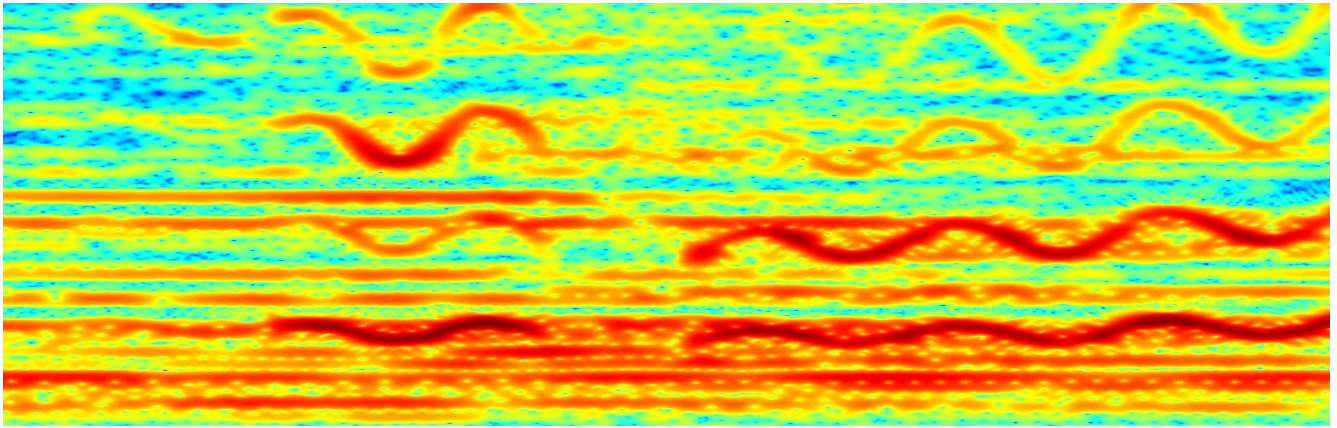




Bestemmelse af melodien i polyfon musik

Et datalogisk speciale om frekvensanalyse og bestemmelse af fundamentalfrekvensen for melodien i polyfon musik

Morten Wendelboe



Datalogisk Institut
Københavns Universitet
1. maj 2009

Vejleder: Søren Olsen

Detecting the Melody in Polyphonic Music

Abstract

This master thesis deals with the task to estimate the melody's fundamental frequency in polyphonic music. It is a problem which is still unresolved, but new and improved methods have been presented within the last few years.

Initially, a number of methods for generation of time/frequency representations of an audio signal is dealt with. The methods include STFT, Morlet wavelet transform (MWT) and chirplet transform. In addition a method called OQSTFT is formulated. This method is suitable for a signal which contains a vibrating harmonic sound, as is the case if a singer uses vibrato.

Some methods for detection of the melody's fundamental frequency are presented. These methods include autocorrelation, cepstrum and SHS. It is proposed that SHS is modified with a spectral normalized factor and a smoothing factor.

An implementation of the modified SHS which use OQSTFT is tested on 2 data sets. One of these data sets was used at the MIREX competitions (year 2006 and 2008) and compared with the methods that participated in the competitions the method implemented in this thesis obtains good results. Non of the methods that participated in the mentioned MIREX competitions achieved as high Raw Chroma score as the implemented method.

Keywords Detection of the melody, estimation of the melody, fundamental frequency, F0, harmonics, autocorrelation, cepstrum, SHS, STFT, wavelet transform, Morlet wavelet transform, MWT, constant Q transform, QSTFT, OQSTFT.

Bestemmelse af melodien i polyfon musik

Resumé

Denne specialeopgave handler om at estimere melodiens fundamentalfrekvens i polyfon musik. Der er tale om et problem, som endnu er uløst, men inden for de seneste par år har der været præsenteret nye og forbedrede metoder.

I første omgang behandles en række metoder for generering af tid/frekvens-repræsentationer af et lydsignal. Metoderne inkluderer STFT, Morlet-wavelet-transformationen (MWT) og chirplet-transformationen. Der bliver desuden formuleret en metode, som kaldes OQSTFT. Denne metode er egnet til et signal, som indeholder en vibrerende harmonisk tone, sådan som det er tilfældet, hvis en sanger anvender vibrato.

Der bliver præsenteret nogle metoder for bestemmelse af melodiens fundamentalfrekvens, og de behandlede metoder inkluderer autocorrelation, cepstrum og SHS. Det bliver foreslået, at SHS modificeres med en spektral normaliseringsfaktor samt en udglattende faktor.

En implementering af den modificerede SHS, som anvender OQSTFT, bliver afprøvet på 2 datasæt. Det ene datasæt blev anvendt ved MIREX-konkurrencerne (år 2006 og 2008) og i forhold til de metoder, som deltog ved disse konkurrencer, opnår den implementerede metode gode resultater. Ingen af de metoder, som deltog ved de omtalte MIREX-konkurrencer, opnåede lige så høj Raw Chroma score, som den implementerede metode.

Nøgleord Bestemmelse af melodien, estimering af melodien, fundamentalfrekvens, F_0 , overtoner, autocorrelation, cepstrum, SHS, STFT, wavelet transformation, Morlet wavelet transformation, MWT, constant Q transformation, QSTFT, OQSTFT.

Forord

Til denne specialeopgave hører 2 bilag.

CD Det ene bilag er en CD, som indeholder en række test-filer, samt kildekoden til den implementerede metode.

Kildekoden er skrevet i Matlab, så programmet Matlab er nødvendigt for kørslen af den implementerede metode. Der er en enkelt fil, som er skrevet i C, og denne skal oversættes igennem Matlab ved kommandoen “mex beregnS.c”.

For at finde melodien i en af testfilerne (eksempelvis den som hedder “daisy1.wav”) køres kommandoen “findMelody(‘daisy1.wav’)”. Derved genereres en fil, som hedder “daisy1.MEL.txt”. For at evaluere hvor godt melodiens fundamentalfrekvens er blevet fundet køres kommandoen “eval2006(‘daisy1.MEL.txt’, ‘daisy1.REF.txt’)”. Hvis man også ønsker grafisk output køres kommandoen “eval2006(‘daisy1.MEL.txt’, ‘daisy1.REF.txt’, 1)”.

Da testfilernes referencerresultat er registreret med forskellige intervaller er det nødvendigt at oplyse hvilket interval referencerresultatet er registreret med. Så hvis man kører en testfil som stammer fra TEST1-datasættet (filerne starter med “train”), er det nødvendigt at skrive: “findMelody(‘train01.wav’, 441/44100)”.

Ud over kildekoden og test-filerne, indeholder CD’en også en pdf-udgave af denne specialeopgave, samt en pdf-udgave af en tidligere opgave, som jeg har lavet. Skulle CD’en være bortkommet, eller hvis den ikke kan læses, kan specialeopgave, den tidligere opgave samt testfilerne alle findes via: <http://www.morwen.dk/university/master-thesis/index.html>. Kildekoden må ikke offentliggøres og er derfor ikke tilgængelig på internettet.

Gammel opgave Det andet bilag er en opgave, som jeg tidligere har skrevet. Da opgavens emne falder sammen med specialeopgavens emne, er der en del begreber og værktøjer, som jeg tidligere har stiftet bekendskab med og skrevet om. Der vil blive henvist en del til den tidligere opgave, for at undgå at skrive det samme indhold 2 gange. Dette har givet mulighed for større fordybelse i denne specialeopgave. Der henvises til den gamle opgave ved at referere til [66].

Det er primært kapitel 2 i den gamle opgave, som er relevant, og det antages at læseren skimmer dette kapitel, hvis læseren ikke i forvejen er bekendt med specialeopgavens emne.

Det skal understreges at den gamle opgave IKKE tæller med til bedømmelsen af denne specialeopgave!

Indhold

1	Introduktion	1
1.1	Målgruppe	2
1.2	Afgrænsning af opgaven	2
1.3	Overordnet indhold og bidrag	4
1.3.1	Tid/frekvens-repræsentation	4
1.3.2	Bestemmelse af fundamentalfrekvensen	5
1.4	Figurer	5
1.5	Forudsætninger	5
1.5.1	Notation og bemærkninger til formler	5
1.5.2	Vigtige funktioner	8
1.5.3	Værktøjskasse	8
1.6	Følgende kapitler	9
2	Test-filer	11
2.1	Datasæt	11
2.1.1	ADC2004	12
2.1.2	TEST1	12
2.1.3	Anvendelse af datasæt	13
2.2	Beregning af scorer	13
2.3	Evaluerings af scorer	14
3	Repræsentation af lyd i tid/frekvens-domæne	16
3.1	Fouriertransformation	16
3.1.1	Beregningskompleksitet	19
3.1.1.1	FFT	19
3.1.2	Vinduesbegrænset fouriertransformation	21
3.2	STFT	21
3.2.1	Vinduer	25
3.2.1.1	Rektangulært vindue	25
3.2.1.2	Foldede rektangulære vinduer	27
3.2.1.3	Gauss-vindue	28
3.2.1.4	Hann-vindue	33
3.2.1.5	Hamming-vindue	34
3.2.1.6	Sammenligning af vinduer	36
3.2.1.7	ENB	37
3.2.2	Beregningskompleksitet	42
3.2.3	Tid/frekvens-repræsentation ved STFT	43
3.3	Wavelet-transformation	43
3.3.1	Definition på den kontinuerte wavelet-transformation	44
3.3.2	Fordele ved wavelet-transformationen	46

3.4	Morlet-wavelet	47
3.4.1	Beregningskompleksitet	49
3.4.2	Tid/frekvens-repræsentation ved MWT	49
3.5	STFT med frekvensafhængig vindueslængde	50
3.5.1	Constant-Q-transformation	50
3.5.1.1	Fordele og problemer	51
3.5.1.2	Sammenligning med MWT	52
3.5.2	FDTF	52
3.5.3	QSTFT	53
3.5.3.1	Optimal Q	54
3.5.3.2	Praktisk afprøvning	58
3.6	STFT med variabel vindueslængde	60
3.6.1	Dresslers multiopløsnings-STFT	66
3.6.1.1	Opdeling af STFT	66
3.6.1.2	Tilføjelse af vindue	67
3.6.1.3	Beregningskompleksitet	68
3.6.1.4	Anvendelse i forbindelse med QSTFT	69
3.7	Chirplet-transformation	69
3.7.1	STFChT	72
3.7.2	Praktisk eksempel	73
3.7.3	Matching Pursuit	74
3.8	Wigner-Ville-fordeling	74
3.8.1	Definition	76
3.8.2	Pseudo Wigner-Ville-fordelingen	77
3.9	Valg af tid/frekvens-repræsentation	78
3.10	Konklusion	79
4	Estimering af melodiens fundamentalfrekvens	80
4.1	Generel opbygning af metoder	82
4.1.1	Preprocessing	82
4.1.1.1	Støj	82
4.1.1.2	Menneskelig hørelse	82
4.1.2	Postprocessing	83
4.1.2.1	Klangfarve	83
4.1.2.2	Menneskestemme	84
4.2	Autocorrelation, cepstrum og lignende metoder	84
4.2.1	Autocorrelation	85
4.2.2	Cepstrum	88
4.2.3	Generalisering	89
4.2.4	Yin	90
4.3	SHS	91
4.3.1	Spektral normaliseringsfaktor	92
4.3.2	Oktavfejl	93
4.3.2.1	Udglattende faktor	95
4.3.3	Ny definition på en modificeret SHS	97
4.4	Andre interessante metoder	97
4.4.1	TWM	98
4.5	Konklusion	99

5	Implementering	100
5.1	Preprocessing	100
5.2	Tid/frekvens-repræsentation	101
5.3	Estimering af melodiens fundamentalfrekvens	102
5.3.1	Postprocessing	104
6	Testresultater	106
6.1	Autocorrelation, cepstrum og generalisering	106
6.1.1	Autocorrelation	107
6.1.2	Cepstrum	107
6.1.3	Generalisering	108
6.1.4	Resultatanalyse	108
6.1.5	Yin	109
6.2	Modificeret SHS	110
6.2.1	Almindelig SHS	110
6.2.2	Modificeret SHS med amplitude	111
6.2.3	Modificeret SHS med styrke	113
6.2.4	Preprocessing	115
6.2.5	Postprocessing	116
6.2.6	Konklusion	116
6.3	Overordnet vurdering af resultater	117
7	Vurdering af den implementerede metode	119
7.1	Problemer	119
7.2	Forslag til forbedringer	121
8	Konklusion	123

Kapitel 1

Introduktion

Denne specialeopgave handler om automatisk estimering af melodien i et stykke musik. Idéen til opgavens emne udspringer af et praktisk problem: Mange mennesker kommer ud for, at de ønsker at høre et bestemt stykke musik, men de kan ikke huske, hvad det hedder, eller hvem der har lavet det. Til gengæld kan de huske, hvordan melodien lyder. Systemer, som anvendes til søgning efter bestemte stykker musik, bruger som regel tekststrengene som søgenøgler. Således er der til hvert stykke musik knyttet metadata så som "artist" og "titel", og det er denne metadata, der søges efter. Problemet er dog, at denne metadata ikke knytter sig direkte til det lydsignal, som musikken består af. Det, som oftest knytter erindringen, om hvordan et stykke musik lyder, sammen med musikstykkets lydbillede, er den abstrakte beskrivelse, som kaldes melodien. Det er oplagt, at det i forbindelse med søgning efter et bestemt musikstykke ville være nyttigt, hvis man kunne foretage søgningen blot ved at synge eller nynne melodien.

Den beskrevne funktionalitet kaldes generelt for "Query by Humming/Singing" (QbHS), og der har specielt i de seneste par år været lagt meget arbejde i udarbejdelsen af en sådan funktionalitet. De eksisterende forslag til QbHS virker kun i ringe grad, og flere af forslagene kræver i høj grad menneskelig arbejdskraft, hver gang et nyt stykke musik føjes til en samling. At de eksisterende forslag til QbHS kun i ringe grad virker og ofte kræver en del menneskelig arbejdskraft, skyldes i høj grad, at de eksisterende metoder for automatisk estimering af melodien i polyfon musik har store problemer med at estimere netop de toner, som mennesker opfatter som melodibærende. Det er dette grundlæggende og endnu uløste problem, som vi her vil beskæftige os med.

Anvendelsesmuligheder Såfremt der engang i fremtiden bliver udviklet en metode, som i høj grad kan estimere melodien i et stykke musik, er der flere anvendelsesmuligheder for metoden. Herunder oplistes nogle forslag til anvendelsesmuligheder:

- Metoden kan (som det blev skitseret herover) anvendes i QbHS.
- Man kan forestille sig metoden anvendt ved automatisk transskribering fra en lydoptagelse til noder.
- I forbindelse med koncerter kan man forestille sig, at metoden kan bruges som en hjælp ved lyssætningen, således at lyssætningen på automatisk vis kan være afhængig af melodien og eventuelt kan følge et partitur.

- Da melodien er det, som binder 2 forskellige fremførelser/fortolkninger af et stykke musik sammen, vil metoden formentlig kunne hjælpe til at finde forskellige fremførelser af et stykke musik.
- Især for musikanalytikere kan metoden muligvis anvendes til at finde melodi-tematiske og motiviske ligheder mellem forskellige stykker musik.
- I forbindelse med karaoke kunne man forestille sig, at metoden kan bruges til på automatisk vis at slette melodien fra et stykke musik.

1.1 Målgruppe

Det forudsættes ikke at læseren kender til emnet i forvejen, men det antages at læseren er bekendt med den matematik, som normalt vil være tilegnet ved begyndelsen til kandidatstudiet på datalogi eller lignende uddannelse. Således kan det være en fordel, men ikke en forudsætning, at læseren har taget introducerende kurser, hvor der anvendes signalbehandling.

1.2 Afgrænsning af opgaven

Da denne opgaves emne er så bredt, at der kan skrives PhD-afhandlinger¹ over emnet, er det nødvendigt at afgrænse de områder, som opgaven arbejder med.

Fundamentalfrekvens Vi vil koncentrere os om estimering af den såkaldte fundamentalfrekvens for de melodibærende toner i et polyfont stykke musik. I de fleste tilfælde afgiver et instrument en lydbølge, som dannes ud fra de frekvenser, der naturligt set opstår i/på instrumentet. Den dybeste af disse frekvenser kaldes for fundamentalfrekvensen, mens de øvrige frekvenser kaldes for overtoner. Samlingen af alle frekvenserne omtales som en tone. En tone siges at have en højde (tonehøjden), og denne er som regel lig fundamentalfrekvensen. En nøjagtig definition på tonehøjden findes ikke, og der kan også være forskel på, hvordan forskellige mennesker opfatter en tones tonehøjde. Men i de fleste tilfælde er der enighed om en tones tonehøjde, og i så fald er tonehøjden som regel lig fundamentalfrekvensen.

Det er altså fundamentalfrekvensen for den melodibærende tone, som vi her vil koncentrere os om at estimere.

Tid/frekvens-repræsentationer Nogle metoder for estimering af en tones fundamentalfrekvens arbejder direkte på en lydbølge eller et "signal"². Men det menneskelige øre er udstyret med en mekanisme, som i en hvis grad kan opdele et signal i frekvenskomponenter. Det er oplagt at undersøge muligheden for at efterligne denne funktionalitet på en eller anden måde.

For at begrænse opgaven er det valgt kun at berøre metoder, som arbejder med en tid/frekvens-repræsentation af signalet. Derudover vil vi kun behandle nogle tid/frekvens-repræsentationer, som kan udregnes på simpel vis. Derfor vil

¹Se f.eks. [9].

²Herefter vil vi i mange tilfælde benævne en lydbølge ved betegnelsen "signal", da analysen af lydbølger ikke nødvendigvis adskiller sig fra analysen af andre bølger og signaler.

vi eksempelvis ikke beskæftige os med modeller, som forsøger at efterligne det menneskelige øres formodede tid/frekvens-repræsentation³.

Metoder for estimering af fundamentalfrekvensen Overordnet set kan metoder for estimering af melodiens fundamentalfrekvens deles i 2 grupper. Metoderne i den ene gruppe fokuserer primært på signalbehandling, og metoderne i den anden gruppe fokuserer primært på statistisk mønstergenkendelse. Metoderne i den første gruppe kan i forhold til metoderne i den anden gruppe opfattes som mere "ligefremme". Der kan dog ikke altid skelnes klart mellem de 2 grupper, for metoderne indenfor statistisk mønstergenkendelse arbejder på en eller flere egenskabs-indikatorer for signalet; og disse egenskabs-indikatorer bliver som oftest fundet ved hjælp af signalbehandling. Således kan man se signalbehandlingen og metoderne i den første gruppe som et grundlag for en efterfølgende statistisk mønstergenkendelse.

For at begrænse opgaven vil vi se på metoder, som hører til den første gruppe, og af disse metoder vil vi nøjes med at se på ganske få.

Eksistens af melodi Vi vil nøjes med at koncentrere os om bestemmelsen af frekvensen for en formodet melodi-tone, og vil altså ikke beskæftige os med, om der overhovedet er en melodi-tone til stedet i et signal. Til brug i et QbHS-system vil det være nødvendigt også at kunne bestemme, hvornår signalet blot indeholder lyde fra akkompagnerende instrumenter, og hvornår der er en melodi-tone til stede.

Da bestemmelsen af, om der er overhovedet er en melodi-tone tilstede i signalet, som regel kræver et meget nøjagtigt estimat af frekvensen for en formodet melodi-tone, vil vi kun beskæftige os med bestemmelsen af frekvensen for en formodet melodi-tone. Det antages altså, at der til enhver tid eksisterer en melodibærende tone, selv om dette kun yderst sjældent er tilfældet.

Evaluerings Som regel er mennesker enige om, hvad der er melodien i et stykke musik, men der er tilfælde, hvor der kan være en del uenighed, og der er visse musikstykker, hvor der snarere er tale om flere melodier samtidig. Hvis man ikke har et klart bud på, hvad der er melodien i et stykke musik, kan det være svært at afgøre, hvor godt en metode for estimering af melodiens fundamentalfrekvens udfører sit job. En måde, hvorpå man kan evaluere en sådan metode, er ved at inddrage en stor gruppe mennesker, hvor hver person udtaler sig om, hvor godt de mener metoden har estimeret melodiens fundamentalfrekvens. Ved at se på den generelle holdning i gruppen, kan man få et billede af, hvor godt metoden efterligner den generelle menneskelige opfattelse af melodier. Denne evalueringsprocedure er dog meget langsommelig og arbejdskrævende. Derudover lider evalueringsproceduren under en manglende mulighed for let at vurdere metoden i forhold til andre metoder.

Der har i 2005, 2006 og 2008 været afholdt konkurrencer for metoder, der estimerer fundamentalfrekvensen for melodien, og ved disse konkurrencer har evalueringen af de deltagende metoder været baseret på basis-svar for melodiens fundamentalfrekvens. Til de musikstykker, som metoderne evalueres på

³En tid/frekvens-repræsentation, som efterligner det menneskelige øres respons på et signal, benævnes ofte som et "cochleagram". Et eksempel på brugen af et cochleagram ved estimering af melodien i polyfon musik findes i [9].

baggrund af, er der altså defineret et “korrekt” svar på melodiens fundamental-frekvens.

Som evalueringsværktøj vil vi nøjes med de musikstykker og tilhørende bases-svar, som er blevet gjort offentligt tilgængelige ved de omtalte konkurrencer, og evalueringsproceduren vil være den samme som ved de omtalte konkurrencer. Det betyder, at der er mulighed for at vurdere en metodes succes i forhold til succes for de metoder, som deltog ved de omtalte konkurrencer.

1.3 Overordnet indhold og bidrag

Denne opgave tjener primært som en introduktion til det at bestemme melodien på automatisk vis. Således vil en læser, som ikke allerede beskæftiger sig med emnet blive klædt bedre på, så det forhåbentlig vil være lettere for læseren at udvælge metoder, finde metoder og eventuelt selv udvikle metoder for estimering af melodiens fundamental-frekvens.

1.3.1 Tid/frekvens-repræsentation

Mens mange skrivelser over emnet kun i begrænset omfang berører valget af en tid/frekvens-repræsentation, vil der i denne opgave lægges væsentlig vægt på at undersøge mulighederne for tid/frekvens-repræsentationer. Dette skyldes dels, at megen anden litteratur netop synes at mangle denne del, således at en uerfaren inden for emnet kan få fornemmelsen af at mangle nogle grundlæggende færdigheder. Derudover kan man sige, at det er vigtigt at have det mest grundlæggende på plads, inden man betyder at bygge ovenpå.

Mange metoder for estimering af melodiens fundamental-frekvens anvender en af de mest simple tid/frekvens-repræsentationer, og denne bliver ofte valgt uden begrundelse. Vi kan på ingen måde dække emnet grundigt, men ved at præsentere en række forskellige tid/frekvens-repræsentationer kan vi gøre læseren opmærksom på deres eksistens, og dette kan gøre læseren bedre i stand til at vælge en egnet repræsentation og eventuelt søge andre mere ukendte, men måske meget brugbare tid/frekvens-repræsentationer.

Et vigtigt argument for at undersøge tid/frekvens-repræsentationer nærmere er, at hvis udgangspunktet, altså tid/frekvens-repræsentationen, er langt fra optimalt, så kan selv den mest avancerede metode vise sig at give dårlige resultater; ikke fordi metoden som sådan er dårlig, men fordi udgangspunktet er dårligt. Det er vigtigt at være bevidst om hvilke begrænsninger, der ligger i en valgt tid/frekvens-repræsentation.

Bidrag: OQSTFT Der bliver formuleret en forholdsvis simpel tid/frekvens-repræsentation, som opkaldes “OQSTFT”. Denne er specielt rettet mod signaler, hvor det antages, at melodiens tone kan svinge hurtigt, som det f.eks. er tilfældet, når en operasanger anvender vibrato. At der specielt er fokus på dannelsen af en tid/frekvens-repræsentation, som er rettet mod sådanne signaler, skyldes, at jeg i [66] afprøvede en metode for bestemmelse af melodiens fundamental-frekvensen, og metoden viste specielt dårlige resultater for et musikstykke, hvor melodiens fundamental-frekvens svinger hurtigt. Det har dog sidenhen vist sig, at den primære årsag til de dårlige resultater skyldtes en fejl i implementeringen.

Det viser sig, at OQSTFT med fordel kan anvendes som en del af denne opgaves forslag til en metode for estimering af melodiens fundamentalfrekvens.

1.3.2 Bestemmelse af fundamentalfrekvensen

Jeg har valgt at koncentrere metoder for bestemmelse af fundamentalfrekvensen omkring et par simple metoder, som kan udtrykkes ved enkle matematiske udtryk. Der er blevet valgt 2 metoder, nemlig "autocorrelation" og en modificeret "SHS". Autocorrelation er yderst simpel, og effektiviteten (altså et mål for hvor godt metoden estimerer den korrekte fundamentalfrekvens) kan derfor beskrive en mindste effektivitet, som det må forventes, at mere komplicerede metoder kan overgå.

Bidrag: Modifikationer til SHS Der bliver foreslået et par modifikationer til "SHS"-metoden. Modifikationerne er en videreudvikling af de modifikationer, som jeg anvendte i [66]. Der bliver også foreslået en meget simpel postprocessing, hvis mål er, at holde bedre fast i en meloditone, når denne klinger ud og derved bliver svagere.

SHS-metoden med de foreslåede modifikationer giver gode resultater, men den foreslåede postprocessing forbedrer kun i ringe grad de generelle resultater.

1.4 Figurer

Det skal bemærkes at alle figurer er egne kreationer, med mindre andet er bemærket i forbindelse med en given figur.

På en del figurer søges det at illustrere en 3-dimensionel overflade. Hvis disse figurer er i gråtoner, så angiver de mørke områder lave værdier, mens de lyse områder angiver høje værdier. Hvis figurerne er i farver angiver mørkeblå en lav værdi, mens farven mørkerød angiver en høj værdi. Der er ingen sammenhæng mellem farver og positive/negative værdier.

1.5 Forudsætninger

I dette afsnit vil vi forberede læseren på de følgende kapitler, som indeholder denne opgaves primære indhold. Først vil vi klarlægge den anvendte notation i formler, og dernæst vil vi give nogle få bemærkninger til den såkaldte "værktøjskasse", som er at finde i [66].

1.5.1 Notation og bemærkninger til formler

Her vil vi opliste de vigtigste symboler, som anvendes i denne specialeopgave. Det første vi vil bemærke er, at vi generelt arbejder med ikke-diskrete værdier. Oversættelsen fra en funktion, som arbejder med kontinuerte variable, til en funktion, som arbejder med diskrete værdier, anses generelt blot som et implementerings-spørgsmål. I tilfælde, hvor der specifikt arbejdes med diskrete værdier, vil dette fremgå ved anvendelsen af indekse, således at hver diskret værdi kan tilgås ved et indeks i en følge af diskrete værdier. Således beskriver $s[n]$ den n 'te værdi i følgen, som kaldes s .

Den anvendte matematiske notation er meget normal, men herunder præciseres anvendelsen af nogle symboler.

Variable og konstanter	
i	er et symbol for den imaginære enhed, således at $i^2 = -1$.
a	er en variabel eller en konstant. Generelt gælder det, at kursiverede bogstaver repræsenterer variable eller konstanter. Hver variabel eller konstant noteres med et enkelt bogstav. Eksempler på konstanter er det naturlige tal e og tallet π .
t	er en variabel og anvendes som symbol for tiden.
τ	er en variabel og anvendes som symbol for tiden, men i modsætning til t , så anvendes τ som den variabel, der integreres over, hvis der skal integreres over tiden.
n	er en diskret variabel og anvendes ved indeksering i tiden.
η	er en diskret variabel og anvendes ved indeksering i tiden, men i modsætning til n , så anvendes η som den variabel, der summeres over, hvis der skal summeres over tiden.
ν	er en variabel og anvendes som symbol for frekvensen. I mange skrivelser anvendes vinkelfrekvensen, men her vil vi anvende frekvensen og notere vinkelfrekvensen som $2\pi\nu$.
υ	er en variabel og anvendes som symbol for frekvensen, men i modsætning til ν , så anvendes υ som den variabel, der integreres over, hvis der integreres over frekvenser.
k	er en diskret variabel og anvendes ved indeksering af frekvenser.
Funktioner	
$\mathbf{a}$	er et funktionsnavn. Generelt er en følge af ikke-kursiverede fede bogstaver navnet på en funktion. En undtagelse er dog fouriertransformationen. Et eksempel på en funktion er cosinus, som noteres ved cos . Generelt anvendes store bogstaver, når en funktion opererer i frekvensdomænet, og små bogstaver, når funktionen opererer i tidsdomænet.
$\mathcal{F}$	er en funktion og anvendes som symbol for fouriertransformationen.
$\mathcal{F}^{-1}$	er en funktion og anvendes som symbol for invers fouriertransformationen.
$\mathbf{s}$	er en funktion og anvendes som symbol for et signal. Da vi skal arbejde med lydsignaler, er der tale om et reelt signal.
$\mathbf{S}$	er en funktion og anvendes som symbol for et signal, som er beskrevet i frekvensdomænet.

- A** er en funktion og anvendes som symbol for amplituden af frekvenserne i et signal. Således er $\mathbf{A} = |\mathbf{S}|$.
- P** er en funktion og anvendes som symbol for styrken af frekvenserne i et signal. Således er $\mathbf{P} = \mathbf{A}^2$.
- w** er en funktion og anvendes som symbol for et vindue.
- W** er en funktion og anvendes som symbol for et vindue, som er beskrevet i frekvensdomænet.

Andre symboler

- ()** angiver argumenter i forbindelse med funktioner. Argumenter noteres mellem (og), og flere argumenter adskilles med et komma. Således betyder $\mathbf{a}(a,b)$ at funktionen **a** beregnes for variablene *a* og *b*.
- []** angiver argumenter i forbindelse med diskrete funktioner. Argumenter noteres mellem [og], og flere argumenter adskilles med et komma. Således betyder $\mathbf{a}[a,b]$ at funktionen **a** beregnes for variablene *a* og *b*.
- [;]** angiver en følge af diskrete værdier. Således betyder $[1;10]$ følgen af værdier fra 1 op til og med 10.
- [a b]** angiver en sammensætning af 2 følger af diskrete værdier. Hvis $\mathbf{a} = [1;3]$ og $\mathbf{b} = [4;10]$ så er $[\mathbf{a} \ \mathbf{b}] = [1;10]$.
- { }** kan eventuelt anvendes i forbindelse med funktionsnavne. { } markerer en form for specialisering af en funktion, hvilket f.eks. kan være tilfældet, når én funktion anvendes på en anden funktion. Således betyder $\mathcal{F}\{\mathbf{s}\}(t)$, at der foretages en fouriertransformation af funktionen **s**, og resultatet afhænger af variabelen *t*. Især i tilfælde, hvor der er en klar adskillelse af en funktion og variable, anvendes { } frem for (). I tilfælde hvor en funktion afhænger af mange variable, kan en fastlåsning af nogle af disse også markeres med { }, således at $\mathbf{s}\{a\}(t)$ betyder specialiseringen af **s**, hvor parametren *a* holdes fast, og hvor *t* er en variabel.
- *** er en operator og anvendes som symbol for foldning. Således betyder $\mathbf{a} * \mathbf{b}$, foldningen af funktionerne **a** og **b**.
- er en operator og anvendes som symbol for kompleks konjugering. Således betyder $\bar{a}$, den komplekst konjugerede af *a*.
- δ** er et symbol for delta-funktionen, men anvendes som en operator og virker på den efterfølgende variabel eller funktion. Delta-funktionen angiver et lille interval.

1.5.2 Vigtige funktioner

Herunder vil vi opliste navne på de mest benyttede funktioner, samt give definitioner på nogle få knap så normale funktioner:

- dirac** Dirac-delta-funktionen **dirac**(x) er lig 0 for alle værdier af x undtagen $x = 0$. Integralet af Dirac-delta-funktionen er lig 1. Ofte noteres Dirac-delta-funktionen som $\delta(x)$, men her er det valgt ikke at anvende δ -symbolet, da dette ofte også anvendes til at angive et lille interval.
- mwt** anvendes som symbol for Morlet-wavelet-transformationen (MWT).
- sinc** **sinc**(x) = **sin**(x)/ x .
- stft** anvendes som symbol for Short Time Fourier Transform (STFT).
- wt** anvendes som symbol for wavelet-transformationen.

1.5.3 Værktøjskasse

Da jeg tidligere har skrevet en opgave, hvis emne overlapper med denne speciale-opgave, er der en række centrale begreber og værktøjer, som jeg allerede har skrevet om. Den omtalte opgave [66], tæller **ikke** med i bedømmelsen af denne speciale-opgave, men da det kan virke omsonst at gentage, hvad jeg allerede har skrevet i [66], vil der blive henvist en del til [66].

I [66, kapitel 2] findes en række centrale begreber og værktøjer. Hvis læseren ikke i forvejen er inde i denne speciale-opgaves emne, antages det, at læseren i det mindste skimmer nævnte kapitel. [66] er vedlagt som bilag til denne speciale-opgave.

Der er et par centrale begreber, som vi her vil præcisere:

Tonedannelse Som udgangspunkt antages det, at lyden fra et instrument består af en række sinusoidale bølger, som det er beskrevet i [66, afsnit 2.1]. Men dette er blot en tilnærmelse til den virkelige verden, hvor materialer ikke nødvendigvis kan svinge som foreskrevet i en pæn matematisk formel. Generelt set ser det dog ud til, at antagelsen er god.

Amplitude og styrke Vi vil i denne speciale-opgave anvende betegnelsen "styrke" P , når der menes kvadratet på en frekvens amplitude: $P = A^2$, hvor A er frekvensens amplitude.

Decibel-skala Når styrken af nogle frekvenser afbildes på figurer, vil styrken som regel være afbildet på en decibel-skala. Decibel-skalaen beregnes som det relative lydtryk via formel 2.11 i [66, afsnit 2.4.1.1]. Som regel vil det kraftigste lydtryk, som gengives på en figur, definere 0 på decibel-skalaen.

1.6 Følgende kapitler

Her vil vi kort give et overblik over de følgende kapitlers indhold.

Kapitel 2 I dette kapitel beskrives de test-filer (musik-stykker), som den endelige implementering af en metode for estimering af melodiens fundamental-frekvens afprøves på. Det beskrives desuden, hvordan det måles, hvor godt metoden estimerer melodiens fundamental-frekvens.

Kapitel 3 Her skal vi se nærmere på forskellige tid/frekvens-repræsentationer. Der startes helt fra grunden af; således begynder kapitlet med en beskrivelse af fouriertransformationen, som muligvis allerede er kendt af læseren.

Da der gennem tiderne er blevet foreslået mange forskellige "vinduer" ved anvendelsen af fouriertransformationen, og da mange skrivelser omhandlende denne opgaves emne ikke begrundet valget af et vindue, vil vi se nærmere på de mest kendte eller interessante vinduer.

Den meget velkendte short-time fouriertransformation "STFT" bliver præsenteret.

Wavelet-transformationen bliver præsenteret, og her viser det sig, at den såkaldte Morlet-wavelet-transformation har nogle egenskaber, som passer godt til den musikalske skala.

Da Morlet-wavelet-transformationen minder meget om STFT, bliver der præsenteret en generalisering af STFT, og dette fører til formulering af OQSTFT. Håbet er, at OQSTFT er bedre end STFT og Morlet-wavelet-transformationen til at beskrive et signal, som indeholder musik. Da vi i denne speciale-opgave ikke skal beskæftige os med at danne et signal ud fra en tid/frekvens-repræsentation, men kun skal analysere en tid/frekvens-repræsentation, vil vi ikke undersøge muligheden for at transformere tilbage fra en OQSTFT-repræsentation til en tid/sted-repræsentation af signalet.

En generalisering af STFT og Morlet-wavelet-transformationen er den såkaldte chirplet-transformation. Denne transformation præsenteres også, men der er tale om en transformation, som er meget mere beregningstung end de andre nævnte transformationer. Vi berører ganske kort Matching Pursuit, som beskriver en effektiv måde, hvorpå man kan generere en tid/frekvens-repræsentation, som baserer sig på chirplet-transformationen.

Endelig vil vi ganske kort stifte bekendtskab med Wigner-Ville-fordelingen, da denne udviser specielt pæne resultater i forbindelse med meget simple signaler.

Kapitel 4 I dette kapitel præsenteres nogle metoder som ud fra en analyse af en tid/frekvens-repræsentation af et signal, giver et estimat for en formodet melodis fundamental-frekvens.

Vi starter med autocorrelation, cepstrum og en generalisering af disse.

Dernæst koncentrerer vi os om SHS, og der gives forslag til et par modifikationer til SHS.

Endelig nævnes ganske kort nogle andre metoder heriblandt TWM.

Kapitel 5 I dette kapitel beskrives implementeringen af denne opgaves forslag til en metode for estimering af melodiens fundamental-frekvens. Metoden er baseret på OQSTFT samt den modificerede SHS.

Kapitel 6 I dette kapitel præsenteres test-resultater for en implementering af autocorrelation, cepstrum og generaliseringen. Der præsenteres også test-resultater for implementeringen af denne opgaves forslag til en metode for estimering af melodiens fundamentalfrekvens.

Kapitel 7 I dette kapitel bemærkes en række problemer ved den implementerede metode, og der gives forslag til forbedringer.

Kapitel 8 Dette kapitel indeholder opgavens konklusion.

Kapitel 2

Test-filer

Det kan være svært på automatisk vis at afprøve en metode for estimering af melodiens fundamentalfrekvens, da der kræves nogle musik-filer, hvortil melodiens fundamentalfrekvens er bestemt med stor tidsmæssig og frekvensmæssig nøjagtighed. Her vil vi nævne 2 måder, hvorpå melodiens fundamentalfrekvens kan bestemmes:

- Hvis man anvender musik, som er genereret ud fra MIDI-lydspor, er der i selve MIDI-filen indirekte indeholdt information om melodiens fundamentalfrekvens, da MIDI-filen beskriver, hvornår hvert enkelt instrument spiller, samt hvilken tone der er tale om. Det er dog vigtigt, at det rigtige lydspor i MIDI-filen bliver valgt ud, da det ikke nødvendigvis fremgår tydeligt, hvilket lydspor (og dermed hvilket instrument) der har melodien.
- Man kan anvende lydoptagelser, hvor melodien er optaget på et selvstændigt lydspor. Der eksisterer gode metoder for estimeringen af fundamentalfrekvensen for et enkelt instrument, og disse metoder kan bruges på det selvstændige lydspor. Dernæst kan man på manuel vis verificere resultatet ved at sammenligne resultatet med et spektrogram af lydsporet og samtidig lytte til lydsporet.

Der findes adskillige offentligt tilgængelig MIDI-filer på internettet, så der er sådan set ikke de store problemer ved at skaffe musik-filer samt referenceresultater. Men da musik fra MIDI-filer kun udgør en meget lille del af den musik, som spilles på radiostationer, i forretninger og i private hjem, kan man sige, at et datasæt, som primært består af MIDI-filer, kun i ringe grad repræsenterer musik generelt.

2.1 Datasæt

For at undgå at bruge ressourcer på indsamling af anvendelige test-filer, og for at det skal være let at sammenligne resultatet af en metode for estimering af melodiens fundamentalfrekvens med resultaterne fra andres metoder, er det valgt at anvende den samling af test-filer, som blev anvendt ved MIREX-2005-, MIREX-2006- og MIREX-2008-konkurrencerne¹. Ved MIREX-konkurrencerne var der en

¹MIREX-konkurrencerne bliver afholdt en gang om året, og konkurrencernes hjemmeside kan findes på <http://www.music-ir.org/mirex/>.

kategori, som netop handlede om estimering af melodiens fundamentalfrekvens i polyfon musik. (Denne kategori var dog ikke med i år 2007.)

Vi vil primært evaluere på baggrund af et datasæt, som blev anvendt ved MIREX-konkurrencerne, og som jeg også anvendte i [66]. Datasættet bliver kaldt “ADC2004”, da denne betegnelse anvendes ved MIREX-konkurrencerne. De deltagende metoder ved MIREX-konkurrencerne blev evalueret på baggrund af datasættet, samt et hemmeligt datasæt. Derudover blev der henvist til et datasæt, som vi her vil kalde “TEST1”, men dette datasæt blev de deltagende metoder ikke afprøvet på.

2.1.1 ADC2004

Dette datasæt består af 20 filer med en længde på mellem 10 og 32 sekunder.² Datasættet indeholder referencerresultater for melodiernes fundamentalfrekvenser, og disse referencerresultater er noteret for hvert ca. 0,006 sekunder — helt præcist er der tale om 256/44.100 sekunder. Filerne er samplet ved 44.100 Hz og er i mono.

Filerne falder i 5 grupper med 4 filer i hver gruppe. Filerne, som hedder “daisy”, indeholder en tydelig melodi, hvor hver tone er meget stabil, dvs. der anvendes kun i ringe grad vibrato. Akkompagnementet er forholdsvis afdæmpet. Filerne, som hedder “jazz”, indeholder også en tydelig melodi, men her er akkompagnementet mere fremtrædende, og i akkompagnementet kan der til tider høres en form for modstemme til melodien. I filerne, som hedder “midi”, er det til tider svært at adskille de forskellige instrumenters lyde, og derfor er melodien muligvis ikke så klar. I nogle af disse filer kan melodien primært identificeres, fordi akkompagnementet har dybere toner end melodien. I filerne, som hedder “opera”, høres melodien tydeligt, og akkompagnementet er lidt mere dæmpet, end det ofte vil være i en opera. Der anvendes i høj grad vibrato i disse filer. I filerne, som hedder “pop”, anvendes en del glissando (der glides op på en tone), og akkompagnementet er mere fremtrædende.

2.1.2 TEST1

Dette datasæt indeholder 13 filer med en længde på mellem 24 og 39 sekunder.³ Datasættet indeholder referencerresultater for melodiernes fundamentalfrekvenser, og disse referencerresultater er noteret for hvert 0,01 sekunder. Også disse filer er samplet ved 44.100 Hz og er i mono.

Filerne falder i 4 genrer, med 2 jazz-melodier, 5 pop/rnb-melodier, 2 singer-song-melodier og 4 midi-melodier. Det har været muligt at identificere 4 af melodierne, således at “train02” er Madonna med sangen Frozen, “train04” er John Legend med sangen Ordinary People, “train07” er 112 med sangen Anywhere og “train08” er Monica med sangen Angel Of Mine. Fælles for de 9 første filer (“train01” til “train09”) er at melodien synges. De sidste 4 filer (“train10” til “train13”) er midi-filer.

²Datasættet ADC2004 er blevet fundet på <http://www.iaa.upf.es/mtg/ismir2004/contest/melodyContest/FullSet.zip>.

³Datasættet TEST1 er blevet fundet på http://www.ee.columbia.edu/~graham/mirex_melody/mirex05TrainFiles.zip.

2.1.3 Anvendelse af datasæt

Begge datasæt vil blive brugt under udarbejdelsen af en metode for estimering af melodiens fundamentalfrekvens. Således kan begge datasæt anvendes til at estimere størrelsen af en række parametre. Men for at sikre, at den udarbejdede metode ikke bliver specialdesignet til ét af de 2 datasæt, vil de 2 datasæt blive brugt hver for sig under estimeringen af en række parametre for metoden. Således kan nogle af de mest optimale parametre anvendt på datasæt TEST1 afprøves på ADC2004, og nogle af de mest optimale parametre anvendt på datasæt ADC2004 kan afprøves på TEST1.

Et problem ved denne adskillelse af de 2 datasæt er dog, at de er meget forskellige. Eksempelvis indeholder datasæt TEST1 ikke musikstykker indefor opera-genren. For at undgå dette kunne man have sammenblandet de 2 datasæt, men dette ville give problemer i forhold til at sammenligne en metodes succes med succes for de metoder, som deltog ved MIREX-konkurrencerne.

Ved MIREX-konkurrencerne blev de deltagende metoder afprøvet på datasæt ADC2004, men desværre ikke på datasæt TEST1. Dette er ærgerligt, specielt fordi de første 9 filer i datasæt TEST1 virker mere "normale" end filerne i datasæt ADC2004. Man kan kritisere filerne i datasæt ADC2004 for at være lidt for simple.

2.2 Beregning af scorer

For at det er muligt at sammenligne resultater fra en implementering med resultaterne fra MIREX-konkurrencerne, bliver evalueringsmetoderne fra MIREX-konkurrencerne anvendt. Ud over muligheden for at bestemme melodiens fundamentalfrekvens var det ved MIREX-konkurrencerne også muligt at give et estimat for, hvornår melodien er tilstede. Men bestemmelsen af og evalueringen af dette sidstnævnte estimat er uafhængig af det førstnævnte estimat. Således kan man ved MIREX-konkurrencerne uden risiko altid give et gæt på en fundamentalfrekvens, selv om det anses for sandsynligt, at der ikke er nogen melodi til stede.

De 2 interessante scorer, som bliver målt ved MIREX-konkurrencerne, er den såkaldte "Raw Pitch" og den såkaldte "Raw Chroma".

Raw Pitch For de tidsrum, hvor der er en melodi til stede, beregnes det hvor stor en procentdel af tiden, den korrekte fundamentalfrekvens er blevet fundet, og dette mål kaldes for Raw Pitch. Det tolereres, at gættet på en fundamentalfrekvens ligger op til en kvart tone⁴ ved siden af det korrekte gæt. Da den mindste enhed i den vestlige musikalske skala er en halv tone, er en kvart tones nøjagtighed tilstrækkelig til bestemmelsen af, om der bliver gættet korrekt.

