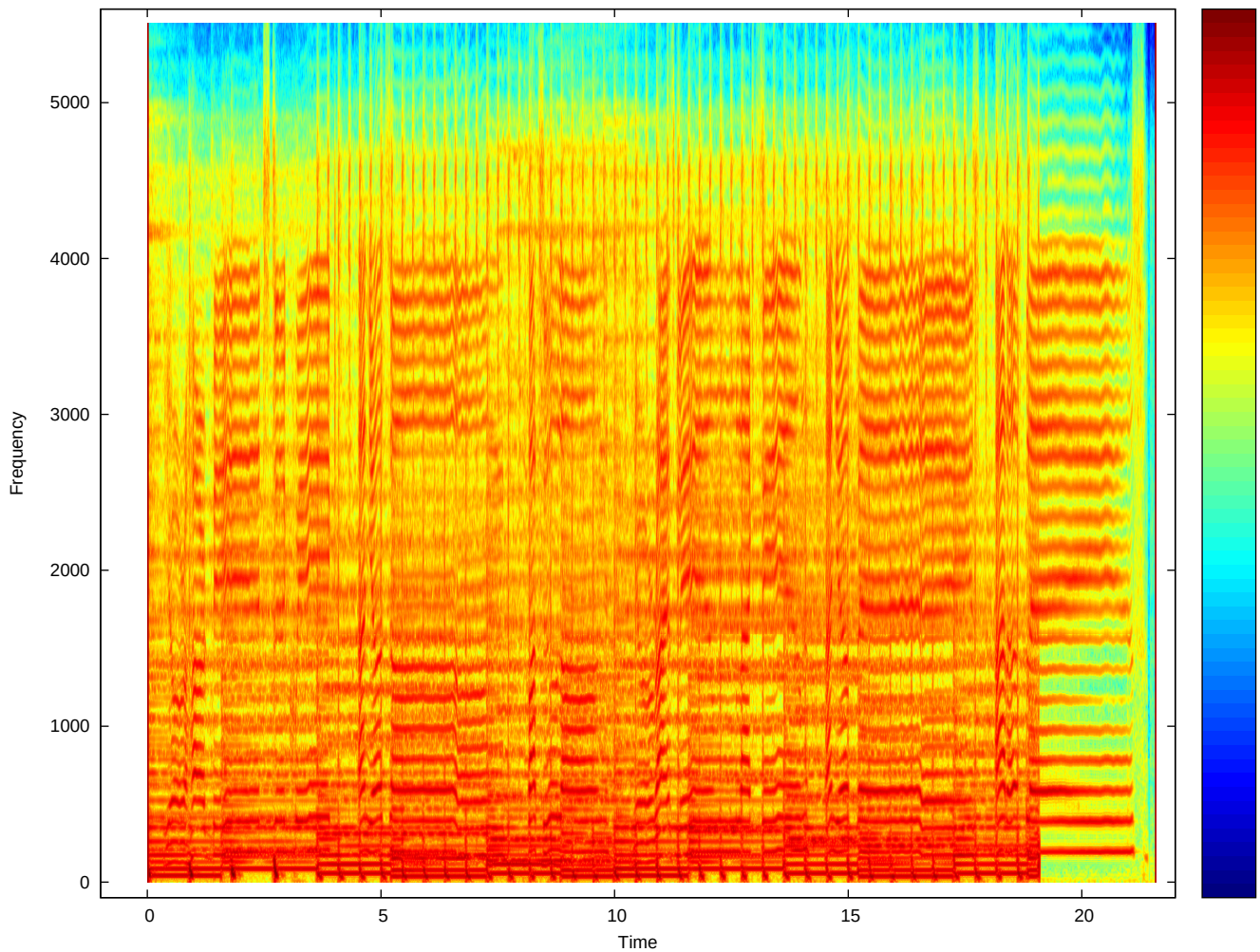
pop4 Vi vil også se nærmere på pop4. Spektrogrammet er vist på figur 6.3. Det ses, at melodiens overtoner er ganske tydelige i dette tilfælde, så man kunne måske forvente, at der ikke burde være de store problemer med at bestemme fundamentalfrekvensen. Af resultatet, som kan ses på figur 6.4, kan man da også se, at der er blevet gættet rigtig de fleste gange. Nogle gange er der gættet på, at den første overtone er fundamentalfrekvensen, hvilket ikke nødvendigvis er noget problem, hvis man kun er interesseret i raw chroma. Det, som har været det største problem, ser ud til at være, at der også er tilfælde, hvor der er blevet gættet på den 2. overtone. Dette er tilfældet omkring 15 og omkring 18 sekunder inde i melodien, hvor der er blevet gættet på en frekvens, som ligger omkring 600 Hz, mens den rigtige frekvens ligger omkring 200 Hz. Ser man igen på figur 6.3, kan man se, at ved disse tidspunkter er der netop en kraftig frekvens omkring 600 Hz. Ved udregningen af vægte for gæt på fundamentalfrekvenser bliver selve fundamentalfrekvensen vægtet mere end de formodede overtoner. Det betyder, at når der gættes på, at fundamentalfrekvensen er på 200 Hz, så tæller den kraftige frekvens ved 600 Hz med, men den bliver ikke vægtet så tungt. Omvendt bliver netop denne frekvens vægtet tungt, når der gættes på, at fundamentalfrekvensen er 600 Hz. Dette viser en klar svaghed ved metoden.