Antag at der til et givent tidspunkt bliver gættet på frekvensen ν_g , og den korrekte frekvens er ν_k . Da kan nedenstående formel angive om gættet ligger tæt nok på den korrekte frekvens. Hver frekvens omregnes til en tone i 12-toneskalaen, og et korrekt gæt noteres med værdien 1, mens et forkert gæt

⁴Se [66, afsnit 2.2] hvor ordet "tone" beskrives.

noteres med værdien 0.

$$\begin{aligned}
 t_g &= 69 + 12 \log_2(\nu_g/440) \\
 t_k &= 69 + 12 \log_2(\nu_k/440) \\
 E &= |t_g - t_k| \\
 \text{RawPitch}(E) &= \begin{cases} 1 & \text{for } E < 0,5 \\ 0 & \text{ellers} \end{cases} \quad (2.1)
 \end{aligned}$$

I formelen angiver t_g midi-tone-nummeret for den givne frekvens, og kammertonen (som har en frekvens på 440 Hz) har midi-tone-nummer 69.

Raw Chroma Ved Raw Chroma forstås det samme mål som Raw Pitch, bortset fra at oktavfejl tillades. Man kan sige at Raw Chroma i nogle tilfælde er tilstrækkelig, da mennesker kan finde på at synge en melodi, som er forskudt med en eller flere oktaver. Desuden kan mennesker undervejs i en melodi finde på at forlægge melodien med en oktav, f.eks. fordi melodien pludselig bliver for høj eller for dyb for sangerens stemme. Følgende formel angiver beregningen af Raw Chroma, hvor ν_g angiver en gæt på en frekvens, og ν_k angiver en korrekt frekvens:

$$\begin{aligned}
 t_g &= 69 + 12 \log_2(\nu_g/440) \\
 t_k &= 69 + 12 \log_2(\nu_k/440) \\
 t_{g\text{Chroma}} &= \text{mod}(t_g, 12) \\
 t_{k\text{Chroma}} &= \text{mod}(t_k, 12) \\
 E_{\text{chroma}} &= |t_{g\text{Chroma}} - t_{k\text{Chroma}}| \\
 E &= \begin{cases} E_{\text{chroma}} & \text{for } E_{\text{chroma}} \leq 6 \\ 12 - E_{\text{chroma}} & \text{ellers} \end{cases} \\
 \text{RawChroma}(E) &= \begin{cases} 1 & \text{for } E < 0,5 \\ 0 & \text{ellers} \end{cases} \quad (2.2)
 \end{aligned}$$

I formelen angiver **mod** modulus-funktionen. $t_{k\text{Chroma}}$ angiver den korrekte tone noteret uafhængig af oktaven. Når tonen noteres uafhængig af oktaven, er den største mulige afstand mellem 2 toner 6.

2.3 Evaluering af scorer

De angivne scorer, Raw Pitch og Raw Chroma, angiver, hvor stor en procentdel korrekte gæt, der er i den tid, hvor der er en melodi til stede. I bedste fald kan der opnås en score på 100%. Spørgsmålet er, om disse scorer kan bruges til andet end blot at placere forskellige metoder på en rangliste. Er det nødvendigvis bedre at have en score på 91% end en score på 90%? Da det netop er procentdelen af korrekte scorer, som angives med scoren, vil det nok i langt de fleste tilfælde være en fordel med en score på 91% set i forhold til en score på 90%. Men set i forhold til den menneskelige opfattelse af melodier, er der nogle steder i melodien, hvor et forkert gæt er værre end andre steder. Således gør det ikke så meget, hvis der kun bliver gættet forkert ved slutningen af en tone — der hvor tonen klinger ud, og det alligevel ikke er så vigtig, hvornår tonen stopper.

Jeg har erfaret, at når en metode scorer over 80% for en given fil, så er melodien generelt set fundet. De fejl, der er, ligger oftest i områder, hvor

melodien er svag, og disse områder kunne man lige så godt have undladt at tage med i betragtning, fordi det også er svært for mennesker at høre melodien, når den er svag. Desværre er der undtagelser, således at fejlene også findes i områder, hvor melodien er kraftig. Det kunne være interessant og brugbart, hvis det for hver melodi også var defineret, hvornår melodien er kraftig, og dermed vigtig at få estimeret korrekt.

Når en metode scorer omkring 70%, er der som regel en del huller imellem de korrekte estimer, og disse huller forstyrrer helheden. Der er dog undtagelser: Hvis melodien består af mange korte toner, kan der opnås en dårlig score omkring 70%, selv om det stort set kun er afslutninger på toner, der bliver estimeret forkert.

For nogle melodier vil stort set alle metoder kunne opnå en høj score. Det betyder at det primært er de få "svære" melodier, som trækker op eller ned i den samlede score for et helt datasæt. Generelt set har det stor betydning at stige blot et enkelt procentpoint, når den gennemsnitlige score for et datasæt ligger over 80%. Det skyldes dels, at det er de få "svære" melodier, som trækker op eller ned, og dels kan det være svært at finde en metode, som virker godt på alle testfilerne i datasættet — testfilerne er simpelthen for forskellige.

Det er vigtig at gøre sig klart om Raw Pitch eller Raw Chroma er vigtigst. Man kan sige, at hvis Raw Pitch er høj, så er alt godt, men hvis Raw Pitch ligger lavt, så kan metoden alligevel godt vise sig at være udmærket, såfremt Raw Chroma ligger højt. Spørgsmålet om, hvorvidt det er tilstrækkeligt at Raw Chroma ligger højt, afhænger af, hvad metoden skal anvendes til. Hvis metoden skal anvendes til QbHS, vil det oftest være tilstrækkeligt at anvende Raw Chroma, men hvis metoden skal bruges til på automatisk vis at slette melodien i et stykke musik, så er det vigtig, at der bliver gættet på den korrekte oktav.

Kapitel 3

Repræsentation af lyd i tid/frekvens-domæne

I dette kapitel skal vi stifte bekendtskab med nogle af de vigtigste repræsentationer af lyde i tid/frekvens-domænet. I første omgang koncentrerer vi os om 2 velkendte teknikker, som har været anvendt til analyse af musik-signaler og tale-signaler. De 2 velkendte teknikker, nemlig STFT og wavelet-transformation, kan ligne hinanden, så vi vil undersøge muligheden for at kombinere disse velkendte teknikker. Til slut vil vi kort berøre nogle alternative teknikker. Da fouriertransformation anvendes i STFT, vil vi først lægge ud med en introduktion til fouriertransformation og dernæst fortsætte med STFT.

3.1 Fouriertransformation

Selv om fouriertransformation blev introduceret i [66, afsnit 2.3.1], vil vi ikke blot henvise til afsnittet om fouriertransformation i [66]. Følgende afsnit kan læses alene, men kan også opfattes som et supplement til omtalte afsnit.

Sinusoidale bølger Fouriertransformation kan ses som en oversættelse fra tid/sted-domænet til "frekvens/amplitude"-domænet. Egentlig er "amplitude" en forkert betegnelse, da der for hver frekvens både beskrives amplituden af en cosinus-bølge og af en sinus-bølge med den aktuelle frekvens. Så "amplituden" skal her forstås som et komplekst tal. Ofte anvendes blot betegnelserne "tid-domæne" og "frekvens-domæne".

Når et signal beskrives i frekvens/amplitude-domænet, så beskrives det ved amplituderne af en række sinusoidale bølger med forskellige frekvenser. Det er vigtigt at bide mærke i, at bølgerne netop er beskrevet ved hjælp af formen på en sinus-bølge. Da bølgerne, som dannes på en guitarstreng eller i en orgelpibe, ideelt set har form af en række sinusoidale bølger (jævnfør [66, afsnit 2.1] om tonedannelse), vil en fouriertransformation af den ideelle guitarstrengs bevægelser netop vise, hvilke frekvenser der optræder på strengen. Signaler, som ikke er dannet ved summen af en række sinusoidale bølger, kan selvfølgelig også fouriertransformeres, men for sådanne signaler er det ikke nødvendigvis en fordel at repræsentere dem i frekvens/amplitude-domænet.

Man skal bemærke, at fouriertransformationen ikke foretages til et bestemt tidspunkt, og at et signal analyseres med stabile bølger¹. Så hvis et signal

¹Vi vil fremover anvende termen "stabil bølge" til at beskrive en bølge, der er dannet af

består af en sinus-bølge med en frekvens, som ikke er stabil, vil fouriertransformationen ikke umiddelbart kunne angive, hvornår frekvensen skifter. Fouriertransformationen angiver blot, hvilke stabile bølger der kan adderes og dermed give signalet.

Kompleks eksponentialfunktion De fleste er nok vandt til de trigonometriske funktioner: sinus og cosinus. Disse funktioner kan ved hjælp af Eulers formel beskrives med den komplekse eksponentialfunktion:

$$\begin{aligned}\cos(x) &= \frac{e^{ix} + e^{-ix}}{2} & \sin(x) &= \frac{e^{ix} - e^{-ix}}{2i} \\ e^{ix} &= \cos(x) + i\sin(x) & e^{-ix} &= \cos(x) - i\sin(x)\end{aligned}$$

Anvendes den komplekse notation, kan amplituden af en frekvens findes som den absolutte værdi, mens faseforskydningen kan findes ved hjælp af forholdet mellem den komplekse og de reelle del - dette er beskrevet i [66]. Vi anvender her den komplekse eksponentialfunktion til beskrivelse af sinusoidale bølger, da det gør notationen enklere.

Definition på fouriertransformation: Her gengives formel 2.1 fra [66] som en definition på fouriertransformationen. Der anvendes dog en anden notation, således at tiden, der integreres over, noteres ved τ , signalet er noteret som $\mathbf{s}$ og frekvensen noteres med ν :

$$\mathbf{S}(\nu) = \mathcal{F}\{\mathbf{s}\} = \int_{-\infty}^{\infty} \mathbf{s}(\tau) e^{-i2\pi\nu\tau} d\tau \quad (3.1)$$

Vi vil bemærke, at vi i det følgende vil benævne det sidste led altså $e^{-i2\pi\nu\tau}$ som en "analyse-bølge", og der vil forekomme betegnelser så som "analyserende bølge". Dette giver bedre mulighed for at skelne mellem signalets bølge, og de bølger, som indirekte anvendes, når signalet fouriertransformeres.

Herunder noteres den diskrete definition på fouriertransformationen, hvor der anvendes et fast interval mellem samples:

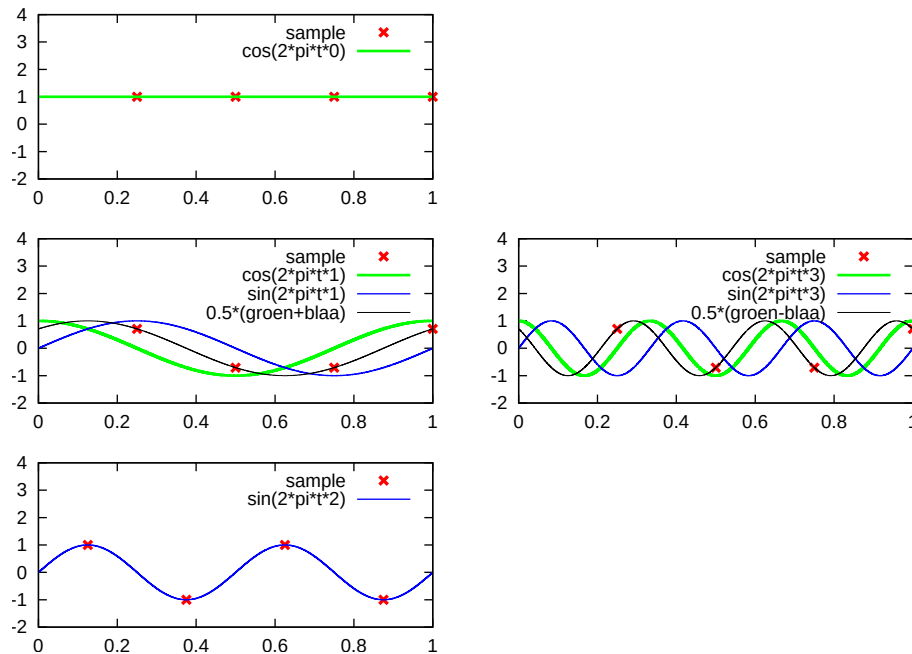
$$\mathbf{S}[k] = \mathcal{F}\{\mathbf{s}\} = \sum_{\eta=0}^{N-1} \mathbf{s}[\eta] e^{-i2\pi k\eta/N} \quad (3.2)$$

k løber fra 0 til $N - 1$ og angiver antallet af svingninger inden for de N samples. Husk på at der for at gengive et signal med $N/2$ svingninger kræves mindst N samples.² Så når man fouriertransformerer et reelt signal, er det egentlig kun de $N/2 + 1$ første frekvenser, som er interessante. (Der er egentlig tale om $N/2 + 1$ frekvenser, når N er lige, og $(N + 1)/2$ frekvenser, når N er ulige.) De sidste $N/2 - 1$ frekvenser svarer på grund af aliasing-effekten blot til en spejling af de første frekvenser. Det bevirker, at $\mathbf{s}[k] = \overline{\mathbf{s}[N - k]}$, for $k > N/2$, således at hver frekvens har en komplekst konjugeret makker. På figur 3.1 er dette forsøgt illustreret.

For at bestemme hvilken frekvens k svarer til, skal der ganges med samplingsfrekvensen og deles med antallet af samples: $\nu = k \frac{f_s}{N}$, hvor f_s er samplingsfrekvensen og angiver antallet af samples pr. tidsenhed.

sinusoidale bølger, som ikke ændrer frekvens i tiden.

²Se [66, afsnit 2.3.1.2] om nyquistfrekvensen.



Figur 3.1: Her er vist eksempler på fouriertransformationen, hvor der anvendes 4 samples. De 4 målinger ligger inden for et sek. og samplingsfrekvensen er altså 4 Hz. 3 gange er 4 samples blevet målt — først er der blevet målt på et signal med en frekvens på 0 Hz, dernæst med en frekvens på 1 Hz, og endelig med den maksimale frekvens på 2 Hz. Fouriertransformationen foretages i de 3 tilfælde, og da der er 4 samples anvendes der også 4 frekvenser, nemlig 0, 1, 2 og 3 Hz.

I det første tilfælde identificeres det, at frekvensen er 0 Hz, da en cosinus-bølge med en frekvens på 0 Hz passer perfekt til signalet. I det nederste tilfælde identificeres det, at frekvensen er 2 Hz, da en sinus-bølge med en frekvens på 2 Hz passer perfekt til signalet. I det midterste tilfælde er der tale om et signal, som er genereret ud fra en faseforskydning af sinus-bølger med en frekvens på 1 Hz. På grund af aliasing kan de 4 samples både passe til en sammensætning af sinusoidale bølger med en frekvens på 1 Hz og til en sammensætning af sinusoidale bølger med en frekvens på 3 Hz. Tilfældet med en frekvens på 3 Hz er vist til højre; her ses det, at sinus-bølgen skal trækkes fra cosinus-bølgen for at give bølgeformen, som går igennem de 4 samples. Dette illustrerer altså, at vægten af bølgen med en frekvens på 1 Hz (til venstre), kan genfindes som den komplekst konjugerede vægtning af bølgen med en frekvens på 3 Hz (til højre).

Den opmærksomme læser vil opdage, at de anvendte vægte i midten er dobbelt så store, som de burde være, men de valgte vægte illustrerer effekten godt.

Invers fouriertransformation: Som nævnt i [66] findes der også en invers fouriertransformation. Denne er faktisk meget intuitiv, da der blot er tale om en sum af sinusoidale bølger. Bølgerne er i princippet uendelig lange. Vægtningen af de enkelte bølger angives ved $\mathbf{S}$, og denne vægtning kunne f.eks. være fremkommet som fouriertransformationen af et signal.

Her noteres definitionen på den inverse fouriertransformation, hvor ν angiver den frekvens der integreres over:

$$\mathbf{s}(t) = \mathcal{F}^{-1}\{\mathbf{S}\} = \int_{-\infty}^{\infty} \mathbf{S}(\nu) e^{i2\pi\nu t} d\nu \quad (3.3)$$

På samme måde gengives den diskrete definition, hvor der anvendes et fast interval mellem samples:

$$\mathbf{s}[n] = \mathcal{F}^{-1}\{\mathbf{S}\} = \frac{1}{N} \sum_{k=0}^{N-1} \mathbf{S}[k] e^{i2\pi kn/N} \quad (3.4)$$

Bemærk at den inverse fouriertransformation (formel 3.3) kun adskiller sig fra fouriertransformationen (formel 3.1) ved fortegnsskiftet i eksponenten til e . I princippet kunne man godt have byttet om på fortegnene, for de har kun betydning for fasen. Vi skal senere se eksempler på en anden transformation (wavelet-transformationen), hvor fortegnene er byttet om.

3.1.1 Beregningskompleksitet

Vi ser på det diskrete tilfælde. Da fouriertransformationen beregnes over et signal af længden N , kræver beregningerne for hver frekvens N multiplikationer og additioner. Og da et signal med længde N giver N frekvenser, er den samlede beregningskompleksitet når alle frekvenser er beregnet $\mathcal{O}(N^2)$.

3.1.1.1 FFT

Der findes en række algoritmer, som betegnes ved FFT (Fast Fourier Transform), og fælles for disse er, at beregningskompleksiteten er $\mathcal{O}(N \log_2(N))$. I de fleste tilfælde henvises der med FFT til den mest kendte algoritme nemlig Cooley-Tukey algoritmen.

Cooley-Tukey Denne algoritme anvender i virkeligheden aliasing-fænomenet til at gøre brug af "del og hersk"-princippet. Vi skal gøre brug af at have lige og ulige samples, så vi kræver, at N (antallet af samples i signalet) er en potens af 2; dvs. $N = 2^a$, hvor a er et heltal. Vi opdeler nu summen i formel 3.2, således at de

lige og de ulige samples summeres hver for sig:

$$\begin{aligned}
 \mathbf{S}[k] &= \sum_{n=0}^{N-1} \mathbf{s}[n] e^{-i2\pi kn/N} \\
 &= \sum_{n=0}^{N/2-1} \left(\mathbf{s}[2n] e^{-i2\pi k(2n)/N} + \mathbf{s}[2n+1] e^{-i2\pi k(2n+1)/N} \right) \\
 &= \sum_{m=0}^{M-1} \mathbf{s}[2m] e^{-i2\pi k(2m)/N} + \sum_{m=0}^{M-1} \mathbf{s}[2m+1] e^{-i2\pi k(2m+1)/N} \quad (\text{hvor } M = N/2) \\
 &= \sum_{m=0}^{M-1} \mathbf{s}[2m] e^{-i2\pi km/M} + \sum_{m=0}^{M-1} \mathbf{s}[2m+1] e^{-i2\pi km/M - i\pi k/M} \\
 &= \sum_{m=0}^{M-1} \mathbf{s}[2m] e^{-i2\pi km/M} + e^{-i\pi k/M} \sum_{m=0}^{M-1} \mathbf{s}[2m+1] e^{-i2\pi km/M} \\
 &= \sum_{m=0}^{M-1} \mathbf{s}_e[m] e^{-i2\pi km/M} + e^{-i\pi k/M} \sum_{m=0}^{M-1} \mathbf{s}_o[m] e^{-i2\pi km/M} \quad (3.5)
 \end{aligned}$$

hvor signalet $\mathbf{s}_e$ består af de lige samples fra $\mathbf{s}$, og signalet $\mathbf{s}_o$ består af de ulige samples fra $\mathbf{s}$.

Lad os nu konstatere at $\mathbf{S}$, som er fouriertransformationen af $\mathbf{s}$, indeholder redundant data. Det reelle signal $\mathbf{s}$ har en længde på N samples, og derfor består $\mathbf{S}$ af N komplekse vægte for de N frekvenser. Men på grund af aliasing-effekten kan vægtene af de $N/2$ højeste frekvenser findes som spejlinger af vægtene for de $N/2$ laveste frekvenser. Som det fremgår af figur 3.1, er spejlingen dog lidt speciel, da det er den komplekst konjugerede som spejles.

Vi kan altså nøjes med at beregne fouriertransformationen for de $N/2$ første frekvenser, og så blot kopiere de fundne vægte til de $N/2$ sidste frekvenser. Så lad os antage, at vi kun er interesserede i at finde vægtene for de $N/2$ første frekvenser, således at k kun løber op til $(N-1)/2 = M-1$. Det ses nu, at de 2 summer i den sidste linje i formel 3.5 kan beregnes som fouriertransformationerne af de 2 signaler $\mathbf{s}_e$ og $\mathbf{s}_o$, da M er antallet af samples i de 2 signaler.

Vi kan opsummere beregningen af $\mathbf{S}$ ved hjælp af de 2 delsignaler $\mathbf{s}_e$ og $\mathbf{s}_o$ i følgende ligning:

$$\mathbf{S}[k] = \begin{cases} \mathbf{S}_e[k] + e^{-i\pi k/M} \mathbf{S}_o[k] & \text{for } k < M \\ \overline{\mathbf{S}[2M-1-k]} & \text{for } k \geq M \end{cases} \quad (3.6)$$

hvor N er antallet af frekvenser i $\mathbf{S}$, $M = N/2$, k løber fra 0 til $2M-1$, $\mathbf{S}_e$ er fouriertransformationen af de lige samples $\mathbf{s}_e$, og $\mathbf{S}_o$ er fouriertransformationen af de ulige samples $\mathbf{s}_o$.

Vi har altså opdelt beregningen af fouriertransformationen i 2 mindre delberegninger, som hver især består af fouriertransformationer af signaler med den halve længde. Hvis vi ser den samlede fouriertransformation som det problem, der skal løses, kan vi altså løse problemet ved at løse delproblemerne rekursivt.

Lad os antage at vi beregner fouriertransformationen ved hjælp af den rekursive løsning. Lad os dernæst antage, at vi har løst rekursionen på et bestemt niveau. Det interessante er nu at undersøge, hvor mange beregninger der kræves, for at gå et niveau op i rekursionen. Vi undersøger antallet af beregninger ved en forgrening. Ifølge formel 3.6 kræves der for de første M frekvenser M additioner og M multiplikationer med ledet $e^{-i\pi k/M}$, og for de

sidste M frekvenser kræves der M kopieringen; så den samlede beregningskompleksitet er $\mathcal{O}(M)$, hvor M er antallet af samples for den aktuelle gren på det aktuelle niveau i rekursionen. Ingen grene deler samples på det samme niveau. På hvert niveau anvendes der i alt N samples, hvor N er det totale antal samples i signalet. Så beregningskompleksiteten på hvert niveau i rekursionen er $\mathcal{O}(N)$. Da antallet af grene fordobles på hvert niveau, man bevæger sig ned i rekursionen, er det samlede antal niveauer $\log_2(N)$, og derfor er den samlede beregningskompleksitet for FFT $\mathcal{O}(N \log_2(N))$.

Praktisk anvendelse I praksis anvendes ofte FFTW-biblioteket³ - FFTW står for “Fastest Fourier Transform in the West”. I biblioteket anvendes algoritmer, som ikke kun har en lav beregningskompleksitet, når N er en potens af 2. I [7, kapitel 14] afprøves effektiviteten af Matlab’s indbyggede FFT, som anvender FFTW-biblioteket, og der sammenlignes med Bluestein’s FFT-algoritme, som også har en beregningskompleksitet på $\mathcal{O}(N \log_2(N))$, når N er et primtal. Konklusionen er, at FFTW-bibliotekets algoritme er hurtigere og mere præcis end Bluestein’s FFT-algoritme.

3.1.2 Vinduesbegrænset fouriertransformation

Her vil vi tilføje et vindue til fouriertransformationen, således at ikke alle tidspunkter vægter lige meget. Vinduet noteres som funktionen $\mathbf{w}$, og for denne gælder det, at den er større end eller lig 0, samt at den er absolut integrabel: $\mathbf{w}(t) \geq 0$ og $\int |\mathbf{w}(t)| dt < \infty$. Dette giver anledning til følgende definitioner på den kontinuerte og den diskrete vinduesbegrænsede fouriertransformation (Windowed Fouriertransform):

$$\mathbf{wf}\{\mathbf{s}\}(\nu) = \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}(\tau) e^{-i2\pi\nu\tau} d\tau \quad (3.7)$$

$$\mathbf{wf}\{\mathbf{s}\}[k] = \sum_{\eta=0}^{N-1} \mathbf{s}[\eta] \mathbf{w}[\eta] e^{-i2\pi k\eta/N} \quad (3.8)$$

I ovenstående diskrete formel for den vinduesbegrænsede fouriertransformation vægter nogle tidspunkter som sagt højere end andre tidspunkter. Som oftest anvendes et vindue, som kun er større end 0 i et lille interval. Det betyder, at der reelt set kun er ganske få samples, som tæller med ved beregningen af den vinduesbegrænsede fouriertransformation. Man kan sige, at et vinduet begrænser hvilket interval i signalet, som den vinduesbegrænsede fouriertransformation beregnes for. Lad os benævne længden af det interval, hvori vinduet er større end 0, med K . Det vil være fornuftigt kun at beregne den vinduesbegrænsede fouriertransformation for K forskellige frekvenser, da der reelt set kun indgår K samples ved beregningerne.

3.2 STFT

STFT står for Short Time Fourier Transform og kan betragtes som en videreudvikling af den vinduesbegrænsede fouriertransformation, hvor der anvendes et

³FFTW har hjemmeside på <http://www.fftw.org>.

vindue, som flyttes i tiden.⁴ Ved at lade vinduet forskyde i tiden, kan man altså bestemme, hvilke frekvenser et signal består af som funktion af tiden. STFT kan defineres som:

$$\text{stft}\{\mathbf{s}\}(t, \nu) = \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}(t - \tau) e^{-i2\pi\nu\tau} d\tau \quad (3.9)$$

hvor t er tiden, ν er frekvensen, $\mathbf{s}$ er signalet og $\mathbf{w}$ er vinduet. Der integreres over tiden, hvilket er angivet med τ . I integrationen spejles vinduet og flyttes med tiden t . Det normale er nok at opfatte vinduet som nært knyttet til signalet, således at vinduet begrænser signalet, men som vi skal se i et senere afsnit, kan det godt være en fordel at opfatte vinduet og den komplekse eksponentialfunktion, som noget der hænger tæt sammen.

Den diskrete STFT kan defineres som:

$$\text{stft}\{\mathbf{s}\}[n, k] = \sum_{\eta=0}^{N-1} \mathbf{s}[\eta] \mathbf{w}[n - \eta] e^{-i2\pi k\eta/N} \quad (3.10)$$

I den diskrete repræsentation angiver n , hvilken sample STFT ønskes beregnet for. k angiver antallet af svingninger og løber op til $N - 1$, hvor N angiver antallet af samples i signalet. Der summeres over η , som løber gennem alle samples. Vinduet flyttes n samples ved beregningen af STFT for den n 'te sample, og vinduets længde kan angives ved M . Vinduets længde er i første omgang uinteressant i forhold til formlen, men da vinduesfunktionen er lig 0 for samples uden for vinduet, har vinduets længde betydning for den faktiske frekvensopløsning og dermed muligheden for klart at adskille 2 frekvenser fra hinanden.

På figur 3.2 og 3.3 er de forskellige dele for udregningen af STFT illustreret. Ved at beregne STFT til forskellige tider og ved forskellige frekvenser kan man danne et 3-dimensionel overflade, hvor styrken af STFT afbildes som funktion af tiden og frekvensen. Et sådant billede kaldes for et spektrogram, og vi vil se eksempler herpå senere.

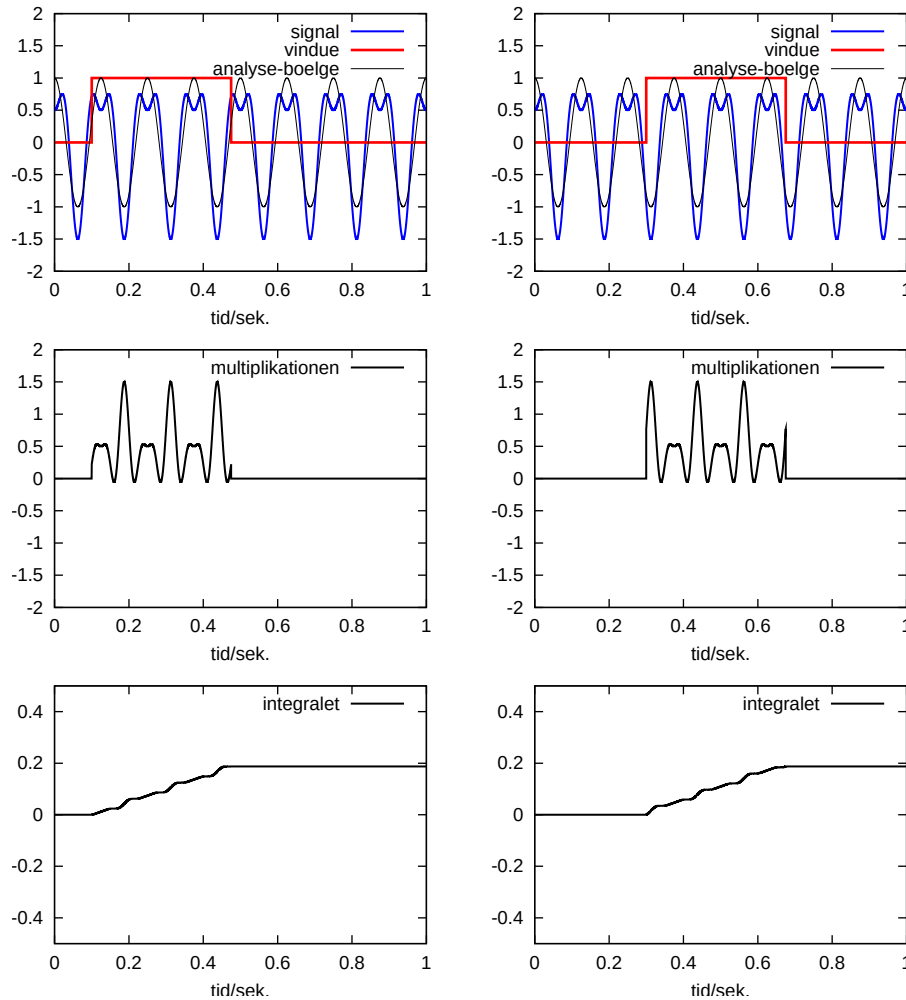
Vindue og frekvensopløsning: Vinduet og frekvensopløsningen hænger tæt sammen. Lad os se på den diskrete definition (formel 3.10) og bemærke, at vinduets længde bestemmer, hvor mange samples der reelt bidrager til analysen af signalet, da alle samples uden for vinduet multipliceres med 0. Selv om der anvendes N forskellige analyserende frekvenser, er den faktiske frekvensopløsning ikke bedre, end hvad der opnås med M analyserende frekvenser, hvor M angiver længden af vinduet. Derfor vil man som regel udregne den diskrete fouriertransformation over det antal samples, som er i vinduet:

$$\text{stft}\{\mathbf{s}\}[n, k] = \sum_{\eta=0}^{M-1} \mathbf{s}[\eta] \mathbf{w}[n - \eta] e^{-i2\pi k\eta/M} \quad (3.11)$$

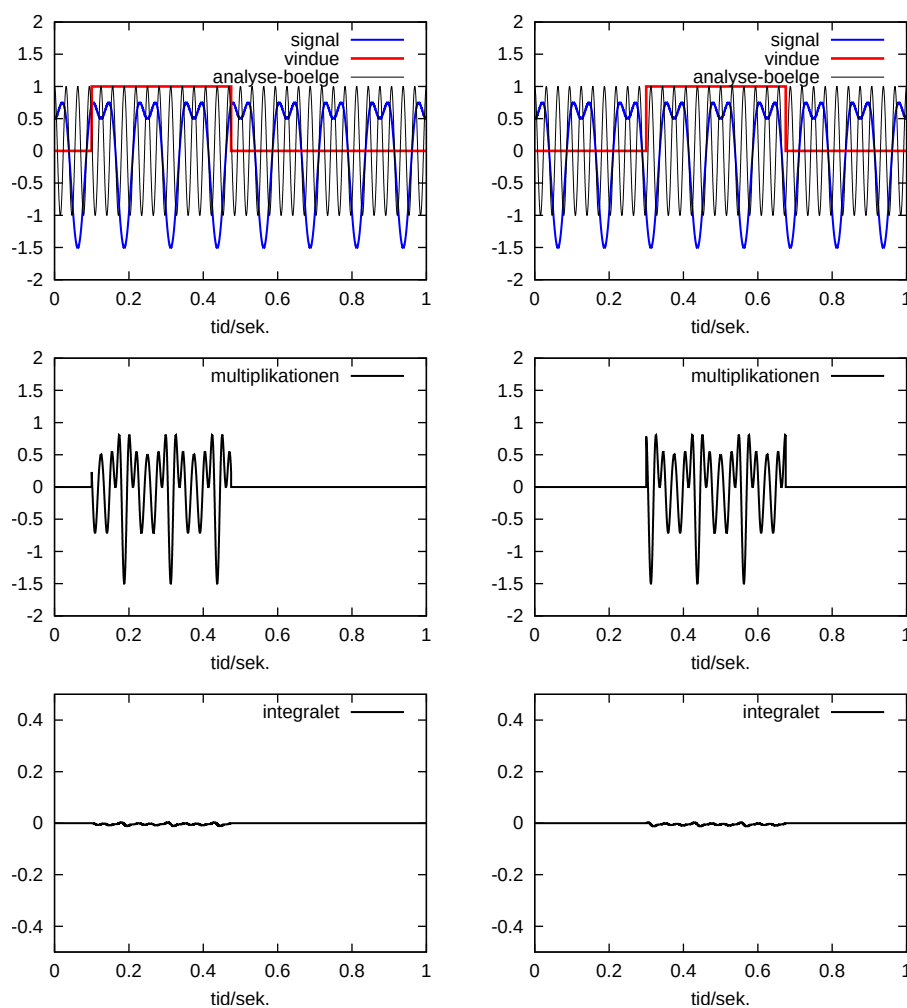
hvor M angiver vinduets længde, og k løber op til $M - 1$.

Da den sidste halvdel af frekvenserne blot er en spejling af den første halvdel af frekvenserne (jævnfør figur 3.1), er det reelt set kun $M/2$ frekvenser, der kan bestemmes. Bemærk at der er lige stor afstand mellem hver frekvens, ligesom der er lige stor afstand mellem hver sample. Man er altså nødt til at

⁴Ifølge [1, side 5] anvendes der også andre navne for denne metode, inklusive navnet "heterodyne filter analysis".



Figur 3.2: De 2 kolonner illustrerer beregningen af STFT med et glidende vindue. I den øverste række er vist 2 figurer, hvor et signal, en analyserende bølge, samt et vindue er afbildet. Forskellen på de 2 figurer er blot, at vinduet i figuren til højre er flyttet et stykke. Det signal, der bliver analyseret, er sammensat af 2 bølger: en bølge med en frekvens på 8 Hz og en bølge med en frekvens på 16 Hz. Den analyserende bølge har en frekvens på 8 Hz. I anden række er vist resultatet af multiplikationen af signalet, vinduet og den analyserende bølge. I tredje række er vist integralet af resultatet fra anden række. Man kan altså se, at resultatet af beregningen af STFT til 2 forskellige tider (men med den samme analyserende bølge) giver et resultat på omkring 0,2. Der er anvendt en vindueslængde, som passer med perioderne af de bølger, som signalet består af. Havde der været anvendt en anden vindueslængde, ville de 2 resultater ikke nødvendigvis have været ens. Men da man kan se, at integralet inden for vinduet tilnærmelsesvist stiger lineært, ville de 2 resultater have været tilnærmelsesvist ens, forudsat at der var anvendt en anden vindueslængde.



Figur 3.3: Denne figur er en kopi af figur 3.2, men her er der anvendt en analyserende bølge med en frekvens på 32 Hz; dvs. der er anvendt en analyserende bølge, hvis frekvens ikke er i blandt de frekvenser, som signalet består af. Resultatet i nederste række viser da også, at integralet ligger og svinger meget nær 0.

gå på kompromis: Hvis det er vigtigt, at den tidsmæssige opløsning skal være god, må vinduet højst dække et lille antal samples - men dette bevirker, at der også kun kan bestemmes et lille antal frekvenser; omvendt kræver en god frekvensopløsning mange samples og derved en dårlig tidsmæssig bestemmelse.

Zero-padding og hop-size I det diskrete tilfælde (formel 3.11) vil man ofte kunstigt forbedre estimeringen af et signals frekvens ved at forlænge vinduet med en række 0'er, således at fouriertransformationen beregnes for flere frekvenser end det antal samples, der er i den anvendte del af signalet. Se [66, afsnit 2.3.1.5] for en forklaring på dette trick, som kaldes for "zero-padding". Til at angive, hvor meget vinduet flytter sig i tiden mellem hver fouriertransformation i STFT anvendes begrebet "hop-size". Se [66, afsnit 2.3.1.4] hvor dette begreb forklares.

3.2.1 Vinduer

I dette afsnit skal vi se nærmere på de mest populære vinduer. Ind til videre har vi blot omtalt vinduer som noget, der begrænser det tidsrum, fouriertransformationen beregnes over. Generelt set må man vælge mellem en god opløsning i tiden eller en god opløsning i frekvens-domænet. I forbindelse med analyse af musik er det især frekvensopløsningen, som er kritisk. Derfor vil vi i det følgende primært koncentrere os om de forskellige vinduers udseende i frekvens-domænet. Lad os begynde med det mest simple, nemlig et rektangulært vindue.

3.2.1.1 Rektangulært vindue

Det rektangulære vindue har den samme vægt inden for det interval, hvori vinduet er åbent. Vi vil nu analysere effekten af at bruge et sådant vindue, og vi beregner derfor fouriertransformationen af et signal, som er begrænset af et rektangulært vindue. For at gøre notationen enkel, lader vi vinduet være centreret omkring $t = 0$. Vi vil betegne vinduet med $\mathbf{w}_R$ for at indikere, at der er tale om det rektangulære vindue.

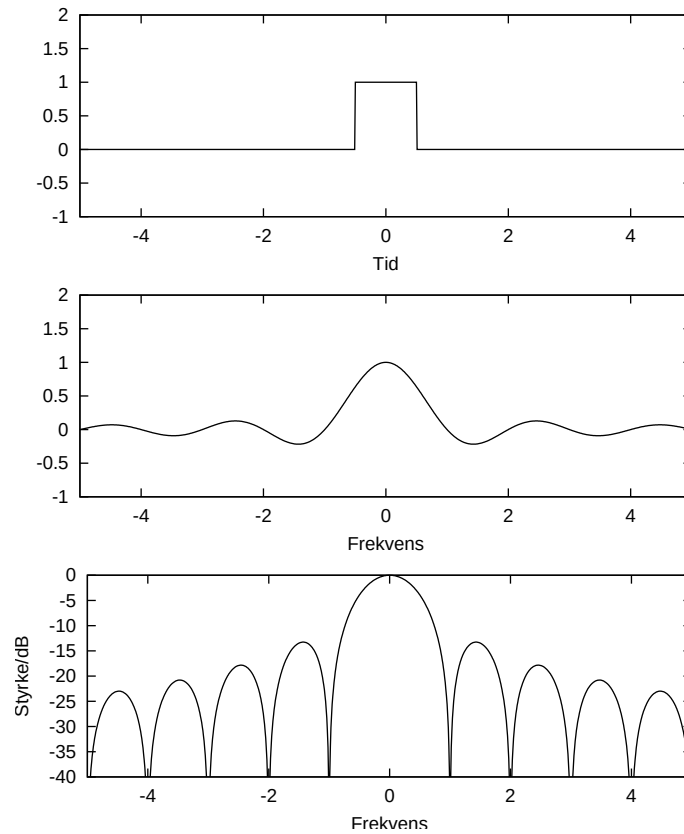
$$\mathbf{S}_{\mathbf{w}_R}(\nu) = \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}_R(\tau) e^{-2\pi i \nu \tau} d\tau \quad \text{hvor} \quad \mathbf{w}_R(\tau) = \begin{cases} 1 & \text{for } |\tau| \leq T/2 \\ 0 & \text{for } |\tau| > T/2 \end{cases} \quad (3.12)$$

Ifølge foldningssætningen⁵ er fouriertransformationen af et produkt lig foldningen af en fouriertransformation af hvert led. Dette anvender vi og får derved:

$$\mathbf{S}_{\mathbf{w}_R}(\nu) = \mathbf{S}(\nu) * \mathbf{W}_R(\nu) \quad (3.13)$$

hvor $\mathbf{S}$ er fouriertransformationen af signalet, og $\mathbf{W}_R$ er fouriertransformationen af vinduet.

⁵Foldningssætningen blev introduceret i [66, afsnit 2.3.3.1].



Figur 3.4: Her er vist fouriertransformationen af et rektangulært vindue. Øverst vises vinduet repræsenteret i tidsdomænet. I midten er vinduet repræsenteret i frekvensdomænet. Vinduet har en længde på 1 tidsenhed, så repræsentationen i frekvensdomænet krydser 0 for hver hele frekvens, undtagen frekvensen 0. Nederst er styrken for de forskellige frekvenser afbildet på en decibel-skala.

Fouriertransformationen af det rektangulære vindue noteres herunder:

$$\begin{aligned}
 \mathbf{W}_R(\nu) &= \mathcal{F}\{\mathbf{w}_R\} = \int_{-\infty}^{\infty} \mathbf{w}_R(\tau) e^{-2\pi i \nu \tau} d\tau \\
 &= \int_{-T/2}^{T/2} e^{-2\pi i \nu \tau} d\tau \\
 &= \left[\frac{1}{-2\pi i \nu} e^{-2\pi i \nu \tau} \right]_{\tau=-T/2}^{T/2} \\
 &= \frac{1}{-2\pi i \nu} e^{-2\pi i \nu T/2} - \frac{1}{-2\pi i \nu} e^{2\pi i \nu T/2} \\
 &= \frac{1}{\pi \nu} \frac{e^{\pi i \nu T} - e^{-\pi i \nu T}}{2i} \quad (\text{ved hjælp af Eulers formel fås nu}) \\
 &= \frac{\sin(\pi \nu T)}{\pi \nu} \\
 &= T \text{sinc}(\pi \nu T) \quad (\text{hvor } \text{sinc}(x) = \sin(x)/x) \tag{3.14}
 \end{aligned}$$

På figur 3.4 er denne funktion vist. Da det mest interessante ved effekten af vinduet er, at det ændrer på vægten af de forskellige frekvensers amplituder, og da det menneskelige øre tilnærmelsesvist opfatter styrken af en lyd afbildet på en logaritmisk skala, er funktionen også vist på den logaritmiske decibel-skala⁶.

Når der foldes med det rektangulære vindue, er det altså ikke kun én

⁶Der henvises til [66, afsnit 2.4.1.1] for en forklaring på decibel.

frekvens, som vil fremstå i resultatet, selv om signalet kun består af én frekvens. Resultatet af foldningen vil være identisk med afbildningen i figur 3.4, blot centreret omkring den frekvens, som signalet består af. Så effekten af at anvende et rektangulært vindue er, at energien, som burde være koncentreret i én frekvens, bliver spredt ud over uendeligt mange frekvenser.

Sinus-funktionen krydser 0 regelmæssigt, og derfor vil frekvensrepræsentationen af det rektangulære vindue også krydse 0 regelmæssigt — se formel 3.14 og figur 3.4. Områderne mellem hver gang frekvensrepræsentationen krydser 0 kaldes på engelsk for “lobe”. Den primære lobe omkring frekvensen 0 kaldes for “main lobe” eller “main beam”, mens de øvrige lobes hver for sig kaldes for en “side lobe”. Generelt ønsker man at maksimum for hver side lobe er så lav som mulig, således at energien bliver spredt minimalt. Det kan især være et problem, hvis maksimum for hver side lobe aftager langsomt, for det bevirker, at en kraftig frekvens kan komme til at “dække” over en svag frekvens, som befinder sig langt fra den kraftige frekvens.

Det er både et problem at main lobe er bred, således at energien lokalt set bliver spredt ud, og at maksimum for hver side lobe er høj, således at energien også spredes globalt set. Ved at anvende et filter, som koncentrerer al energien i netop én frekvens, kan vi i den ideelle verden undgå problemet med, at energien bliver spredt ud. Vi ønsker at anvende den omvendte operation af en foldning — altså en invers foldning, eller som det kaldes på engelsk: deconvolution. Lad os nu se på et eksempel, hvor en sinusbølge er blevet multipliceret med et rektangulært vindue, og resultatet er blevet fouriertransformeret. Vi ønsker ud fra denne fouriertransformation at finde fouriertransformationen af sinusbølgen. Vi har altså $\mathbf{W}_R * \mathbf{S}$ og ønsker at finde $\mathbf{S}$. Vi starter med at anvende foldningssætningen:

$$\begin{aligned}\mathbf{W}_R * \mathbf{S} &= \mathcal{F}^{-1} \{ \mathcal{F} \{ \mathbf{W}_R \} \mathcal{F} \{ \mathbf{S} \} \} \\ \mathcal{F} \{ \mathbf{W}_R * \mathbf{S} \} &= \mathcal{F} \{ \mathbf{W}_R \} \mathcal{F} \{ \mathbf{S} \} \\ \frac{\mathcal{F} \{ \mathbf{W}_R * \mathbf{S} \}}{\mathcal{F} \{ \mathbf{W}_R \}} &= \mathcal{F} \{ \mathbf{S} \} \\ \mathcal{F}^{-1} \left\{ \frac{\mathcal{F} \{ \mathbf{W}_R * \mathbf{S} \}}{\mathcal{F} \{ \mathbf{W}_R \}} \right\} &= \mathbf{S}\end{aligned}\tag{3.15}$$

Det ser altså umiddelbart ud som om det skulle være muligt at genskabe $\mathbf{S}$ ved hjælp af 2 fouriertransformationer og en invers fouriertransformation, men der kan være et problem, hvis der divideres med 0. Da $\mathbf{W}_R$ er en lige funktion, dvs. $\mathbf{W}_R(\nu) = \mathbf{W}_R(-\nu)$, er fouriertransformationen lig den inverse fouriertransformation; og da vi kender $\mathbf{w}_R$ og dermed $\mathcal{F}^{-1} \{ \mathbf{W}_R \}$, kender vi altså også $\mathcal{F} \{ \mathbf{W}_R \}$, som dermed er lig $\mathbf{w}_R$. Og om $\mathbf{w}_R$ gælder det, at $\mathbf{w}_R = 0$ for alle værdier uden for vinduet. Derfor er det ikke muligt at foretage divisionen, og derfor er det ikke muligt at genskabe $\mathbf{S}$. I praksis kunne man måske fastsætte at det rektangulære vindue ikke er lig 0, men har en positiv værdi nær 0. Da ville divisionen kunne lade sig gøre, men der ville stadig være et problem, for ved divisionen bliver små numeriske usikkerheder ved $\mathcal{F} \{ \mathbf{W}_R \}$ blæst op og kan komme til at overskygge signalet.

3.2.1.2 Foldede rektangulære vinduer

Vi så i forrige afsnit, at det ikke er muligt at slippe af med side lobes, når der anvendes et vindue, som ikke er uendeligt bredt. Det, som skaber de markante side lobes ved det rektangulære vindue, er diskontinuiteten ved randen til

vinduet. Så hvis man minimerer diskontinuiteten, eller helt fjerner denne, er der håb om at maksimum for hver side lobe vil aftage hurtigere.

Vi vil starte med at lave det rektangulære vindue om til et vindue, som har form af en triangel, således at der ikke opstår en diskontinuitet ved randen til vinduet. Bemærk at det triangulære vindue kan dannes som en foldning af det rektangulære vindue med sig selv. Det betyder, at det er meget let at bestemme det triangulære vindues udseende i frekvensdomænet, for ved at anvende foldningssætningen ses det, at der blot er tale om det rektangulære vindue ganget med sig selv. Det triangulære vindue $\mathbf{W}_T$ kan i frekvensdomænet skrives som:

$$\begin{aligned}\mathbf{W}_T(\nu) &= \mathcal{F}\{\mathbf{w}_T(t)\} \\ &= \mathcal{F}\{\mathbf{w}_R(t) * \mathbf{w}_R(t)\} \\ &= \mathcal{F}\left\{\mathcal{F}^{-1}\left\{\mathcal{F}\{\mathbf{w}_R(t)\}^2\right\}\right\} \\ &= \mathcal{F}\{\mathbf{w}_R(t)\}^2 = \mathbf{W}_R^2(\nu)\end{aligned}\tag{3.16}$$

Problemet er blot, at det triangulære vindue (set i tidsdomænet) er dobbelt så bredt som det rektangulære vindue. For at skabe et triangulært vindue, som er lige så bredt som det oprindelige rektangulære vindue, er det altså nødvendigt at generere det triangulære vindue ud fra foldningen af 2 rektangulære vinduer, som har den halve bredde af det oprindelige rektangulære vindue. Da den halve bredde i tidsdomænet fører til den dobbelte bredde i frekvensdomænet, vil main lobe for det triangulære vindue altså være dobbelt så bred som for det rektangulære vindue. På figur 3.5 er det triangulære vindue afbildet i den anden række. Her ses det, at når der er tale om det triangulære vindue set i forhold til det rektangulære vindue, så har den første side lobe maksimum på et lavere niveau og maksimum for de følgende side lobes aftager hurtigere.

Det triangulære vindue kan selvfølgelig også foldes med det rektangulære vindue og dermed danne et vindue, som i frekvensdomænet fremstår som det rektangulære vindue opløftet i tredje. I nederste række på figur 3.5 er det resulterende vindue vist. Det ses, at den første side lobe har maksimum på et endnu lavere niveau, men bredden af main lobe er også steget til det tredobbelte.

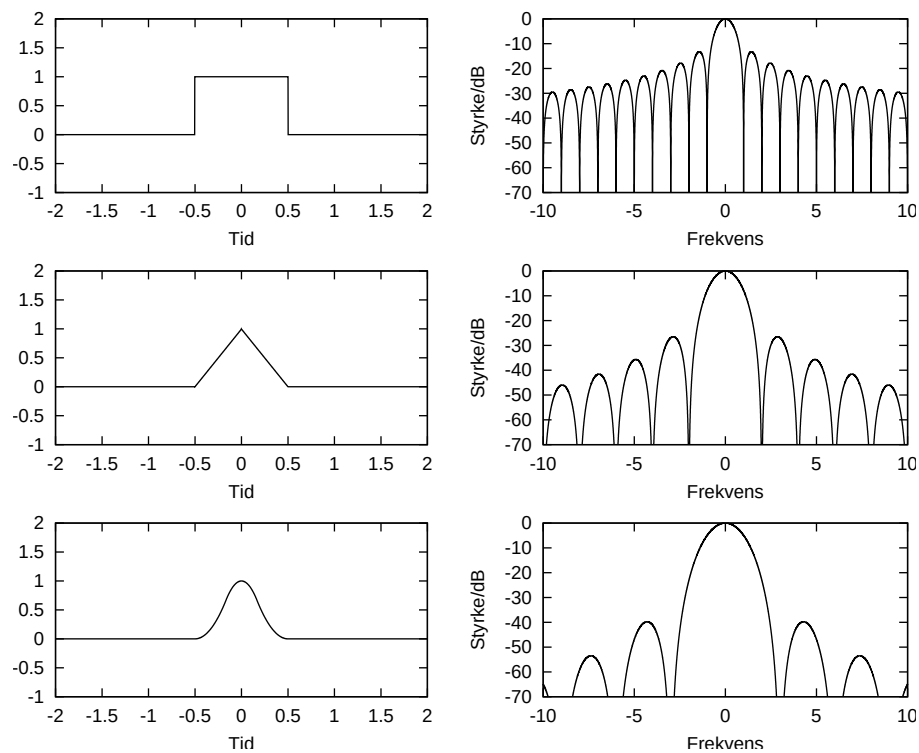
I tidsdomænet kan vinduer, som er dannet ud fra foldninger med det rektangulære vindue, beskrives ved B-splines. Formel 3 i [64] angiver ved hjælp af B-splines udseendet af et vindue af grad m (hvor m angiver antallet af foldninger). Generelt set kan man fortsætte foldningerne uendeligt og på denne måde generere et vindue, som kan beskrives ved uendeligt mange B-splines. Når antallet af foldninger går mod uendelig, tilnærmer funktionen sig en meget velkendt funktion nemlig Gauss-funktionen.

3.2.1.3 Gauss-vindue

Vi lader en unormaliseret Gauss-funktion, som er centreret omkring 0 være defineret ved:

$$\mathbf{g}(t) = e^{-t^2/(2\sigma^2)}\tag{3.17}$$

hvor σ bestemmer bredden på funktionen. Normalt er integralet af en Gauss-funktion lig 1, men her har vi ikke brug for denne egenskab og anvender derfor



Figur 3.5: I øverste række vises det rektangulære vindue i tids- og frekvens-domænet. I anden række vises et triangulært vindue, som kan dannes ved foldningen af 2 rektangulære vinduer, der er halvt så brede som det rektangulære vindue i øverste række. I den nederste række er et triangulært vindue foldet med et rektangulært vindue, hvilket giver en blødere kurve, der kan beskrives ved sammensætningen af 3 andengradspolynomier. Det ses, at main-lobe i det nederste tilfælde er blevet meget bred, hvilket kan være et problem, hvis der ønskes høj frekvensopløsning.

ovenstående forenklede definition på en Gauss-funktion. Fra sandsynlighedsregningens normalfordeling er det velkendt, at omkring 95% af integralet af Gauss-funktionen findes i intervallet fra $t = -2\sigma$ til $t = 2\sigma$.

Den fouriertransformerede af en Gauss-funktion er også en Gauss-funktion. Herunder noteres den fouriertransformerede til den Gauss-funktion, som blev defineret i formel 3.17:

$$\begin{aligned} \mathbf{G}(\nu) &= \mathcal{F}\{\mathbf{g}(t)\}(\nu) = \int_{-\infty}^{\infty} e^{-t^2/(2\sigma^2)} e^{-i2\pi\nu t} dt \\ &= \sqrt{2\pi\sigma} e^{-2\pi^2\nu^2\sigma^2} \end{aligned} \quad (3.18)$$

Gauss-funktionen kan i teorien godt bruges som et vindue i sig selv, men da den ikke er begrænset i tiden, er det i praksis ikke muligt at bruge Gauss-funktionen alene som vinduesfunktion.⁷ I praksis vil man være nødt til at gøre noget ved Gauss-funktionen, således at den bliver lig 0 uden for et vist interval, fordi det kun er muligt at regne på signaler med endelige længder. Der kan gives flere forslag til, hvordan Gauss-funktionen kan modificeres så den bliver lig 0 uden for et vist interval. Den metode, som nok er mest normal, er at Gauss-funktionen ganges med et rektangulært vindue. Vi vil her definere et Gauss-vindue, som en Gauss-funktion der ganges med et rektangulært vindue:

$$\mathbf{w}_G(t) = \mathbf{g}(t)\mathbf{w}_R(t) \quad (3.19)$$

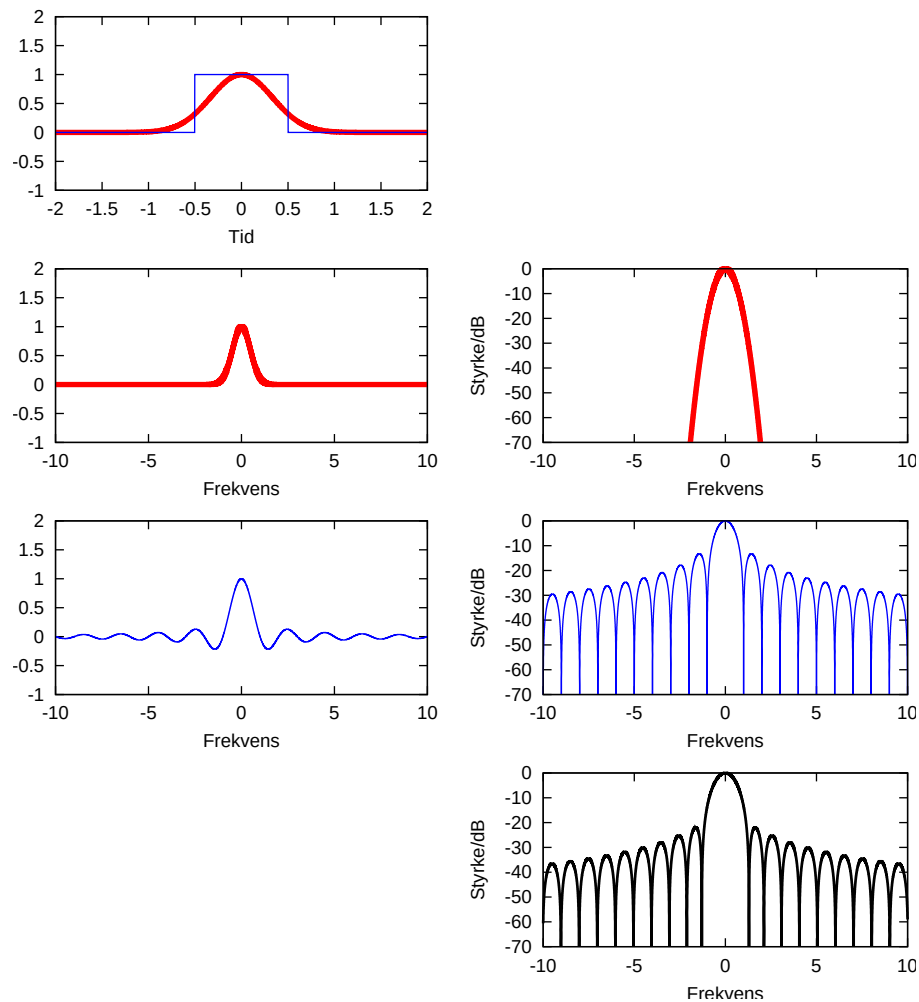
Når Gauss-funktionen ganges med det rektangulære vindue, ændres integralet, men det er ligegyldigt i forbindelse med frekvensanalysen. Med det Gauss-vindue, som er defineret i formel 3.19, opstår der en diskontinuitet ved grænsen til vinduet. Ved at minimere bredden af Gauss-funktionen, kan man minimere effekten af denne diskontinuitet, men en Gauss-funktion, som er smal i tidsdomænet, er bred i frekvensdomænet, så bredden af main lobe blive stor. På figur 3.6 og 3.7 er der vist eksempler, hvor der er anvendt Gauss-funktioner med forskellige bredder. Afhængig af anvendelsen må man vælge den bredde, som passer bedst.

Det kan gives mange bud på, hvordan problemet med diskontinuiteten ved kanten til vinduet kan mindskes; her vil vi se på et enkelt simpelt forsøg i form af et såkaldt sænket Gauss-vindue. Man kan sænke Gauss-funktionen, således at den krydser 0 ved kanten til det rektangulære vindue. Da fouriertransformationen af en sum er lig summen af fouriertransformationen af hvert led, er det enkelt at bestemme det sænkede Gauss-vindue i frekvensdomænet:

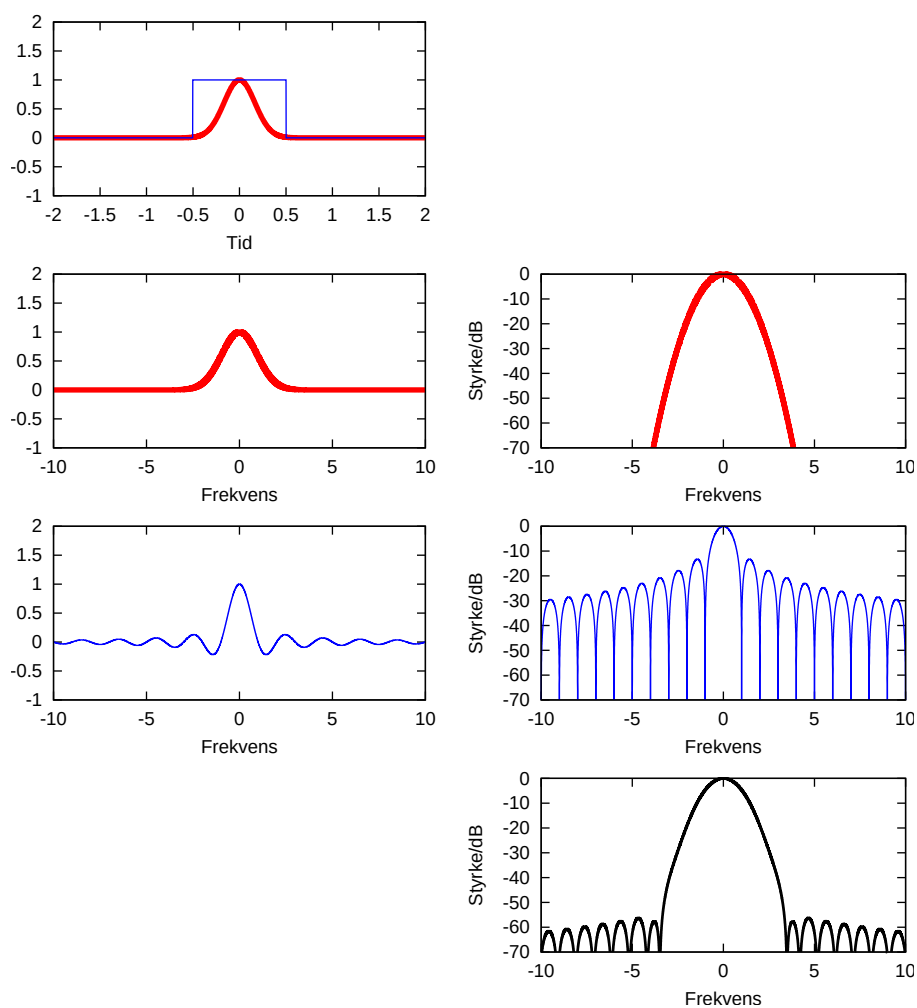
$$\begin{aligned} \mathbf{W}_G(\nu) &= \mathcal{F}\{(\mathbf{g}(t) - a)\mathbf{w}_R(t)\}(\nu) \\ &= \mathcal{F}\{\mathbf{g}(t)\mathbf{w}_R(t) - a\mathbf{w}_R(t)\}(\nu) \\ &= \mathcal{F}\{\mathbf{g}(t)\mathbf{w}_R(t)\}(\nu) - a\mathcal{F}\{\mathbf{w}_R(t)\}(\nu) \\ &= (\mathbf{G}(\nu) * \mathbf{W}_R(\nu)) - a\mathbf{W}_R(\nu) \end{aligned} \quad (3.20)$$

hvor $a = \mathbf{g}(T/2)$ og T er bredden af det rektangulære vindue.

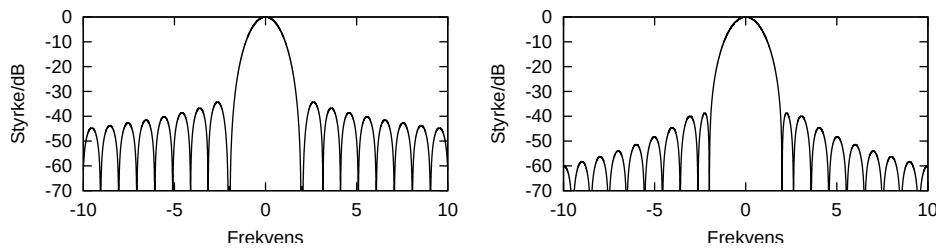
⁷En STFT, som bruger Gauss-funktionen som et uendeligt langt vindue, kaldes også for en Gabor-transformation. Ifølge [46] blev Gabor-transformationen introduceret i 1940'erne af Dennis Gabor. Der er kilder, som ikke bider mærke i, at Gauss-funktionen bruges ved Gabor-transformationen; således bemærkes det i [67], at STFT og "windowed Fourier transform" bruges som andre navne for Gabor-transformationen. I denne opgave har vi altså valgt at anvende betegnelsen vinduesbegrænset fouriertransformation, når vinduet ikke flyttes i tiden.



Figur 3.6: Her illustreres effekten af brugen af en gaussisk funktion sammen med det rektangulære vindue. Øverst er et rektangulært vindue samt en gaussisk funktion afbildet i tidsdomænet. Det rektangulære vindue har en bredde på 1 tidsenhed, mens den gaussiske funktion (forstået som en normalfordeling) har en standard-afvigelse på $1/3$. I den anden række vises det gaussiske vindue i frekvens-domænet — på figuren til højre er der anvendt en decibel-skala. Bemærk at den gaussiske funktion har den samme form, når den afbildes i tidsdomænet og i frekvensdomænet, samt at den gaussiske funktion aldrig er lig 0, så den er ikke tidsbegrænset. I den tredje række er det rektangulære vindue repræsenteret i frekvensdomænet. Nederst er den gaussiske funktion ganget med det rektangulære vindue og vises i frekvensdomænet. Det ses, at effekten af at gange det rektangulære vindue med den gaussiske funktion er, at main lobe bliver lidt bredere, samt at den første side lobe har et lavere maksimum.



Figur 3.7: Denne figur svarer til figur 3.6, men her er anvendt en gaussisk funktion med en standard-afvigelse på $1/6$. En sammenligning af de 2 nederste figurer til højre viser, at maksimum for hver side lobe ligger meget lavere end i tilfældet med det rektangulære vindue, men bredden af main lobe er steget betydeligt.



Figur 3.8: Her vises et eksempel på anvendelsen af det sænkede Gauss-vindue. Til venstre vises et almindeligt Gauss-vindue i frekvensdomænet. Til højre er Gauss-vinduet blevet sænket, således at det er lig 0 ved kanten til vinduet. Det ses, at main lobe er knap så spids på figuren til højre, og at maksimum for hver side lobe aftager hurtigere, når der anvendes en sænket gaussisk funktion.

På figur 3.8 vises det sænkede Gauss-vindue i frekvensdomænet. Det ses, at den primære effekt i frekvensdomænet af anvendelsen af det sænkede Gauss-vindue er, at maksimum for hver side lobe falder hurtigere, end hvis der anvendes et almindeligt Gauss-vindue.

Optimal tid/frekvens-opløsning Gabor [18] henviser til kvantefysikken, når han berører usikkerhedsprincippet, som fortæller, at nøjagtigheden, hvormed en relation mellem tid og frekvens kan bestemmes, er begrænset således, at $\Delta t \Delta f \leq \frac{1}{2}$. Gabor konstaterer, at Gauss-funktionen (som han kalder sandsynlighedsfunktionen) optimerer nøjagtigheden, og han bemærker, at denne egenskab ved Gauss-funktionen ikke ser ud til at have været registreret tidligere. Gauss-funktionen ganget med en sinusoidal bølge kalder Gabor for et elementærsignal, og med henvisning til det minimale tids/frekvens-område ved brug af dette elementærsignal omtaler Gabor elementærsignalet som en naturlig basis for tids/frekvens-analyse. En tids/frekvens-repræsentation, hvor et signal repræsenteres ved hjælp af disse elementærsignaler, foreslår Gabor at kalde for en "logon". Så Gabors logon svarer stort set til det, som vi nu om stunder kalder for et spektrogram.

3.2.1.4 Hann-vindue

Vi så i forrige afsnit, at det kunne være en fordel at anvende det sænkede Gauss-vindue, da det ikke er diskontinuert ved kanten til vinduet. Her præsenteres det vindue, som nok er det mest populære, når der beregnes en STFT, og dette vindue er heller ikke diskontinuert ved randen. Vinduet består af en Hann-funktion h , som ganges med det rektangulære vindue. Hann-funktionen kan tilpasses, så den netop er lig 0 ved grænsen til vinduet.

Hann-funktionen er defineret i tidsdomænet som:

$$h(t) = \frac{1}{2} \left(1 + \cos \left(\frac{\pi t}{a} \right) \right) = \cos^2 \left(\frac{\pi t}{2a} \right) \quad (3.21)$$

hvor a angiver den halve afstand mellem 2 nul-punkter. Ved at sætte $a = T/2$, hvor T angiver bredden på det rektangulære vindue og gange Hann-funktionen med det rektangulære vindue, fås et nyt vindue, som kaldes Hann-vinduet. Ligesom i tilfældet med Gauss-vinduet, vil vi nu undersøge Hann-vinduet ved at se på foldningen af Gauss-funktionen og det rektangulære vindue.

Først fouriertransformerer vi Hann-funktionen — dette er enkelt, da Hann-funktionen består af en forhøjet cosinus-funktion:

$$\begin{aligned}
 \mathbf{H}(\nu) &= \mathcal{F}\{\mathbf{h}\} = \int_{-\infty}^{\infty} \mathbf{h}(\tau) e^{-2\pi i \nu \tau} d\tau \\
 &= \int_{-\infty}^{\infty} \cos^2\left(\frac{\pi \tau}{2a}\right) e^{-2\pi i \nu \tau} d\tau \\
 &= \int_{-\infty}^{\infty} \left(\frac{e^{i\pi \tau/(2a)}}{2} + \frac{e^{-i\pi \tau/(2a)}}{2} \right)^2 e^{-2\pi i \nu \tau} d\tau \quad (\text{Eulers regel}) \\
 &= \int_{-\infty}^{\infty} \left(\frac{e^0}{2} + \frac{e^{i\pi \tau/a}}{4} + \frac{e^{-i\pi \tau/a}}{4} \right) e^{-2\pi i \nu \tau} d\tau \\
 &= \frac{1}{2} \mathbf{dirac}(\nu) + \frac{1}{4} \mathbf{dirac}\left(\nu - \frac{1}{2a}\right) + \frac{1}{4} \mathbf{dirac}\left(\nu + \frac{1}{2a}\right) \quad (3.22)
 \end{aligned}$$

hvor **dirac** er Dirac-delta-funktionen⁸.

Lad os nu igen notere fouriertransformationen af det rektangulære vindue, og anvende at $a = T/2$:

$$\begin{aligned}
 \mathbf{W}_R(\nu) &= T \mathbf{sinc}(\pi \nu T) \\
 &= 2a \mathbf{sinc}(2a \pi \nu) \quad (3.23)
 \end{aligned}$$

Da foldning med Dirac-delta-funktionen blot svarer til en translation, er det enkelt at udføre foldningen med de 3 Dirac-delta-funktioner, som den fouriertransformerede Hann-funktion består af:

$$\begin{aligned}
 \mathbf{W}_H(\nu) &= \mathbf{H}(\nu) * \mathbf{W}_R(\nu) \\
 &= \frac{1}{2} 2a \mathbf{sinc}(2a \pi \nu) + \frac{1}{4} 2a \mathbf{sinc}(2a \pi (\nu - \frac{1}{2a})) + \frac{1}{4} 2a \mathbf{sinc}(2a \pi (\nu + \frac{1}{2a})) \\
 &= a \left(\mathbf{sinc}(2\pi \nu a) + \frac{1}{2} \mathbf{sinc}(2\pi a \nu - \pi) + \frac{1}{2} \mathbf{sinc}(2\pi a \nu + \pi) \right) \quad (3.24)
 \end{aligned}$$

På figur 3.9 er Hann-vinduet vist i både tid- og frekvens-domænet. Her ses det, at maksimum for hver side lobe aftager meget hurtigt; så når der anvendes et Hann-vindue, kan man næsten sige, at effekten af én stærk frekvens i et signal kun har meget lokal betydning i frekvensdomænet — hvilket altså ikke er tilfældet med det rektangulære vindue.

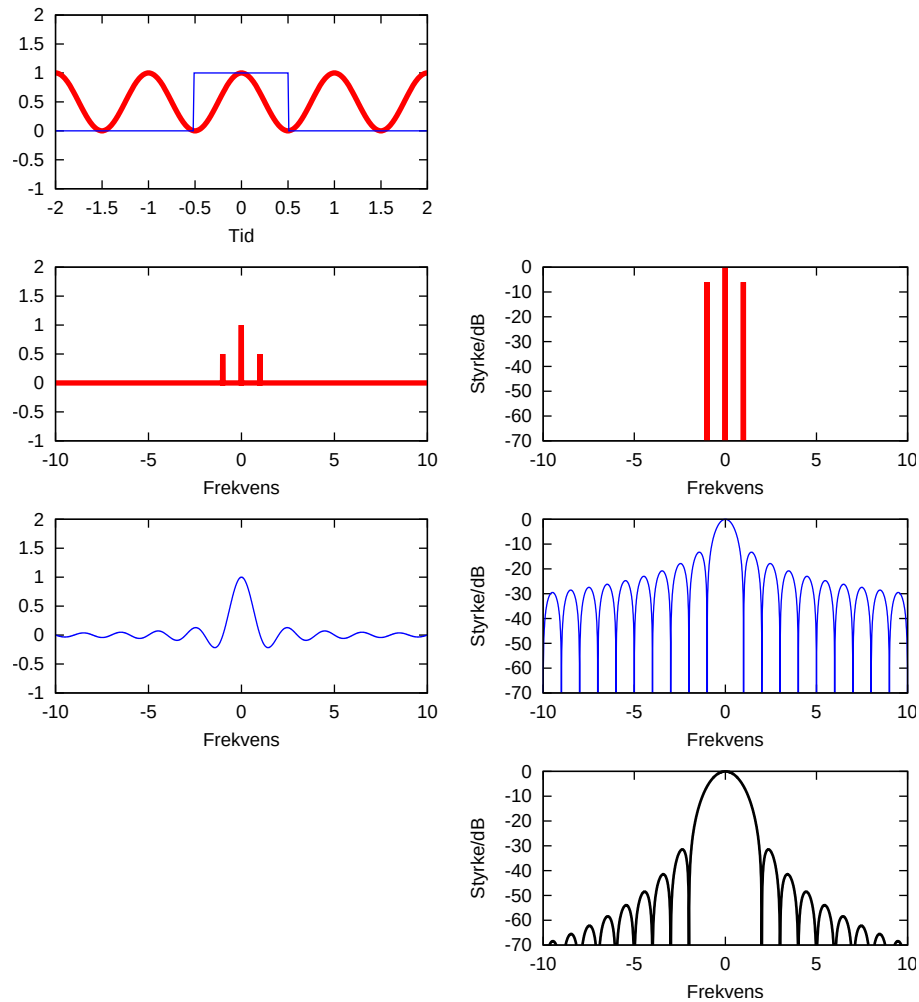
3.2.1.5 Hamming-vindue

Hamming-vinduet vil vi kun ganske kort notere, da det ifølge formelen for Hamming-vinduet minder om Hann-vinduet, men grafisk set kan det minde mere om Gauss-vinduet. Hamming-vinduet kan defineres som:

$$\mathbf{w}_{\text{hamming}}(t) = \mathbf{w}_R \left(\alpha + (1 - \alpha) \cos\left(\frac{\pi t}{a}\right) \right) \quad \text{hvor: } \alpha = 0,53836 \quad (3.25)$$

Ligesom med Hann-funktionen består Hamming-funktionen af 3 Dirac-delta-funktioner, og derfor består også Hamming-vinduet i frekvensdomænet af 3 **sinc**-funktioner.

⁸Se forklaringen af **dirac**-funktionen i afsnit 1.5.2.



Figur 3.9: Øverst er det rektangulære vindue samt Hann-funktionen vist i tidsdomænet. På anden række er Hann-funktionen vist i frekvensdomænet — til højre er der anvendt en decibel-skala. På tredje række er det rektangulære vindue vist i frekvensdomænet. På nederste række vises det frekvensmæssige udseende af Hann-vinduet, som i frekvensdomænet fremkommer som foldningen af Hann-funktionen og det rektangulære vindue. Det ses, at Hann-vinduet har en bred main lobe, men maksimum for hver side lobe aftager meget hurtigt. I tilfældet med Hann-vinduet er energifordelingen altså i højere grad lokal.

3.2.1.6 Sammenligning af vinduer

Der findes et hav af kendte vinduer, som hver især har deres specielle egenskaber. Men generelt må man vælge mellem en bred main lobe eller høje maksimum for hver side lobe. Nogle vinduer fokuserer på, at maksimum for hver side lobe aftager så hurtigt som muligt, mens andre fokuserer på, at maksimum for hver side lobe befinder sig på et konstant lavt niveau.⁹

Gauss-vinduet er nok primært populært inden for billed-analyse, men det bliver også anvendt indenfor lyd-analyse, og nogle af de metoder, vi senere skal se på, anvender vinduet.

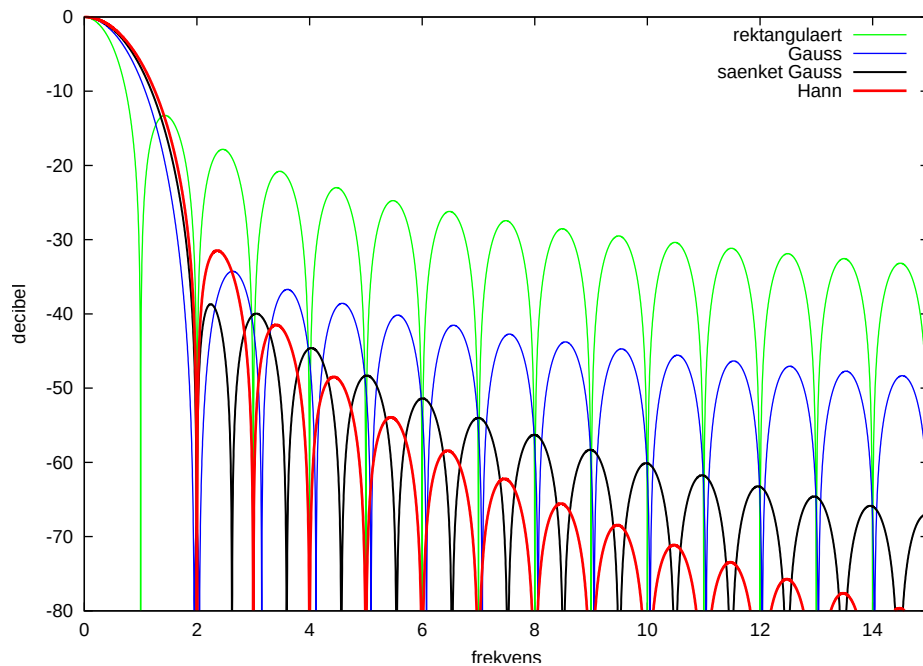
Hann-vinduet udmærker sig ved at maksimum for hver side lobe hastigt aftager, og derudover er det enkelt at folde med vinduet i frekvensdomænet. Hamming-vinduet er lige så enkelt at folde med i frekvensdomænet, men maksimum for hver side lobe aftager langsommere. Til gengæld er Hamming-vinduets første side lobes maksimum lavere end Hann-vinduets første side lobes maksimum.

På figur 3.10 er vist eksempler på nogle af de vinduer, vi i de foregående afsnit stiftede bekendtskab med. Her ses forklaringen på, hvorfor Hann-vinduet er så populært, da maksimum for hver side lobe falder markant hurtigere, end det er tilfældet for de øvrige vinduer.

På figur 3.11, 3.12 og 3.13 illustreres brugen af Hamming-vinduet, Hann-vinduet, Gauss-vinduet og det sænkede Gauss-vindue. På den første figur (figur 3.11) er et signal, som kun indeholder en frekvens på 666 Hz, blevet fouriertransformeret. Signalet har en længde på 350 samples, og der anvendes en samplingsfrekvens på 44.100 Hz. De forskellige kurver viser for hvert vindue styrken af hver frekvens efter fouriertransformationen. Det ses, at når der anvendes et Hamming-vindue eller et Gauss-vindue, så ser det ud som om signalet indeholder frekvenser med en styrke på omkring -60 dB i nærheden af 10.000 Hz. Frekvenser, som befinder sig langt væk fra de 666 Hz, bliver altså i en vis grad påvirket.

På den anden figur (figur 3.12) er en optagelse af en operasanger blevet fouriertransformeret med de 4 forskellige vinduer. Optagelsen har også en længde på 350 samples og operasangeren synger en tone, hvis fundamental-frekvens er på omkring 666 Hz. Ud over fouriertransformationen af det aktuelle signal, vises også fouriertransformationen af et signal, som kun indeholder en frekvens på 666 Hz. På denne måde illustreres det, i hvor høj grad den meget kraftige fundamental-frekvens påvirker muligheden for at identificere de forskellige overtoner. På figuren kan man se, at overtonerne som ligger omkring 5.000 - 6.000 Hz i tilfældet med Hamming-vinduet og Gauss-vinduet ligger på samme niveau som side lobes fra fundamental-frekvensen. Det betyder, at det i disse 2 tilfælde kan være meget svært at afgøre, om der er tale om overtoner, eller om hvert lokalt maksimum blot skyldes en side lobe fra fundamental-frekvensen. Man kan sige, at den kraftige fundamental-frekvens med de anvendte vinduer kommer til at overskygge de højere overtoner. I tilfældet med Hann-vinduet og det sænkede Gauss-vindue er der ikke de samme problemer, og her er det meget

⁹Som eksempel på vinduer, hvor maksimum for hver side lobe aftager hurtigt, kan $\cos^\alpha(x)$ -familien nævnes. Et medlem af denne familie har vi stiftet bekendtskab med, da tilfældet, hvor $\alpha = 2$, også kaldes for Hann-vinduet. Som eksempel på et vindue, hvor maksimum for hver side lobe holder sig på et konstant lavt niveau, kan nævnes Dolph-Chebyshev vinduet, hvor maksimum for hver side lobe aftager med 0 dB. Se i øvrig [25] for en oversigt over hvor hurtigt maksimum for hver side lobe aftager for forskellige vinduer.



Figur 3.10: Her vises 4 vinduer i frekvensdomænet. Der er tale om det rektangulære vindue, et almindeligt Gauss-vindue, et sænket Gauss-vindue og det populære Hann-vindue. Hamming-vinduet er ikke taget med, da det minder meget om et Gauss-vindue med den bredde, som er anvendt her. De 2 Gauss-vinduer er tilpasset således, at bredden af main lobe er 2 ligesom for Hann-vinduet.

Hvis vi ser bort fra det rektangulære vindue, så klarer det sænkede Gauss-vindue sig bedst ved den første side lobe. Desuden ligger maksimum for side lobe for det sænkede Gauss-vindue under maksimum for det normale Gauss-vindues tilsvarende side lobe. Hvis man sammenligner main lobe for de 2 Gauss-vinduer, ses det, at det sænkede Gauss-vindues main lobe er knap så spids.

Hann-vinduet klarer sig i (forhold til de 2 Gauss-vinduer) dårligst ved den første side lobe, men allerede ved den anden side lobe klarer Hann-vinduet sig lige så godt som det sænkede Gauss-vindue. Derefter falder maksimum for hver side lobe markant hurtigst for Hann-vinduet. Hann-vinduets main lobe har næsten samme udseende som det sænkede Gauss-vindues main lobe, men er knap så spids.

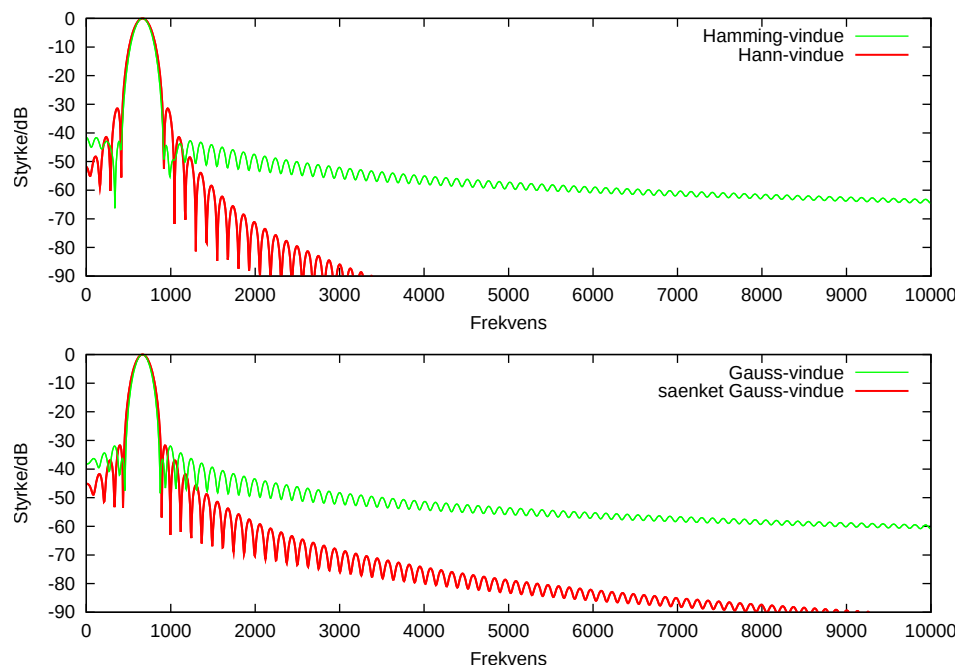
lettere at identificere hver overtone.

Den tredje figur (figur 3.13) er en fortsættelse af den foregående figur, og her vises frekvenserne fra 10.000 til 20.000 Hz. Det ses, at overtonerne over 14.000 Hz i tilfældet med Hamming-vinduet og Gauss-vinduet forsvinder fuldstændig i "støj" fra fundamentalfrekvensens side lobes.

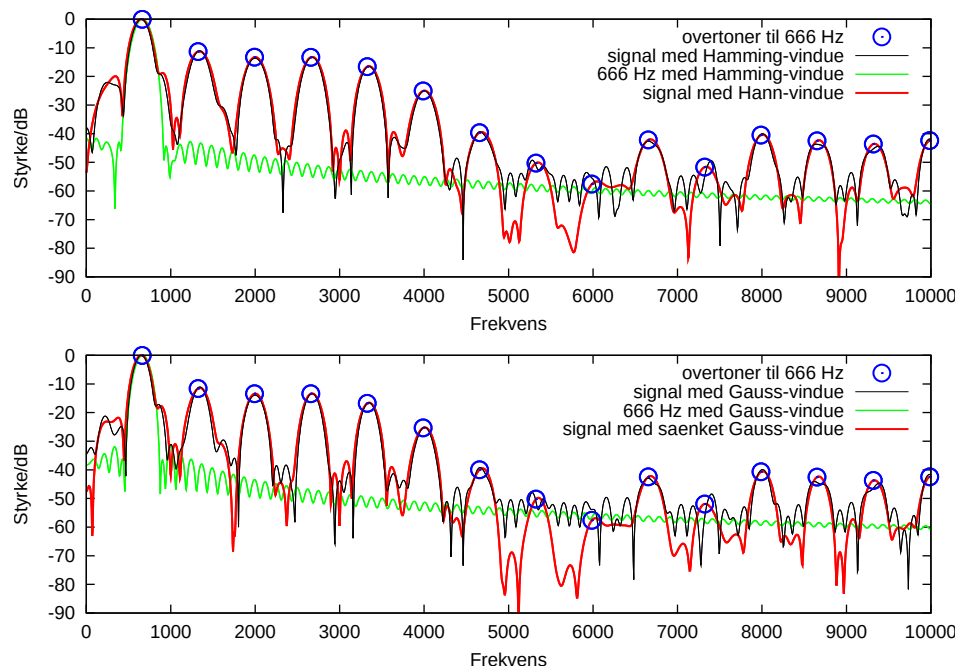
Et andet populært vindue er Blackman-Harris-vinduet. Dette har vi desværre ikke stiftet nærmere kendskab til, men ligesom Hann-vinduet kan Blackman-Harris-vinduet tilføjes i frekvensdomænet ved at folde det rektangulære vindue med nogle Dirac-delta-funktioner. I [25] foreslås Blackman-Harris-vinduet sammen med det såkaldte Kaiser-Bessel-vindue som de bedste vinduer.

3.2.1.7 ENB

Ved sammenligning af vinduer repræsenteret i frekvensdomænet anvendes ofte ENB eller ENBW, som står for Equivalent Noise Bandwidth. Set ud fra frekvensdomænet ønsker man, at styrken af hver frekvenskomponent i et signal forstyrrer de øvrige frekvenskomponenter så lidt som muligt. Set ud fra én frekvenskomponent medfører de øvrige frekvenskomponenter blot støj. Det kan



Figur 3.11: Her vises forskellen på brugen af Hann-vinduet og Hamming-vinduet, samt Gauss-vinduet og det sænkede Gauss-vindue. Øverst vises fouriertransformationen af en sinus-bølge med en frekvens på 666 Hz. Der er anvendt 350 samples ved en samplingsfrekvens på 44.100 Hz. Det ses, at Hamming-vinduet i frekvensdomænet flader hurtigere ud end Hann-vinduet og stadig ligger omkring -60 dB ved 10.000 Hz. Nederst er der anvendt et Gauss-vindue og et sænket Gauss-vindue. Bredden på Gauss-vinduerne er valgt således, at deres maksimale side løbe er lige så høj, som den maksimale side løbe hos Hann-vinduet. Med de valgte bredder er Gauss-vinduernes main lobe smallere end Hann-vinduets og Hamming-vinduets.



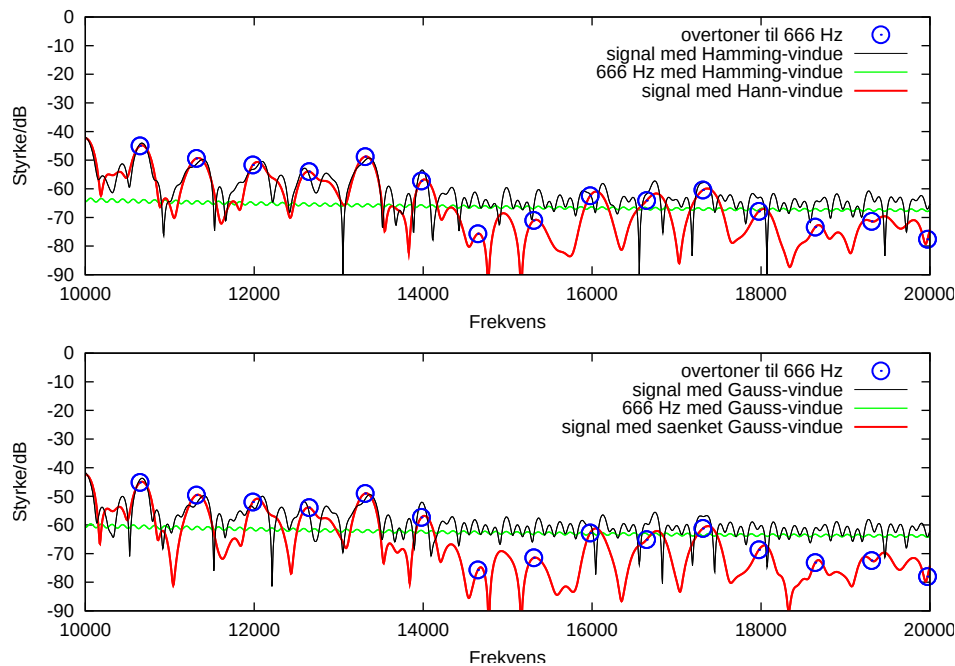
Figur 3.12: Denne figur er en fortsættelse af den foregående figur. Her anvendes de vinduer, som blev vist på foregående figur, men i dette tilfælde er signalet en lydoptagelse af en operasanger, som synger en tone med en fundamentalfrekvens på omkring 666 Hz. Der er også et svagt akkompagnement i baggrunden. (Lydklippet, som varer i 0,008 sekunder svarende til 350 samples ved 44.100 Hz, starter 1,36 sekunder inde i den testfil, som hedder "opera_fem2", som er en del af ADC2004 datasættet.