Lad os nu se på figur 6.5, som viser vægtningen af hvert gæt på en fundamentalfrekvens. Hvis vi fokuserer på tilfældet 15 sekunder inde i melodien, kan vi se, at der er 4 frekvenser, som har fået større vægte end de andre. Det drejer sig om 200, 300, 400 og 600 Hz. Desuden er det 200 og 600 Hz, som har fået størst vægtning blandt disse frekvenser. Så den rigtige frekvens er iblandt de frekvenser, som har fået størst vægtning, og den er endda i top 2.

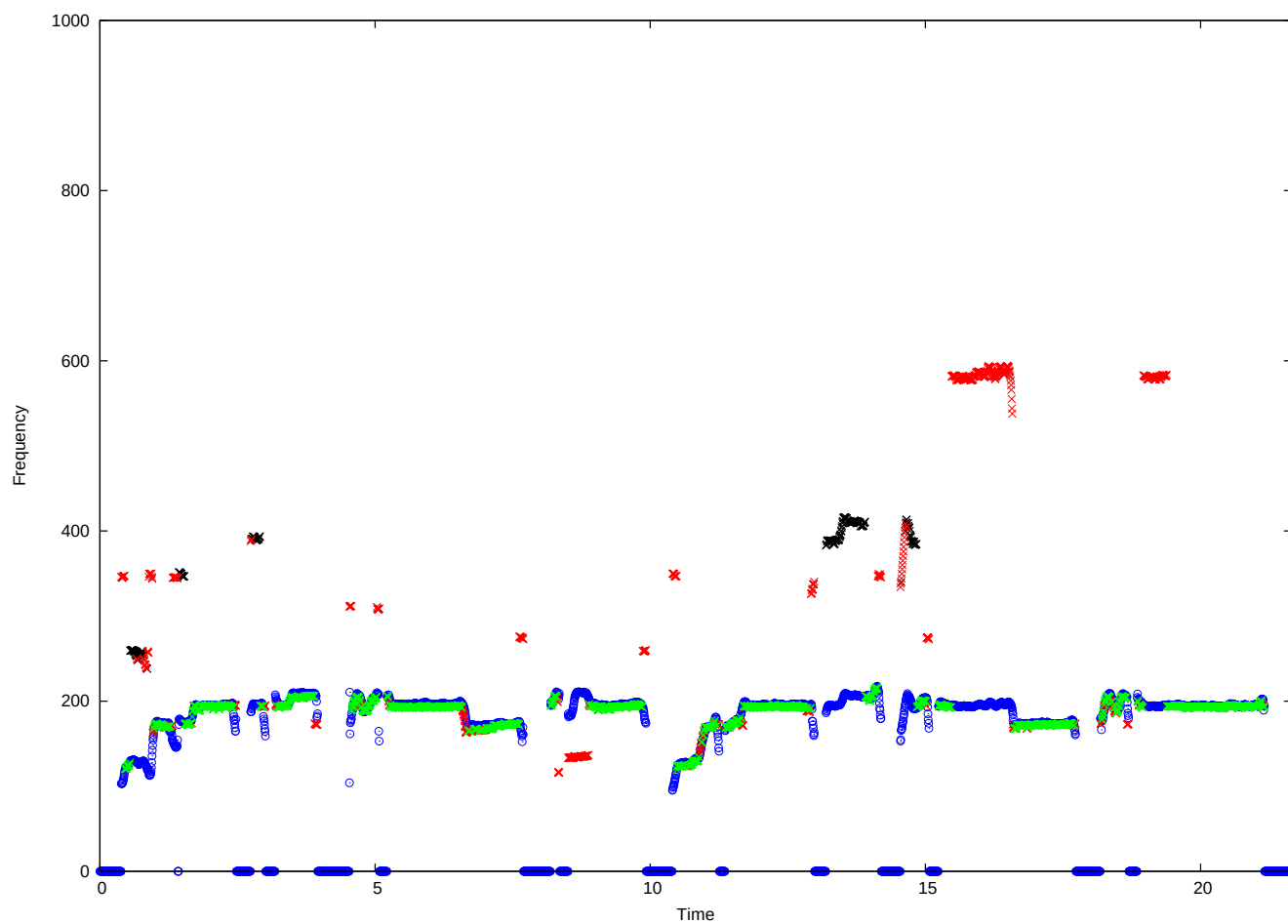
Umiddelbart er der 2 ændringer, som måske kan få metoden til at vælge 200 Hz som fundamentalfrekvensen: Enten kan man ændre metodens parametre, således at overtonerne vægter tungere, for på den måde at give større vægt til 200 Hz, set i forhold til 600 Hz; eller også kan man forbedre den meget simple post-processing. Man kunne f.eks. anvende HMMs og bruge de 4 nævnte frekvenser som kandidater. Ved hjælp af HMMs ville det sikkert kunne lade sig gøre at undgå problemet med, at melodien tilsyneladende pludselig hopper 1,5 oktaver - man kunne nemlig vægte et sådant meget usædvanligt hop meget lavt.