Øverst viser den røde kurve fouriertransformationen, hvor der er anvendt et Hann-vindue. Den sorte tyde kurve viser fouriertransformationen, hvor der er anvendt et Hamming-vindue. Sangeren synger en tone med en fundamentalfrekvens på omkring 666 Hz, og for denne samt hver teoretisk overtone er der vist en blå ring på den røde kurve. Det ses, at de blå ringe generelt findes på toppunkter på den røde kurve. Med den grønne kurve er vist fouriertransformationen, hvor der er brugt et Hamming-vindue, af en sinus-kurve med samme frekvens som sangerens fundamentalfrekvens. Det ses, at fundamentalfrekvensen er så kraftig, at den kommer til at dække over frekvenser i området omkring 6.000 Hz. Så når der i forbindelse med det anvendte signal bruges et Hamming-vindue, kan man ikke drage nogen konklusioner angående styrken af frekvenserne omkring 6.000 Hz, da det ikke er muligt at skelne signalet fra den støj, som skyldes den kraftige fundamentalfrekvens på 666 Hz. Det ses også, at resultaterne af de 2 transformationer er meget forskellige i netop det berørte område, men meget ens uden for dette område.

Nederst er de 2 Gauss-vinduer anvendt. Det ses, at det almindelige Gauss-vindue har samme problemer som Hamming-vinduet. Ved at gøre Gauss-vinduet bredere, kan man mindske problemet, men det skal være meget bredt, før problemet næsten kan siges at være forsvundet. Det sænkede Gauss-vindue udviser ligesom Hann-vinduet ikke nogen problemer med side lobes i området omkring 6.000 Hz.

Fouriertransformationerne ved hjælp af Hann-vinduet og det sænkede Gauss-vindue er meget ens, men det ser ud, som om der er lidt flere detaljer i tilfældet med det sænkede Gauss-vindue.

På næste figur vises frekvensintervallet fra 10.000 til 20.000 Hz.



Figur 3.13: Denne figur er en forlængelse af den foregående. Her vises frekvensintervallet fra 10.000 Hz til 20.000 Hz. Det ses, at Hamming-vinduet og det almindelige Gauss-vindue ikke er brugbare for frekvenserne over 14.000 Hz, da støjen fra fundamentalfrekvensen på 666 Hz overskygger signalet.

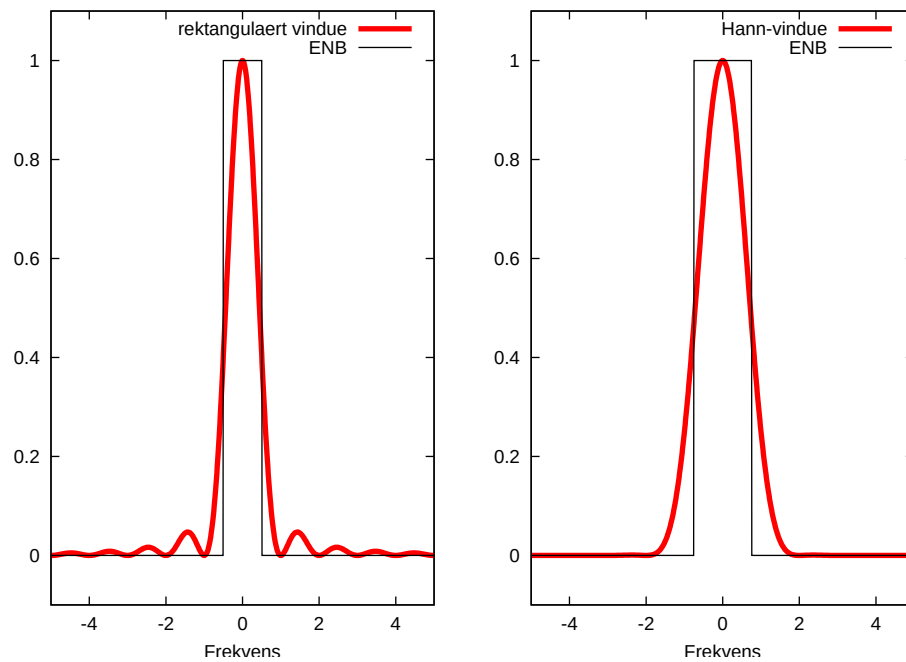
være brugbart med et mål, som gør det muligt at sammenligne forskellige vinduers evne til frekvensmæssigt set at adskille støj og signal.

ENB angiver bredden af et frekvensbånd, hvori alle frekvenserne er lige kraftige, og uden for dette frekvensbånd er frekvensernes styrke lig 0. Frekvensbåndet kan være fremkommet ved anvendelsen af et vindue, som i frekvensdomænet er rektangulært, og vinduet anvendes på et signal, som kun indeholder én frekvens. Når man anvender et sådant vindue, kan man altså kun adskille 2 frekvenskomponenter klart, hvis de befinder sig i en frekvensmæssig afstand, som er større end båndbredden af vinduet. Jo smallere vinduets båndbredde er, des bedre frekvensopløsning opnås.

Til sammenligning af forskellige vinduer i frekvensdomænet samles den samlede styrke af vinduet i et rektangulært område, som er lige så højt som vinduets main lobe. Når styrken er samlet i det rektangulære område, angiver bredden af rektanglet netop ENB. Eller sagt på en anden måde: For et aktuelt vindue angiver ENB båndbredden af et vindue, som er rektangulært i frekvensdomænet, og i frekvensdomænet er dette vindues højde lig maksimum for det aktuelle vindues main lobe, og integralet af vinduets styrke er lig integralet af styrken af det aktuelle vindue.

Alle de vinduer, vi har set på, har i princippet uendelig udstrækning i frekvensdomænet, så det kan forekomme lidt mystisk at tale om en båndbredde. Dette kan afhjælpes ved anvendelsen af ENB.

Figur 3.14 illustrerer i frekvensdomænet sammenhængen mellem et vindue og dets tilhørende ENB. På figuren vises ENB for det rektangulære vindue samt for Hann-vinduet. De 2 vinduers normaliserede styrke er vist som funktion af frekvensen. I stedet for blot at angive ENB for de 2 vinduer, er der vist nogle vinduer, som er rektangulære i frekvensdomænet, og som har en båndbredde på ENB, samt en højde, som er lig maksimum for det aktuelle vindues main lobe.



Figur 3.14: Her vises ENB for det rektangulære vindue og Hann-vinduet. ENB er bredden af de 2 kasser, som skal forestille vinduer, der er rektangulære i frekvensområdet.

Det rektangulære vindue, som vises til venstre, har en smal main lobe og samtidig en smal ENB. Hann-vinduet, som vises til højre, har derimod en bred main lobe og en bred ENB. Men selv om ENB er større for Hann-vinduet, så påvirker det stort set ikke frekvenser, som ligger langt væk. På figuren ses det, at Hann-vinduet er næsten lig 0 for frekvenser, som ligger over 2 frekvenser væk, mens dette ikke er tilfældet for det rektangulære vindue. For det rektangulære vindue er en stor del af støjen henlagt i de side lobes, som ligger langt væk fra main lobe.

Som det noteres i [25], kan det omtalte integral samt maksimum for main lobe findes i tidsdomænet. Her vil vi notere en definition på ENB, hvor vi i modsætning til i [25] ikke anvender diskrete tidsskidt:

$$\text{enb}\{\mathbf{w}\} = \frac{\int_{-\infty}^{\infty} \mathbf{w}^2(\tau) d\tau}{\left(\int_{-\infty}^{\infty} \mathbf{w}(\tau) d\tau\right)^2} \quad (3.26)$$

I tilfældene med de 2 vinduer, som er vist på figur 3.14, kan man anvende ovenstående formel og derved finde at ENB for det rektangulære vindue er 1 og at ENB for Hann-vinduet er 1,5.¹⁰ I [25] kan man finde en tabel, som blandt andet angiver ENB for mange af de mest anvendte vinduer.

ENB kan være ganske anvendelig som mål for et vindues båndbredde, men er primært nyttig, hvis de forskellige frekvenskomponenter i et signal har stort set samme styrke. Dette er ikke tilfældet med frekvenskomponenterne i polyfon musik. Det er ganske normalt, at frekvenskomponenterne, som befinder langt fra de kraftigste frekvenskomponenter, er meget svage. Derfor er det meget vigtig, at en kraftig frekvenskomponent ikke overskygger en svag frekvenskomponent, som ligger lang væk. ENB fortæller ikke noget om, hvordan et vindue påvirker frekvenserne langt væk fra en aktuel frekvens, og derfor er ENB et knap så nyttigt mål ved sammenligning af vinduer, som skal bruges til analyse af polyfon musik.

3.2.2 Beregningskompleksitet

Vi så i afsnit 3.1.1, at beregningskompleksiteten for fouriertransformationen er $\mathcal{O}(N \log_2(N))$. Når man beregner en STFT, svarer det til at tage adskillige fouriertransformationer af korte signaler, og derfor må den samlede beregningskompleksitet som udgangspunkt være $\mathcal{O}(MN \log_2(N))$, hvor M er antallet af tidspunkter, som fouriertransformationen udregnes til, og N er antallet af samples, som indgår i hver fouriertransformation, og dermed længden af det anvendte vindue.

Men da der ofte anvendes overlap i forbindelse med STFT, således at afstanden mellem hvert tidspunkt, som fouriertransformationerne beregnes for, er mindre en længden af det tidsrum, fouriertransformationerne beregnes over, er der en del samples, som indgår ved beregningen af flere fouriertransformationer. Det kunne være anvendeligt, hvis dette genbrug af samples kunne føre til færre beregninger.

I [65] nævnes det, at der er blevet udviklet hurtige rekursive algoritmer (dvs. algoritmer, som anvender beregningerne fra ét tidspunkt til beregningerne ved næste tidspunkt). De rekursive algoritmer er specielt anvendelige, når hop-size er lig 1. Algoritmerne har en beregningskompleksitet på $\mathcal{O}(MN)$, hvor M er længden af signalet og dermed også angiver antallet af tidspunkter, som fouriertransformationen beregnes til, og N angiver antallet af samples, som anvendes ved beregningen af fouriertransformationen. I [65] anvendes et vindue, som er baseret på B-splines og tilnærmet en Gauss-funktion, men der findes også algoritmer for f.eks. Hamming-vinduet.

¹⁰Lad os anvende formel 3.26 til at bestemme ENB for Hann-vinduet. Vi integrerer fra $-1/2$ til $1/2$, da vi sætter $a = 1/2$ i formel 3.21. Først udregner vi tælleren i formel 3.26: $\int_{-1/2}^{1/2} \cos(\pi\tau) d\tau = [\tau/2 + \sin(2\pi\tau)/(4\pi)]_{-1/2}^{1/2} = 1/2$. Dernæst udregner vi nævneren: $\int_{-1/2}^{1/2} \cos^2(\pi\tau) d\tau = [\cos^3(\pi\tau)\sin(\pi\tau)/(4\pi) + 3(\tau/2 + \sin(2\pi\tau)/(4\pi))/4]_{-1/2}^{1/2} = 3/16$. Endelig kan vi derved udregne ENB: $\text{enb}\{\mathbf{w}_H\} = (3/8)/(1/2)^2 = 3/2$.

3.2.3 Tid/frekvens-repræsentation ved STFT

En tid/frekvens-repræsentation ved brug af STFT kaldes for et spektrogram. Det er styrken af hver frekvens, som afbildes, så vi vil her notere spektrummet beregnet ved STFT:

$$\mathbf{sp}_{stft}\{\mathbf{s}\}(t, \nu) = |\mathbf{stft}\{\mathbf{s}\}(t, \nu)|^2 \quad (3.27)$$

Normalt afbildes styrken på en logaritmisk skala og angives i decibel. Der henvises til [66, afsnit 2.3.2] for eksempler på brugen af spektrogrammer med forskellige vindueslængder. (Det skal bemærkes at der en enkelt gang fejlagtigt bliver nævnt "amplituden" i stedet for "styrken" i [66, afsnit 2.3.2].)

En vigtig konklusion i forhold til figurerne i [66, afsnit 2.3.2] er, at lange vinduer passer bedst til dybe toner, mens korte vinduer passer bedre til høje toner, fordi der blandt de dybe toner findes mange klare frekvenser, som er forholdsvis stabile, og de højere toner kan ofte være noget ustabile. Det er et problem ved STFT, fordi STFT anvender den samme tid/frekvens-opløsning for alle frekvenser.

3.3 Wavelet-transformation

Vi erfarede i forrige afsnit, at STFT har et problem i forbindelse med den faste tid/frekvens-opløsning. Man kan sige, at det måske vil være en fordel at have en frekvensopløsning, som passer til den logaritmiske skala, der anvendes i musik, da det jo netop er musik, vi ønsker at analysere. Mens STFT er bundet op på en frekvensopløsning, som er fast, er det med wavelet-transformation muligt at anvende en logaritmisk frekvensakse. Vi vil i de følgende afsnit undersøge muligheden for at anvende wavelet-transformation til at generere en tid/frekvens-repræsentation af lydsignaler. Først vil vi dog introducere wavelet-transformationen mere generelt.

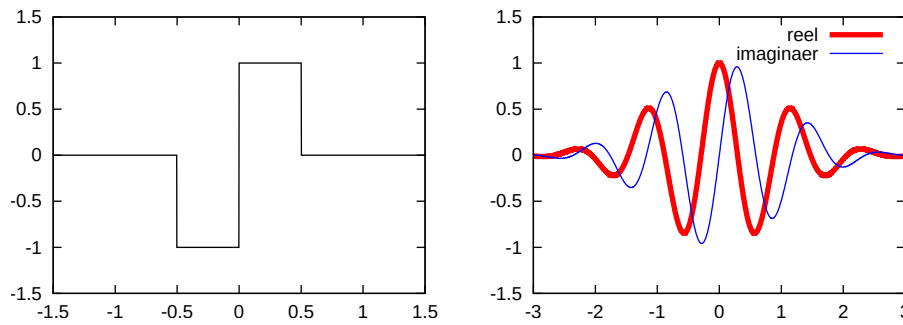
Historisk Wavelet-transformations oprindelse er noget uklar, da der gennem tiderne er givet flere uafhængige bud på metoder, som minder om dem, der anvendes ved wavelet-transformation. Et eksempel på, at sporene går langt tilbage i tiden, er den såkaldte Haar-wavelet, som stammer fra 1909¹¹. I begyndelsen og midten af 1980'erne blev den endelige wavelet-teori udviklet fra forskellige sider, men det var Grossman og Morlet, som i 1984 foreslog ordet wavelet, der stammer fra den franske betegnelse "ondelette", som betyder "lille bølge".¹²

Generelt overblik Mens der i fouriertransformation anvendes analyserende bølger¹³, som ikke er lokaliserede i tiden, så anvendes der i wavelet-transformation bølger, som netop er lokaliserede i tiden. En bølge er lokaliseret i tiden, hvis der er et tidsrum, hvori bølgen opnår sit maksimum. Dette er ikke

¹¹Begrebet wavelet blev ikke anvendt i 1909, men Alfred Haar præsenterede dette år et ortogonalt funktionssystem af diskontinuerte funktioner, som netop har udseende af den funktion, som nu kaldes for en Haar-wavelet. (Se [23, kapitel 3].)

¹²Se f.eks. [32] for en kort gennemgang af wavelet'ens historie.

¹³Som det blev nævnt efter definitionen på fouriertransformationen (afsnit 3.1), så kalder vi i denne opgave det sidste led i integralet for fouriertransformationen for en "analyse-bølge", eller "analyserende bølge". Da man i wavelet-transformationen også analyserer et signal ved hjælp af en bølgeform, vil vi anvende de samme betegnelser for denne bølge.



Figur 3.15: Her vises 2 eksempler på wavelets. Til venstre vises den simple Haar-wavelet, som er diskontinuert og kun antager værdierne 0, -1, 1 og endelig 0 igen. Til højre vises en Morlet-wavelet, som består af en kompleks bølge, som er ganget med en Gauss-funktion. Den komplekse bølges real-del er en cosinus-bølge, mens den imaginære del er en sinus-bølge. Morlet-wavelet'en kaldes også for en Gabor-filter-funktion, da det er denne funktion, som anvendes i Gabor-transformation. Som det blev nævnt i afsnit 3.2.1.3, har Gabor benævnt funktionen som et elementærsignal og har omtalt den som en naturlig basis for tid/frekvens-analyse.

tilfældet for den analyserende bølge i fouriertransformationen, da denne bølge opnår sit maksimum til uendeligt mange tidspunkter.

For at en analyserende bølge kan bruges til wavelet-transformation, uden at der mistes information ved transformationen, skal den opfylde en række krav. Der er mange bølger, som opfylder kravene, og der har igennem tiden være lagt megen energi i at finde wavelets, som egner sig til forskellige signaler. Den simpleste wavelet kaldes for Haar-wavelet'en og består blot af en flad fordybning efterfulgt af en flad forhøjning (se til venstre på figur 3.15). Som sagt kan wavelet-transformation se meget forskellige ud, men da vi her analyserer signaler, som formodes at være dannet ud fra sinusoidale bølger, er det Morlet-wavelet'en (se til højre på figur 3.15), som vi ender med at koncentrere os om. Morlet-wavelet'en anvender nemlig en analyserende bølge, som består af en sinusoidal bølge.

Det generelle princip bag wavelet-transformation er, at et signal analyseres med forskellige udgaver af en bestemt wavelet. Den wavelet, der tages udgangspunkt i, kaldes ofte for en "mother wavelet". Denne wavelet strækkes eller sammenpresses i tiden, og de wavelets, der genereres ved denne process, kan anvendes til analyse af signalet til et bestemt tidspunkt. Ved at flytte de genererede wavelets kontinuert i tiden, kan signalet analyseres til ethvert tidspunkt. Man kan opfatte processen som en form for zoom: Når wavelet'en (den analyserende bølge) er strakt ud, er det de overordnede strukturer i signalet, der analyseres, mens det er de fine detaljer, som analyseres, når wavelet'en er presset sammen. Man kan altså vælge hvilken skala, man ønsker at anskue et signal på.

Lad os nu bevæge os fra denne lidt løse snak til den matematiske definition på wavelet-transformation.

3.3.1 Definition på den kontinuerte wavelet-transformation

Vi vil i det følgende benævne et reelt signal med s og en wavelet med ψ . Inspireret af notationerne i [4] og [63] vil vi her notere definitionen for den kontinuerte wavelet-transformation, således at den minder om notationen for

STFT i formel 3.9:

$$\mathbf{wt}_{\psi}\{\mathbf{s}\}(b,a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} \mathbf{s}(\tau) \psi\left(\frac{\tau-b}{a}\right) d\tau \quad (3.28)$$

I ovenstående definition er wavelet-transformationen noteret som $\mathbf{wt}_{\psi}$ for at indikere, at den er afhængig af, hvilken mother wavelet der anvendes. Ligesom med STFT er der 2 variable, som i tilfældet med STFT er t for tiden og ν for frekvensen. I wavelet-transformationen er variablene: b , som angiver en transposition, og a , som angiver en strækning. Da vi arbejder med signaler, som er definerede i tiden, kan vi lige så godt opfatte b som en tidsangivelse, og derfor vil vi senere erstatte b med t . Hvad angår parametren a , så angiver den hvilken skala signalet analyseres på, fordi den bestemmer i hvor høj grad mother wavelet'en strækkes. Hvis der anvendes en mother wavelet, som er beregnet til at analysere frekvenser for sinusoidale bølger, kan man "oversætte" a til en frekvens. Senere vil vi at lave en formulering, hvori a bliver erstattet af ν .

Som det blev nævnt i forrige afsnit, er der nogle krav, som skal overholdes, for at en bølge kan anvendes som en wavelet i wavelet-transformationen. Der er ifølge [41] følgende krav: (1) En wavelet $\psi(t)$ skal være absolut integrabel og kvadratisk integrabel, hvilket betyder, at den har endelig energi: $\int |\psi(t)| dt < \infty$ og $\int |\psi(t)|^2 dt < \infty$. (2) En wavelet skal have 0 som gennemsnitlig værdi — dette andet krav benævnes også med betegnelsen "admissibility condition" i f.eks. [16]. Kravene sikrer, at bølgen er lokaliseret, og at et signal kan genskabes efter wavelet-transformationen. Der findes også en invers wavelet-transformation, men denne vil vi ikke beskæftige os med, da vi ikke har brug for at genskabe et signal.

Lad os nu se nærmere på formel 3.28. Vi ser, at der integreres over tiden ligesom i formlen for STFT (formel 3.9). Der integreres over et produkt, hvor den ene komponent er signalet. I tilfældet med STFT er den anden komponent en sinusoidal bølge, som er ganget med et begrænsende vindue. Vi lægger mærke til, at bølgen ligger fast, mens vinduet afhænger af tiden t . I tilfældet med wavelet-transformationen er den anden komponent en wavelet, hvis form ikke er fastlagt, og som flyttes afhængigt af b . Wavelet'en kan opfattes som en kombination af en uendelig bølge og et begrænsende vindue. Wavelet'ens udstrækning afgøres af variabelen a . Hvis vi opfatter wavelet'en som kombinationen af den uendelige bølge og et begrænsende vindue, har a altså betydning for begge disse komponenter, for a strækker den uendelige bølge lige så meget som vinduet. I tilfældet med STFT, hvor variabelen ν afgør, hvor hurtigt den sinusoidale bølge svinger, kan vi også tale om en strækning af en uendelig lang bølge. Her er det altså blot ν , som indirekte angiver, hvor meget en bølge strækkes eller sammenpresses i forhold til en fiktiv reference-bølge.

Vi kan altså se, at der er flere ligheder mellem STFT og wavelet-transformation: Der foretages et integral over signalet og en bølgeform, som er begrænset i tiden; og bølgen kan strækkes og sammenpresses. Lad os nu se på hvilke forskelligheder der er mellem STFT og wavelet-transformationen.

Først vil vi notere os, at STFT analyserer et signal ved hjælp af sinusoidale bølger, mens wavelet-transformationen ikke har denne begrænsning. Dermed ikke sagt at man kan bruge en hvilken som helst bølge, men der er flere muligheder, da bølgen blot skal overholde de krav, der tidlige blev fremsat. Dernæst kan vi notere os, at den sinusoidale bølge i STFT ligger fast i forhold til tiden, mens bølgen i wavelet-transformationen flytter med tiden - her henvises

til variablen b i formel 3.28. Endelig kan vi notere, at vinduerne i begge tilfælde flytter med tiden, men i STFT ligger vinduets størrelse fast og er dermed uafhængig af strækningen af bølgen; i wavelet-transformation derimod strækkes vinduet i takt med, at bølgen strækkes. Vi bemærker også, at der i wavelet-transformation anvendes en normalisering af integralet. Integralet udføres over tiden, og for at en kort bølge kan vægte lige så meget som en lang udgave af den samme bølge, er det nødvendigt at foretage normaliseringen. Dette er ikke nødvendigt i STFT, da vinduets længde ligger fast, og dermed integreres der altid over den samme længde.

3.3.2 Fordele ved wavelet-transformationen

Selv om vi på ingen måde har dækket emnet wavelet-transformation, vil vi nu undersøge om wavelet-transformation overhovedet ser ud til at være anvendelig i forhold til frekvensanalyse.

Vi så i forrige afsnit at wavelet-transformationen giver mulighed for at anvende en række forskellige bølgeformer. Men har vi overhovedet brug for at have denne mulighed? Som udgangspunkt er vi interesserede i at analysere bølger, som antages at bestå af sinusoidale bølger. Derfor vil det nok være mest nærliggende at anvende sinusoidale bølger ved analysen.

Mens STFT kun giver mulighed for at analysere med én sinusoidal bølge af gangen, giver wavelet-transformationen mulighed for at analysere med en bølge, som er sammensat af flere sinusoidale bølger. Man kunne f.eks. forestille sig, at et signal blandt andet indeholder lyde fra et instrument med et helt bestemt overtonemønster. I så fald kunne det være en fordel at analysere signalet med en wavelet, som netop er sammensat af sinusoidale bølger, hvis vægtning afgøres af instrumentets overtonemønster. Da ville wavelet'en være specialdesignet til at finde det aktuelle instrument i et sammensat lydbillede. Men de fleste instrumenter har ikke et fast overtonemønster, og selv om de havde, ville det være nødvendigt at kende det aktuelle overtonemønster på forhånd.

STFT anvender stabile sinusoidale bølger¹⁴, men ofte er de sinusoidale bølger på et instrument ikke stabile - dette er f.eks. tilfældet ved glissando på en violinstreng, hvor fingeren flyttes i en glidende bevægelse, således at strengens længde ændres, og fundamentalfrekvensen også ændres glidende. Den analyserende bølge i wavelet-transformationen kan have udseende af en sinusoidal bølge, som ændrer frekvens i tiden. Det betyder, at der ved brugen af wavelet-transformationen er bedre muligheder for at bestemme frekvensen, såfremt man kender hastigheden, hvormed frekvensen ændrer sig. Når man anvender en bølge, som glidende ændrer frekvensen, og som har lokal udstrækning, kaldes det også for en "chirplet". I afsnit 3.7 vil vi se nærmere på den såkaldte "chirplet-transformation".

I forrige afsnit så vi også, at den analyserende bølge i STFT ikke flytter sig med tiden, mens wavelet'en flytter sig i tiden. Dette har ikke nogen betydning for muligheden for at bestemme frekvensen for et signal og påvirker udelukkende måden, hvorpå man kan bestemme fasen. Så hvad angår denne forskel på STFT og wavelet-transformation, er der blot tale om et design-spørgsmål.

Vi skal i næste afsnit se, at den variabel, som i wavelet-transformationen

¹⁴Som tidligere nævnt anvendes ordet "stabil" for at beskrive at frekvensen ikke ændrer sig i tiden.

bestemmer vinduets længde, kan opfattes som en angivelse af frekvensen. Dette kan muligvis være en fordel, da afhængigheden bevirker at jo højere frekvens des kortere vindue og dermed også større båndbredde, og dette bevirker at frekvensaksen bliver logaritmisk, og dermed fint passer til det logaritmiske udseende, som den musikalske skala har, hvor der gælder en logaritmisk sammenhæng mellem hvert trin i skalaen.

Tilbage har vi altså, at wavelet-transformationen kan være en fordel, hvis man ønsker at anvende chirplets, eller hvis man anvender den logaritmiske frekvensakse. Hvad angår den logaritmiske frekvensakse, så er det dog kun ganske få ændringer, der skal foretages, før STFT også kan få denne egenskab — dette skal vi se på i afsnit 3.5.1.

3.4 Morlet-wavelet

Vi skal nu se på en wavelet, som er oplagt at anvende til frekvensanalyse af signaler, som består af sinusoidale komponenter, da wavelet'en selv består af en kompleks sinusoidal bølge. Morlet-wavelet'en kan defineres ved:

$$\psi(t) = e^{-t^2/2} \cdot e^{i2\pi\nu t} \quad (3.29)$$

hvor ν angiver frekvensen. Til højre på figur 3.15 er der vist et eksempel på en Morlet-wavelet, hvor den reelle del og den imaginære del er afbildet i samme koordinatsystem.

Vi bemærker, at integralet af en Gauss-funktionen er begrænset, så Morlet-wavelet'en opfylder det første krav til en wavelet. Men det andet krav, som angiver at den gennemsnitlige værdi skal være lig 0, kan ikke opfyldes; omend den gennemsnitlige værdi nærmer sig 0 når ν forøges.¹⁵ Som det nævnes i [63] er det i praksis ikke noget problem, når der blot skal foretages en analyse af signalet.

Ifølge Smith [62] har Daubechies og Holschneider foreslået at frekvensen for mother wavelet'en sættes til $\nu = 1/\sqrt{2\log(2)}$, da dette bevirker, at cosinus-kurven har gennemført sin første periode når eksponentialfunktionen er halveret.¹⁶ På figur 3.15 er denne anvisning anvendt. Til analyse af musiksignaler er anvisningerne angående ν dog alt for lave, da der kræves meget nøjagtige frekvensbestemmelser. Hvilken ν , der er bedst, afhænger af signalet.

Da mother wavelet'en skal skaleres, er det i princippet ligegyldigt, hvilken frekvens der anvendes (så længe den er større end 0); det, som har betydning, er forholdet mellem bredden på Gauss-funktionen og frekvensen. Vi vil nu omdefinere Morlet-wavelet'en, hvor vi fastsætter frekvensen $\nu = 1$ og anvender β til at styre Gauss-funktionens bredde:

$$\psi(t) = e^{-\beta t^2/2} e^{i2\pi t} \quad (3.30)$$

¹⁵Dette ses let ved anvendelse af foldningssætningen: Vi fouriertransformerer Morlet-wavelet'en og undersøger om resultatet er lig 0 for en frekvens på 0. (Ved at undersøge styrken af frekvensen 0, kan vi finde funktionens gennemsnit - se desuden [41].) Ifølge foldningssætningen kan vi finde fouriertransformationen ved at folde fouriertransformationerne af Gauss-funktionen og den komplekse eksponentialfunktion. Fouriertransformationen af den komplekse eksponentialfunktion er lig 0 for en frekvens på 0 (da $\nu > 0$). Men Fouriertransformationen af en Gauss-funktion er også en Gauss-funktion, så efter foldningen vil styrken af en frekvens på 0 altid være større end 0. Så når ν forøges falder styrken af frekvensen på 0, og dermed nærmer den gennemsnitlige værdi sig 0.

¹⁶ $e^{-t^2/2} = 1/2 \Rightarrow t = \sqrt{2\log(2)}$ så $\nu = 1/\sqrt{2\log(2)}$.

Ovenstående formel angiver en mother-wavelet, men vi skal kunne generere wavelets ud fra denne wavelet. Disse wavelets skal være skalerede og translaterede udgaver af mother-wavelet'en. Dette klares ved at tilføje parametrene a og b fra formelen på wavelet-transformationen, formel 3.28:

$$\psi(t) = e^{-\beta(t-b)^2/(2a^2)} e^{i2\pi(t-b)/a} \quad (3.31)$$

Lad os nu notere wavelet-transformationen igen og binde den til Morlet-wavelet'en, så vi ender ud med en Morlet-wavelet-transformation. For at gøre notationen lettere vil vi notere Gauss-funktionen som et vindue og kalde den $\mathbf{w}_G$. Derved fremkommer nedenstående midlertidige definition for MWT (Morlet-Wavelet-Transformation):

$$\mathbf{mwt}\{\mathbf{s}\}(b, a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}_G\left(\frac{\tau - b}{a}\right) e^{i2\pi((\tau - b)/a)} d\tau \quad (3.32)$$

Det ses, at den komplekse eksponentialfunktion har 2 led: ét hvor τ indgår, og ét hvor τ ikke indgår. Det led, hvor τ ikke indgår, fungerer som et faseskift og sikrer, at den komplekse bølge følger med vinduet i takt med translationsparametren b . Men da der blot er tale om et faseskift, og τ derfor ikke indgår i ledet, kan det trækkes uden for integrationen:

$$\mathbf{mwt}\{\mathbf{s}\}(b, a) = \frac{1}{\sqrt{|a|}} e^{-i2\pi b/a} \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}_G\left(\frac{\tau - b}{a}\right) e^{i2\pi\tau/a} d\tau \quad (3.33)$$

Lad os stadig fokusere på den komplekse eksponentialfunktion. Vi konstaterer at $1/a$ angiver frekvensen, og vi vil derfor erstatte $1/a$ med ν , således at transformationens anden parameter bliver ν . Derudover kan vi erstatte b med t , da vi anvender transformationen på signaler, som er definerede i tiden:

$$\mathbf{mwt}\{\mathbf{s}\}(t, \nu) = \sqrt{|\nu|} e^{-i2\pi\nu t} \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}_G(\nu(\tau - t)) e^{i2\pi\nu\tau} d\tau \quad (3.34)$$

Vi konstaterer, at en Gauss-funktionen er lige, og vi kan derfor uden problemer bytte fortegn i parametren til $\mathbf{w}_G$. Vi ønsker også at bytte fortegn i eksponenten til e . Den reelle del ændrer sig ikke af dette fortegnsskift, da cosinus er lige, men den imaginære del ændrer fortegn. Derfor kan vi kompensere for et fortegnsskift i eksponenten til e ved at konjugere. Derved får vi følgende endelige formel for MWT:

$$\begin{aligned} \mathbf{w}_G(x) &= e^{-\beta x^2/2} \\ \mathbf{mwt}\{\mathbf{s}\}(t, \nu) &= \sqrt{\nu} e^{-i2\pi\nu t} \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}_G(\nu(t - \tau)) e^{-i2\pi\nu\tau} d\tau \end{aligned} \quad (3.35)$$

En sammenligning med definitionen for STFT (formel 3.9) viser store ligheder. Lad os først koncentrere os om indholdet af integrationen. Den eneste forskel mellem ovenstående integrationen og integrationen i STFT er, at parametren til vinduet bliver ganget med en variabelen ν , altså frekvensen. Dette sikrer, at vinduet bliver smallere, når frekvensen stiger. Lad os nu se på, hvad der står uden for integrationen. I formelen for STFT står der ikke noget, men i formel 3.35 bliver der ganget med $\sqrt{\nu}$, hvilket (som tidligere beskrevet) sikrer at høje frekvenser kompenseres for den korte tidsmæssige udstrækning. Det sidste, som mangler at blive beskrevet, er konjugeringen og faseskiftet. Begge dele har kun betydning for selve fasen. Der skal selvfølgelig tages højde for dette, hvis man

f.eks. ønsker at anvende fase-informationerne til at estimere den øjeblikkelige frekvens (instantaneous frequency). Men det er en uvæsentlig forskel på STFT og MWT.

Så konklusionen er, at der i bund og grund kun er én væsentlig forskel på MWT og STFT: I MWT er vinduets længde er afhængig af og bundet til den logaritmiske frekvensopløsningen. En knap så væsentlig forskel er, at MWT'en er bundet til det gaussiske vindue, mens vinduesfunktionen kan vælges frit i STFT. Til analyse af signaler er det dog ikke noget problem at anvende MWT'en sammen med et andet vindue.

3.4.1 Beregningskompleksitet

Når man skal bedømme beregningskompleksiteten for MWT'en, kan man umiddelbart få den idé, at transformationen kan være en fordel set i forhold til STFT. I forbindelse med STFT så vi, at det er nødvendigt at have en høj frekvensopløsning, for at dybe toner kan skelnes fra hinanden. Det betyder dog, at frekvensopløsningen også er høj ved høje frekvenser, og dette er måske ikke nødvendigt. Med MWT'en er frekvensaksen logaritmisk, så en god adskillelse af dybe toner, giver ikke anledning til en unødvendig høj frekvensopløsning for høje frekvenser. Problemet er, at den varierende vindueslængde giver problemer i forhold til optimering af beregningerne.

I en uoptimeret beregning af MWT'en er beregningskompleksiteten ifølge [4] $\mathcal{O}(M \cdot N^2)$, hvor N angiver længden af signalet, og M angiver antallet af frekvenser. Der findes også en hurtig version, som kaldes "Fast Morlet wavelet transform", og som anvender fouriertransformationen af signalet. Det giver en beregningskompleksitet på $\mathcal{O}(M \cdot N \cdot \ln(N))$.

I [4] bliver der dog fremvist en optimering, som blandt andet anvender det faktum, at nøjagtigheden på en computer er begrænset, samt at det i de fleste tilfælde ikke er nødvendigt med stor nøjagtighed. Derudover begrænses antallet af tidspunkter, hvortil transformationen beregnes for de lave frekvenser, da tidsopløsningen er lav for disse. I den præsenterede beregningsmetode reduceres beregningskompleksiteten til $\mathcal{O}(M \cdot N)$.

3.4.2 Tid/frekvens-repræsentation ved MWT

En repræsentation af wavelet-transformationen kaldes generelt for et scalogram, da signalet anskues på forskellige skalaer. Her har vi dog med en MWT at gøre, og derfor giver det mere mening at repræsentere transformationen i et tid/frekvens-domæne. Et sådant domæne kaldes generelt for et spektrum, så vi vil også anvende denne betegnelse her. Spektrummet beregnet ud fra MWT'en kan bestemmes på samme måde, som det var tilfældet med STFT:

$$\mathbf{sp}_{mwt}(t, \nu) = |\mathbf{mwt}(t, \nu)|^2 \quad (3.36)$$

I [61] bemærkes det dog, at der er et problem ved vægten af de forskellige frekvenser. Da den effektive længde af det gaussiske vindue halveres, når frekvensen fordobles (se formel 3.31, hvor frekvensen ν ganges på tiden), vil integralet også halveres. Som det bliver foreslået i [61], kan der kompenseres for dette ved at gange med ν i stedet for at gange med $\sqrt{\nu}$. I stedet for at omdefinere MWT'en, vil vi i stedet for definere et nyt styrkespektrum:

$$\mathbf{sp}_{mwt}(t, \nu) = |\sqrt{\nu} \mathbf{mwt}(t, \nu)|^2 \quad (3.37)$$

Vi vil ikke vise nogen eksempler på brugen af MWT'en til generering af en tid/frekvens-repræsentation, da der i afsnit 3.5.3.2 gives et eksempel på brugen af den såkaldte Q-transformation, som på figur 3.22 er tilpasset, så den i høj grad ligner MWT'en.

3.5 STFT med frekvensafhængig vindueslængde

I dette afsnit vil vi præsentere 2 bud på modifikationer af STFT. Det første bud blev kort introduceret i [66], men kan godt trænge til at blive gennemgået nærmere her. Det andet bud er blot en generalisering af det første bud.

3.5.1 Constant-Q-transformation

Her skal vi se på den såkaldte constant-Q-transformation, som blev præsenteret af Brown i [2] i 1991. Der er tale om en transformation, hvor vindueslængden, ligesom i wavelet-transformationen, afhænger direkte af frekvensen, således at båndbredden er lineært afhængig af frekvensen og frekvensaksen er logaritmisk. Constant-Q transformationen er taget med, fordi den er blevet tildelt en del opmærksomhed, fordi den kan ses som en simplificeret formulering af MWT'en, og endelig fordi den bygger på STFT.

Constant-Q dækker over forholdet mellem frekvens og båndbredden. Med båndbredden menes blot en relativ båndbredde, således at et vindue, som er dobbelt så langt som et andet vindue med samme udseende, vil have den halve båndbredde af det andet vindue. Generelt beregnes en tid/frekvens-repræsentation for visse frekvenser, og båndbredden af de vinduer, som anvendes, er lineært afhængig af afstanden mellem hver af de anvendte frekvenser. I tilfældet med MWT anvendes en logaritmisk frekvensakse, således at afstanden mellem hver frekvens er større jo højere frekvensen er. Da båndbredden af vinduerne afhænger lineært af afstanden mellem frekvenserne, er båndbredden altså også større for høje frekvenser.

Da båndbredden er lineært afhængig af afstanden mellem frekvenserne, kan constant-Q beskrives ved forholdet mellem frekvens og frekvensafstand, således at $\nu/\delta\nu = Q$, hvor ν er frekvensen og $\delta\nu$ er afstanden mellem 2 frekvenser ved frekvensen ν . Mens det forhold, som Q beskriver, ikke er konstant i STFT, så er forholdet konstant i tilfældet med constant-Q-transformationen. I [2] foreslås det at $Q = 34$, hvilket svarer til, at der er 2 frekvenser pr. halvtone i den musikalske skala.¹⁷

Den diskrete formulering af constant-Q-transformationen, som den er defineret i [2], er:

$$S[k] = \frac{1}{N[k]} \sum_{\eta=0}^{N[k]-1} w[\eta, k] s[\eta] e^{-i2\pi Q\eta/N[k]} \quad (3.38)$$

¹⁷Der er 12 halvtone pr. oktav i den musikalske skala, så ud fra én tones frekvens kan den efterfølgende halvtone frekvens bestemmes til at være $2^{1/12}$ gange højere. Det betyder, at den relative afstand mellem 2 halvtone er $2^{1/12} - 1$. Hvis der skal være plads til 2 frekvenser pr. halvtone, er den relative afstanden mellem hver frekvens altså $2^{1/24} - 1$. Det betyder at Q kan fastsættes til $Q = 1/(2^{1/24} - 1) \approx 34$.

hvor k angiver frekvensindekseringen, η angiver tidindekseringen, N bestemmer længden af vinduet, w bestemmer vinduets udseende, s er signalet og konstanten Q angiver den relative frekvensopløsning. Vinduets længde er fastsat således, at der for hver frekvens er det samme antal bølger inden for vinduet. I [2] er $N[k] = N_{\max}/(2^{k/24})$, hvor $N_{\max}$ er den maksimale vindueslængde, som bliver anvendt.

Ovenstående formel (formel 3.38) er noget upræcis, hvad angår vinduet. Det er vigtigt at bide mærke i, at vinduets udseende snarere afhænger af $N[k]$ end af k . På grund af begrænsninger ved diskrete definitioner, vil vi her foreslå, at man i stedet for det diskrete vindue anvender et ikke-diskret vindue, så sammenhængen mellem vinduets udseende og $N[k]$ fremstår tydelig. Vi noterer derfor, at $w[n, k] = w(nQ/N[k])$. Det ikke-diskrete vindue skal da være defineret inden for intervallet 0 til $Q - Q/N$. Dette giver følgende definition for constant-Q-transformationen:

$$\text{cq}\{s\}[k] = \frac{1}{N[k]} \sum_{n=0}^{N[k]-1} w\left(\frac{nQ}{N[k]}\right) s[n] e^{-i2\pi Qn/N[k]} \quad (3.39)$$

En sammenligning af formel 3.39 for constant-Q-transformationen og formel 3.8 for den vinduesbegrænsede fouriertransformation viser, at de eneste forskelle er, at vinduet i constant-Q-transformationen afhænger af frekvensen, samt at frekvensaksen i constant-Q-transformationen er logaritmisk. Vægten af hver frekvens i constant-Q-transformationen modsvarer blot den variable vindueslængde.

Ved beregningen af STFT anvendes ofte en hop-size¹⁸ på f.eks. halvdelen af vinduets længde, således at hop-size afhænger af vindueslængden. Hvis man bruger det samme princip ved anvendelsen af constant-Q-transformationen i stedet for fouriertransformationen, skal constant-Q-transformationen beregnes med kortere tidsmæssige intervaller for høje frekvenser end for lave frekvenser. Brown har dog valgt, at hop-size ikke er afhængig af vinduets længde, men derimod konstant. Det betyder, at det kun er nogle få samples, som anvendes ved beregningen af constant-Q-transformationen for de høje frekvenser. Man kan med andre ord sige, at der i de høje frekvenser opstår større og større "sorte huller", der hvor samples ikke medtages i beregningerne. Ved at lade vinduet forskyde med den mindste enhed, således at hop-size er lig 1, kunne dette "problem" have været undgået; men denne løsning er meget beregningstung.

3.5.1.1 Fordele og problemer

Hvis hop-size er lille, kan det være en fordel at anvende constant-Q-transformationen frem for fouriertransformationen, da der derved gives mulighed for bedre temporal opløsning for de høje frekvenser. Hvis det omvendt gælder, at hop-size er stor, kan det måske snarere opfattes som et problem at anvende constant-Q-transformationen, for da kan der temporalt set opstå "sorte huller" i de høje frekvenser, fordi ikke alle samples indgår i beregningerne.

Hvis det kun er den logaritmiske akse, som man er interesseret i, kan denne også genereres ud fra fouriertransformationen ved at samle frekvensbåndene i de høje frekvenser, så i dette tilfælde er constant-Q-transformationen ikke den store fordel at anvende.

¹⁸Hop-size angiver afstanden i samples mellem hvert tidspunkt, som fouriertransformationen i STFT beregnes for. Se desuden [66, afsnit 2.3.1.4], hvor dette begreb er forklaret.

Beregningsmæssigt set melder Brown om, at en constant-Q-transformation med 156 frekvenser bruger omkring det samme antal multiplikationer, som en (uoptimeret) fouriertransformation af 512 samples, hvilket giver 256 frekvenser. Brown antager dog, at de 156 frekvenser er bedre egnet til analyse af et musikalsk signal. Problemet er blot, at man som regel vil anvende en FFT til beregning af fouriertransformationen, og det betyder, at den hastighedsmæssige fordel ved constant-Q-transformationen forsvinder.

Optimeret beregning I [3] beskriver Brown en optimeret version af constant-Q-transformationen, og denne version gør brug af FFT. Brown anvender ganske enkelt Parsevals sætning.¹⁹ Signalet fouriertransformeres ved hjælp af FFT og det samme gør multiplikationen af vinduerne og de analyserende bølger. For hver enkelt frekvens ganges de 2 resultater sammen. Hvis man kan acceptere at beregningerne ikke bliver helt nøjagtige, kan der på denne måde spares megen beregningskraft, fordi den fouriertransformerede til multiplikationen af vinduet og den analyserende bølge er nær 0 for næsten alle frekvenser.²⁰ I det eksempel, som Brown anvender i [3], tilføjer beregningen af constant-Q-transformationen under 3% ekstra beregninger til beregningen af FFT.

3.5.1.2 Sammenligning med MWT

Hvis vi sammenligner en MWT med en STFT, som anvender constant-Q-transformationen, og dertil medtager definitionen af styrkespektrummet for MWT, kan vi konstatere, at der i virkeligheden er tale om den samme transformation, såfremt vi ser bort fra fase-forskellen og ignorerer at MWT er bundet til det gaussiske vindue. Så hvis vi kun er interesserede i at lave en tid/frekvens-repræsentation ved styrkespektrummet af MWT, er en STFT, som anvender constant-Q-transformationen, en simpel og effektiv måde, hvorpå repræsentationen kan beregnes.

3.5.2 FDTF

Browns constant-Q-transformation er bundet til den konstante Q , som fastsætter frekvensopløsningen. Men generelt set kan frekvensopløsningen også fastsættes af en funktion, som er mere fleksibel. På samme måde, som det bliver gjort af Zhou i [68]²¹, vil vi her definere en frekvens-afhængig tid/frekvens-analyse og kalde den for FDTF (Frequency-Dependent Time-Frequency analysis), men i

¹⁹Meget kort fortalt, så går Parsevals sætning ud på, at man kan beregne summen af produktet af 2 komplekse funktioner, ved at tage summen af produktet af deres fouriertransformationer. Formuleret på et signal af længde N kan sætningen skrives som: $\sum_{n=0}^{N-1} \mathbf{x}[n]\overline{\mathbf{y}[n]} = \frac{1}{N} \sum_{k=0}^{N-1} \mathbf{X}[k]\overline{\mathbf{Y}[k]}$.

²⁰Husk på at den fouriertransformerede til multiplikationen af vinduet og den analyserende bølge blot svarer til en translation af vinduets frekvensrepræsentation. For de mest anvendte vinduer ligger størsteparten af styrken for hver frekvens inden for main lobe, så frekvenserne, som ligger uden for main lobe, kan sættes lig 0, uden at det har den store betydning for resultatet.

²¹I [68] anvendes FDTF til udviklingen af en såkaldt RTFI (Resonator Time-Frequency Image), hvor udseendet af vinduesfunktionen lægges fast. Her vil vi ikke beskæftige os mere med denne RTFI, da vi ikke ønsker at fastsætte en bestemt vinduesfunktion. Men grundidéen bag RTFI, hvor en funktion $\mathbf{r}(\nu)$ bestemmer bredden af vinduet, kan være anvendelig.

forhold til definitionen i [68] skifter vi her fortegn i eksponenten til e :

$$\begin{aligned}\mathbf{fdtf}\{\mathbf{s}, \mathbf{w}\}(t, \nu) &= \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}(\tau - t, \nu) e^{i2\pi\nu(\tau - t)} d\tau \\ &= e^{-i2\pi\nu t} \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}(\tau - t, \nu) e^{i2\pi\nu\tau} d\tau\end{aligned}\quad (3.40)$$

En sammenligning af FDTF med STFT (formel 3.9) og med MWT (formel 3.34) viser, at FDTF deler mange kendetegn med de 2 andre transformationer. Ligesom med MWT så flytter den analyserende bølge med vinduet, men i modsætning til MWT, så er vinduets udseende ikke låst, men dog afhængigt af frekvensen. FDTF kan faktisk opfattes som en generalisering af MWT, da det er muligt at vælge en vinduesfunktion, således at FDTF med det anvendte vindue er lig MWT. Dette er tilfældet, hvis det anvendte vindue er defineret som:

$$\mathbf{w}_{mwt}(t, \nu) = \sqrt{|\nu|} \mathbf{w}_G(\nu t) \quad (3.41)$$

Således kan MWT defineres ved hjælp af FDTF som:

$$\mathbf{mwt}\{\mathbf{s}\}(t, \nu) = \mathbf{fdtf}\{\mathbf{s}, \mathbf{w}_{mwt}\}(t, \nu) \quad (3.42)$$

$$= e^{-i2\pi\nu t} \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}_{mwt}(\tau - t, \nu) e^{i2\pi\nu\tau} d\tau \quad (3.43)$$

$$= \sqrt{|\nu|} e^{-i2\pi\nu t} \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}_G(\nu(\tau - t)) e^{i2\pi\nu\tau} d\tau \quad (3.44)$$

3.5.3 QSTFT

Her vil vi præsentere et forslag til en modifikation af STFT. Forslaget er inspireret både af Zhous FDTF og Browns constant-Q-transformation. Vi vil kalde forslaget for QSTFT, hvilket står for "Q-STFT". Mens faktoren Q i Browns constant-Q-transformation er konstant, vil vi her foreslå en Q , som ikke er konstant. I stedet for Q vil vi herefter anvende funktion $Q(\nu)$ og dermed angive, at den er afhængig af frekvensen. Af notationsmæssige årsager er det valgt, at funktionen Q ikke har samme betydning som faktoren Q .

Afstanden mellem 2 frekvenser i fourier-transformation og i constant-Q-transformation ligger fast i forhold til båndbredden²² af de anvendte vinduer. Eksempelvis vil man udregne MWT for frekvenser, som ligger på en logaritmisk frekvensakse, således at overlappet mellem hvert vindue er det samme, selv om der bruges vinduer med lille båndbredde for lave frekvenser og vinduer med stor båndbredde for høje frekvenser. Det kan dog være en fordel, hvis afhængigheden mellem vinduernes båndbredder og frekvensaksen ikke er tilstede. Hvis begrænsningen ikke er tilstede, betyder det blot, at vinduerne kan have et større frekvensmæssigt overlap i nogle frekvensområder end i andre. Hvis man tillader dette, kan man vælge frekvensaksen frit uafhængig af måden, hvorpå vinduets længde ændrer sig i forhold til frekvensen. Her vil vi antage, at dette er tilfældet, og vi kan derfor koncentrere os om måden, hvorpå vinduets længde ændrer sig i forhold til frekvensen.

Vi kan definere et vindues båndbredde B_w ved $B_w = a/L_w$, hvor L_w angiver vinduets længde, og a er en konstant for den aktuelle vinduestype. Vi ønsker nu

²²Bemærk at vi her ikke har fastsat hvorledes båndbredden af et vindue bestemmes — båndbredden må altså ikke forveksles med ENB (Se afsnit 3.2.1.7). Båndbredden bliver blot brugt ved sammenligning af 2 vinduer med samme udseende, og derfor er det nøjagtige mål underordnet.

at beskrive båndbredden ved hjælp af frekvensen og $\mathbf{Q}$. $\mathbf{Q}$ defineres til at bestemme forholdet mellem vinduets længde og frekvensen. Da kan båndbredden for en bestemt frekvens bestemmes ved: $B_w = a\nu/\mathbf{Q}(\nu)$. $\mathbf{Q}$ er ikke nogen bestemt funktion, så den kan fastlægges fuldstændig frit, men den beskriver som sagt forholdet mellem vindueslængden og frekvensen.

Den konstante båndbredde i fourier-transformationen fås ved at sætte $\mathbf{Q} = \mathbf{Q}_f$, hvor $\mathbf{Q}_f(\nu) = \nu$. På samme måde kan vi finde en $\mathbf{Q}$, som passer til constant-Q-transformationen ved at sætte $\mathbf{Q} = \mathbf{Q}_c$, hvor $\mathbf{Q}_c = 1$.

Nu vil vi definere QSTFT og anvende $\mathbf{Q}$ til denne definition. $\mathbf{Q}$ fungerer som en skalering af et vindue og indtager derfor den samme rolle, som funktionen $\mathbf{r}$ har i Zhous RTFI (se [68]). Den væsentligste forskel på QSTFT og RTFI er, at RTFI fastsætter et bestemt vindue, hvilket ikke er tilfældet i QSTFT. Definitionen på QSTFT er:

$$\text{qstft}\{\mathbf{s}\}(t, \nu) = \mathbf{Q}(\nu) \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}(\mathbf{Q}(\nu)(\tau - t)) e^{i2\pi\nu(\tau - t)} d\tau \quad (3.45)$$

Beregning af QSTFT Jeg foreslår, at det trick, som Brown anvender ved den optimerede beregning af constant-Q-transformationen, også anvendes ved beregningen af QSTFT. På denne måde kan beregningskompleksiteten holdes på samme niveau som beregningskompleksiteten for FFT. Dette medfører dog, at resultatet vil indeholde en del unøjagtighed, da der foldes med et tilnærmet vindue i frekvensdomænet.

Der skal foretages en FFT med den længste vindueslængde, så hvis $\mathbf{Q}$ har et udseende så vindueslængden for lave frekvenser er markant længere end for høje frekvenser, vil det give nogle tunge beregninger, set i forhold til en STFT med en gennemsnitlig vindueslængde. Dette skyldes for det første, at FFT skal beregnes med den længste vindueslængde. Dernæst kræves der mange beregninger for de korte vinduer, da de har en stor båndbredde. Man kan kompensere for dette problem, ved at foretage FFT'er ved forskellige vindueslængder. Hvis vi starter med at beregne frekvenserne fra toppen af (altså de højeste frekvenser først), kan der først foretages en FFT med en længde, som svarer til den dobbelte af den mindste vindueslængde. Hver gang vindueslængden stiger til det dobbelte foretages en resampling med den halve samplefrekvens og FFT beregnes igen. Forslaget om at lave en FFT for hver dobbelte vindueslængde er ikke nødvendigvis optimal; det kunne måske være bedre at anvende den 4-dobbelte vindueslængde. Det optimale forhold kan findes i en endelig implementering, når der tages højde for, hvor mange beregninger der præcis kræves for implementeringens FFT set i forhold til implementeringens beregningen af QSTFT på baggrund af FFT'erne.

3.5.3.1 Optimal $\mathbf{Q}$

Nu melder spørgsmålet sig: Hvordan skal funktionen $\mathbf{Q}$ helst se ud? Vi har allerede set på 2 forslag, nemlig $\mathbf{Q}_f$ som fører til en almindelig STFT, og $\mathbf{Q}_c$ som fører til en STFT med constant-Q-transformation. Hvis vi skal analysere en stabil harmonisk bølge, altså en bølge som indeholder mange overtoner til fundamentalfrekvensen, så er den almindelige STFT god at anvende; og her vil det være en fordel med så lange vinduer som muligt, da det giver bedre muligheder for at estimere frekvenserne. Men i musik finder man sjældent stabile harmoniske bølger i et længere tidsrum. Man kan godt finde korte tidsrum med

stabile harmoniske bølger, men specielt i de højere frekvenser bliver disse bølger ofte i korte perioder generet af støj eksempelvis fra slagtojsinstrumenter. Dette samt det logaritmiske udseende af den musikalske skala samt målinger af den menneskelige opfattelse af lyde²³ taler for, at Q er konstant. Men som vi skal se på de følgende figurer, er der et klart problem ved den konstante Q , hvis der anvendes vibrato. Vibrato er et meget udbredt virkemiddel til at give en tone en bestemt karakter eller til at fremhæve tonen.

Vibrerende toner Normalt antages det, at frekvenserne inden for et kort tidsrum er stabile, men da der i musik ofte anvendes vibrato, vil vi her undersøge hvilken Q -funktion, som bedst egner sig til vibrerende toner.

Først undersøger vi tilfældet hvor $Q = Q_f$, således at båndbredden er konstant. På figur 3.16 ses et eksempel på en vibrerende harmonisk bølge. Med kurverne er de teoretiske overtoner samt fundamentalfrekvensen vist som funktion af tiden. Bemærk at figuren ikke viser styrken af hver frekvens, men hver kurve angiver en overtones frekvens som funktion af tiden. Det ses, at et lille udsving i tonens fundamentalfrekvens har betydelig effekt for de høje overtoners frekvens. Med rektanglerne vises en teoretisk placering af nogle af vinduerne i en normal STFT. Vinduerne er smalle nok til, at fundamentalfrekvensen kan bestemmes med rimelig præcision. Men i de højere overtoner kommer vinduet tidsmæssigt til at dække over flere forskellige overtoner. Det bevirker, at det kan være svært at få en brugbar analyse ud af en STFT — i hvert fald for de højere frekvenser.

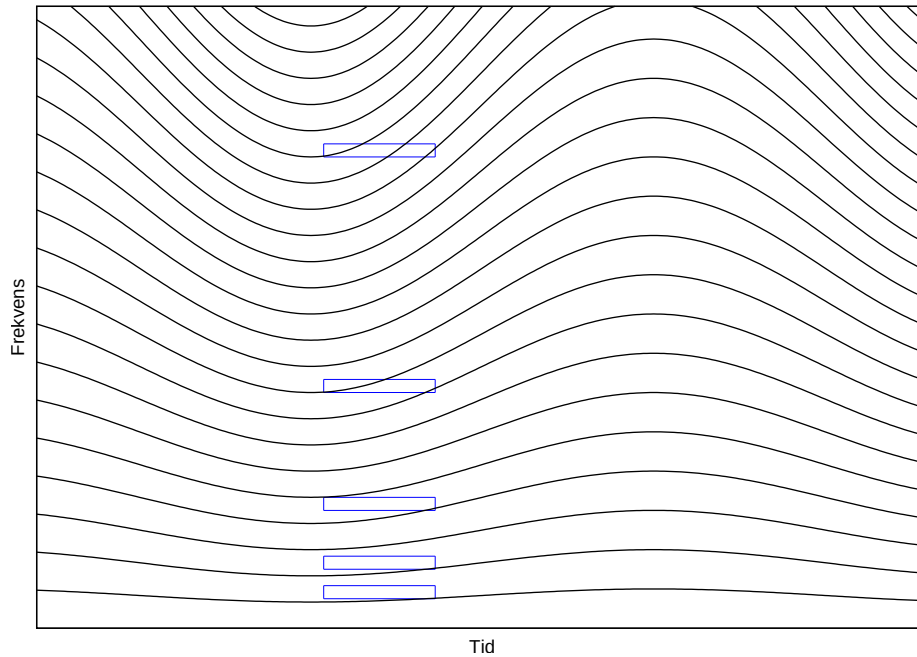
På figur 3.17 gengives eksemplet fra figur 3.16, men her er vist vinduer fra en constant- Q -transformation, således at $Q = Q_c$. Da MWT også kan betragtes som en constant- Q -transformation, er eksemplet også gældende for MWT. Det ses, at det vindue, som ligger i det laveste frekvensområde, er for langt, således at 2 adskilte dele af fundamentalfrekvensens kurve ligger inden for vinduet. Går man op i de højere frekvenser ses det, at vinduet kommer til at have noget nær en optimal længde. For de højeste frekvenser opstår der et nyt problem: Vinduet er blevet så højt, at det dækker flere forskellige overtoner samtidig.

Af de 2 figurer kan vi konkludere, at der ikke er så meget at gøre i forhold til de højeste overtoner — vinduets areal er simpelthen for stort i forhold til disse overtoners hastigt skiftende frekvens. Men vi kan se, at figur 3.16 klarer sig bedst i de lave frekvenser, og 3.17 klarer sig bedst for frekvenserne i midten. Det vi søger er altså en Q , som ligger mellem Q_f og Q_c .

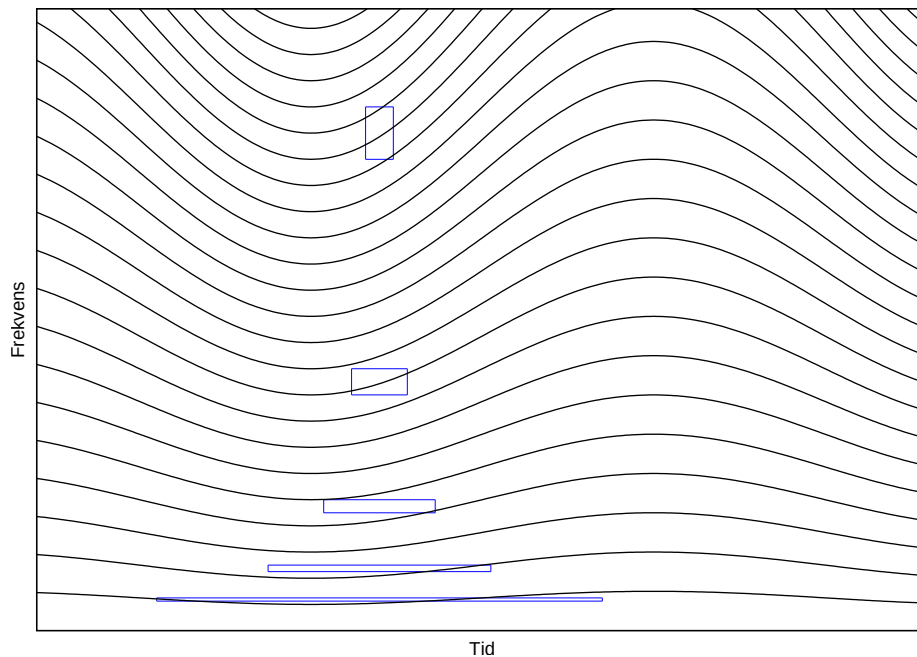
Udledning af en "optimal" Q Lad os undersøge det scenarie, som blev afbildet på figur 3.16 og 3.17. Det største problem i forhold til en vibrerende tone er, at tonen ændrer frekvens i løbet af kort tid. Der, hvor kurverne har den maksimale hældning, er det særligt svært at finde et passende vindue.

Lad os se isoleret på situationen, hvor frekvensen stiger med den største hastighed. Vi kan tilnærme tonens bevægelse på figurerne med en lineær bevægelse. Så vi antager, at vi har en tone, som har en lineært stigende fundamentalfrekvens. Vi har altså situationen, hvor fundamentalfrekvensen er f_1 til tidspunktet t_1 . Fundamentalfrekvensen stiger lineært ind til tidspunktet t_2 , og til dette tidspunkt er fundamentalfrekvensen $f_2 = cf_1$, hvor c angiver stigningen. Lad os nu undersøge, hvor meget en overtone stiger i det samme

²³Her tænkes f.eks. på Bark-skalaen og Mel-skalaen.



Figur 3.16: Her er en vibrerende tones fundamentalfrekvens samt overtoner illustreret på skematisk vis. Hver kurve svarer til en overtone og angiver altså overtonens frekvens som funktion af tiden. Rektanglerne illustrerer vinduer repræsenteret i tid/frekvens-domænet. For at illustrere de vinduer, som bruges i en STFT, har de angivne vinduer samme størrelse uafhængig af frekvensen. Det ses, at vinduet i forhold til de lave frekvenser ikke dækker over flere frekvenskomponenter, og derfor egner sig til adskillelsen af de forskellige frekvenskomponenter. I de højere frekvenser er vinduets længde et problem, fordi det kommer til at gå ind over frekvenskomponenterne fra flere forskellige overtoner.



Figur 3.17: Denne figur svarer til figur 3.16, men her illustreres vinduesstørrelser, som fastlægges efter constant-Q. Det ses, at fundamentalfrekvensen krydser ind over det lange vindue på 2 forskellige tidspunkter. Dette fører til en meget unøjagtig og til dels fejlagtig frekvensbestemmelse til et givent tidspunkt. I de høje frekvenser kommer vinduet til at dække ind over flere forskellige overtoners frekvenskomponenter.

tidsrum. Vi ser på en overtone, som til tidspunktet t_1 har frekvensen xf_1 . Til tidspunktet t_2 må denne overtone da have en frekvens på $xf_2 = cxf_1$. Så mens fundamentalfrekvensen stiger med faktoren c , stiger overtonen med faktoren cx .

Vi ønsker at finde en optimal vindueslængde som funktion af frekvensen, og begrænsningen er, at alle vinduerne skal have et bestemt areal (ligesom vinduerne på figur 3.17, hvor alle vinduerne har det samme areal). Lad os antage, at vi har fastlagt et vindue, som passer til fundamentalfrekvensen, og vinduets båndbredde er netop større end fundamentalfrekvensens udsving indenfor vinduet. Vi kalder vinduets længde for l_1 og vinduet båndbredde for B_1 . Da er arealet af vinduet: $A_1 = l_1 B_1$. Vi bemærker, at båndbredden netop er så stor, at fundamentalfrekvensens udsving kan være inden for vinduet. Nu vil vi angive vinduets areal ved hjælp af fundamentalfrekvensens hældning og længden af vinduet: $A_1 = cl_1^2$.

Lad os fokusere på en overtone, hvis frekvens er x gange fundamentalfrekvensen. For overtonen ønsker vi også, at vinduet netop har en bredde, så det dækker fundamentalfrekvensens udsving. Vi kan på samme måde som før angive vinduets areal ved hjælp af vinduets længde l_x . Da hældningen af frekvensen for overtonen er cx , kan arealet skrives som: $A_x = cxl_x^2$. Da alle vinduerne skal have samme areal, kan vi skrive l_x ved hjælp af l_1 :

$$\begin{aligned} A_1 &= A_x \\ \Downarrow \\ cl_1^2 &= cxl_x^2 \\ \Downarrow \\ l_x &= \sqrt{\frac{1}{x}} l_1. \end{aligned} \quad (3.46)$$

Konklusionen er, at når frekvensen stiger med en faktor x , så bliver vinduets længde delt med $\sqrt{x}$. Nu kan vi altså udtrykke vinduets længde som funktion af frekvensen ved:

$$l = \frac{k}{\sqrt{\nu}} \quad (3.47)$$

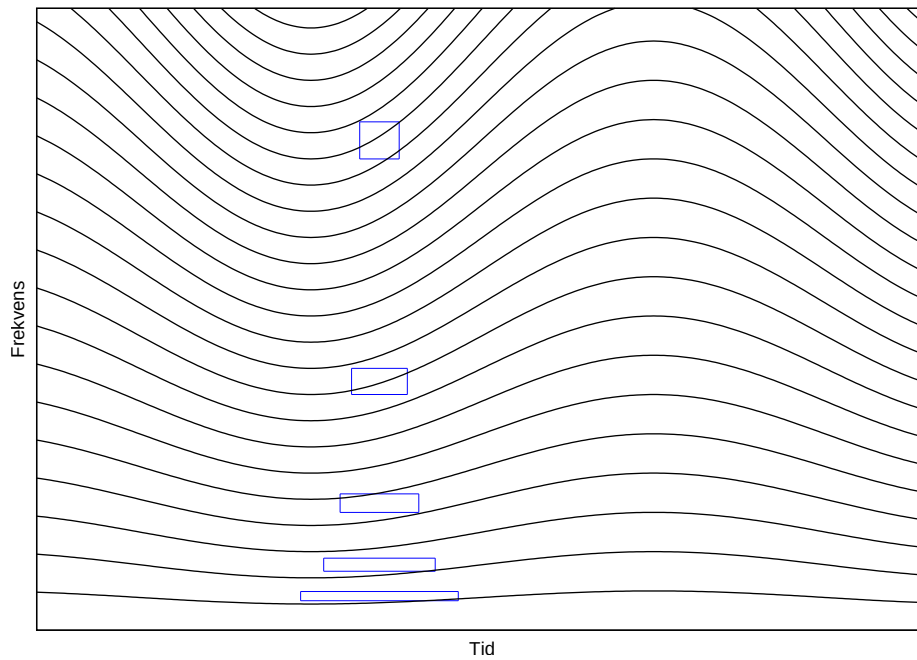
hvor k er en konstant. Den optimale $\mathbf{Q}_o$ kan altså defineres som:

$$\mathbf{Q}_o(\nu) = \sqrt{\nu} \quad (3.48)$$

Vi indsætter den optimale $\mathbf{Q}_o$ i definitionen for QSTFT og får derved en OQSTFT (Optimal-Q-STFT):

$$\text{oqstft}\{\mathbf{s}\}(t, \nu) = \sqrt{\nu} \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}(\sqrt{\nu}(\tau - t)) e^{i2\pi\nu(\tau - t)} d\tau \quad (3.49)$$

På figur 3.18, som svarer til figur 3.16 og 3.17, anvendes $\mathbf{Q}_o$. Det næstøverste vindue har samme form som det næstøverste vindue på figur 3.17, og det næstnederste vindue har samme form som vinduerne på figur 3.16. I modsætning til på figur 3.16 og 3.17 så passer alle vinduerne undtagen det øverste fint til signalet. Selv om der er problemer med det øverste vindue, så er problemerne mindst ved anvendelsen af $\mathbf{Q}_o$.



Figur 3.18: Denne figur svarer til figur 3.16 og figur 3.17, men her illustreres størrelsen af vinduerne i QSTFT. I forhold til de vinduer, som blev benyttet ved de 2 omtalte figurer, er vinduet her bedre tilpasset signalet. Figurerne kan let sammenlignes, da det næstøverste vindue har samme form som det næstøverste vindue på figur 3.17, og det næstnederste vindue har samme form som vinduerne på figur 3.16.

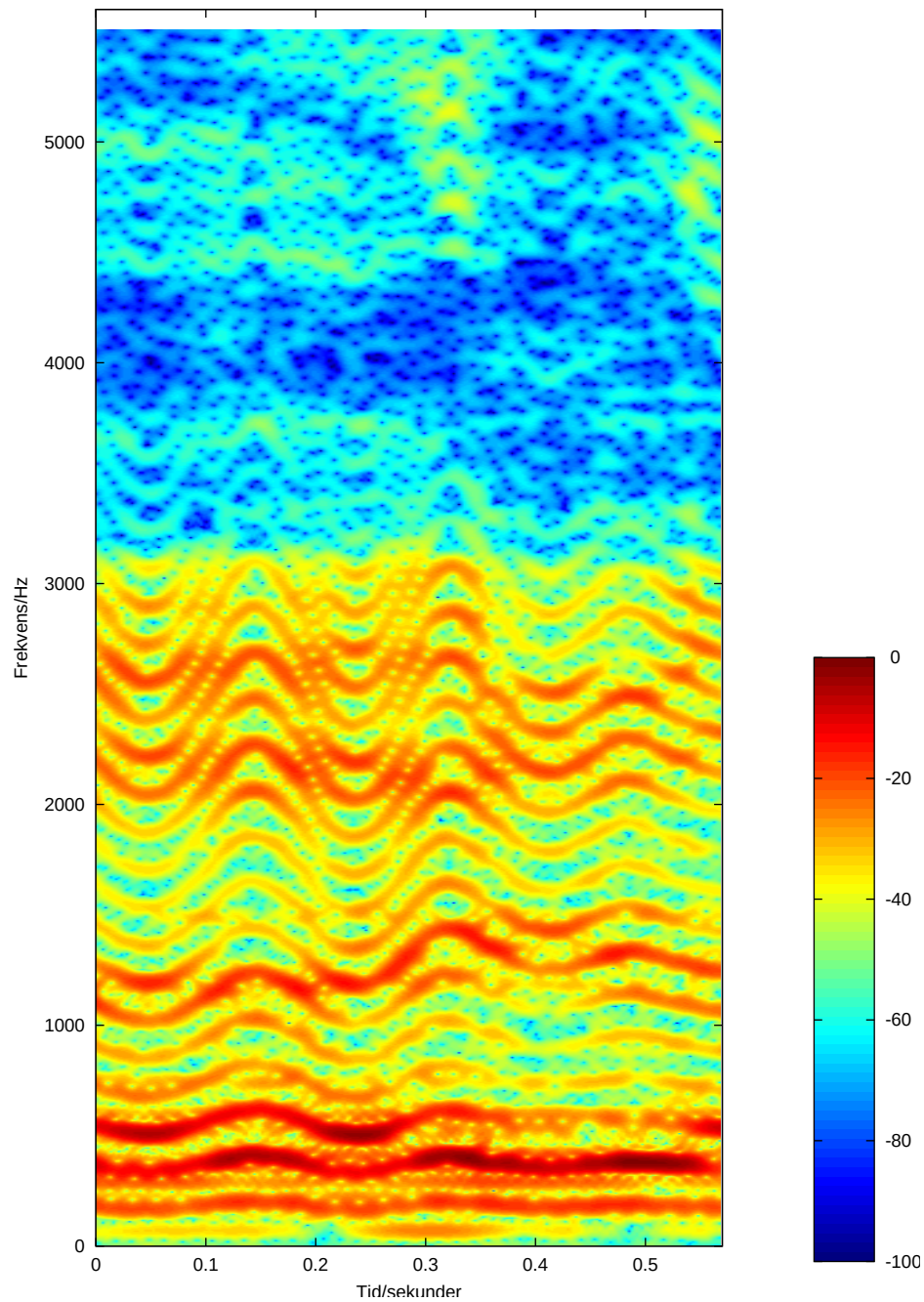
3.5.3.2 Praktisk afprøvning

Jeg har implementeret en QSTFT, som anvender et Hann-vindue. Det trick, som Brown benyttede ved den optimerede constant-Q-transformation er blevet anvendt — se afsnit 3.5.1.1. Det bemærkes, at en sinusoidal bølge, som ganges med Hann-vinduet, vil have det udseende, som blev vist på figur 3.9, så de frekvenser, der vægter mest, ligger inden for main-lobe. Det er derfor kun frekvenser inden for main-lobe, som her medtages ved beregningen af QSTFT. På figur 3.23 og 3.24 kan man derfor også se nogle svage artefakter, specielt i de høje frekvenser.

Der tages udgangspunkt i en lydoptagelse af en mandlig operasanger, som synger en tone på omkring 190 Hz, og tonen vibrerer med en maksimal afvigelse fra udgangsfrekvensen på omkring 10%. En almindelig STFT af lydoptagelsen kan ses på figur 3.19. Det ses af figuren, at vibratoet i de første 0,3 sekunder ligner en sinusbølge. Dette har inspireret til genereringen af et kunstigt signal, som indeholder en grundfrekvens og tilhørende overtoner. Hver frekvenskomponent moduleres med en sinusbølge. Vægten af overtonerne er fastlagt, så det kunstige signal ligner signalet fra lydoptagelsen. I det følgende afprøves QSTFT på det kunstige signal.

Først vil vi se på det kunstige signal, hvor grundfrekvensen ligger omkring 190 Hz, således at signalet minder om lydoptagelsen. De følgende figurer er genereret ud fra QSTFT, hvor Q sættes til 3 forskellige funktioner; nemlig $Q_f(\nu) = 1$, som svarer til STFT, $Q_o(\nu) = \sqrt{\nu}$, som svarer til QSTFT og $Q_c(\nu) = \nu$, som svarer til en STFT med constant-Q-transformation, eller en MWT med et Hann-vindue. På figur 3.20, 3.21 og 3.22 er vindueslængden på 0,034 sekunder for en frekvens på 1.000 Hz.

Lad os først se på figur 3.20, som anvender Q_f . Det ses, at frekvenserne



Figur 3.19: Her vises en STFT af en lydoptagelse af en mandlig operasanger som synger en tone på omkring 190 Hz. Operasangeren anvender vibrato, og dette får tonen til at have et frekvensmæssigt udsving på omkring 10%. Lyden findes i filen, som hedder "test2.wav". Den maksimale styrke definerer 0 dB.

over 2.000 Hz er noget utydelige, der hvor frekvensen skifter hurtigt. For frekvenserne over 4.000 Hz har der indsnæget sig "huller" ind i overtonernes bølger, således at en overtone knapt kan følges som en fuldt optrukket kurve.

På figur 3.21, som anvender Q_o , kan overtonerne bedre adskilles ind til omkring 3.000 Hz. For frekvenserne omkring 4.000 Hz er der, set i forhold til på figur 3.20, i højere grad tale om fuldt optrukne kurver.

På figur 3.22, som anvender Q_c , kan overtonerne adskilles ind til omkring 3.000 Hz. Frekvenserne under 1.000 Hz lider i højere og højere grad under en uklar tidsmæssig bestemmelse. Desuden ser det ud, som om der længst højre på figuren er tale om 2 frekvenser nær fundamentalfrekvensen. Det skal bemærkes, at QSTFT i dette tilfælde kun er beregnet ned til 150 Hz, mens QSTFT på de øvrige blev beregnet ned til 50 Hz. Dette skyldes de ekstremt lange vinduer, som skal anvendes under 150 Hz.

Af de figurer, vi netop har set på, er det klart, at specielt QSTFT med Q_c har meget store problemer, både for lave frekvenser og for høje frekvenser. Lad os nu se på et eksempel, hvor det kunstige signals fundamentalfrekvens ligger omkring 95 Hz. Vi kasserer allerede Q_c og ser derfor kun på Q_f og Q_o . På figur 3.23 er Q_f anvendt, og her ses det, at frekvenserne akkurat kan adskilles op til omkring 2.000 Hz. På figur 3.24, hvor Q_o anvendes, er der også problemer med adskillelsen af frekvenserne over 2.000 Hz, men specielt frekvenserne under 500 Hz er klart bedre adskilte.

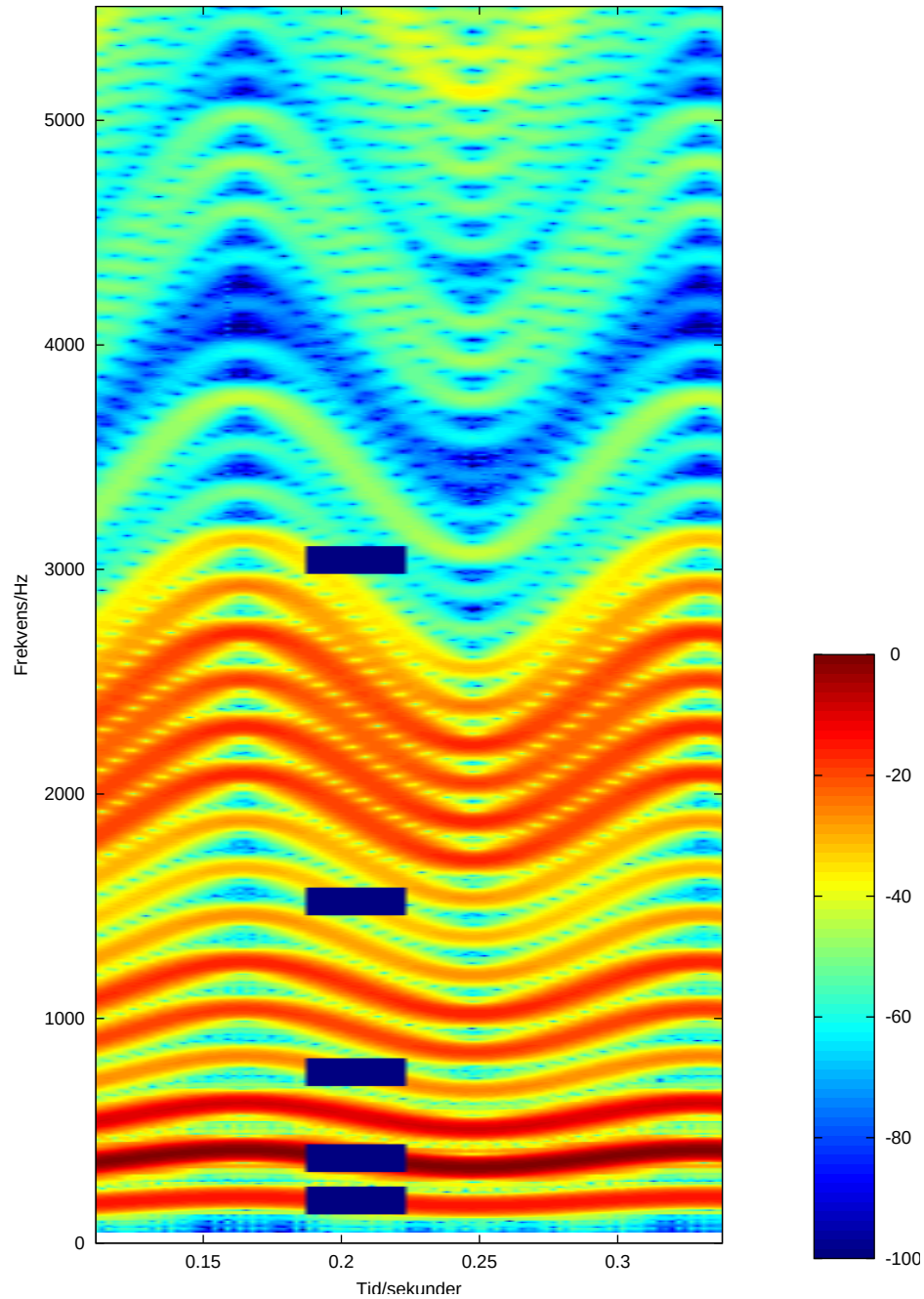
Eksperimentet bekræfter altså hypotesen om, at Q_o bedre end både Q_f og Q_c gengiver indholdet af et signal med en svagt vibrerende harmonisk tone. Der har kun været udført et ganske simpelt eksperiment her. I forhold til en given metode, som estimerer melodiens fundamentalfrekvens på baggrund af en tid/frekvens-repræsentation, vil det være nødvendigt at foretage adskillige afprøvninger ved hjælp af både Q_o og Q_f , før det kan vurderes, om den givne metode nyder godt af at anvende Q_o .

3.6 STFT med variabel vindueslængde

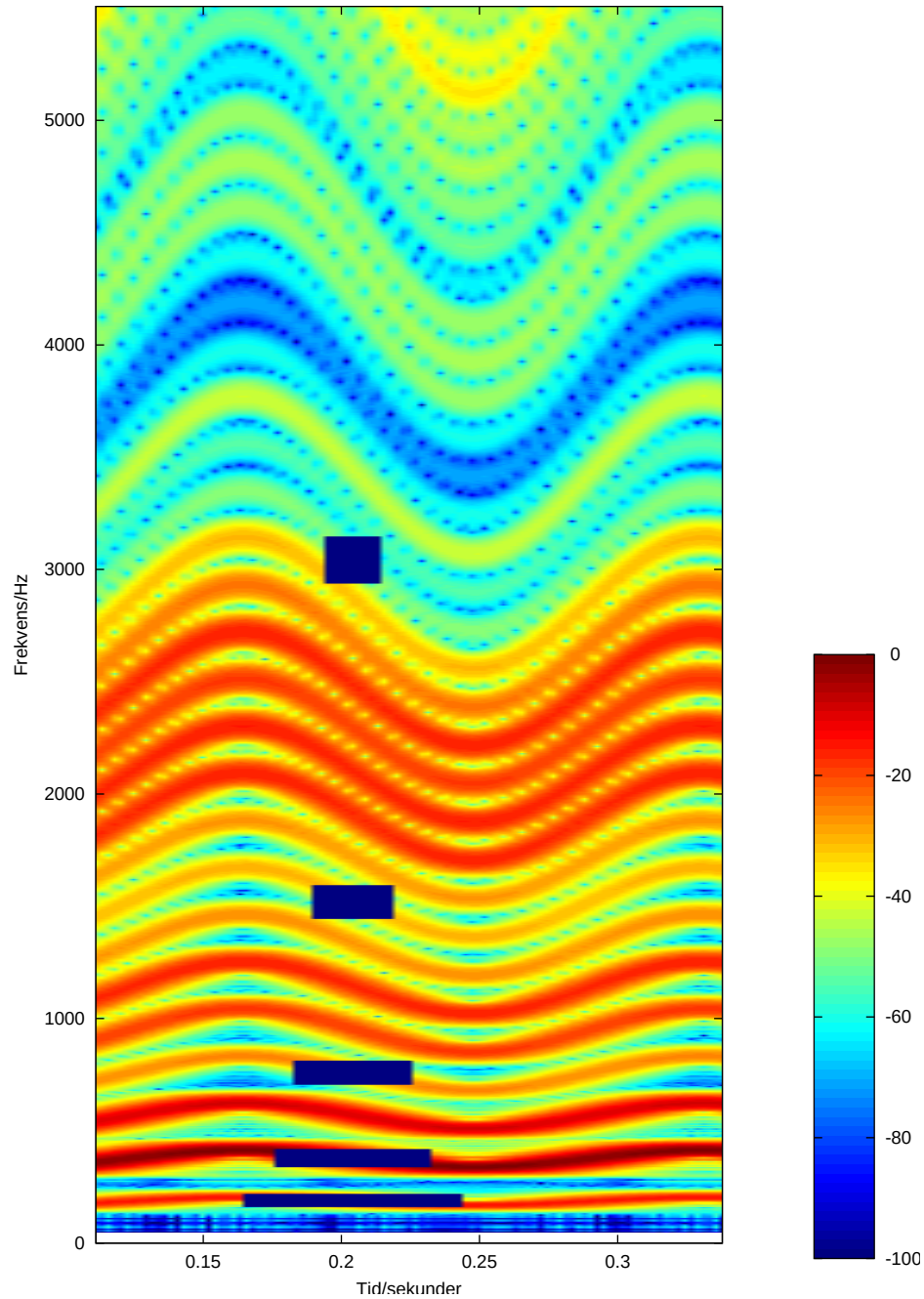
Der findes flere eksempler på brugen af STFT'er, som anvender variable vindueslængder eller multiple vindueslængder. Man kan beregne STFT for flere vindueslængder og så udvælge netop den længde, som på baggrund af et eller andet kriterie, fremstår som den bedste til et bestemt tidspunkt og for en bestemt frekvens.

I det følgende er det valgt kun at se på et enkelt eksempel på en multiopløsnings-STFT, altså en STFT beregnet for flere forskellige vindueslængder. Men der findes flere eksempler på anvendelsen af multiopløsnings-STFT'er. Som eksempler herpå kan nævnes: I [35] anvendes STFT'er med korte vinduer til bestemmelse af, hvornår en ny tone begynder, og STFT'er med lange vinduer anvendes til bestemmelse af toners frekvens. I [47] anvendes en multiopløsnings-STFT, hvor vindueslængderne til hvert tidspunkt og for hver frekvens bestemmes ud fra lokale forhold i signalet. Den anvendte adaptive multiopløsnings-STFT viser sig at forbedre støjreduktion af nogle musik-signaler.

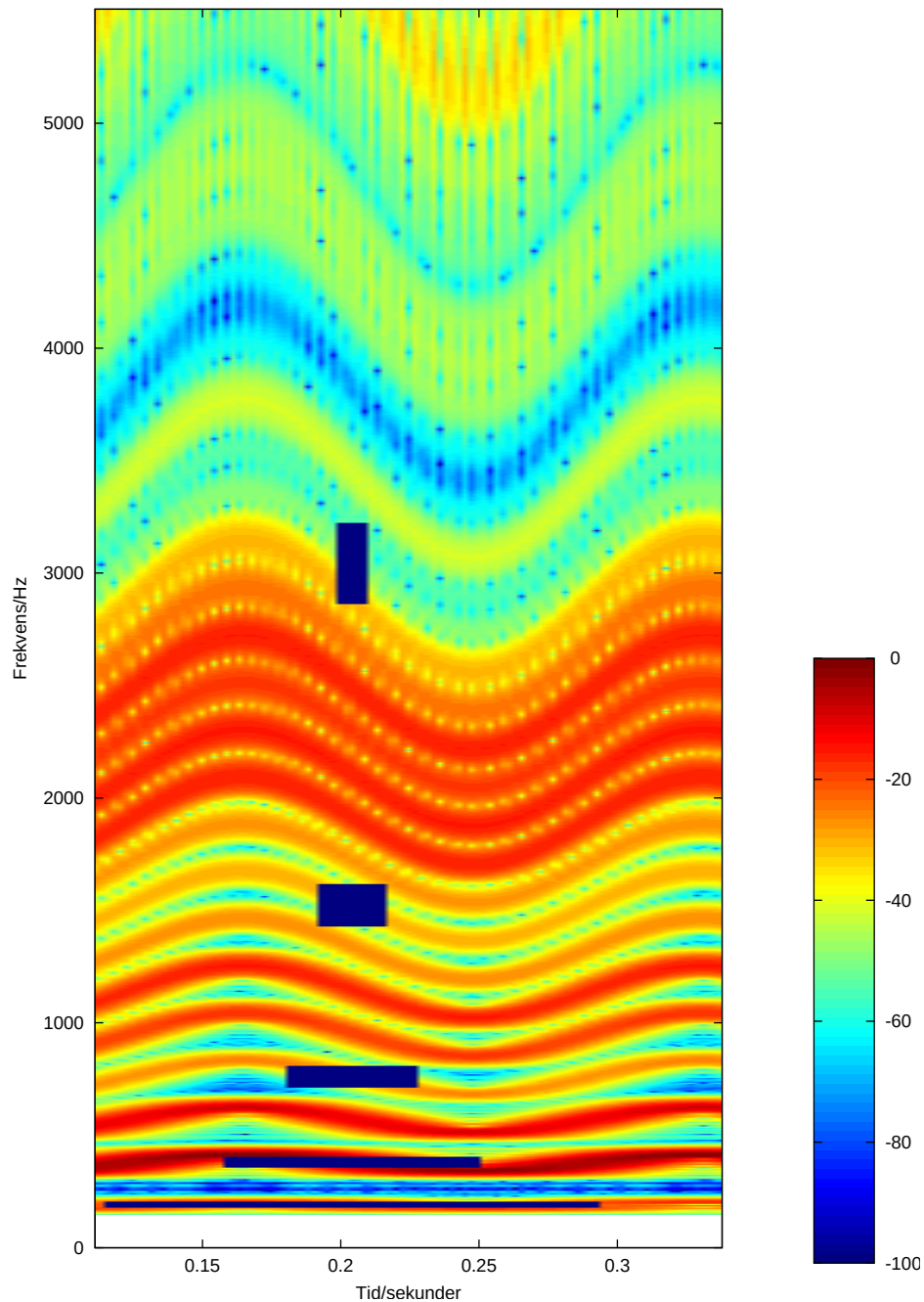
Metoden, som behandles i næste afsnit, angiver blot en smart måde, hvorpå man kan udregne multiple STFT'er. Metoden, som blev præsenteret af Dressler i [11], blev anvendt til at generere en tid/frekvens-repræsentation, hvor forskellige frekvensbånd anvender forskellige vindueslængder; vindueslængderne er



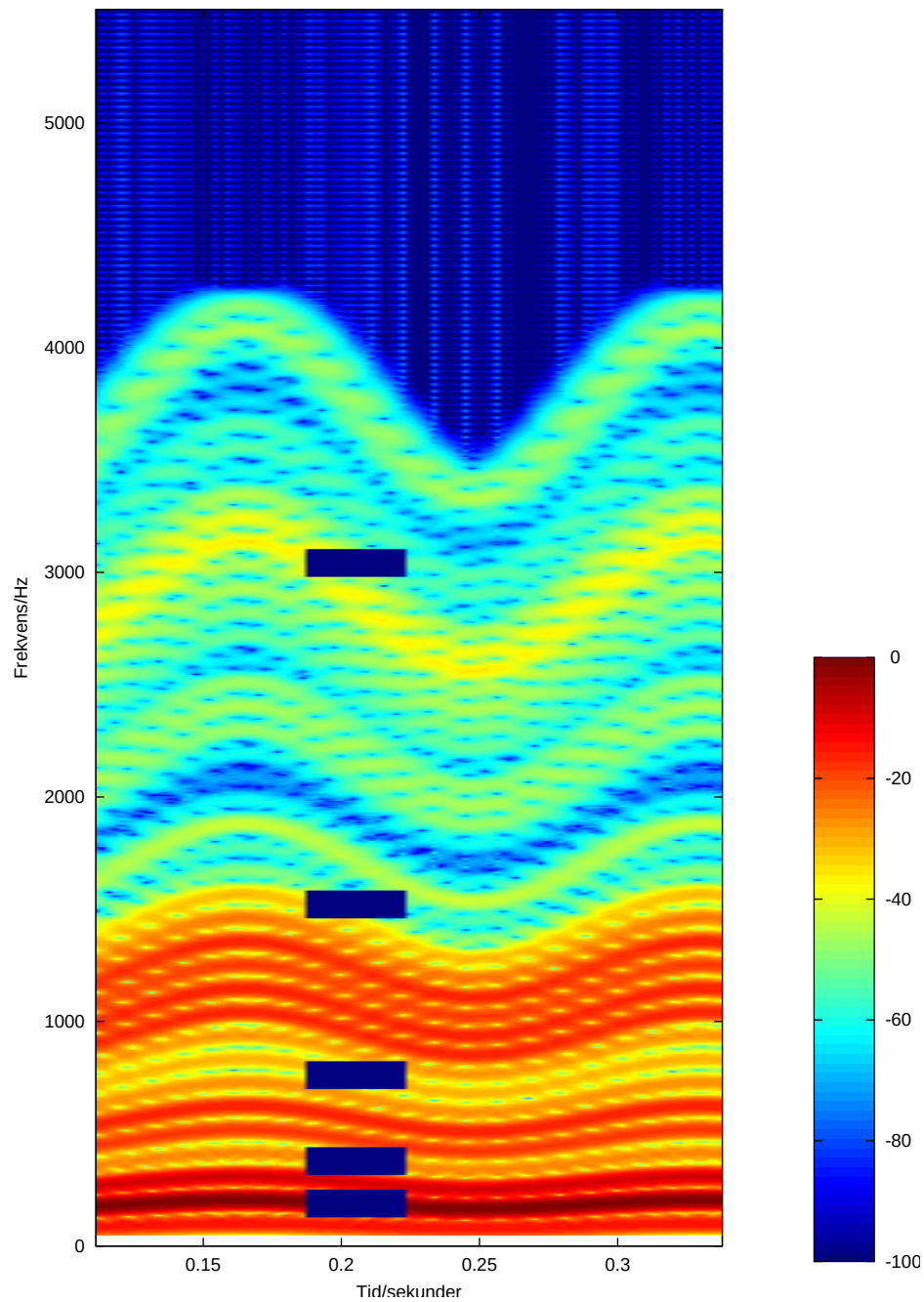
Figur 3.20: Her ses en QSTFT med Q_f af det kunstige signal, hvor fundamentalfrekvensen ligger omkring 190 Hz. De mørke kasser angiver en skematisk størrelse for de anvendte vinduer. Det ses, at der opstår problemer med klart at adskille 2 overtoner, når frekvensen ligger over ca. 2.000 Hz. Den maksimale styrke definerer 0 dB.