6.2 Flere idéer til forbedringer

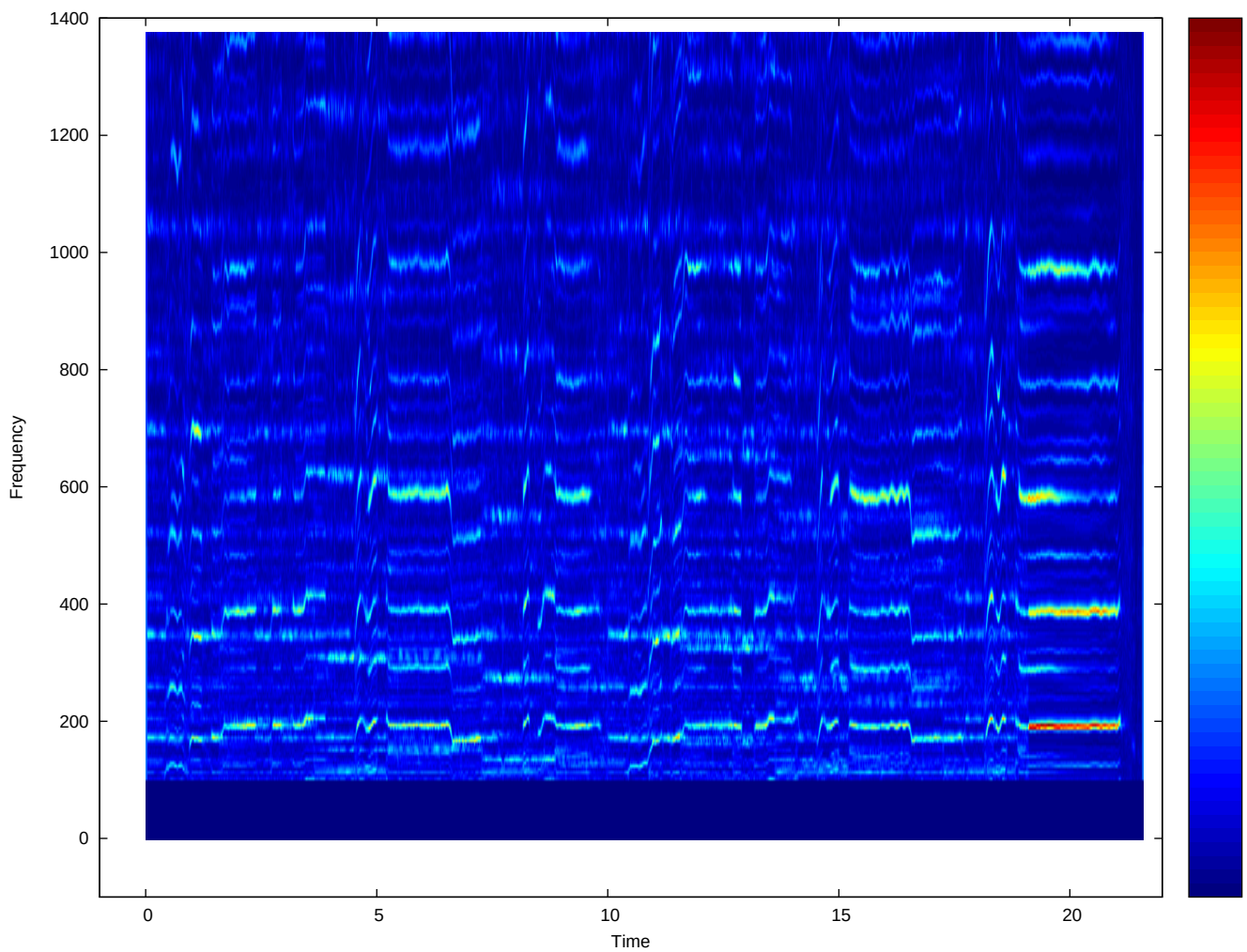
Som det blev nævnt i forrige afsnit under analysen af resultatet for pop4, kunne det måske være en fordel, at vægte overtonerne højere. Men meget ofte findes melodiens tone også i akkompagnementet, og i akkompagnementet kan den sagtens blive spillet en oktav eller 2 dybere. Det tæller ikke negativt, at nogle formodede overtoner er meget kraftige, mens andre måske slet ikke er til stede.



Figur 6.3. Her vises spektrogrammet for pop4. Melodiens overtoner er ganske tydelige her. Hvis man fokuserer på tidspunktet omkring 15 sekunder inde i melodien, kan man se, at melodiens 2. overtone, som ligger ved 600 Hz, er meget kraftig, hvilket er årsagen til, at der på dette tidspunkt er blevet gættet på en forkert fundamentalfrekvens.



Figur 6.4. Her vises melodien for pop4 samt hvilke frekvenser, der er blevet gættet på. Melodien er vist med de blå markeringer. Hvis gættet er korrekt, er dette markeret med grøn, mens et ukorrekt gæt er markeret med rød. Hvis et gæt er korrekt, når der anvendes chroma, er det markeret med sort.



Figur 6.5. Her vises vægtene af de forskellige gæt på fundamentalfrekvenser for pop4. Det kan ses, at det rigtige gæt næsten altid er blandt de frekvenser, som har fået størst vægt, så en forbedret post-processing kunne måske løse problemet med, at den 2. over-tone nogle gange bliver valgt.

Så der er en risiko for, at en større vægtning af overtonerne bevirker, at der oftere vil blive gættet på en tone, som er for dyb, og som findes i akkompagnementet. For at mindske dette problem kunne man indføre en form for negativ vægtning af manglende eller meget svage overtoner.