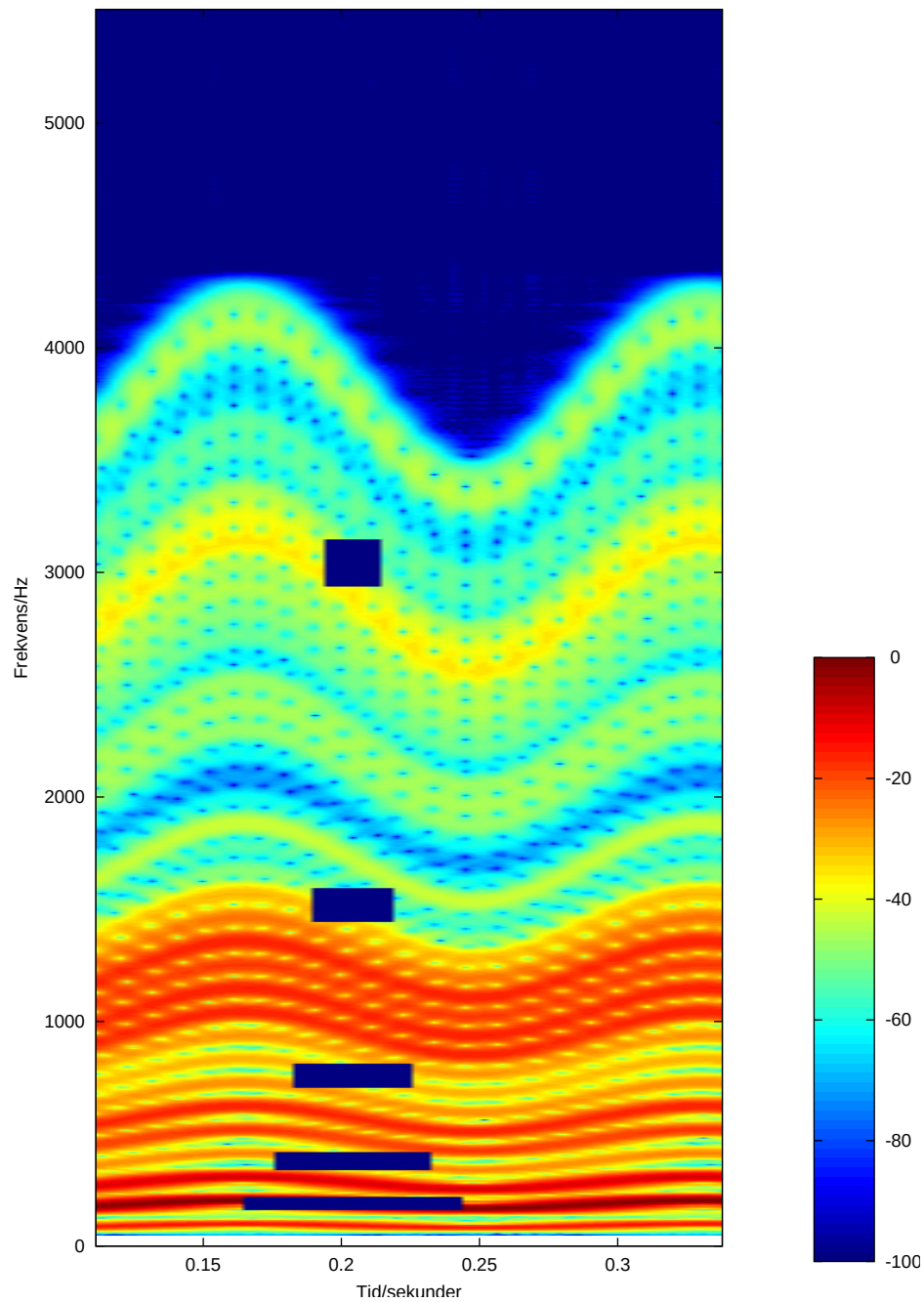
Figur 3.21: Her ses en QSTFT med Q_0 af det kunstige signal, hvor fundamentalfrekvens ligger omkring 190 Hz. De mørke kasser angiver en skematisk størrelse for de anvendte vinduer. Det ses, at overtonerne set i forhold til på figur 3.20 bedre kan adskilles op til omkring 3.000 Hz. Den maksimale styrke definerer 0 dB.



Figur 3.22: Her ses en QSTFT med Q_c af det kunstige signal, hvor fundamentalfrekvensen ligger omkring 190 Hz. De mørke kasser angiver en skematisk størrelse for de anvendte vinduer. Overtonerne kan adskilles til omkring 3.000 Hz, men der er store tidsmæssige ændringer i nærheden af 3.000 Hz. For fundamentalfrekvensen opstår der et problem med det lange vindue, hvilket får det til at se ud som om, der er 2 frekvenser til stede i nærheden af fundamentalfrekvensen — se længst til højre. Den maksimale styrke definerer 0 dB.



Figur 3.23: Her ses en QSTFT med Q_f af det kunstige signal, hvor fundamentalfrekvensen ligger omkring 95 Hz. De mørke kasser angiver en skematisk størrelse for de anvendte vinduer. Det ses, at der er problemer med adskillelsen af frekvenserne over 2.000 Hz. Den maksimale styrke definerer 0 dB.



Figur 3.24: Her ses en QSTFT med Q_0 af det kunstige signal, hvor fundamentalfrekvensen ligger omkring 95 Hz. De mørke kasser angiver en skematisk størrelse for de anvendte vinduer. Her er der også problemer over 2.000 Hz, men frekvenserne under 500 Hz står tydeligere her end på forrige figur. Den maksimale styrke definerer 0 dB.

dog ikke afhængige af signalet, men fastlagt på forhånd som funktion af frekvensen. Men blot fordi metoden blev anvendt til at generere en repræsentation, hvor vindueslængden var afhængig af frekvensen, kan metoden sagtens anvendes i repræsentationer, hvor vindueslængden er variabel.

3.6.1 Dresslers multiopløsnings-STFT

Dresslers multiopløsnings-STFT blev præsenteret i [12]. I stedet for at gange vinduet på i tidsdomænet og så foretage fouriertransformationen, foldes der med multiple vinduer i frekvensdomænet. Dette giver mulighed for effektivt at beregne fouriertransformationen med forskellige vindueslængder. Vi vil i det følgende forklare, hvordan Dresslers multiopløsnings-STFT beregnes, men notationen og forklaringen adskiller sig væsentligt fra den, som Dressler anvender i [12]. Den notation og forklaring, som følger herefter, passer bedre til den notation, som har været anvendt i de foregående afsnit.

3.6.1.1 Opdeling af STFT

Lad os først se på formel 3.10 for den diskrete STFT. Vi bemærker, at det anvendte vindue spejles og flyttes, som det er tilfældet med foldning. Men nu vil vi anvende vinduer, som ikke er centrerede omkring 0, og det vil derfor være en fordel, hvis STFT ikke er formuleret som en foldning. Derfor vil vi notere definitionen for den diskrete STFT igen, og her bytte om på n og η :

$$\text{stft}\{\mathbf{s}\}[n,k] = \sum_{\eta=0}^{N-1} \mathbf{s}[\eta] \mathbf{w}[\eta - n] e^{-i2\pi k\eta/N} \quad (3.50)$$

De vinduer, vi hidtil har set på, er centrerede omkring 0, og det bevirker, at fouriertransformationen, beregnet til sample n , primært beregnes på baggrund af samples rundt om n . Vi skal her se på en beregningsmetode, hvor fouriertransformationen, beregnet til sample n , primært beregnes på baggrund af samples, som ligger i et interval efter n . Dette kan virke noget ulogisk, men kan være praktisk i forhold til notation.

Lad os nu definere 2 rektangulære vinduer, som har den egenskab, at de er åbne i et interval efter 0. Det ene vindue $\mathbf{w}_1$ har den halve effektive længde af det andet vindue $\mathbf{w}_2$, og vindue $\mathbf{w}_2$ kan beskrives ved 2 korte vinduer af typen $\mathbf{w}_1$, som ligger i forlængelse af hinanden. Vi kalder den effektive længde af $\mathbf{w}_1$ for l_1 og den effektive længde af $\mathbf{w}_2$ for l_2 ; således er $l_2 = 2l_1$. De 2 vinduer kan defineres som:

$$\mathbf{w}_1(n) = \begin{cases} 1 & \text{for } 0 \leq n < l_1 \\ 0 & \text{ellers} \end{cases} \quad (3.51)$$

$$\mathbf{w}_2(n) = \mathbf{w}_1(n) + \mathbf{w}_1(n - l_1) \quad (3.52)$$

Vi antager, at et signal $\mathbf{s}$ har en længde på Ll_1 , således at vinduet $\mathbf{w}_1$ kan være L gange i signalets længde. Dette anvender vi til at opdele summen i formel 3.50, således at der opstår L summer over tidsintervaller på længden l_1 , og vi kalder

disse summer for $\mathbf{d}$:

$$\mathbf{stft}\{\mathbf{s}\}[n, k] = \sum_{j=0}^{L-1} \left(\sum_{\eta=0}^{l_1-1} \mathbf{s}[\eta + jl_1] \mathbf{w}[(\eta + jl_1) - n] e^{-i2\pi k(\eta + jl_1)/N} \right) \quad (3.53)$$

$$= \sum_{j=0}^{L-1} \mathbf{d}[j] \quad (3.54)$$

Lad os nu antage, at vi kun ønsker at beregne STFT til tidspunkterne $n = ql_1$, hvor q er et heltal, som løber fra 0 til $L - 1$. I så fald kan vi erstatte n med ql_1 i formel 3.53. Bemærk nu, at hver delsum $\mathbf{d}[j]$ kun anvender l_1 samples. Hvis vi indsætter $\mathbf{w}_1$ i formel 3.53 og betragter eksemplet, hvor $q = j$, så vil vinduet $\mathbf{w}_1$ være lig 1 for alle værdier af η løbende fra 0 til $l_1 - 1$. Og da vinduet netop har en effektiv længde på l_1 må vinduet være lig 0 for alle $q \neq j$. Det betyder, at beregningen af STFT reduceres til:

$$\mathbf{stft}_{w_1}\{\mathbf{s}\}[ql_1, k] = \mathbf{d}[q] \quad (\text{hvor } q \text{ løber fra } 0 \text{ til } l_1 - 1) \quad (3.55)$$

og $\mathbf{d}$ reduceres til:

$$\mathbf{d}[j] = \sum_{\eta=0}^{l_1-1} \mathbf{s}[\eta + jl_1] e^{-i2\pi k(\eta + jl_1)/N} \quad (3.56)$$

Det interessante er nu at betragte det samme scenarie igen, men erstatte vinduet med $\mathbf{w}_2$, som har den dobbelte effektive længde. Da $\mathbf{w}_2$ er dobbelt så bred som $\mathbf{w}_1$, vil $\mathbf{w}_2$ være åben i det dobbelte interval, hvilket medfører at:

$$\mathbf{stft}_{w_2}\{\mathbf{s}\}[ql_1, k] = \mathbf{d}[q] + \mathbf{d}[q + 1] \quad (3.57)$$

Så hvis man først har udregnet delsummerne $\mathbf{d}$ ved anvendelse af vinduet $\mathbf{w}_1$, kan man hurtigt finde STFT for $\mathbf{w}_2$. Både i tilfældet med $\mathbf{w}_1$ og i tilfældet med $\mathbf{w}_2$ er den anvendte hop-size lig med l_1 , således at der ikke er noget overlap i tilfældet med $\mathbf{w}_1$, men i tilfældet med $\mathbf{w}_2$ er der et overlap på halvdelen af vinduets effektive bredde.

Bemærk at overlappet i tilfældet med $\mathbf{w}_2$ netop bevirker, at resultatet af STFT beregnet til tidspunktet ql_1 kan anvendes til beregningen af STFT til tidspunktet $(q + 1)l_1$, således at:

$$\begin{aligned} \mathbf{stft}_{w_2}\{\mathbf{s}\}[(q + 1)l_1, k] &= \mathbf{d}[(q + 1)] + \mathbf{d}[(q + 1) + 1] \\ &= \mathbf{d}[q] - \mathbf{d}[q] + \mathbf{d}[q + 1] + \mathbf{d}[(q + 1) + 1] \\ &= \mathbf{d}[q] + \mathbf{d}[q + 1] - \mathbf{d}[q] + \mathbf{d}[(q + 1) + 1] \\ &= \mathbf{stft}_{w_2}\{\mathbf{s}\}[ql_1, k] - \mathbf{d}[q] + \mathbf{d}[(q + 1) + 1] \end{aligned} \quad (3.58)$$

Det antal frekvenser, som benyttes ved udregningen af $\mathbf{d}$, har betydning for bestemmelsen af $\mathbf{stft}_{w_1}$, $\mathbf{stft}_{w_2}$ o.s.v., så antallet af frekvenser skal helst fastsættes, så det passer til det største vindues effektive længde. Det bevirker blot, at der indirekte anvendes zero-padding for de øvrige vinduer. Så jo kortere et vindue er, des mere zero-padding, og jo længere et vindue er, des mere overlap.

3.6.1.2 Tilføjelse af vindue

Den metode, vi indtil videre har gennemgået, anvender et rektangulært vindue, hvilket er problematisk i forhold til side lobes. Vi ønsker at tilføje et mere

fordelagtigt vindue, således at maksimum for vinduets side lobes mindskes. I afsnit 3.2.1.6 om sammenligning af vinduer så vi, at især Hann-vinduet og det sænkede Gauss-vindue udviser hastigt aftagende maksimum for side lobes. Da vi er nødt til at tilføje et vindue i frekvensdomænet (og derfor skal folde med den fouriertransformerede til vinduet), er det vigtigt at undersøge beregningskompleksiteten for denne foldning. Her træder Hann-vinduet frem som det vindue, der helt klart har en fordel, da den fouriertransformerede til Hann-vinduet er lig 0 for alle frekvenser, undtagen 3 frekvenser (se formel 3.24). Gauss-vinduet derimod har et gaussisk udseende i frekvensdomænet og vil derfor være meget beregningstungt at folde med.

3.6.1.3 Beregningskompleksitet

Lad os nu se på beregningskompleksiteten for beregningen af en multiopløsnings-STFT ved hjælp af Dresslers metode, set i forhold til multiple STFT'er. I begge tilfælde ønsker vi, at beregne en multiopløsnings-STFT, hvor den maksimale vindueslængde er l_{max} , og l_{max} angiver samtidig antallet af frekvenser i hver STFT. Den minimale vindueslængde, som anvendes, sættes til at være l_1 og er længden af vinduet w_1 , som blev defineret i formel 3.51.

Først konstaterer vi, at det i begge tilfælde (multiple STFT'er og Dresslers metode) er nødvendigt at beregne en STFT med den minimale vindueslængde l_1 , og beregningskompleksiteten herved er den samme. Til at beregne STFT for den minimale vindueslængde anvendes FFT, så beregningskompleksiteten er som udgangspunkt $\mathcal{O}(l_1 \log_2(l_1))$, men da der skal beregnes for l_{max} frekvenser ender beregningskompleksiteten med at blive $\mathcal{O}(l_{max} \log_2(l_1))$.

Vi vil undlade at se på beregningskompleksiteten for beregningen af hver enkelt STFT, men nøjes med at se på beregningen af den STFT, hvor den anvendes den maksimale vindueslængde l_{max} . Hvis vi ikke bruger Dresslers metode er beregningskompleksiteten $\mathcal{O}(l_{max} \log_2(l_{max}))$ for beregningen af den STFT, hvor vindueslængden er l_{max} .

Lad os nu se på anvendelsen af Dresslers metode og en vindueslængde på l_{max} . Ifølge formel 3.57 kræves der l_{max}/l_1 additioner pr. frekvens, fordi det længste vindue er l_{max}/l_1 gange længere end det korteste. Hvis STFT for det længste vindue beregnes på denne måde, vil det være mere krævende at anvende Dresslers metode end at beregne STFT direkte. Men da hop-size er lig l_{min} , kan formel 3.58 anvendes for en effektiv beregning af STFT. Hvis formel 3.58 anvendes, kræves der 2 additioner pr. frekvens for at beregne STFT for det længste vindue (forudsat at STFT for det næstlængste vindue allerede er beregnet). Foldningen i frekvensdomænet kræver 3 additioner (samt 2 multiplikationer og et antal operationer for indeksering) pr. frekvens. Konklusionen er altså, at ved hjælp af Dresslers metode har beregningen af STFT for det længste vindue en beregningskompleksitet på $\mathcal{O}(l_{max})$, da antallet af operationer afhænger lineært af l_{max} .

Da beregningskompleksiteten er af en lavere orden ved anvendelse af Dresslers metode, vil denne metode være en fordel, når multiopløsnings-STFT'en anvender mange vinduesstørrelser. Det er dog ikke nødvendigvis en fordel at bruge Dresslers metode, når der kun anvendes få vinduesstørrelser, da den konstante omkostning pr. frekvens i dette tilfælde kan være for stor. Om Dresslers metode er en fordel afhænger primært af, hvor effektivt der kan indekseres i lagrede data.

3.6.1.4 Anvendelse i forbindelse med QSTFT

Vi har nu set, at Dresslers metode beregningsmæssigt set kan være en fordel set i forhold til multiple STFT'er. Derfor kan det være interessant at undersøge, om beregningen af QSTFT kan nyde godt af Dresslers metode. En sådan undersøgelse kræver dog at det nøjagtige antal beregninger i en faktisk implementering findes, da valget af programmeringssprog og platform kan have stor betydning for effektiviteten. Der har ikke været foretaget en sådan undersøgelse.

3.7 Chirplet-transformation

Her vil vi se på en transformation, som blev introduceret af Mann *et al.* i 1991 i [51]. Ordet "chirp" dækker over en bølge, hvis frekvens ændrer sig konstant i tiden. Vi har i eksemplerne, som blev præsenteret i afsnit 3.5.3.2, set, at STFT og QSTFT begge har problemer når en tone gradvist ændrer frekvens - omend QSTFT gav bedre mulighed for at kompensere for dette. Problemet med de nævnte transformationer er, at et signal med vekslende frekvens sammenlignes med et signal, som har konstant frekvens. Chirplet-transformationen skal kompensere for dette problem ved at tillade sammenligningen med et signal, som ikke har konstant frekvens.

Chirplet-transformationen er en generalisering af MWT, men kan også ses som en generalisering af Gabor-transformation. Således kan de 2 transformationer ses som specialtilfælde af chirplet-transformationen.

Både Gabor-transformationen (og STFT) og MWT afhænger af 2 variable; nemlig en tidsmæssig variabel og en frekvensmæssig variabel. Den frekvensmæssige variabel i MWT fungerer samtidig som en skaleringsvariabel. Chirplet-transformationen derimod afhænger af mindst 4 variable; nemlig en tidsmæssig variabel b , en frekvensmæssig variabel d , en "chirpy-ness" variabel c , som angiver, hvor hurtigt frekvensen ændrer sig, samt en skalamæssig variabel a , som angiver, hvor meget en chirplet er strakt, og dermed angives også chirplet'ens tidsmæssige længde. (Variablenes navne er valgt, så a og b har samme betydning som i wavelet-transformationen.) For at simplificere og begrænse forklaringerne har vi her valgt kun at se på de 4 nævnte variable, selv om der, som det nævnes senere, kan tilføjes flere variable.

Lad os nu definere en chirp som et signal ved:

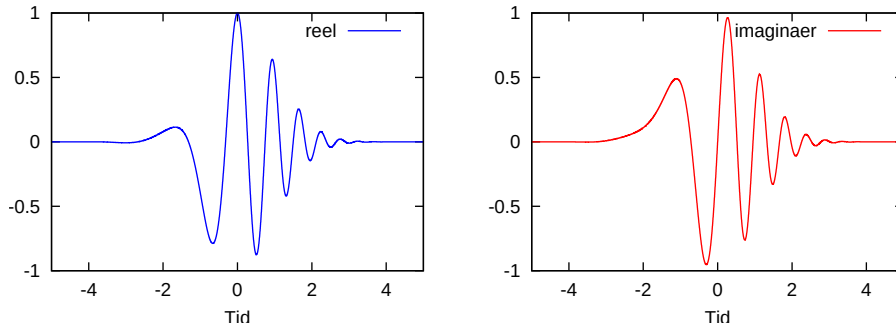
$$\mathbf{s}_{chirp}\{c, d\}(t) = e^{-i2\pi(ct+d)t} = e^{-i2\pi ct^2} e^{-i2\pi dt} \quad (3.59)$$

således at den umiddelbare frekvens (instantaneous frequency) er lineært afhængig af t . Tilfældet hvor $c = 0$ svarer til et signal med konstant frekvens, hvor frekvensen ν angives med variabelen d . Bemærk at vi i stedet for parametren d også kan angive, at en chirp til tiden t_1 har en umiddelbar frekvens på ν_1 , og frekvensen ændrer sig med hældningen c . Vi kan opfatte ovenstående definition på en chirp, som en multiplikation af en "chirpy-ness"-funktion $\mathbf{c}$ og en sinusoidal bølge:

$$\mathbf{s}_{chirp}\{c, d\}(t) = \mathbf{c}\{c\}(t) e^{i2\pi dt} \quad \text{hvor: } \mathbf{c}\{c\}(t) = e^{i2\pi ct^2} \quad (3.60)$$

En chirplet er en chirp, som er lokaliseret i tiden, på samme måde som en wavelet er en bølge, der er lokaliseret i tiden. Så ved at gange en chirp med et gaussisk vindue fås en chirplet:

$$\psi_{chirp}\{c, d\}(t) = \mathbf{g}(t) \mathbf{c}\{c\}(t) e^{i2\pi dt} \quad \text{hvor: } \mathbf{g}(t) \text{ er en gaussisk funktion} \quad (3.61)$$



Figur 3.25: Her vises et eksempel på en chirplet. Til venstre ses den reelle del og til højre ses den imaginære del.

Vi mangler dog at give chirplet'en en tidsmæssig og en skalamæssig variabel. Begge variable påvirker tiden og har derfor betydning for både det gaussiske vindue og chirp'en. Først giver vi chirplet'en den tidsmæssige variabel b :

$$\psi_{chirp}\{b, c, d\}(t) = \psi_{chirp}\{c, d\}(t - b) \quad (3.62)$$

$$= \mathbf{g}(t - b) \mathbf{c}\{c\}(t - b) e^{i2\pi d(t-b)} \quad (3.63)$$

Dernæst giver vi chirp'en den skalamæssige variabel:

$$\psi_{chirp}\{a, b, c, d\}(t) = \frac{1}{\sqrt{a}} \psi_{chirp}\{b/a, c, d\}(t/a) \quad (3.64)$$

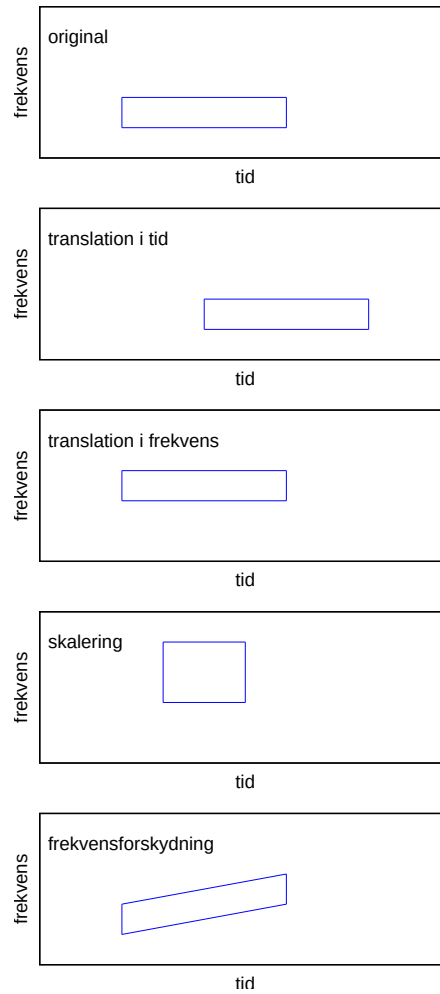
$$= \frac{1}{\sqrt{a}} \mathbf{g}((t - b)/a) \mathbf{c}\{c\}((t - b)/a) e^{i2\pi d(t-b)/a} \quad (3.65)$$

På figur 3.25 er vist et eksempel på en chirplet, og på figur 3.26 er effekten af de 4 parametre vist. På den øverste figur er en tidsmæssigt begrænset sinusoidal bølge, som vi her vil kalde den originale bølge, vist på skematisk vis i en tid/frekvens-repræsentation. Denne bølge kan påvirkes af de 4 parametre, som bestemmer en chirplets udseende. På den næstøverste figur er den originale bølge translateret i tiden ved hjælp af parametren b . På den mellemste figur er den originale bølge translateret i frekvensen ved hjælp af parametren d . På den næstnederste figur er den originale bølge skaleret ved hjælp af parametren a . På den nederste figur er den originale bølge forskudt i frekvensen ved hjælp af parametre c .

Vi vil ikke give en klar definition på chirplet-transformationen, da det ikke ligger fast, hvilke parametre der er frie, og hvilke der holdes fast. Derfor vil vi nøjes med at angive, at en chirplet-transformation blot er et indre produkt af en chirplet og et signal:

$$\mathbf{CT}_{\psi_{chirp}}\{\mathbf{s}\}(a, b, c, d) = \int_{-\infty}^{\infty} \mathbf{s}(\tau) \psi_{chirp}\{a, b, c, d\}(\tau) d\tau \quad (3.66)$$

Det ses, at chirplet-transformationen svarer til en MWT, når vi fastholder d og sætter $c = 0$. Vi kan også fastholde a og sætte $c = 0$, og på denne måde have de frie parametre b og d , således at chirplet-transformationen svarer til Gabor-transformationen. Chirplet-transformationen er interessant, hvis vi ikke sætter $c = 0$, for da vil transformationen være god til at identificere chirps, hvis chirpy-ness er lig eller tæt på c . Ved at fastlåse c og enten a eller d kan vi generere en tid/frekvens-repræsentation, hvor frekvensen angiver den gennemsnitlige frekvens for en chirplet med de anvendte parametre.



Figur 3.26: Her illustreres det, hvordan en tidsbegrænset sinusoidal bølge påvirkes af en chirplets 4 parametre. På næstøverste figur er den originale bølge blevet translateret i tiden. På den mellemste figur er bølgen translateret i frekvensen. På den næstnederste figur er bølgen skaleret. På den nederste figur er bølgen forskudt i frekvensen.

Flere parametre Vi har her præsenteret 4 parametre, som kan modificere en enkel lille bølgeform. Som det bliver foreslået i [52], kan en bølge også forskydes i tidsdomænet (parallelle til den frekvensmæssige forskydning c), og der kan foretages perspektivisk modulering i både tid- og frekvensdomænet, således at firkanterne på figur 3.26 kan antage en vilkårlig form, som består af 4 sider, og altså ikke er begrænset til rektangler og parallelogrammer.

Multiple dimensioner Gabor-transformationen (og STFT) og MWT kan beregnes for forskellige tid/frekvens-opløsninger, og dette giver anledning til at repræsentere transformationerne i 3 dimensioner. (I tilfældet med STFT er de 3 dimensioner: Tiden, frekvensen og vindueslængden.) Med chirplet-transformationen tilføjes mindst 1 ekstra dimension i form af chirpy-ness parametren. Spørgsmålet om hvor mange dimensioner, man ønsker at arbejde med, afhænger nok mest af, hvor hurtigt beregningerne kan foretages. Selv om det med chirplet-transformationen er muligt at anvende adskillige dimensioner, vil man i praksis ofte begrænse sig til 3 dimensioner; og ofte vil man udvælge netop de 3 dimensioner, som bliver valgt i [43], hvor tid (b), frekvens (d) og chirpy-ness (c) er udvalgt.

Tilpasset chirplet-transformation En tilpasset chirplet-transformation (adaptive chirplet transform) kan give en tid/frekvens-repræsentation, som beregnes på baggrund af en multi-dimensionel chirplet-transformation. Til enhver tid og frekvens søges der i en multi-dimensionel chirplet-transformation efter den kombination af parametre for de øvrige dimensioner, som bedst opfylder et eller andet kriterie.

3.7.1 STFChT

Her vil vi ganske kort præsentere et bud på hvordan chirplet-transformationen kan bruges. STFChT, som blev præsenteret i [44], står for "Short-Time Fan-Chirp Transform" og er en generalisering af STFT. Mens STFT bruger sinusoidale bølger med konstant frekvens som analyserende bølger, anvender STFChT chirps som analyserende bølger.

Hvis vi opfatter STFChT som en form for tilpasset chirplet-transformation, er det en chirplet-transformation, hvor variabelen a (skalering) er fast, variabelen c (chirpy-ness) er næsten fast, da den er bestemt på forhånd men frekvensafhængig, og variabelen d (vinduets størrelse) er fast. Det betyder, at STFChT arbejder i et 2-dimensionelt rum, men anvender multiple c afhængig af frekvensen. Dette står i kontrast til en generel tilpasset chirplet-transformation, hvor der ofte arbejdes i et rum på mindst 3 dimensioner.

STFChT er beregnet for identifikation af overtonerne i menneskelig tale, og er derfor rettet mod analysen af et signal, som primært består af en fundamental-frekvens og tilhørende overtoner. Antagelsen er, at fundamental-frekvensen inden for et kort tidsrum (vinduets længde) ændrer sig med en konstant hastighed og altså opfører sig som en chirp. Vi så i afsnit 3.5.3.1, at en overtones chirpy-ness kan bestemmes, hvis man kender fundamental-frekvensens chirpy-ness. Lad os se på en overtone, som har en frekvens på x gange fundamental-frekvensens frekvens. Fundamental-frekvensen har en chirpy-ness på c . Da har overtonen en chirpy-ness på cx . Dette anvender STFChT, da chirpy-ness for en tilfældig frekvens fastsættes som forholdet mellem frekvensen og fundamental-frekvensens

frekvens ganget med fundamentalfrekvensens chirpy-ness.

I nogle metoder, som anvender chirplet-transformationen, beregnes transformationen for flere forskellige chirpy-ness værdier for hver frekvens, og den chirpy-ness-værdi, som ser ud til at give det mest detaljerede spektrum, vælges. På samme måde kunne der i STFChT anvendes flere gæt på fundamentalfrekvensens chirpy-ness. Men antagelsen er i stedet for, at der kan gives et godt bud på fundamentalfrekvensens frekvens og chirpy-ness til tiden t allerede inden STFChT er beregnet til dette tidspunkt. Hvis blot chirpy-ness-værdien ligger tæt på fundamentalfrekvensens chirpy-ness, er spektret detaljeret, og fundamentalfrekvensen antages derfor at være let at finde. STFChT indeholder en metode for estimering af fundamentalfrekvensen ud fra spektret. Måden, hvorpå denne metode anvendes, beskrives i det følgende: Vi antager, at metoden har været anvendt til bestemmelse af fundamentalfrekvensen til n tidspunkter før tidspunkt t , og vi ønsker et gæt på fundamentalfrekvensen til tidspunkt t . De n målinger af fundamentalfrekvensen anvendes til dannelse af en andengradsfunktion, som tilnærmer målingernes forløb. Ud fra den fundne andengradsfunktion gættes der på fundamentalfrekvensens frekvens og chirpy-ness til tidspunkt t .

STFChT er altså afhængig af en god bestemmelse af fundamentalfrekvensen. Dette er måske ikke noget stort problem i forbindelse med lyden fra en enkelt menneskelig stemme, men i polyfon musik, hvor nogle toner kan være på vej ned, mens andre går op, kan det være et stort problem at anvende transformationen. Anvendelsen af transformationen kræver i hvert fald, at man kun er interesseret i at finde én fundamentalfrekvens til enhver tid, samt at den metode, som finder fundamentalfrekvensen, også kan finde denne, når andre instrumenter er kraftigere end det instrument, som har melodien. Hvis man ønsker at finde flere fundamentalfrekvenser for at følge flere instrumenter samtidig, kan det være et stort problem at spektret genereres på baggrund af én funden fundamentalfrekvens.

3.7.2 Praktisk eksempel

Vi vil ikke afprøve den fulde STFChT, men i stedet for blot se på, hvilke muligheder der ligger i den grundlæggende transformation, som anvendes i STFChT, og som kaldes for FChT (Fan-Chirp Transform). Herunder gives en definition på FChT, hvor der ikke anvendes nøjagtig den notation, som er angivet i [44].

$$\text{fcht}_c\{\mathbf{s}\}(\nu) = \int_{-\infty}^{\infty} \mathbf{s}(\tau) \mathbf{w}(\tau) \overline{\xi_{\{c\}}(\tau, \nu)} d\tau \quad (3.67)$$

hvor $\mathbf{s}$ er signalet, $\mathbf{w}$ er et vindue, c angiver chirpy-ness og ξ er den specielle fan-chirp, som afhænger af tiden og frekvensen.

I [44] er den diskrete definition på FChT noteret, og det er beskrevet, hvorledes FChT kan beregnes effektivt ved hjælp af FFT. Disse beskrivelser har jeg anvendt til implementeringen af en funktion, som kan beregne FChT. Den effektive beregning som anvender FFT kræver interpolation af signalet, og denne interpolation er blevet udført med kubisk interpolation, da en lineær interpolation giver for støjfyldte resultater.

På figur 3.27 vises et eksempel på anvendelsen af FChT. Figuren kan sammenlignes med figur 3.20, 3.21 og 3.22. FChT er blevet anvendt på hele signalet, og det er således den samme chirpy-ness c , som anvendes til enhver

tid. c er fastsat således, at den passer til den maksimale positive chirpy-ness i det kunstige signals tone, og man må derfor forvente, at transformationen klarer sig godt til de tidspunkter, hvor signalet tone stiger hurtigst. Der er blevet anvendt den samme vindueslængde, som på figur 3.20. Det ses af figur 3.27, at FChT klarer sig godt, når tonen stiger hurtigst, men specielt når tonen falder hurtigst klarer den sig meget dårligt. Specielt til tiden 0,2 sek. i området mellem 2.000 og 3.000 Hz ser det ud, som om signalet både indeholder en langsomt faldende tone og samtidig indeholder en noget lavere stigende tone — det ser ud, som om de tykke linjer krydses af tynde linjer.

Så længe signalet kun består af én tone til hver tid, og den fulde implementering af STFChT anvendes, er ovenstående ikke noget problem, for til tiden 0,2 sek. ville STFChT have valgt en chirp-ness c , som passer til signalet. Men hvis signalet rent faktisk består af flere toner med hver deres chirpy-ness, kan det omtalte problem ikke undgås.

3.7.3 Matching Pursuit

Dette afsnit fungerer blot som en appetitvækker til Matching Pursuit-metoden, og derfor dækkes metoden kun meget overfladisk.

Vi så i forrige afsnit et eksempel, som generelt set anvendte chirplet-transformationen, men som reelt set kun arbejdede i 2 dimensioner, selv om chirplet-transformationen giver mulighed for at arbejde i adskillige dimensioner. Fordelen ved at arbejde i et multidimensionelt rum er, at det er muligt til enhver tid og for enhver frekvens at repræsentere signalet med netop den chirplet, som bedst passer. Problemet er, at det kan være svært at finde de optimale chirplets.

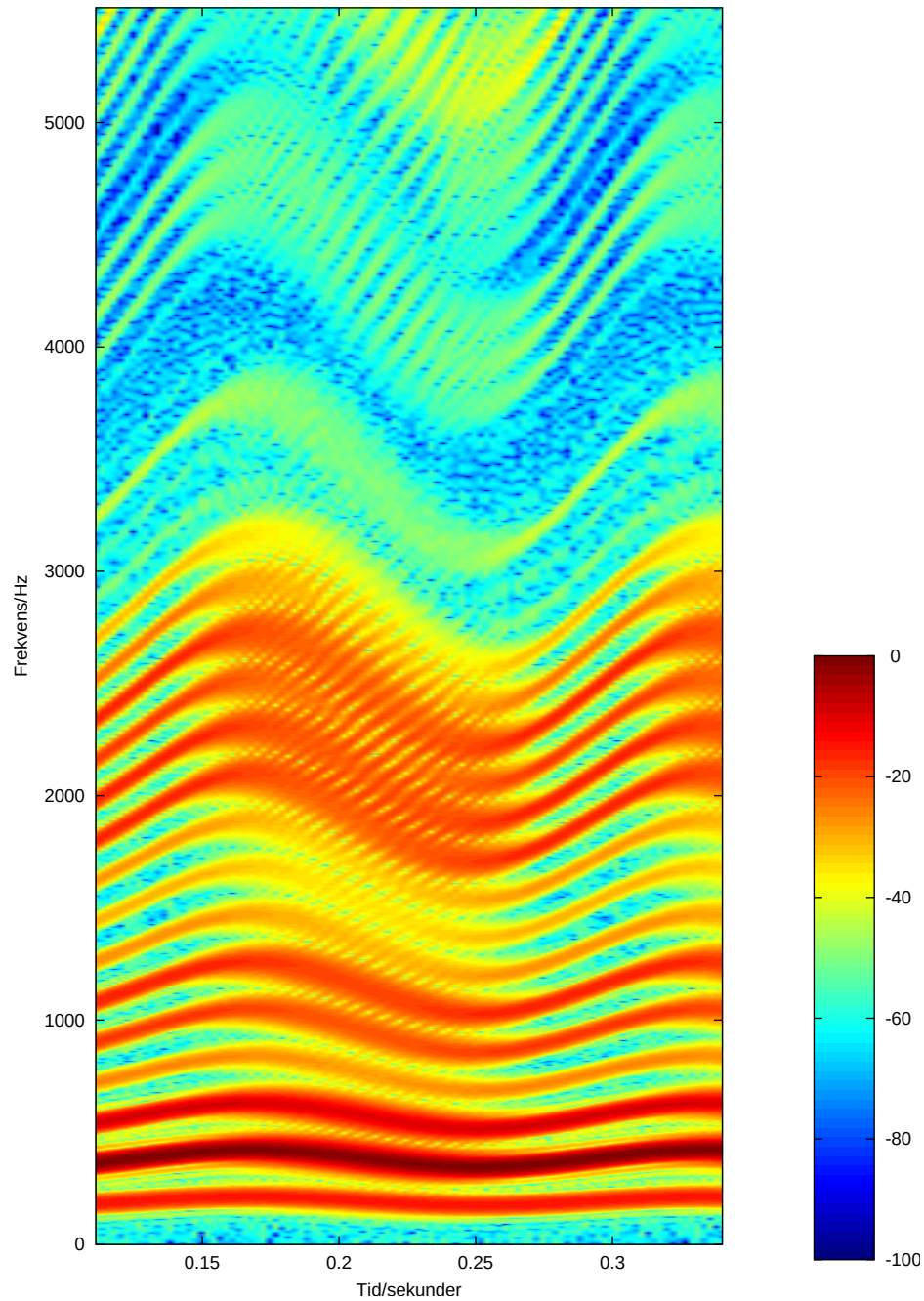
For et givent antal chirplets (eller tid-frekvens-atomer, som de også kan kaldes) vil det i teorien være nødvendigt at afprøve samtlige kombinationer af de forskellige chirplets før en kombination, som er optimal i forhold til et givent signal, kan findes. Der tale om en ikke-polynomiell (NP-hard) beregningsomkostning, og det vil derfor være en næsten umulig opgave at løse, hvis antallet af chirplets er stort. Matching Pursuit, som giver en suboptimal løsning af problemet, anvender en grådig algoritme og mindsker derved beregningsomkostningerne.

Matching Pursuit blev introduceret af Mallat og Zhang i [50] i 1993. Meget kort fortalt, så genereres der først et bibliotek af chirplets. Biblioteket gennemses for det bedste match i forhold til et givent signal. Det givne match tilføjes gruppen af match, som beskriver signalet, og det givne match trækkes fra signalet. Således fortsættes ind til energien i det tilbageværende signal falder under en vis grænse. De fundne matchs antages at give en god beskrivelse af signalets vigtigste strukturer. Som regel kan man nøjes med at udtrække få chirplets, da signalets energi ofte er koncentreret i få frekvenskomponenter.

Det nævnes i [22] at Matching Pursuit har problemer med, at der bliver introduceret pre-echo. High Resolution Matching Pursuit, som behandles i [22] er et forsøg på at undertrykke dette problem.

3.8 Wigner-Ville-fordeling

Her vil vi kort introducere Wigner-Ville-fordeling, da den giver en god tid/frekvens repræsentation, når der er tale om et pænt signal så som en chirp. I det følgende vil vi anvende termen Wigner-Ville-fordeling, mens nogle artikler blot anvender termen Wigner-fordeling.



Figur 3.27: Her vises et eksempel på brugen af den interne transformation i STFChT, altså FChT. Den analyserende bølges chirpy-ness er tilpasset, så den passer til den maksimale positive chirpy-ness i signalet. Det ses, at FChT giver et godt resultat for alle frekvenser ved tiden 0,3 sek., men til tiden 0,2 sek. giver FChT et meget dårligt resultat i området mellem 2.000 og 3.000 Hz. Den maksimale styrke definerer 0 dB.

I nogle artikler (f.eks. [54]) gives den samme definition på Wigner-Ville-fordelingen som andre artiklers definition på Wigner-fordelingen (f.eks. [29]); og i [24] omtales de 2 termer som synonymer. Men i [57] og [59] nævnes det, at Wigner-fordelingen beregnes ud fra det originale signal, mens Wigner-Ville-fordelingen beregnes ud fra det "analytiske signal"²⁴. Mens der i fouriertransformationen opstår aliasing ved den halve frekvens af samplingsfrekvensen, opstår der aliasing allerede ved 1/4 af samplingsfrekvensen, når Wigner-fordelingen beregnes. Ifølge [59] kan aliasing-effekten fjernes ved at samplingsfrekvensen fordobles. Men en anden mulighed er, at anvende det analytiske signal, og det er denne metode, der anvendes i Wigner-Ville-fordelingen. Her vil vi som sagt anvende Wigner-Ville-fordelingen, og vi bruger det analytiske signal, selv om nogle kilder anvender det originale signal.

3.8.1 Definition

Hvis vi har 2 signaler s_1 og s_2 og benævner deres tilsvarende analytiske signaler ved $\tilde{s}_1$ og $\tilde{s}_2$, kan kryds-Wigner-Ville-fordelingen ifølge [29] noteres som:

$$\mathbf{xwv}\{\tilde{s}_1, \tilde{s}_2\}(t, \nu) = \int_{-\infty}^{\infty} \tilde{s}_1(t + \frac{\tau}{2}) \overline{\tilde{s}_2(t - \frac{\tau}{2})} e^{-i2\pi\nu\tau} d\tau \quad (3.68)$$

I det tilfælde, hvor de 2 signaler er ens ($s_1 = s_2$), kan Wigner-Ville-fordelingen, også kaldet auto-Wigner-Ville-fordelingen, beregnes som:

$$\mathbf{wv}\{\tilde{s}\}(t, \nu) = \int_{-\infty}^{\infty} \tilde{s}(t + \frac{\tau}{2}) \overline{\tilde{s}(t - \frac{\tau}{2})} e^{-i2\pi\nu\tau} d\tau \quad (3.69)$$

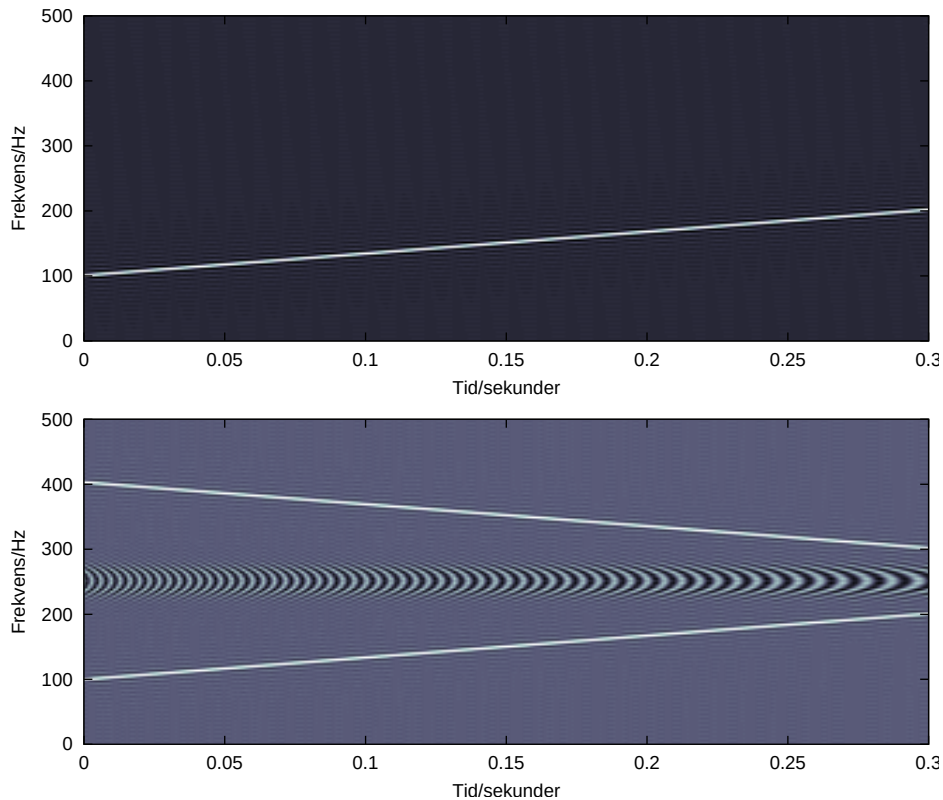
Hvis et signal er dannet ud fra 2 signaler kan Wigner-Ville-fordelingen findes ved:

$$\begin{aligned} \mathbf{wv}\{\tilde{s}_1 + \tilde{s}_2\}(t, \nu) = \\ \mathbf{wv}\{\tilde{s}_1\}(t, \nu) + \mathbf{wv}\{\tilde{s}_2\}(t, \nu) + 2\Re(\mathbf{xwv}\{\tilde{s}_1, \tilde{s}_2\}(t, \nu)) \end{aligned} \quad (3.70)$$

hvor det sidste led er 2 gange den reelle del af kryds-Wigner-Ville-fordelingen af de 2 delsignaler.

Der er 2 store problemer med Wigner-Ville-fordelingen. Det ene er, at den ikke nødvendigvis er positiv (dvs. nogle frekvenser skal trækkes fra andre for at give signalet), hvilket kan virke ulogisk set ud fra et fysisk perspektiv. Det andet problem findes i form af det sidste led i formel 3.70, altså kryds-Wigner-Ville-fordelingen. Det omtalte led giver fejlagtige positive og negative styrker primært til frekvenser, som befinder sig midt mellem frekvenserne af de 2 signaler. Det største problem ved det fejlagtige krydsled er, at det kan give en respons, som ifølge [40] er dobbelt så kraftig som den virkelige styrke af de 2 delsignaler. For en nærmere behandling af interferensen i Wigner-Ville-fordelingen henvises der til [30].

²⁴Vi kalder et signal for s . Det analytiske signal kan da noteres som: $\tilde{s} = s + i\hat{s}$, hvor $\hat{s}$ er Hilbert-transformationen af s . Relationen mellem s og $\tilde{s}$ forklarer faktisk bedst i frekvensdomænet. Fouriertransformationen til signalet kaldes S og fouriertransformationen til det analytiske signal kaldes $\tilde{S}$. Ifølge [24] gælder det for alle frekvenser $\nu < 0$, at $\tilde{S}(\nu) = 0$. For frekvensen $\nu = 0$ er $\tilde{S}(\nu) = S(\nu)$. For alle frekvenser $\nu > 0$ er $\tilde{S}(\nu) = 2S(\nu)$.



Figur 3.28: Her vises et eksempel på Pseudo Wigner-Ville-fordelingen for 2 signaler, som er samplet ved 1.000 Hz. Øverst består signalet af en enkelt chirp, som starter ved 100 Hz og slutter ved 200 Hz. Nederst indeholder signalet både den omtalte chirp og en anden chirp, som starter ved 400 Hz og slutter ved 300 Hz. Det ses, at Pseudo Wigner-Ville-fordelingen giver et godt resultat i det øverste tilfælde, men i det nederste tilfælde opstår der interferens mellem de 2 chirps. Interferensen antager både positive og negative værdier.

3.8.2 Pseudo Wigner-Ville-fordelingen

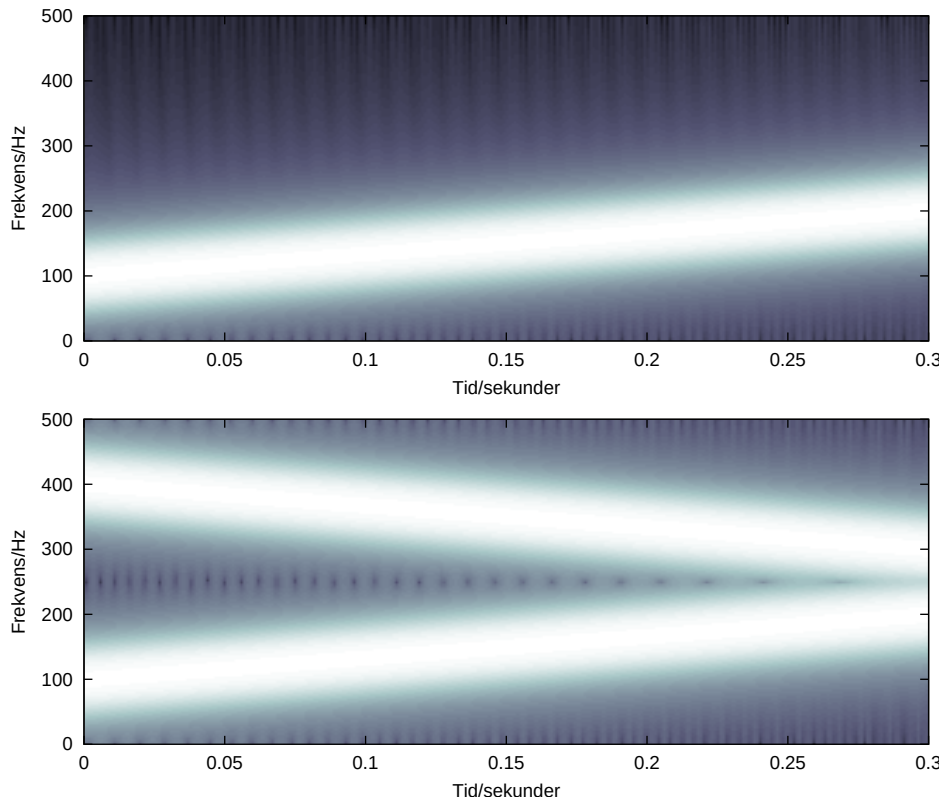
Wigner-Ville-fordelingen beregnes på baggrund af et uendelig langt integral. For at begrænse det interval, som integralet beregnes over, og for at begrænse interferensen anvendes der i praksis et vindue, og der integreres over vinduets længde. Således kommer definitionen på Pseudo Wigner-Ville-fordelingen til at se ud som følger:

$$\text{pwv}\{\tilde{\mathbf{s}}\}(t, \nu) = \int_{-T/2}^{T/2} \mathbf{w}(\tau) \tilde{\mathbf{s}}(t + \frac{\tau}{2}) \overline{\tilde{\mathbf{s}}(t - \frac{\tau}{2})} e^{-i2\pi\nu\tau} d\tau \quad (3.71)$$

hvor $\mathbf{w}$ er et vindue med længden T .

Der findes også en Smoothed Pseudo Wigner-Ville-fordeling, hvor interferensen begrænses af vinduer i både tid- og frekvens-domænet. Jo mere udglatning der foretages, des mere kommer fordelingen til at ligne en Gabor-transformation, og spektrogrammet er ifølge [31] en udglattet Wigner-Ville-fordeling.

På figur 3.28 er vist Pseudo Wigner-Ville-fordelingen for 2 signaler. Øverst er der anvendt en chirp, som starter ved 100 Hz og slutter ved 200 Hz. Nederst er den samme chirp anvendt, men der er derudover tilføjet en chirp, som starter ved 400 Hz og slutter ved 300 Hz. Det ses at Pseudo Wigner-Ville-fordelingen giver gode resultater, når signalet kun består af en enkelt chirp, men hvis der er flere chirps, opstår der interferens imellem de chirps, som signalet består af.



Figur 3.29: Her vises eksempler på anvendelsen af STFT på 2 signaler. Figuren skal ses i forhold til figur 3.28, da det er de samme signaler, der anvendes. Signalerne er samplet ved 1.000 Hz. Øverst består signalet af en enkelt chirp, som starter ved 100 Hz og slutter ved 200 Hz. Nederst indeholder signalet både den omtalte chirp og en anden chirp, som starter ved 400 Hz og slutter ved 300 Hz. Det ses, at STFT i forbindelse med det øverste eksempel klarer sig meget dårligt, da det ser ud, som om der er mange forskellige frekvenser til stede i et bånd omkring den frekvens, som chirp'en har. I det nederste eksempel har STFT også problemer, men set i forhold til Pseudo Wigner-Ville-fordelingen på figur 3.28, er STFT at foretrække, fordi der er ikke er nogen interferens mellem de 2 chirps.

For en sammenligning af Pseudo Wigner-Ville-fordelingen med en STFT er der på figur 3.29 vist en STFT af de signaler, som blev anvendt på figur 3.28. Det ses, at STFT i det øverste tilfælde ikke giver det samme klare billede som Pseudo Wigner-Ville-fordelingen gav. Hvis der er flere chirps tilstede i signalet (se nederst på figur 3.28 og 3.29), er rollerne byttet om for de 2 metoder. I tilfældet med STFT er billedet blot noget uskarpt, men i tilfældet med Pseudo Wigner-Ville generer interferensen mellem de 2 chirps i høj grad.

3.9 Valg af tid/frekvens-repræsentation

Når man står over for at skulle vælge en tid/frekvens-repræsentation af et signal med det mål for øje, at repræsentationen skal bruges som udgangspunkt for en analyse af signalets indhold, er det håbet, at den valgte repræsentation giver et godt billede af, hvad signalet blev dannet ud fra.

Ved at generere kunstige signaler, kan man have kontrol over signalernes indhold, og ved at sammenligne teoretiske ideelle tid/frekvens-repræsentationer af de kunstige signaler med forskellige metoders tid/frekvens-repræsentationer af signalerne kan man i grove træk vurdere, at nogle tid/frekvens-repræsentationer

bedre end andre tid/frekvens-repræsentationer gengiver indholdet i et signal. I forrige afsnit så vi et eksempel, hvor tid/frekvens-repræsentationen reflekterer signalets indhold relativt klart. Dette var tilfældet med Wigner-Ville-transformationen af et signal, som består af en enkelt chirp — se øverst på figur 3.28. I nævnte tilfælde gav en STFT af det samme signal en noget uklar tid/frekvens-repræsentation — se figur 3.29.

På grund af interferensen er der store problemer ved Pseudo Wigner-Ville-fordelingen, og derfor er Pseudo Wigner-Ville-fordelingen ikke anvendelig i forhold til polyfon musik.

Chirplet-transformationen ser lovende ud, fordi det ved anvendelse af denne transformation er muligt at repræsentere et signal med tid/frekvens-atomer (chirplets), som passer til det lokale indhold i signalet. Ved anvendelse af Matching Pursuit er det muligt at gøre brug af chirplet-transformationen, uden at der kræves for mange beregninger.

Vi har set, at QSTFT giver mulighed for at anvende en vindueslængde, som er afhængig af frekvensen, og vi så, at OQSTFT anvender en vindueslængde, som passer godt til en vibrerende tone. Dette giver OQSTFT en fordel, set i forhold til en almindelig STFT og MWT.

Anvendelsen af en multiopløsnings-STFT giver mulighed for at vælge vindueslængde uafhængig af frekvensen, men til gengæld kræves der beregninger med flere vinduer for hvert tid/frekvens-punkt, og det skal på en eller anden måde udvælges, hvilke vinduer der skal anvendes til de forskellige tid/frekvens-punkter.

3.10 Konklusion

Vi har i dette kapitel stiftet bekendtskab med de vigtigste og mest anvendte tid/frekvens-repræsentationer af signaler, nemlig STFT, MWT, chirplet-transformationen og Wigner-Ville-fordelingen. Vi så at QSTFT kunne anvendes som en fleksibel STFT og samtidig kunne fungere som ækvivalent til MWT. OQSTFT udviste særligt gode egenskaber i forhold til en vibrerende tone.

Angående vinduesfunktioner så vi, at den frekvensmæssige støjfordelingen specielt for Hann-vinduet og det sænkede Gauss-vindue primært er lokal. Vi så desuden at Hann-vinduet i nogle tilfælde kan have en fordel, fordi det er let at folde med dette vindue i frekvensdomænet.

I forbindelse med multiple opløsninger så vi, at multiple STFT'er med forskellige vindueslængder kan beregnes effektivt, hvis vinduet tilføjes i frekvensdomænet. Vi så desuden at Matching Pursuit angiver en effektiv metode for generering af en tid/frekvens-repræsentation ved hjælp af chirplets i et multidimensionelt rum.

Kapitel 4

Estimering af melodiens fundamentalfrekvens

I dette kapitel skal vi se på metoder, som arbejder på tid/frekvens-repræsentationer af signaler. Metoderne er ikke afhængige af en bestemt tid/frekvens-repræsentation af et signal, så det burde være muligt at kombinere metoderne i dette kapitel med forskellige tid/frekvens-repræsentationer. Der kan dog være forhold, som gør nogle tid/frekvens-repræsentationer mere oplagte end andre.

Som det blev nævnt i indledningens afgrænsning af denne specialeopgave (afsnit 1.2), kan metoderne for estimering af fundamentalfrekvensen generelt deles i 2 grupper, hvor grupperne primært baserer sig på henholdsvis signalbehandling og statistisk mønstergenkendelse. Det blev fastslået, at vi her koncentrerer os om metoder inden for den første gruppe, da man kan se signalbehandlingen som et grundlag for en efterfølgende statistisk mønstergenkendelse. Men inden vi ser nærmere på et par metoder, som baserer sig på signalbehandling, vil vi kort nævne et par eksempler på metoder, som primært er baserede på statisk mønstergenkendelse.

Statistisk mønstergenkendelse Et meget kendt eksempel er Gotos PreFEST-metode, som blev præsenteret i 1999 i [19]. Der blev berettet om gode resultater i forhold til estimering af melodiens fundamentalfrekvens i almindelig polyfon musik så som pop-musik. Tidligere havde metoder for estimering af melodiens fundamentalfrekvens primært været afprøvet på monofon musik eller på specielle musikstykker. PreFEST-metoden deltog ved MIREX-2005 og var blandt de bedste metoder det år.¹ Blandt de metoder, som deltog ved MIREX-2006 og MIREX-2008, er der dog en del metoder, som har opnået meget bedre resultater. PreFEST-metoden er sidenhen blevet foreslået forbedret på en række punkter, og således berettes der i [17] om væsentligt bedre resultater. Ryynänen og Klapuri medvirkede ved MIREX-2005 [60], MIREX-2006 og MIREX-2008 med en metode, som også gør brug af statistisk mønstergenkendelse, og denne metode fremviste gode resultater. For flere detaljer i forbindelse med metoden henvises der til [39].

Harmoniske toner I de fleste af de metoder, vi i det følgende vil beskæftige os med, antages det indirekte, at den melodibærende tone er en harmonisk²

¹Angående MIREX-konkurrencerne se afsnit 2.1 samt hjemmesiden for MIREX-konkurrencerne: http://www.music-ir.org/mirexwiki/index.php/Main_Page.

²Se [66, afsnit 2.1.1] for en forklaring på harmoniske og inharmoniske lyde.

tone, således at overtonerne ligger med et fast interval. Derfor vil metoderne formentlig have problemer, hvis den melodibærende tone eksempelvis spilles på en tromme, og der kan endda opstå problemer, hvis melodien spilles på et klaver — se [66, figur 2.2 i afsnit 2.1.1]. Et eksempel på en metode, som tager højde for den inharmoniske effekt, som f.eks. opstår på en klaverstreng, kan findes i [37], hvor der gives en modifikation til Meddis' og Hewitts metode [53].

Tid/sted-repræsentation Det er valgt kun at se på metoder, som anvender en tid/frekvens-repræsentation af signalet, men der er talrige eksempler på metoder, som baserer sig på en tid/sted-repræsentation af signalet — de baserer sig altså på den umiddelbare repræsentation af signalet. Det skal dog ikke forstås således, at metoderne på ingen måde kan anvende en form for frekvens-repræsentation. Eksempelvis kan autocorrelation, som behandles i afsnit 4.2.1, både udregnes i tid/sted-domænet og i tid/frekvens-domænet.

Flere af metoderne, som primært er baserede på en tid/sted-repræsentation af signalet, er bygget op omkring en form for model for det menneskelige øre. Et eksempel herpå er [9], hvor et signal filtreres i en række frekvensbånd, for på denne måde at efterligne den frekvensfiltrering, som foregår i det menneskelige øre. Der estimeres en periode for det frekvensfiltrede signal i hvert bånd, ved at udføre autocorrelation på de filtrerede signal. Resultatet heraf er et såkaldt cochleagram, som kan anvendes til at give et bud på, hvilke fundamentalfrekvenser et signal indeholder.

Der er andre metoder, som i en vis grad forsøger at efterligne det menneskelige øre, og som også anvender en tid/sted-repræsentation af signalet til at finde perioder for flere båndfiltreringer af signalet. Der kan eksempelvis henvises til Meddis [53].

Følgende afsnit I de følgende afsnit vil vi lægge ud med en beskrivelse af den generelle fremgangsmåde for estimering af fundamentalfrekvensen, og dernæst vil vi se nærmere på et par af de centrale metoder, som baserer sig på signalbehandling.

Det er valgt at begrænse antallet af præsenterede centrale metoder, således at vi starter med autocorrelation og lignende metoder, og dernæst vil vi se nærmere på SHS samt se, hvordan SHS kan modificeres. Endelig vil vi kort nævne et par metoder, som er forholdsvis kendte eller væsentlige.

Det blev vist i [66], at hvis man antager, at den kraftigste frekvens til enhver tid er melodiens fundamentalfrekvens, så har man i mange tilfælde gættet rigtig. I forbindelse med 3 af test-filerne fra datasættet ADC2004³ viste det sig endda, at denne metode fik mindst lige så høj Raw Chroma⁴ værdi som de deltagende metoder ved MIREX2006. Metoden blev i [66] anvendt til at skabe et mindstemål, som helst skal overgås af mere komplicerede metoder.

I denne opgave er det blevet valgt at se på autocorrelation som en metode, der kan skabe et lignende mindstemål. Autocorrelation kan egentlig kun bestemme perioden af signalet, og dermed har autocorrelation som udgangspunkt ikke mulighed for at estimere flere toner.

SHS er et eksempel på en metode, som giver mulighed for at danne et detaljeret billede af, hvilke toner der formodes at være tilstede i et lydsignal.

³Se afsnit 2.1 for en beskrivelse af de anvendte datasæt.

⁴Se afsnit 2.2 for en forklaring på hvordan Raw Pitch og Raw Chroma beregnes.

En tones fundamentalfrekvens kan estimeres, hvis blot tonen har en kraftig fundamentalfrekvens eller nogle kraftige overtoner.

4.1 Generel opbygning af metoder

Som det også blev nævnt i [66, afsnit 3.1], så kan metoder for bestemmelse af fundamentalfrekvensen deles i 3 dele: en preprocessing, en central del og en postprocessing. I dette afsnit vil vi ganske kort først se på preprocessing og dernæst postprocessing. Dækningen af metoderne i den centrale del kommer i selvstændige afsnit.

4.1.1 Preprocessing

I preprocessing-delen modificeres signalet således, at det forhåbentlig er lettere at finde fundamentalfrekvensen.

4.1.1.1 Støj

Et generelt problem ved estimeringen af fundamentalfrekvensen er, at lyde, som opstår på/i et instrument, påvirkes af selve instrumentet, således at der opstår resonansstøj. Derudover kan der blive tilføjet støj fra andre lydkilder. Det ønskes at undertrykke de 2 former for støj, som hver især kaldes “convolutive noise” og “additive noise” [9]. Et eksempel på en metode, som behandler begge former for støj, er RASTA-metoden [26] [27]. Denne metode er blevet anvendt af Paiva [9] og Klapuri [36].

4.1.1.2 Menneskelig hørelse

Da det i bund og grund er målet at efterligne den menneskelige hørelse, kan det måske være en god ide at se nærmere på nogle akustiske egenskaber, som gør sig gældende både fysisk set og psykologisk set.

Skalaer Størrelsen af øregangen og cochlea⁵ (øresneglen) har betydning for, i hvor høj grad forskellige frekvenser fremhæves eller undertrykkes. Derudover spiller psykologiske effekter ind på opfattelsen af lyde. Som det blev nævnt i [66, afsnit 2.4.1.2 og 2.4.1.3], er der blevet lavet undersøgelser af den menneskelige opfattelse af lydstyrker for forskellige frekvenser (phon-skalaen), samt af den menneskelige opfattelse af afstanden mellem forskellige frekvenser (Mel-skalaen og Bark-skalaen). I en preprocessing kan man eksempelvis repræsentere styrkerne af hver frekvens på phon-skalaen, således at frekvensområdet omkring 3.000 Hz bliver fremhævet, og man kan anvende en frekvensskala, som minder om det overvejende logaritmiske udseende, som Mel- og Bark-skalaerne benytter.

Maskering Den menneskelige hørelse undertrykker svage frekvenser, som ligger nær en kraftig frekvens, således at de kraftige frekvenser siges at dække over de svage frekvenser. På samme måde dækker en kraftig lyd til ét tidspunkt over svagere lyde i et stykke tid derefter. Maskeringen dækker begge former for undertrykkelse af svage frekvenser og anvendes i høj grad ved kompression

⁵Se [66, afsnit 2.4] for en kort introduktion til det menneskelige øre.

af lyd. Der kan eksempelvis henvises til Ehmers maskeringskurver⁶ i [14] og [15] eller til nyere studier hos Moore *et al.* i [55]. Dresslers metode [11], som medvirkede ved MIREX-2006, anvendte en maskeringsmetode [12], som er simplere end den, der eksempelvis anvendes i AAC-lydkompressionen.

4.1.2 Postprocessing

Overgangen mellem den centrale del og postprocessing er ofte noget uklar, men mange metoder indeholder en form for udglatning, eller elimination af små "fejl". Som regel vil den centrale del generere en form for vægtning af nogle formodede fundamentalfrekvenser. Hvis der ikke foretages postprocessing, er det blot den formodede fundamentalfrekvens, som har fået størst vægt, der bliver valgt ud, som et bud på melodiens fundamentalfrekvens. Men hvis der foretages en postprocessing, forsøges det på en eller anden måde at vælge en fundamentalfrekvens ud fra information om vægtningen af de formodede fundamentalfrekvenser til forskellige tidspunkter.

Man kan eksempelvis forestille sig et scenarie, hvor melodiens fundamentalfrekvens stort set er konstant over et længere stykke tid, og den centrale del har tildelt melodiens fundamentalfrekvens størst vægt blandt alle de formodede fundamentalfrekvenser. Der er dog i et ganske kort tidsrum et slagtpøjsinstrument, som spiller en kraftig tone, og denne korte men kraftige tone bliver ikke opfattet som en del af melodien. Da tonen er kraftig, er der risiko for at den centrale del har tildelt slagtpøjsinstrumentets fundamentalfrekvens den største vægt i det ganske korte tidsrum. For i det angivne scenarie at finde den korrekte fundamentalfrekvens, kan man eksempelvis søge efter tidsrum, hvor den formodede fundamentalfrekvens er stabil i ganske kort tid. Hvis der findes sådanne korte tidsrum, kan de aktuelle formodede fundamentalfrekvenser fjernes fra listen af formodede fundamentalfrekvenser, og det vil derfor kun være stabile formodede fundamentalfrekvenser, som er tilbage. Hvis melodiens fundamentalfrekvens antages at være forholdsvis stabil, burde den stadig være blandt de tilbageværende formodede fundamentalfrekvenser.

4.1.2.1 Klangfarve

Her vil vi ganske kort beskæftige os med klangfarven af lyde. Dette emne i sig selv er et helt studie for sig, og begrebet klangfarve er, på trods af den megen forskning inden for området endnu kun delvist forstået. I forbindelse med klangfarve kan der generelt henvises til [33].

Klangfarven er central, fordi vi mennesker opfatter 2 forskellige instrumenter som klingende forskelligt. Hvis man kan få en god forståelse for begrebet klangfarve og kan give et godt estimat for klangfarven af en tone, vil det måske være muligt at identificere forskellige instrumenter i et stykke musik, og derved bliver det måske muligt at følge de forskellige instrumenters toneforløb. Hvis man på en eller anden måde kan estimere hvilket instrument, der spiller melodien, og man samtidig kan følge hvert instruments toneforløb, vil det være let at bestemme melodien korrekt — såfremt melodien holder sig til ét instrument.

⁶Maskeringskurverne i referenceimplementeringen af lydkompressionsformatet Vorbis var oprindeligt baseret på Ehmers maskeringskurver, men de er nu erstattet med maskeringskurver, som er baseret på udviklernes egne målinger. Kildekoden, som indeholder maskeringskurverne, kan findes i libvorbis på <http://www.xiph.org>.

Som Jensen nævner det i [33], er klangfarve generelt anset for at være multidimensional. Nogle af dimensionerne anses for at kunne beskrives som tonehøjde, lydstyrke og tonelængde, og disse dimensioner er i en vis grad forståede. Andre dimensioner kan måske beskrives ved det spektrale hylster⁷ (envelope) og ved det temporale hylster. Det spektrale hylster beskriver det generelle spektrale udseende, mens det temporale hylster beskriver, hvordan lydstyrken generelt ændrer sig i tiden. De 2 nævnte begreber har muligvis stor betydning for opfattelsen af en klangfarve, men der mangler stadig en god forståelse for disse begreber.

Problemet med klangfarven er, at det endnu ikke er fuldt forstået, hvordan mennesker hører forskel på forskellige instrumenter og samtidig opfatter 2 forskellige toner fra det samme instrument som tæt forbundne. Som det nævnes i [33], er det et problem at mange parametre for klangfarve er mere similære for toner med samme tonehøjde, og som spilles på forskellige instrumenter, end for toner med forskellige tonehøjde, og som spilles på det samme instrument.

I søgningen efter redskaber, som kan identificere similaritet mellem klangfarver, har der været gjort flere undersøgelser, hvor personer er blevet bedt om at bedømme similariteten mellem forskellige lyde. Grey fandt i [21] og [20], at den mest betydningsfulde klangfarve-dimension er det spektrale hylster (eller den spektrale energifordeling). Et meget anvendt mål for den spektrale energifordeling er MFCC⁸ (Mel Frequency Cepstral Coefficient), som blandt andet bliver anvendt af [17] til at bestemme, om et signal indeholder menneskelig stemme.

4.1.2.2 Menneskestemme

Da melodien ofte synges af et menneske, kan det være en god idé at søge efter formodede fundamentalfrekvenser, som ser ud til at høre til en menneskelig stemme. Det kan dog være et stort problem at identificere den menneskelige stemme, specielt fordi den kan lyde på mange forskellige måder. Der kan f.eks. henvises til [17], hvor der anvendes en metode, som forsøger at finde de formodede fundamentalfrekvenser, som antages at stamme fra en menneskelig stemme. I [17] rapporteres der om gode forbedringer ved anvendelse af denne metode sammen med PreFEST-metoden. Metoden anvender en modificeret MFCC samt en detektering af, hvor meget fundamentalfrekvensen ændrer sig i tiden. Det bliver antaget, at menneskelig sang ofte anvender vibrato, og derfor antages det, at dette sidstnævnte mål er en indikator for, at der er menneskelig sang til stede i signalet.

4.2 Autocorrelation, cepstrum og lignende metoder

Vi skal i dette afsnit se, at autocorrelation og cepstrum begge er metoder, som arbejder på det styrkemæssige forhold mellem frekvenserne i et signal. Vi skal desuden se, hvordan metoderne kan generaliseres.

⁷Ved hylster (engelsk: envelope) forstås en kurve, som følger en anden kurve men med lavere detaljeringsgrad. Hylsteret beskriver altså det generelle udseende af en kurve.

⁸MFCC (Mel Frequency Cepstral Coefficient) kan kort fortalt beregnes som et signals cepstrum beregnet på en Mel-frekvens-skala. Cepstrum behandles i afsnit 4.2.2 og Mel-skalaen er ligesom Bark-skalaen en skala, som skal beskrive den menneskelige opfattelse af afstande mellem frekvenser. Mel-skalaen har overvejende et logaritmisk udseende. Der kan henvises til [8].

4.2.1 Autocorrelation

Autocorrelation bliver traditionelt set præsenteret som en metode, der arbejder direkte på et signal, og altså ikke som en metode, der arbejder med en tid/frekvens-repræsentation af signalet. Dette er tilfældet i [66], hvor autocorrelation også blev præsenteret. Men vi skal her se, at autocorrelation også kan bestemmes i tid/frekvens-domænet.

Autocorrelation giver et mål for, hvor meget et signal ligner sig selv efter en given periode τ . I forhold til definitionen i [66], vil vi her bringe en mere korrekt definition, som blandt andet kan findes i [10].

$$\mathbf{auto}\{\mathbf{s}\}[t, \tau] = \sum_{j=t}^{t+T-1} \mathbf{s}[j] \mathbf{s}[j + \tau] \quad (4.1)$$

hvor $\mathbf{s}$ angiver et signal, t angiver det tidspunkt, som autocorrelation bestemmes for, τ angiver perioden, og T angiver længden af det vindue, der summeres over. For alle perioder τ summeres der over det samme antal samples.

Et problem ved den angivne definition på autocorrelation er, at autocorrelationen udregnes over større intervaller når perioden τ stiger. Derfor anvendes en modificeret definition ofte:

$$\mathbf{auto}\{\mathbf{s}\}[t, \tau] = \frac{1}{T - \tau} \sum_{j=t}^{t+T-1-\tau} \mathbf{s}[j] \mathbf{s}[j + \tau] \quad (4.2)$$

således at autocorrelationen udregnes over det samme interval uafhængig af perioden. For at kompensere for, at autocorrelationen på denne måde bliver udregnet over et faldende antal samples, når perioden stiger, er der tilføjet normaliseringen $1/(T - \tau)$. Når perioden er stor, er der kun få samples, som indgår i beregningerne. Dette fører til mere usikre resultater, på grund af støj i signalet og på grund af numeriske afrundinger.

Som det også bliver nævnt i [66], kan autocorrelation ses som en foldning, hvor signalet foldes med det spejlede signal. Ved brug af foldningssætningen kan det ses, at autocorrelation kan beregnes effektivt via fouriertransformationen:

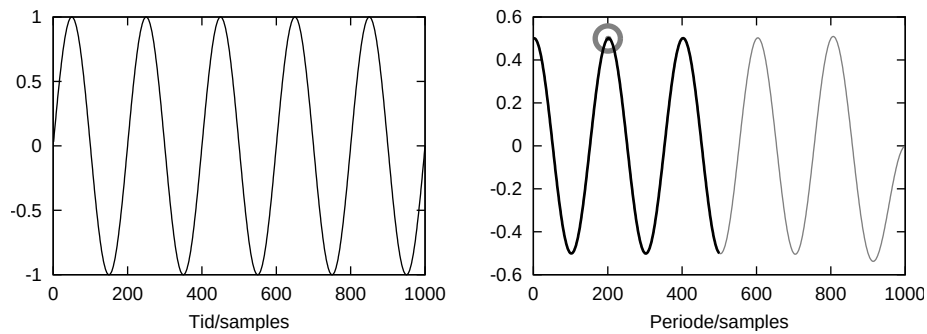
$$\mathbf{auto}\{\mathbf{s}\}[t, \tau] = \mathcal{F}^{-1}\{\mathbf{g} \bar{\mathbf{g}}\} \quad \text{hvor: } \mathbf{g} = \mathcal{F}\{\mathbf{s}[t; t + T - 1]\} \quad (4.3)$$

Fouriertransformationen ved udregningen af $\mathbf{g}$ beregnes over indeksene t op til indeks $t + T - 1$. $\mathbf{g}$ ganges med den konjugerede til $\bar{\mathbf{g}}$ for at foretage spejlingen af $\mathbf{s}$. Det skal bemærkes, at fouriertransformationen indirekte antager at signalet er periodisk, og dermed gentager sig selv med perioden T . Denne antagelse er ikke nødvendigvis korrekt for signalet $\mathbf{s}$, og derfor bør man foretage fouriertransformationen med en zero-padding⁹ på længden T . I så fald skal man altså se bort fra de sidste T indekse i resultatet.

At gange den komplekst konjugerede til x med x selv, svarer til at tage den absolute værdi af x og opløfte resultatet i anden. Så den endelige formel for udregningen af autocorrelationen via fouriertransformationen, med inddragelse af zero-padding, og hvor normaliseringen fra formel 4.2 medtages, kan skrives som:

$$\mathbf{auto}\{\mathbf{s}\}[t, \tau] = \frac{1}{T - \tau} \mathcal{F}^{-1}\left\{\left|\mathcal{F}\left\{\begin{bmatrix} \mathbf{s}[t; t + T - 1] & 0[1; T] \end{bmatrix}\right\}\right|^2\right\} \quad (4.4)$$

⁹Zero-padding betyder blot at signalet forlænges med en række nuller. Ellers henvises der til [66, afsnit 2.3.1.5], hvor begrebet er nærmere forklaret.



Figur 4.1: Til venstre vises et signal, som består af en sinusbølge med en periode på 200 samples. Der er 1.000 samples i alt i signalet. Til højre vises en autocorrelation af signalet udregnet efter metoden i formel 4.4. Vi bruger kun den første halvdel af værdierne for perioden, fordi beregningerne bliver gradvist mere unøjagtige når perioden stiger. Derfor er de første 500 værdier for perioden markeret med en fed kurve. Signalets periode kan identificeres ved den første bølgetop, og dette er markeret med en ring.

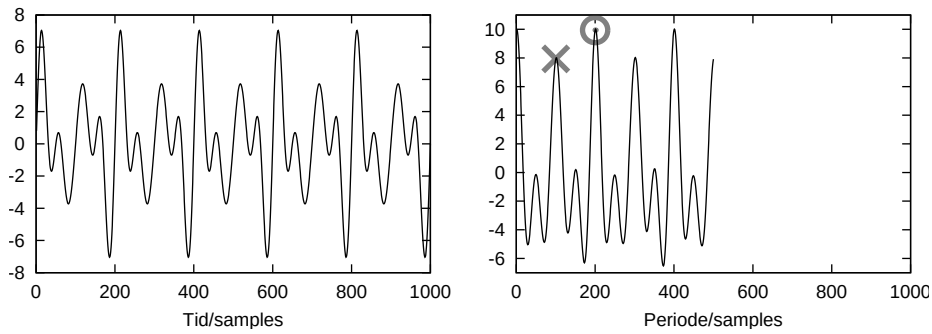
således at fouriertransformationen først beregnes for et signal, som består af T samples fra s samt T samples, som er lig 0; dernæst tages den absolutte værdi og resultatet sættes i anden; dernæst foretages den inverse fouriertransformation; endelig tages de første T værdier i dette resultat, og disse T værdier repræsenterer det endelige resultat, som ganges med den faktor, der kompenserer for at lange perioder vægtes svagere end korte perioder.

Et signal, som gentager sig med en periode på τ , vil også gentage sig med en periode på $x\tau$, hvor x er et heltal. Det betyder, at en autocorrelation af signalet vil give maksimal respons for alle perioder på $x\tau$, hvor τ er et heltal. Den mindste af disse perioder er den periode, vi er interesserede i, da den i tilfældet, hvor signalet kun består af en enkelt sinusbølge, netop svarer til sinusbølgens bølgelængde. På figur 4.1 er autocorrelationen af en sinusbølge illustreret. Den interessante bølgetop er den første af bølgetoppene, og her ses det, at signalet har en periode på 200 samples. Bemærk at autocorrelationen altid giver maksimal respons for $\tau = 0$, da signalet altid vil ligne sig selv maksimalt, når det ikke er flyttet.

Et simpelt eksempel Vi skal nu se et simpelt eksempel på brugen af autocorrelation, sådan som autocorrelation er defineret i formel 4.4. Vi så på figur 4.1, at det ved hjælp af autocorrelation ikke var noget problem at finde perioden for den sinusbølge, der blev brugt i eksemplet. Den anvendte sinusbølge havde en periode på 200 samples. Lad os betragte det omtalte signal som en tone¹⁰, og denne tone indeholder altså kun én frekvens, nemlig fundamentalfrekvensen.

Vi genererer nu et nyt signal, som består af de første 4 overtoner til den før-omtalte tone. Det ny signal indeholder altså ikke selve fundamentalfrekvensen. Vi lader den 1. og den 3. overtone have en amplitude, som er 3 gange højere end den 2. og den 4. overtones amplitude. På figur 4.2 vises det genererede signal og dets autocorrelation. Det ses, at autocorrelation fremhæver en periode på 200 samples, hvilket svarer til fundamentalfrekvens periode. Autocorrelation kan altså finde den korrekte fundamentalfrekvens, selv om fundamentalfrekvensen slet ikke er til stede i signalet. Det ses samtidig, at en periode på 100 samples fremhæves. Dette skyldes at 1. og 3. overtone er meget kraftigere end 2. og

¹⁰Der henvises til [66, afsnit 2.1] for en kort beskrivelse af begrebet "tone" og tonedannelse.



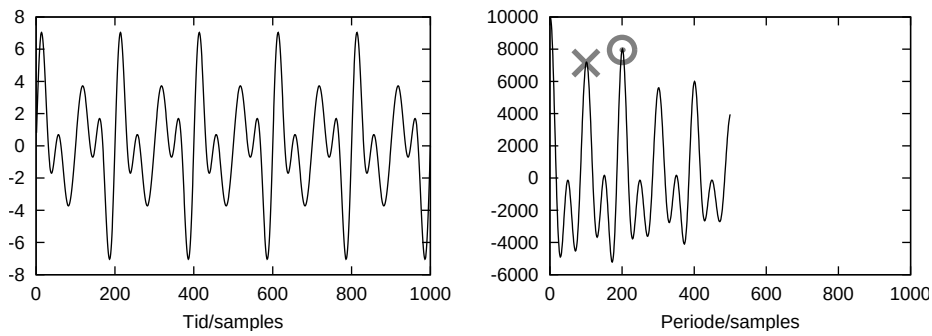
Figur 4.2: Til venstre vises et signal, som består af de 4 første overtoner til en tone med en fundamentalfrekvens, hvis periode er på 200 samples. Den 1. og 3. overtone er kraftigere end den 2. og den 4. overtone. Der er 1.000 samples i alt i signalet. Til højre vises en autocorrelation af signalet udregnet efter metoden i formel 4.4, og det er kun de første 500 værdier for τ som vises. Det ses, at en periode på 200 samples fremhæves, og denne periode svarer til fundamentalfrekvensens periode. Fundamentalfrekvensen for tonen kan altså findes, selv om fundamentalfrekvensen ikke er til stede i signalet. Den bølgetop, som markerer perioden i signalet er markeret med en ring. Der ser også ud til at være en noget svagere periode på 100 samples, og dette er markeret med et kryds.

4. overtone. Som udgangspunkt kan autocorrelation kun anvendes til at finde et signals periode, men som eksemplet illustrerer det, kan der være tilfælde, hvor kortere perioder får tildelt næsten lige så stor vægt som signalets egentlige periode.

Undladelse af normaliseringen Da vi kun er interesserede i det første maksimum i resultatet for autocorrelation, kan det være en fordel at undlade brugen af den normalisering, som anvendes i formel 4.4. Hvis et signal har en periode på x samples, vil autocorrelation for signalet have maksimum for sample x samt hver ax samples, hvor a er et heltal, som er større end 1. Ved at undlade normaliseringen i formel 4.4 vil autocorrelation ved sample ax ikke være lige stor som ved sample x . Det betyder at signalets periode ikke skal findes blandt en række kandidater, men i stedet for kan findes som det eneste maksimum for autocorrelation (når vi ser bort fra samples helt tæt på 0). Der er selvfølgelig stadigvæk mulighed for, at der kan være flere maksima, eller at der er fremkommet et nyt maksimum, men dette behandler vi senere. Lad os nu notere autocorrelation uden normaliseringen:

$$\text{auto}\{s\}[t, \tau] = \mathcal{F}^{-1} \left\{ \left| \mathcal{F} \left\{ [s[t; t + T - 1] \quad 0[1; T]] \right\} \right|^2 \right\} \quad (4.5)$$

På figur 4.3 er der vist et eksempel på brugen af formel 4.5 på det signal, som blev anvendt på figur 4.2. Som det ses af figur 4.3, er der kun ét maksimum (markeret med en ring), mens der på figur 4.2 var et maksimum for hvert hele multiplum af perioden. Det ses dog også, at bølgetoppen ved en periode på 100 samples (markeret med et kryds) nu er næsten lige så høj som maksimum. Hvis den relative styrke af 1. og 3. overtone forøges, kan bølgetoppen ved en periode på 100 samples gå hen og blive maksimum, således at det vil se ud som om signalets periode er på 100 samples. Dette er dog ikke nødvendigvis et problem, men måske snarere en fordel. Godt nok er det dermed ikke signalets egentlige periode, der bliver fundet; men muligvis er det signalets oplevede fundamentalfrekvens, som derved bliver fundet, fordi forskellen mellem overtoners styrker



Figur 4.3: Til venstre vises det signal, som blev anvendt på figur 4.2. Til højre vises autocorrelationen af signalet udregnet efter metoden i formel 4.5. Det ses, at det ved anvendelse af formel 4.5 er lettere at bestemme fundamentalfrekvensen, da der i det viste eksempel kun er ét maksimum, mens der på den forrige figur var flere maksima.

er så stor, at vi mennesker opfatter signalet som bestående af 2 toner: En kraftig tone, hvis fundamentalfrekvens' periode er på 100 samples, og en svagere tone, hvis fundamentalfrekvens' periode er på 200 samples.

4.2.2 Cepstrum

Vi har nu set, hvordan autocorrelation kan anvendes til at finde fundamentalfrekvensen for en tone. Vi har desuden set, at autocorrelation kan beregnes på baggrund af en tid/frekvens-repræsentation. Vi skal nu se, at autocorrelation minder meget om cepstrum-analyse, som også blev kort behandlet i [66, afsnit 3.1.2.1]. Cepstrum-analyse består af den inverse fouriertransformation af logaritmen til den absolutte værdi af et signals spektrum. Så hvis vi beregner et signals cepstrum for korte tidsrum på samme måde, som vi gjorde med autocorrelation, kan vi definere et signals tidsafhængige cepstrum ved:

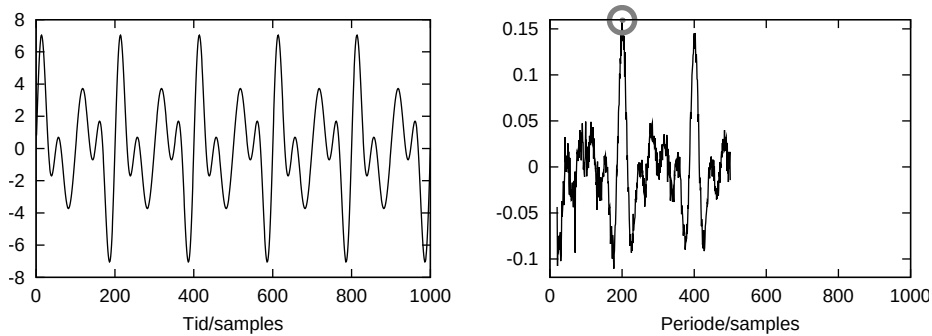
$$\mathbf{ceps}\{\mathbf{s}\}[t, \tau] = \mathcal{F}^{-1} \left\{ \log \left(\left| \mathcal{F} \{ \mathbf{s}[t; t + T - 1] \} \right| \right) \right\} \quad (4.6)$$

Vi kan, ligesom det var tilfældet med autocorrelation, forlænge signalet med nuller, således at fouriertransformationen beregnes for et signal med den dobbelte længde, og ligesom ved autocorrelation kan vi nøjes med de første T værdier i resultatet, da de efterfølgende T værdier blot er en spejling af de første. Således kommer beregningen af cepstrum til at ligne beregningen af autocorrelation, sådan som autocorrelation blev defineret i formel 4.5, hvor normaliseringen ikke anvendes:

$$\mathbf{ceps}\{\mathbf{s}\}[t, \tau] = \mathcal{F}^{-1} \left\{ \log \left(\left| \mathcal{F} \left\{ \begin{bmatrix} \mathbf{s}[t; t + T - 1] & 0[1; T] \end{bmatrix} \right\} \right| \right) \right\} \quad (4.7)$$

På figur 4.4 vises cepstrum for det signal, som også blev anvendt på figur 4.2. Cepstrum er beregnet via formel 4.7, og det er ligesom ved autocorrelation på figur 4.3 kun de første 500 værdier, som vises. Det ses, at cepstrum ligesom autocorrelation fremhæver en periode på 200 samples.