Nu vil vi se på nogle mere drastiske ændringer af metoden. Noget af det mest oplagte er måske, at forsøge at implementere den fulde SHS i stedet for den simplificerede og modificerede udgave, som er blevet præsenteret her. For det første kan det være interessant at se, hvordan den fulde SHS klarer sig i polyfon musik; for det andet er der blevet rapporteret om gode erfaringer med SHS i forbindelse med de 4 pop-melodier fra træningssættet [4]. I [4] opnås der yderligere gode resultater ved brug af en post-processing, som anvender information om den formodede klang for hvert gæt på en fundamentalfrekvens. Klangens beskrives som styrkeforholdet mellem overtonerne. En kandidat til en fundamentalfrekvens vil blive givet større vægt, hvis den har en klang, der minder om den klang, der sidst blev fundet. Introduktionen af et mål for klangen vil sikkert kunne bidrage positivt, da det måske vil gøre det muligt at adskille forskellige instrumenter, hvorved det kan afgøres, hvilke fundamentalfrekvenser der hører til hvilke instrumenter. Hvis man dernæst kan give et kvalificeret bud på, hvilket instrument der har melodien, kan alle fundamentalfrekvenser, som stammer fra det pågældende instrument udvælges, og man vil sikkert få en bedre identifikation af melodien. Dog er der den risiko, at man vælger et forkert instrument, hvilket bevirker, at metoden slet ikke finder de rigtige frekvenser.

Med udgangspunkt i den menneskelige opfattelse af lyd, kunne man måske ændre metoden, så den virker på en phon-skala (se afsnit 2.4.1.2). Dette vil dog muligvis slet ikke kunne lade sig gøre med den gældende metode, fordi den er baseret på en vægtning af overtonerne; og denne vægtning giver måske ikke længere så meget mening, hvis man anvender phon-skalaen, da denne skala specielt fremhæver frekvenserne fra 2000 til 4000 Hz, og især undertrykker de lave frekvenser. Hvis man forsøger sig med en phon-skala, skal man desuden finde ud af, om det er godt at anvende denne skala direkte, eller om der skal foretages en form for tilbagekonvertering, så det er filtrerede amplituder der anvendes af metoden.

I tilfældet med version 0a og 0b viste resultaterne, at det var en fordel at anvende et multiopløsningsspektrogram. Måske kunne det være en fordel at anvende et bedre multiopløsningsspektrogram end det, der allerede bruges i metoden. Her tænkes især på enten at bruge constant Q (se afsnit 3.2.1.2) eller måske en wavelettransformation.

Kapitel 7

Konklusion

Denne opgave har i første omgang givet en introduktion til problemstillingen omkring dét at finde melodien i et stykke. Det blev nævnt, at dette ikke er et let problem, fordi der som regel er flere stemmer, som blander sig med melodien og giver et samlet lydbillede.

I kapitel 2 blev der præsenteret et par værktøjer så som spektrogrammet, som kan bruges til analyse af lydsignaler. Desuden blev der introduceret et par begreber fra den musikalske analyse; begreber som kan være nyttige til beskrivelsen af lyde og de fænomener, som kan observeres i et spektrogram. Som et eksempel blev begrebet formant introduceret, så det var muligt at beskrive dét, at et område af overtoner træder tydeligt frem. I kapitel 2 blev der desuden givet et kort overblik over den menneskelige opfattelse af lyde. Som det blev nævnt i kapitel 6, kan man måske med fordel efterligne nogle af egenskaberne ved den menneskelige opfattelse af lyde. Der tænkes blandt andet på, at anvendelsen af phon-skalaen kan være interessant. Som det kort blev bemærket i kapitel 5, har f.eks. Paiva præsenteret en metode, som anvender en efterligning af den menneskelige opfattelse af lyde - og denne metode så ud til at være ganske succesfuld.

I kapitel 3 blev der præsenteret en række forskellige metoder for bestemmelse af melodien. Mange af dem var enten afhængig af en kraftig fundamentalfrekvens eller af nogle tydelige overtoner, eller også var der tale om mere komplicerede metoder. SHS skilte sig ud som en metode, der ikke er afhængig af en kraftig fundamentalfrekvens, og heller ikke er afhængig af nogle tydelige overtoner, hvis blot en af delene er tilstede. En simplificeret udgave af denne metode, tilføjet en lille modifikation kom til at danne grundlag for den endelige foreslåede metode, som blev præsenteret i kapitel 4.

Der blev præsenteret 3 versioner af et forslag til en metode i kapitel 4, hvor de 2 første versioner tog udgangspunkt i hypotesen fra problemformuleringen til opgaven. Det viste sig at hypotesen generelt var god, og at metoderne især havde succes for visse numre. Men som de overordnede resultater i kapitel 5 viste, er de 2 første metoder dårligere end de metoder, som deltog ved MIREX-2006. Dette var ikke tilfældet for den sidste noget mere komplicerede version, som kunne anvende styrken af formodede overtoner til at give et bedre bud på melodiens fundamentalfrekvens. Det viste sig, at metoden scorede 3. højst blandt deltagerne ved MIREX-2006.

En nærmere analyse af resultaterne blev foretaget i kapitel 6, og her viste det sig, at den foreslåede metode især havde problemer med nogle af operamelodierne. Men de øvrige deltagere til MIREX-2006 havde også problemer med

disse melodier. Det felt, som metoden klarede sig relativt dårligst inden for, var popmelodier. Her var metoden markant dårligere end de 2 metoder, som havde fået højest score i alt.

Som et forsøg på at forbedre resultaterne blev det i kapitel 6 foreslået at implementere den fulde SHS, da der har været rapporteret om gode resultater ved brug SHS i forbindelse med de 4 popmelodier. Desuden blev det foreslået at inddrage et mål for klangen, da dette mål måske kan bruges til at adskille instrumenterne, hvorved der kan gives et bedre bud på hvilken frekvens, der hører til det instrument, som har melodien.

Samlet set må det betragtes som meget succesfuldt, at den sidste foreslåede version er i top 3 blandt deltagerne ved MIREX-2006; specielt når man tænker på, hvor simpel metoden er. Der er dog helt klart plads til forbedringer, for der er tydeligvis problemer med visse melodier. Hvis det skulle lykkes at forbedre metoden i en grad, så der ikke længere er problemer med disse melodier, er der brug for en meget omfattende afprøvning, som inkluderer mange flere genrer, før man kan danne sig et bedre overblik over, hvordan metoden generelt virker på virkelige lydoptagelser.

Litteratur

- [1] Judith C. Brown. Calculation of a constant q spectral transform. *Journal of Acoustic of Society of America*, 89(1):425–434, 1991. <http://www.wellesley.edu/Physics/brown/pubs/cq1stPaper.pdf>.
- [2] Judith C. Brown and Miller S. Puckette. An efficient algorithm for the calculation of a constant q transform. *Journal of Acoustic of Society of America*, 92(5):2698–2701, 1992. <http://www.wellesley.edu/Physics/brown/pubs/effalgV92P2698-P2701.pdf>.
- [3] Chuan Cao, Ming Li, Jian Liu, and Yonghong Yan. Multiple f_0 estimation in polyphonic music (mirex 2007). Technical report, Thinkit Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, 2007. http://www.music-ir.org/mirex/2007/abs/F0_cao.pdf.
- [4] Chuan Cao, Ming Li, Jian Liu, and Yonghong Yan. Singing melody extraction in polyphonic music by hamonic tracking. Technical report, Thinkit Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, 2007. http://ismir2007.ismir.net/proceedings/ISMIR2007_p373_cao.pdf.
- [5] R. I. Damper, J. R. Thorpe, and C. H. Shadle. Separation of speech from simultaneous talkers. In *XIIIth International Congress of Phonetic Sciences*, 1995. <http://eprints.ecs.soton.ac.uk/328/2/paper.pdf>.
- [6] Rui Pedro Pinto de Carvalho e Paiva. *Melody Detection in Polyphonic Audio*. PhD thesis, Universidade de Coimbra, Faculdade de Ciências e Tecnologia, Departamento de Engenharia Informática, 2006. http://cisuc.dei.uc.pt/dlfile.php?fn=1349_pub_Thesis.pdf&get=1&idp=1349&ext=.
- [7] Alain de Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *Acoustical Society of America*, 2002. http://recherche.ircam.fr/equipes/pcm/cheveign/pss/2002_JASA_YIN.pdf.
- [8] Karin Dressler. Extraction of the melody pitch contour from polyphonic audio. Technical report, Fraunhofer Institute for Digital Media Technology, 2005. <http://www.music-ir.org/evaluation/mirex-results/articles/melody/dressler.pdf>.
- [9] Karin Dressler. An auditory streaming approach on melody extraction. Technical report, Fraunhofer Institute for Digital Media Technology, 2006. http://www.music-ir.org/evaluation/MIREX/2006_abstracts/AME_dressler.pdf.
- [10] Karin Dressler. Sinusoidal extraction using an efficient implementation of a multi-resolution fft. Technical report, Fraunhofer Institute for Digital Media Technology, 2006. http://www.dafx.ca/proceedings/papers/p_247.pdf.
- [11] B. Meudic E. Gómez, A. Klapuri. Melody description and extraction in the context of music content processing. *Journal of New Music Research*, 32(1), 2003. <http://www.cs.tut.fi/sgn/arg/klap/gomez.pdf>.
- [12] David Gerhard. *Ph.D. Depth Paper: Audio Signal Classification*. PhD thesis, School of Computing Science, Simon Fraser University, Burnaby, 2000. <http://www2.cs.uregina.ca/~gerhard/publications/depth.pdf>.
- [13] David Gerhard. Pitch extraction and fundamental frequency: History and

- current techniques. Technical Report Technical Report TR-CS 2003-6, University of Regina Department of Computer Science, 2003. <http://www2.cs.uregina.ca/~gerhard/publications/TRdbg-Pitch.pdf>.
- [14] Masataka Goto. A robust predominant-f0 estimation method for real-time detection of melody and bass lines in cd recordings. In *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000)*, pages II-757–760, 2000. <http://staff.aist.go.jp/m.goto/PAPER/ICASSP2000goto.pdf>.
 - [15] Masataka Goto and Satoru Hayamizu. A real-time music scene description system: Detecting melody and bass lines in audio signals. *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pages 31–40, 1999. <http://staff.aist.go.jp/m.goto/PAPER/CASA99.pdf>.
 - [16] Dik J. Hermes. Measurement of pitch by subharmonic summation. *Journal of Acoustic of Society of America*, 83:257–264, 1988.
 - [17] Anssi P. Klapuri. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. In *IEEE Transactions on Speech and Audio Processing*, volume 11, pages 804–816, 2003. <http://www.cs.tut.fi/sgn/arg/klap/multiplef0.pdf>.
 - [18] Anssi P. Klapuri. *Signal processing methods for the automatic transcription of music*. PhD thesis, Tampere University of Technology, Finland, 2004. http://www.cs.tut.fi/sgn/arg/klap/phd/klap_phd.pdf.
 - [19] Philippe Martin. Comparison of pitch detection by cepstrum and spectral comb analysis. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP apos*, volume 7, pages 180–183, 1982.
 - [20] James A. Moorer. The optimum comb method of pitch period analysis of continuous digitized speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 22(5):330–338, 1974.
 - [21] Martin Piszczalski and Bernard A. Galler. Predicting musical pitch from component frequency ratios. *Journal of the Acoustical Society of America*, 66(3):710–720, 1979.
 - [22] Graham E. Poliner and Daniel P. W. Ellis. A classification approach to melody transcription. Technical report, LabROSA, Dept. of Electrical Engineering, Columbia University, New York NY 10027 USA, 2005. <http://www.music-ir.org/evaluation/mirex-results/articles/melody/poliner.pdf>.
 - [23] Graham E. Poliner and Daniel P. W. Ellis. A classification approach to melody transcription. In *Proceedings of the 6th International Conference on Music Information Retrieval*. Queen Mary, University of London and Goldsmiths College, 2005. <http://ismir2005.ismir.net/proceedings/1107.pdf>.
 - [24] M. R. Schroeder. Period histogram and product spectrum: New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America*, 43(4):829–834, 1968.