Den eneste forskel på formel 4.5 for autocorrelation og formel 4.7 for cepstrum analyse er, at spektrummet i tilfældet med autocorrelation opløftes i anden, mens det i tilfældet med cepstrum er logaritmen til spektrummet, som anvendes.



Figur 4.4: Til venstre vises det signal, som blev anvendt på figur 4.2. Til højre vises signalets cepstrum udregnet efter metoden i formel 4.7, og det er kun de første 500 værdier for τ , som vises. Det ses, at en periode på 200 samples fremhæves, og denne periode svarer til perioden for fundamentalfrekvensen for den tone, som signalet indeholder.

Ved at anvende logaritmen bringes kraftige og svage frekvenser på samme niveau, og det er således en periode i frekvensmønsteret som søges med cepstrum, mens autocorrelation søger en periode i signalet ved at fremhæve de kraftige frekvenskomponenter.

4.2.3 Generalisering

Her vil vi nu foreslå en generalisering af autocorrelation og cepstrum. Der er blot tale om et simpelt tankeeksperiment, og vi vil derfor ikke give en stærk teoretisk funderet baggrund for generaliseringen. Vi definerer generaliseringen ved:

$$\mathbf{g}\{\mathbf{s}, \mathbf{f}\}[t, \tau] = \mathcal{F}^{-1} \left\{ \mathbf{f} \left(\left| \mathcal{F} \left\{ \begin{bmatrix} \mathbf{s}[t; t + T - 1] & 0[1; T] \end{bmatrix} \right\} \right| \right) \right\} \quad (4.8)$$

hvor $\mathbf{f}$ er en funktion. Vi kan generere autocorrelation **auto** ud fra $\mathbf{g}$ ved at sætte $\mathbf{f}(x) = x^2$, og cepstrum **ceps** kan genereres ved at sætte $\mathbf{f}(x) = \log(x)$.

Valg af funktionen $\mathbf{f}$ I sidste ende ønsker vi at finde fundamentalfrekvensen for melodiens tone. Vi har set, at både autocorrelation og cepstrum kan finde fundamentalfrekvensen for en enkelt tone, selv om det kun er nogle af tonens overtoner, som er til stede i signalet. Mens autocorrelation benytter sig af at fremhæve de kraftige frekvenser, anvender cepstrum den omvendte fremgangsmetode og bringer i stedet for kraftige og svage frekvenser på omtrent samme niveau. Det kunne derfor være interessant at undersøge om en funktion, som “befinder sig imellem” de 2 nævnte funktioner også kan anvendes, og om det ved anvendelse af en sådan funktion i højere grad er muligt at estimere melodiens fundamentalfrekvens. Ifølge [13] er metoder, som baserer sig på autocorrelation relativt immune overfor støj, men påvirkes af spektrale fænomener så som formanter¹¹. Omvendt rapporterer [13] om, at metoder, som baserer sig på cepstrum, påvirkes af støj, men giver gode resultater selv om signalet indeholder spektrale fænomener så som formanter. Når cepstrum påvirkes af støj, skyldes det, at cepstrum i modsætning til autocorrelation fremhæver de svage frekvenser, hvor støjen vil være mere dominerende.

¹¹Formanten beskriver en akustisk resonans. I forbindelse med den menneskelige stemme kan formanten ofte ses i et spektrum som et bredt frekvensområde, der generelt er kraftigere end de omliggende frekvensområder.

Der er selvfølgelig uendeligt mange måder, hvorpå man kan designe en funktion, som “befinder sig mellem” kvadrat-funktionen fra autocorrelation og logaritme-funktionen fra cepstrum. Man kunne eksempelvis kombinere den kvadratiske funktion og den logaritmiske funktion, således at f følger den kvadratiske funktion i ét interval og følger den logaritmiske funktion i et andet interval. Et meget simple eksempel kunne være funktionen $f(x) = x$. I [37] samt i [34] anvendes indirekte den generalisering, som blev foreslået her. I [37] sættes f indirekte til $f(x) = x^{1.4}$, mens $f(x) = x^{1.5}$ i [34].

Tid/frekvens-repræsentation Vi vil desuden foreslå endnu en generalisering. Idéen er, at den frekvensrepræsentation, som f arbejder på i formel 4.8, ikke behøver at være fouriertransformationen af det signal, som ønskes analyseret. Vi så i kapitel 3, at visse tid/frekvens-repræsentationer bedre illustrerer indholdet i et signal end andre tid/frekvens-repræsentationer. Så hvis man kan finde en god tid/frekvens-repræsentation, kan man måske med fordel benytte denne i stedet for fouriertransformationen:

$$\mathbf{g}\{\mathbf{s}, \mathbf{f}\}[t, \tau] = \mathcal{F}^{-1} \left\{ \mathbf{f} \left(|\mathbf{tfr}\{\mathbf{s}\}[t]| \right) \right\} \quad (4.9)$$

I ovenstående generelle formel repræsenterer $\mathbf{tfr}$ en tid/frekvens-repræsentation af signalet.

4.2.4 Yin

Yin-metoden er taget med her som en meget kort bibemærkning, fordi den er en videreudvikling af autocorrelation. Yin-metoden arbejder dog ikke på en tid/frekvens-repræsentation af signalet.

Metoden, som blev introduceret af Cheveigné i [10], blev fremvist som en metode for estimering af fundamentalfrekvensen i tale og musik, og skal ses som et forsøg på at mindske fejlraten ved autocorrelation. Der er 6 skridt i metoden, og i [10] fremvises fejlratene for hvert skridt i metoden. Fejlratene er udregnet på baggrund af en database med stemmeoptagelser. Der er kun tale om én stemme af gangen, og derfor er databasen meget forskellig fra den samling af test-filer, som anvendes i denne opgave.

Det første skridt i Yin adskiller sig ikke fra en almindelig autocorrelation — bortset fra at det er formel 4.4, som benyttes og altså ikke formel 4.5. Ifølge de målinger, som vises i [10], så giver 2. skridt den største forbedring, da dette skridt nedbringer fejlraten fra 10% til omkring 2%. Efter det 6. skridt er fejlraten bragt ned på 0.5%. Da 2. skridt indeholder den største forbedring, vil vi nøjes med at beskrive dette skridt.

2. skridt kan sådan set godt udregnes for sig selv, men kan også beregnes ved hjælp af autocorrelation. Her vil vi nøjes med udregningen uden hjælp af autocorrelation. I 2. skridt beregnes en differensfunktion. Denne bør være lig 0 eller nær 0 for hele multipla af perioden for fundamentalfrekvensen. Differensfunktionen¹² kan beregnes ved:

$$\mathbf{d}\{\mathbf{s}\}[t, \tau] = \sum_{j=t}^{t+W-1} (\mathbf{s}[j] - \mathbf{s}[j + \tau])^2 \quad (4.10)$$

¹²Differensfunktionen svarer til Optimum Comb funktionen (præsenteret i [56]), som blev nævnt i [66, afsnit 3.1.2.4]. Her blev Optimum Comb dog fejlagtigt omtalt som en anden betegnelse for autocorrelation — dette er altså forkert.

Som det nævnes i [10], så falder et minimum i differensfunktionen ikke nødvendigvis sammen med et maksimum i autocorrelation, og dette er årsagen til, at der opnås forskellige resultater med de 2 metoder.

4.3 SHS

Subharmonic Summation (SHS) blev også præsenteret i [66, afsnit 3.1.2.5], og derfor vil vi ikke gå i dybden med den fulde algoritme, men i stedet for kort gentage den centrale ligning, som angiver beregningen af SHS. Dernæst vil vi se på de problemer der forbinder sig til SHS, samt se på nogle forslag til løsninger.

Beregningen af SHS, som SHS er beskrevet i den originale artikel [28], foregår på baggrund af et amplitude spektrum og arbejder altså alene ud fra en frekvensrepræsentation. I den følgende formel skal $\mathbf{A}$ forstås som amplituderne i en frekvensrepræsentation af et signal til et givent tidspunkt. Amplituderne $\mathbf{A}$ fås ved $\mathbf{A} = |\mathbf{S}|$, og kan altså være fremkommet som den absolute værdi af en søjle i en af tid/frekvens-repræsentationerne i kapitel 3.

$$\text{shs}\{\mathbf{S}\}(\nu_0) = \sum_{n=1}^N \mathbf{h}(n)\mathbf{A}(n\nu_0) \quad (4.11)$$

hvor $\mathbf{S}$ er en frekvensrepræsentation af et signal, $\mathbf{A}$ angiver amplituderne, ν_0 er en formodet fundamentalfrekvens, N angiver det maksimale antal overtoner, som tages med i beregningerne, og $\mathbf{h}$ bestemmer vægten af hver overtone.

I [28] beskrives beregningen af SHS med frekvenserne repræsenteret på en logaritmisk frekvensakse, og det er kun de kraftigste frekvenser, som anvendes i beregningerne.

Der kan være beregningsmæssige fordele ved anvendelsen af en logaritmisk frekvensakse, men der kan også være bagdele, fordi der kan ligge vigtig information i de høje frekvenser, som på en logaritmisk frekvensakse bliver dårligt repræsenteret. De gæt på fundamentalfrekvenser, som SHS i [28] bliver beregnet for, ligger på den samme logaritmiske frekvensakse, som signalet er repræsenteret på. Men man kan sagtens forestille sig, at signalet er repræsenteret på én frekvensakse med én opløsning, og at de formodede fundamentalfrekvenser ligger på en anden frekvensakse med en anden opløsning. Således kan signalet godt være repræsenteret på en lineær frekvensakse, mens de formodede fundamentalfrekvenser ligger på en logaritmisk frekvensakse, således at de følger den musikalske skala.

Det, at SHS ifølge [28] kun beregnes på baggrund af de kraftigste frekvenser, således at svage frekvenser, som ligger nær kraftige frekvenser, sættes til at have en amplitude, som er lig 0, kan ses som et forsøg på at fjerne støj fra signalet eller som et forsøg på at reducere signalet, således at det kun er de vigtigste frekvenser, som lades tilbage. Det kan dog være et problem at anvende denne form for støjundertrykkelse, især i forbindelse med polyfon musik, hvor overtoner fra forskellige instrumenter kan ligge meget tæt. Men da den menneskelige hørelse påvirkes af maskeringseffekten¹³, kan man også sige, at filtreringen blot er et forsøg på at efterligne maskeringseffekten.

¹³Se afsnit 4.1.1.2

4.3.1 Spektral normaliseringsfaktor

Det blev nævnt i [66], at forfatterne til [5] foreslog anvendelsen af en spektral normaliseringsfaktor. I stedet for at basere summen i formel 4.11 på amplituden af hver frekvens, foreslås det at anvende styrken for hver frekvens delt med den gennemsnitlige styrke i et bånd omkring den aktuelle frekvens. I [5] berettes om gode erfaringer ved brugen af denne spektrale normaliseringsfaktor set i forhold til at anvende amplituden. Den spektrale normaliseringsfaktor i [5] noteret som:

$$\mathbf{P}_{norm}(\nu_0, \nu) = \mathbf{P}(\nu) / \left(\frac{1}{2\rho\nu_0} \int_{\nu-\rho\nu_0}^{\nu+\rho\nu_0} \mathbf{P}(\xi) d\xi \right) \quad (4.12)$$

hvor $\mathbf{P}$ angiver styrken for hver frekvens, ν_0 angiver en formodet fundamental-frekvens, ρ er en konstant, som bestemmer bredden af det interval, der integreres over, og ν er den frekvens, som den spektrale normaliseringsfaktor $\mathbf{P}_{norm}$ beregnes for.

Den spektrale normaliseringsfaktor hjælper på 2 problemer anvendelsen af amplituden: 1) Hvis et signal består af nogle kraftige dybe frekvenser og nogle svage høje overtoner, så er der risiko for, at de svage høje overtoner stort set ikke har nogen betydning for beregningen af SHS. 2) Hvis det derudover gælder, at de dybe frekvenser generelt set ligger på stort set samme niveau, således at nogle dybe frekvenser er kraftigere end andre, men kun lidt kraftigere, så er det stadigvæk overvejende styrken af de dybe frekvenser, som bestemmer udregningen af SHS — også selv om det mest er tilfældigheder, som afgør om den ene dybe frekvens til et bestemt tidspunkt er lidt kraftigere end en anden dyb frekvens.

Et problem ved den foreslåede spektrale normaliseringsfaktor i formel 4.12 er, at det interval, som den gennemsnitlige styrke beregnes over, er afhængig af den formodede fundamental-frekvens. Det betyder, at der ikke kan beregnes en enkelt vægt for hver frekvens. Hver gang metoden skal bruge vægten af en bestemt frekvens, skal gennemsnittet udregnes på ny, fordi den formodede fundamental-frekvens er ny. For at nedbringe mængden af beregninger er det i [66, afsnit 4.3.1] foreslået at beregne et enkelt gennemsnit for hver frekvens, og det interval, som gennemsnittet beregnes over, afhænger kun af den aktuelle frekvens.

Et andet problem ved den foreslåede normaliseringsfaktor er, at små udsving i styrken i et bånd af svage frekvenser, kan få stor betydning for beregningen af SHS. For den spektrale normaliseringsfaktor angiver kun, hvor kraftig en frekvens er set i forhold til de omliggende frekvenser, og tager altså ikke højde for, at alle de omliggende frekvenser måske er meget svage, og derfor højst sandsynlig blot indeholder støj. For at mindske dette problem vil jeg her foreslå, at den spektrale normaliseringsfaktor bliver ganget med en vægtning af den aktuelle frekvens' styrke.

Forslaget til en ny spektral normaliseringsfaktor kan noteres som:

$$\begin{aligned} \mathbf{P}_{middel}(\nu) &= \frac{1}{2\rho(\nu)} \int_{\nu-\rho(\nu)}^{\nu+\rho(\nu)} \mathbf{w}(\nu, \xi) \mathbf{P}(\xi) d\xi \\ \mathbf{P}_{norm}(\nu) &= \frac{(\mathbf{P}(\nu))^c}{\mathbf{P}_{middel}(\nu)} \end{aligned} \quad (4.13)$$

Lad os først se på beregningen af en gennemsnitlig styrke $\mathbf{P}_{middel}$: Der foretages et integral over frekvensstyrkerne for frekvenser i nærheden af frekvensen

ν . Funktionen ρ bestemmer bredden af det interval der integreres over. Det foreslås, at ρ defineres som en kvadratisk Bézier-kurve¹⁴, således at det er let at styre længden af det interval, der integreres over. Funktionen w fastsætter en vægtning af frekvensstyrkerne indenfor det interval, der integreres over, og kan f.eks. defineres som en gaussisk funktion centreret omkring ν og med en bredde, som fastsættes af $\rho(\nu)$:

$$w(\nu, \xi) = \exp\left(\frac{-(\xi - \nu)^2}{2(\rho(\nu)/3)^2}\right) \quad (4.14)$$

Selv om jeg altså foreslår at ovenstående gaussiske funktion anvendes som vægtning af de forskellige frekvenser, kan man i praksis sagtens anvende tilnærmede funktioner, som er lettere at foretage beregninger med. Forsøg har vist, at hvis man tilnærmer den gaussiske funktion med en diskret funktion, som antager 8 forskellige værdier med samme interval mellem sig, så er resultaterne næsten identiske.

Lad os nu fokusere på den sidste linje i formel 4.13. Når den gennemsnitlige styrke P_{middel} er beregnet, kan den ny spektrale normaliseringsfaktor bestemmes ved forholdet mellem styrken og den gennemsnitlige styrke. Som det fremgår af formel 4.13 opløftes styrken i konstanten c . c bør være en værdi, som er lidt større end 1. Hvis $c = 1$, kan der opstå problemet med, at små udsving blandt meget svage frekvenser giver anledning til store udsving i den spektrale normaliseringsfaktor. Hvis c omvendt sættes højt, forsvinder normaliseringseffekten.

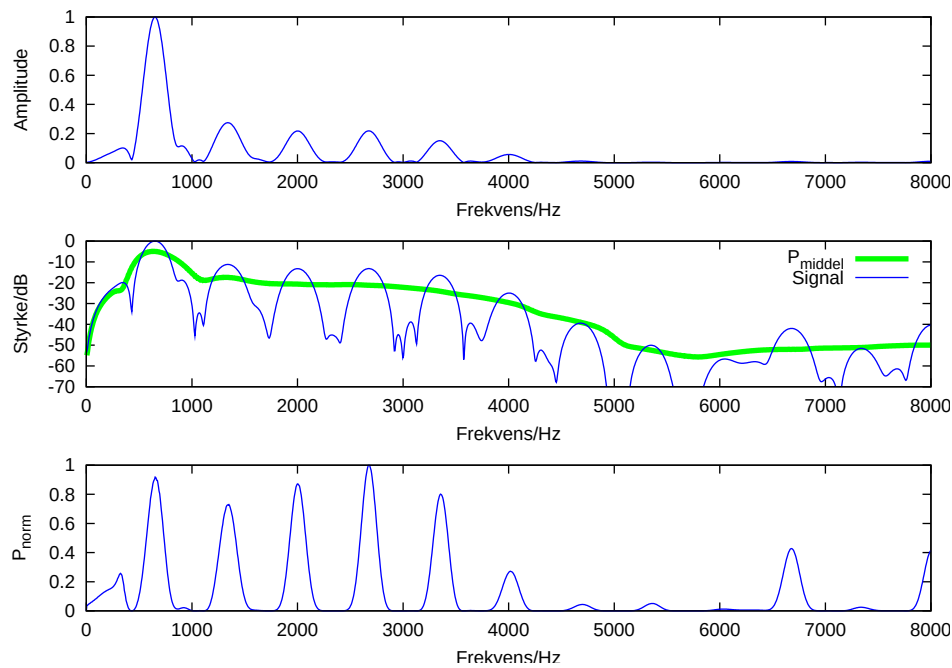
På figur 4.5 vises et eksempel på brugen af den ny spektrale normaliseringsfaktor. Øverst vises amplituden for hver frekvens i det signal, som blev anvendt på figur 3.11. I midten er styrken af hver frekvens i signalet afbildet sammen med den gennemsnitlige styrke P_{middel} . Nederst er den ny spektrale normaliseringsfaktor afbildet. Den spektrale normaliseringsfaktor er beregnet efter formel 4.13 og $c = 1,2$. ρ er fastsat således, at ρ tilnærmelsesvist er lig ν de første 500 Hz, og derefter stiger ρ langsomt, således at ρ ligger omkring $\nu/4$ ved 5.000 Hz. Det ses, at den spektrale normaliseringsfaktor set i forhold til den normale amplitude giver større vægt til de svage frekvenser, såfremt de fremstår som højderykke i det gennemsnitlige styrkebillede.

4.3.2 Oktavfejl

Lad os tage udgangspunkt i formel 4.11. Det centrale i formlen er, at der for hver formodet fundamentalfrekvens foretages en sum af overtonernes amplituder. Amplituderne kan ses som et bud på, hvordan man kan angive styrkefordelingen mellem de forskellige frekvenser. Et andet bud kan f.eks. være den spektrale normaliseringsfaktor.

Funktionen h , som angiver en vægtning af hvert led i summen, har primært til opgave at sørge for, at en dyb formodet fundamentalfrekvens ikke bliver vægtet for højt. Antag at et signal kun indeholder en tone med fundamentalfrekvens på ν_0 . Tonen indeholder kun $N/2 - 1$ overtoner, og alle frekvenserne over den højeste overtone har en amplitude på 0. Lad os beregne SHS for signalet ved hjælp af formel 4.11, og antag at funktionen h ikke tages med. Hvis vi dernæst beregner SHS for en formodet fundamentalfrekvens på $\nu_0/2$,

¹⁴En kvadratisk Bézier-kurve kan defineres ved hjælp af 3 punkter (P_1 , P_2 , P_3) på en flade. Kurven afhænger af en variabel t , som løber i intervallet $[0;1]$ og kan defineres ved: $B(t) = (1 - t)^2 P_1 + 2(1 - t)t P_2 + t^2 P_3$.



Figur 4.5: Her illustreres forskellen på at bruge amplituden set i forhold til at bruge den ny spektrale normaliseringsfaktor. Det anvendte signal blev også anvendt på figur 3.11. Øverst vises amplituden for hver frekvens i signalet. I midten vises styrken af hver frekvens sammen med den gennemsnitlige styrke P_{middel} . Nederst vises den ny spektrale normaliseringsfaktor P_{norm} . Det ses, at den ny spektrale normaliseringsfaktor i forhold til amplituden giver større vægt til de høje overtoner, såfremt de skiller sig ud fra omgivelserne.

og signalet indeholder lidt støj ved en frekvens på $\nu_0/2$, så vil SHS for denne formodede grundfrekvens få en større vægtning end SHS for den korrekte grundfrekvens ν_0 .

Ved at lade $\mathbf{h}$ være en aftagende funktion kan ovenstående scenarie i mange situationer undgås. Det er dog et problem, hvis $\mathbf{h}$ falder for hurtigt, for da vil de høje overtoner kun i ringe grad bidrage ved beregningen af SHS, og dette kan være et problem, hvis en stor del af en tones energi ligger i de høje overtoner.

Der er 2 problemer ved funktionen $\mathbf{h}$. På den ene side skal funktionen sørge for, at der ikke bliver gættet på en for dyb oktav. På den anden side, angiver funktionen også en vægtning af overtonerne og kan dermed enten opprioritere overtonerige toner, eller toner som ikke er overtonerige. Den funktion, som er foreslået i den originale artikel [28], $f(n) = 0.84^{n-1}$, aftager hurtigt til at begynde med, og flader derefter ud, således at de lave overtoner vægter mere end de høje overtoner, og de højeste overtoner vægter stort set lige meget. Denne funktion kan ses som et kompromis, som på den ene side kan give en stor vægtning til en tone, hvis energi stort set kun ligger i de første par overtoner, og på den anden side kan en tone også få en stor vægt hvis den indeholder energi i mange af de højere overtoner. Men for at undgå, at der bliver valgt for dybe oktaver, vil det generelt være en fordel, hvis $\mathbf{h}$ falder hurtigt, og en sådan funktion opprioriterer toner, som ikke er overtonerige. Det kan altså være problematisk at opprioritere overtonerige toner uden samtidig at finde for dybe oktaver.

Jeg har studeret de afprøvninger, der blev lavet i [66] i forbindelse med en modificeret SHS. Det ser ud, som om det i mange situationer ville være en fordel, hvis den anvendte funktion $\mathbf{h}$ havde givet større vægt til de højere overtoner. For at give mulighed for at opprioritere de overtonerige toner uden samtidig at

finde for dybe oktaver, vil jeg her foreslå, at der tilføjes en udglattende faktor. Antagelsen er, at styrken af en tones overtoner følger en blød kurve, forstået således, at overtone n og overtone $n + 1$ har stort set samme styrke.

4.3.2.1 Udglattende faktor

Den udglattende faktors formål er at mindske styrken for en overtone, hvis de omkringliggende overtoner er svagere. Det foreslås, at en vægtet styrke for en overtone beregnes ved en vægtning af styrkerne for selve overtonen og de omkringliggende overtoner:

$$\begin{aligned} \mathbf{P}_{vægt}\{\mathbf{P}\}(\nu_0, n) &= u_- \mathbf{P}((n-1)\nu_0) + (1 - u_- - u_+) \mathbf{P}(n\nu_0) + u_+ \mathbf{P}((n+1)\nu_0) \\ \mathbf{P}_{udglat}\{\mathbf{P}\}(\nu_0, n) &= \min(\mathbf{P}_{vægt}(\nu_0, n), \mathbf{P}(n\nu_0)) \end{aligned} \quad (4.15)$$

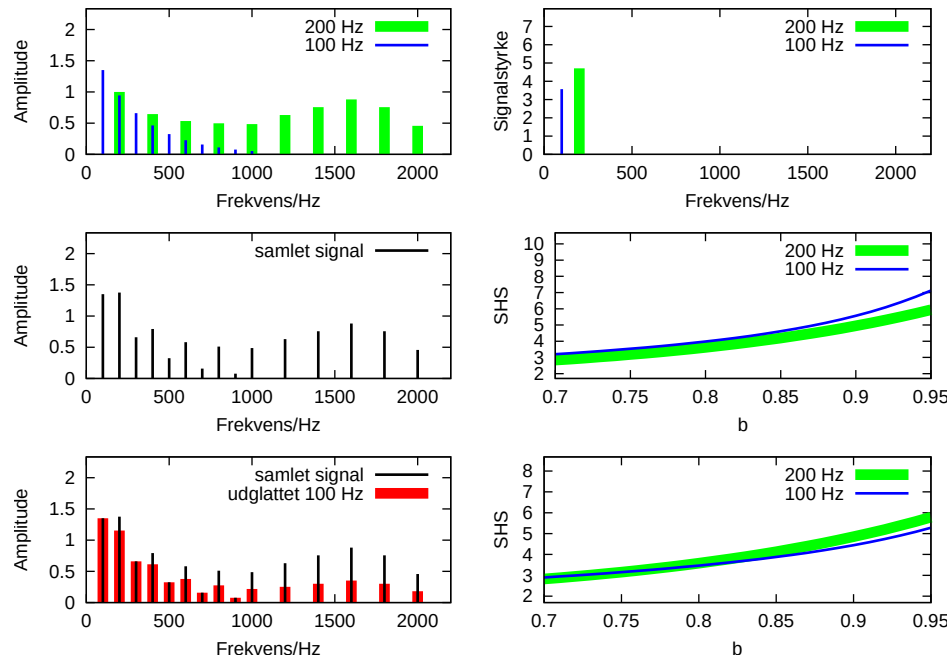
I ovenstående formel angiver $\mathbf{P}$ styrken for frekvenserne i et signal, og $\mathbf{P}_{vægt}$ angiver et vægtet middel af styrken, og dette vægtede middel beregnes på baggrund af styrkerne for overtone nummer $n - 1$, n og $n + 1$. Vægtene u_- og u_+ bestemmer vægten af henholdsvis overtone nummer $n - 1$ og nummer $n + 1$. $\mathbf{P}_{udglat}$ bestemmes som minimum af den vægtede middelstyrke og styrken for overtone n , således at en overtone kun kan beholde sin styrke eller få mindsket sin styrke.

Bemærk at vi i ovenstående formel har anvendt styrken $\mathbf{P}$, men vi kunne lige så godt have beregnet en udglattet amplitude eller et udglattet normaliseret styrkespektrum. Dette ville være muligt ved at erstatte $\mathbf{P}$ med $\mathbf{A}$ eller med $\mathbf{P}_{norm}$.

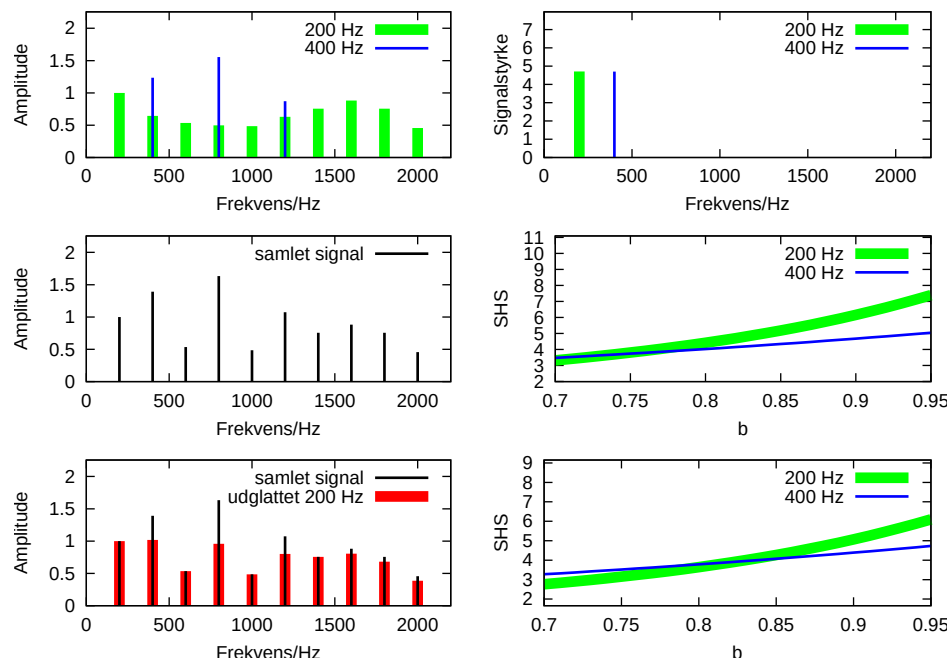
Lad os nu se på anvendelsen af den udglattende faktor i 2 scenarier. Først vil vi se på et scenarie, hvor den udglattende faktor kan forbedre succesraten for SHS, og dernæst vil vi se et scenarie, hvor den udglattende faktor har ringe betydning. De 2 scenarier er illustreret på figur 4.6 og 4.7.

Først ser vi på figur 4.6, hvor der øverst til venstre er vist amplituden for 2 signaler. Det ene signal består af en tone med en fundamentalfrekvens på 200 Hz, samt de første 9 overtoner. Det andet signal, som er svagere består af en tone med en fundamentalfrekvens på 100 Hz, samt de første 9 overtoner. De 2 toners styrker er vist øverst til højre, og her ses det, at tonen med en fundamentalfrekvens på 200 Hz er den kraftigste. Vi vil her antage, at den kraftigste tone er meloditonen, og det er dermed denne tone, som SHS helst skal finde. I midten til venstre vises amplituden for et mix af de 2 signaler. Det er dette mix, som SHS arbejder på. I midten til højre vises værdien af en almindelig SHS, hvor der er anvendt forskellige værdier for funktionen $\mathbf{h}(n) = b^{n-1}$, således at SHS beregnes som en funktion af b . Det ses, at SHS giver et forkert resultat for alle de anvendte værdier for b . Nederst til venstre vises det mixede signal igen, men her vises også $\mathbf{A}_{udglat}$ for en fundamentalfrekvens på 100 Hz — vi anvender her amplituden i stedet for styrken, da den normale SHS arbejder på amplituden. Nederst til højre vises værdien af SHS, hvor den udglattende faktor er anvendt, og hvor der anvendes forskellige værdier for b . Det ses, at SHS i dette tilfælde giver et korrekt resultat, når b er større end ca. 0,75.

På figur 4.7 vises et andet eksempel, hvor den almindelige SHS kan have et problem. Øverst til venstre vises 2 signaler, hvor det ene signal er meget overtonerigt, mens det andet signal kun indeholder 3 kraftige frekvenser. Som det ses øverst til højre, har de 2 signaler omtrent samme styrke. Et mix af de



Figur 4.6: Her illustreres effekten af at bruge den udglattende faktor. Der er tale om 2 toner, hvor den kraftigste tone har en fundamentalfrekvens på 200 Hz, og en svagere tone har en fundamentalfrekvens på 100 Hz. Amplituderne for de 2 toner er vist øverst til venstre, mens amplituderne for det mixede signal vises i midten til venstre. Nederst til venstre vises amplituderne igen, men her vises også anvendelsen af den udglattende faktor for en fundamentalfrekvens på 100 Hz. I midten til højre, samt nederst til højre er SHS beregnet på baggrund af henholdsvis amplituden og amplituden med udglattende faktor. Det ses, at SHS kun giver et korrekt resultat ved anvendelse af den udglattende faktor.



Figur 4.7: Denne figur svarer til den foregående, men her er de 2 toner på henholdsvis 200 Hz og 400 Hz omtrent lige kraftige. Som det ses af de 2 beregninger af SHS, så kan b sættes højere ved anvendelse af den udglattende faktor. Dette giver mulighed for bedre at finde en overtonerig tone, uden at der samtidig opstår for dybe oktav-fejl.

2 signaler vises i midten til venstre og i midten til højre vises værdien af SHS beregnet for forskellige værdier af b . Det ses, at SHS giver et korrekt resultat, når b ligger i nærheden af 0,73, hvilket er meget lavt. Hvis b generelt sættes så lavt, vil overtonerne have meget lille betydning, hvilket kan være problematisk, hvis en stor del af en tones styrke ligger i overtonerne. Nederst til venstre vises $\mathbf{A}_{udglat}$ for en fundamentalfrekvens på 200 Hz. Nederst til højre vises værdien af SHS som funktion af b , og her anvendes den udglattende faktor. Det ses, at SHS giver et korrekt resultat for b omkring 0,83, hvilket giver en bedre vægtning af de høje overtoner.

Generelt set gælder det, at hvis den udglattende faktor anvendes, og der gives stor vægt til de omkringliggende overtoner, så skal b sættes højt. Det kan være et problem at anvende den udglattende faktor, hvis antagelsen om, at styrkerne af overtonerne følger en blød kurve, er forkert. Derfor kan det være en god idé at afprøve forskellige vægte i forbindelse med den udglattende faktor, for at finde en optimal vægtning.

4.3.3 Ny definition på en modificeret SHS

Jeg vil nu bringe en ny definition på en modificeret SHS, hvor den ny spektrale normaliseringsfaktor samt den udglattende faktor bliver anvendt.

$$\text{shs}_{\text{norm}}\{\mathbf{P}\}(\nu_0) = \sum_{n=1}^N \mathbf{h}(n) \mathbf{P}_{udglat}\{\mathbf{P}_{\text{norm}}\}(\nu_0, n) \quad (4.16)$$

I ovenstående formel udregnes først $\mathbf{P}_{\text{norm}}$, og dette resultat anvendes dernæst ved beregningen af $\mathbf{P}_{udglat}$ for hver frekvens. Ved udregningen af $\mathbf{P}_{\text{norm}}$ i formel 4.13 foreslås det, at ρ er næsten lige så stor som ν op til en frekvens på omkring 500 Hz og dernæst stiger langsommere, således at ρ har en størrelse på omkring $\nu/4$ for en frekvens omkring 5.000 Hz. c i formel 4.13 foreslås sat til en værdi mellem 1 og 1,2. Ved beregningen af $\mathbf{P}_{udglat}$ foreslås det, at u_- og u_+ sættes mellem 0 og 0,3. Endelig foreslås det, at b ved beregningen af $\mathbf{h}(n) = b^{n-1}$ sættes til en værdi mellem 0,8 og 0,99, afhængig af værdierne for u_- og u_+ .

Bemærk at det i formel 4.16 blev foreslået at anvende $\mathbf{P}_{udglat}$ på $\mathbf{P}_{\text{norm}}$, men som eksemplerne på figur 4.6 og 4.7 angav, kan man eksempelvis også anvende $\mathbf{P}_{udglat}$ direkte på $\mathbf{A}$.

4.4 Andre interessante metoder

Her vil vi kort nævne et par interessante metoder, som baserer sig på en analyse af en tid/frekvens-repræsentation af et signal. Der er kun plads til at nævne ganske få metoder — for en mere udførlig oversigt over diverse metoder henvises der til [13].

Det er valgt at behandle den såkaldte TWM-metode i et afsnit for sig (afsnit 4.4.1). Mens det i SHS og Component Frequency Ratios¹⁵ antages, at overtonerne ligger med et fast interval (hvilket f.eks. ikke er tilfældet med en tone fra et klaver), så er TWM knap så afhængig af det faste interval mellem overtonerne.

Et eksempel på en gruppe af metoder, som ikke er så afhængig af et fast interval mellem overtonerne, er metoder som baserer sig på en autocorrelation af

¹⁵Component Frequency Ratios blev introduceret i [58], og blev kort behandlet i [66, afsnit 3.1.2.2].

spektrummet. Dette er f.eks. tilfældet hos Lahat [45], hvor spektrummet inddeles i en række frekvensbånd, og autocorrelation beregnes for spektrummet inden for hvert frekvensbånd, således at perioden inden for hvert frekvensbånd kan findes. På baggrund af de fundne perioder, som antages at angive afstanden mellem hver overtone, gives et bud på signalets fundamentalfrekvens. På denne måde er metoden rimelig robust overfor støj og manglende overtoner i visse frekvensbånd. Et andet eksempel, hvor autocorrelation anvendes, er metoden ACLOS [42], hvor autocorrelation beregnes for et logaritmisk spektrum. De svagere frekvensområder i spektrummet er blevet løftet for at udglatte det spektrale hylster¹⁶.

Der kan også henvises til Klapuri [38] for en metode, som er knap så afhængig af et fast interval mellem overtonerne. Også i denne metode undersøges en række frekvensbånd hver for sig. For hvert frekvensbånd udregnes et estimat for den frekvensmæssige afstand mellem overtonerne inden for frekvensbåndet. Dette foregår ved hjælp af en kam-metode¹⁷. På baggrund af estimaterne inden for hvert frekvensbånd gives et estimat for de mest fremtrædende fundamentalfrekvenser i signalet.

4.4.1 TWM

TWM blev introduceret af Maher i [48] i 1990, men i [49] beskrives detaljerne nærmere.

TWM står for “Two-Way Mismatch”. Metoden estimerer for hver formodet fundamentalfrekvens 2 mål, som beskriver, i hvor høj grad en tone ikke passer til et givent spektralt mønster, og i hvor høj grad det givne spektrale mønster ikke passer til tonen. Der tildeles en straf til hvert gæt på en fundamentalfrekvens, og den frekvens, som har fået tildelt den laveste straf, antages at være den væsentligste fundamentalfrekvens i spektrummet.

Metoden analyserer et spektrum og finder lokale maksima i spektrummet. Disse lokale maksima antages at stamme fra diverse toners overtoner, og vi vil derfor kalde dem for “formodede overtoner”. For en given harmonisk tone med tilhørende overtoner, findes for hver af tonens overtoner den formodede overtone i spektrummet, som frekvensmæssigt set ligger tættest på. Dernæst tildeles en straf til hver af den harmoniske tones overtoner afhængig af, hvor godt de tilhørende formodede overtoner opfylder 2 mål. Det ene mål handler om, hvor langt der frekvensmæssigt er mellem en overtone og den tilhørende formodede overtone. Det andet mål handler om, hvor kraftig den tilhørende formodede overtone er. Der tildeles minimal straf, hvis den tilhørende formodede overtone har samme frekvens som den harmoniske tones overtone og samtidig er den kraftigste frekvens i spektrummet. Omvendt tildeles maksimal straf, hvis den tilhørende formodede overtones frekvens er den kraftigste i spektrummet og samtidig ligger langt fra harmoniske tones overtones frekvens. Når der er tildelt straf til hver af den harmoniske tones overtoner, angiver gennemsnittet

¹⁶Det spektrale hylster (spectral envelope) angiver et udglattet spektrum, således at det er de overordnede strukturer, som fremvises. Ved at løfte de svagere frekvensområder opnås et spektralt hylster, som tilnærmelsesvist er konstant.

¹⁷I en kam-metode ganges et signal (eller signalets frekvens-repræsentation) med en funktion som har udseende af en kam. En kamfunktion antager med regelmæssige mellemrum et maksimum. Kamfunktionens udseende kan bestemmes ved bredden af hver tand i kammen og afstanden mellem hver tand. Når signalet er ganget med kammen, kan man eventuelt summere resultatet for på denne måde at få et mål for, hvor godt signalet ligner kammen.

af overtonernes straffe et mål for, hvor dårligt den harmoniske tone passer til spektrummet.

Det måles også, hvor dårligt spektrummet passer til den harmoniske tone. Proceduren, som blev beskrevet herover, gentages, men der byttes om på rollerne for den harmoniske tones overtoner og de formodede overtoner, således at der findes en gennemsnitlig straf og dermed et mål for, hvor dårligt spektrummet passer til den harmoniske tone.

En vægtet sum af de 2 gennemsnitlige straffe angiver den endelige straf for den aktuelle tones fundamentalfrekvens. På grund af anvendelsen af de 2 straffe, er metoden god til at bestemme, hvilken oktav en tone befinder sig i.

4.5 Konklusion

Vi har i dette kapitel stiftet bekendtskab med et par metoder, som arbejder på en tid/frekvens-repræsentation. Der blev formuleret en generalisering af autocorrelation og cepstrum. Autocorrelation og cepstrum har hver deres fordele, og det kan tænkes, at en metode, som "ligger imellem" de 2 metoder, kan være bedre til at estimere fundamentalfrekvensen.

I forhold til SHS blev der formuleret 2 forslag til modifikationer, nemlig den spektrale normaliseringsfaktor og den udglattende faktor. Med den spektrale normaliseringsfaktor angives det for en aktuel frekvens, hvor meget frekvensens styrke afviger fra den gennemsnitlige styrke i et bånd omkring frekvensen. Med den udglattende faktor søges det at mindske antallet af fejl, som skyldes, at der bliver gættet på en for dyb oktav. Med de 2 modifikationer er det håbet, at melodiens fundamentalfrekvens bedre kan estimeres.

Kapitel 5

Implementering

I dette kapitel præsenteres en samlet metode for estimering af melodiens fundamentalfrekvens, og metodens implementering bliver behandlet. I næste kapitel bliver implementeringen afprøvet på de datasæt, som blev præsenteret i afsnit 2.1.

Som det blev beskrevet i afsnit 4.1, er den generelle opbygning af en metode for bestemmelse af melodiens fundamentalfrekvens som regel delt i 3 dele, nemlig en preprocessing, en central del og til sidst en postprocessing. Den centrale del indeholder både et valg af en tid/frekvens-repræsentation samt en analyse af den valgte repræsentation. Vi behandler disse 2 dele i hvert sit afsnit.

5.1 Preprocessing

Jeg har valgt at give mulighed for, at der foretages en filtrering af signalet, således at frekvenskomponenternes styrker bliver repræsenteret på en phon-skala¹. Efter phon-filtreringen vil især frekvenserne omkring 3.000 Hz blive fremhævet, og især frekvenserne under 100 Hz vil blive undertrykket. Den maksimale styrke for en frekvens i signalet antages at være på 60 dB, hvilket svarer til kraftig tale.

Når der gives mulighed for at foretage denne filtrering, skyldes det, at man måske på denne måde kan komme nærmere det signal, som opfattes i det menneskelige øre.

Der er ikke givet mulighed for at foretage en filtrering, som har til hensigt at undertrykke støj eller at efterligne psykoakustiske effekter, så som maskering². Det ville være besværligt at implementere denne funktionalitet. Men da nogle lydkompressionsalgoritmer i høj grad benytter sig af den psykoakustiske maskeringseffekt til at fjerne uhørbare frekvenskomponenter, kan man eventuelt foretage en indirekte preprocessing ved at indkode en aktuel fil med en af de lydkompressionsalgoritme, som anvender psykoakustiske egenskaber. Et eksempel på et godt valg af lydkompressionsalgoritme er "libvorbis", som ifølge kildekoden benytter sig af maskeringseffekten.³

¹Der kan henvises til [66, afsnit 2.4.1.2], hvor sammenhængen mellem phon-værdier og lydstyrken vises.

²Se afsnit 4.1.1.2

³Der henvises til libvorbis på <http://xiph.org>.

5.2 Tid/frekvens-repræsentation

Jeg har valgt at anvende OQSTFT, som blev defineret i formel 3.49. OQSTFT har fordele af at være egnet til et signal, som indeholder svagt vibrerende toner, og derudover er OQSTFT meget simpel, så det er enkelt at implementere OQSTFT.

I afsnittet om STFT (afsnit 3.2) behandlede vi en række forskellige vinduer, og her så vi, at specielt Hann-vinduet og det sænkede Gauss-vindue gav gode resultater i forbindelse med et signal, hvor fundamentalfrekvensen er meget kraftig i forhold til de højere overtoner. Da overtonerne spiller en væsentlig rolle i den menneskelige opfattelse af musik, er det vigtig at vælge et vindue, som gør det muligt at identificere de højere overtoner. Jeg har valgt at anvende Hann-vinduet, men det sænkede Gauss-vindue kunne formentlig have været lige så godt at anvende.

I implementeringen antages det, at samplingsfrekvensen er 44.100 Hz, og for at nedbringe antallet af beregninger nedsamples der 3 gange til en samplingsfrekvens på 14.700 Hz, således at den maksimale frekvens, der arbejdes videre på, er 7.350 Hz. De højeste frekvenser i det resulterende signal bliver dog kraftigt undertrykket, da der inden nedsamplingen foretages en lavpas-filtrering for at undgå aliasing-problemer⁴ ved selve nedsamplingen. Således kan der arbejdes videre på frekvenser op til omkring 7.000 Hz. Ofte vælger man kun at arbejde på frekvenser op til 4.000 eller 5.000 Hz, men her er det valgt at gå helt op til 7.000 Hz. Dette skyldes, at der senere er brug for at finde en udglattet frekvensrepræsentation, og denne beregnes over et bredt vindue af frekvenser. Man kunne også have valgt slet ikke at nedsample signalet, men frekvenserne over ca. 5.000 Hz har generelt set ingen betydning for opfattelsen melodiens tonehøjde. De højere frekvenser har kun betydning for opfattelsen af klangen, og det bør derfor være uden betydning, at de højere frekvenser ikke tages med og dermed belaster beregningerne.

Den aktuelle implementering tager udgangspunkt i de optimerede beregninger, som blev beskrevet i afsnit 3.5.1.1. Der er på forhånd beregnet en frekvensrepræsentation af hvert vindue ganget med den tilhørende analyserende bølge. For at nedbringe antallet af beregninger er det kun frekvenser inden for main-lobe, som tages med i betragtning, når foldningen beregnes.

Der anvendes en maksimal vindueslængde på omkring 0,09 sekunder, således at den maksimale vindueslængde er på 4.096 samples ved en samplingsfrekvens på 44.100 Hz.⁵ Vindueslængderne i OQSTFT er lagt således, at der for en frekvens på 1.000 Hz anvendes en vindueslængde på 1.700 samples ved en samplingsfrekvens på 44.100 Hz. Man kunne også have valgt at bruge f.eks. 1.500 samples eller 1.900 samples — forsøg har vist at den optimale vindueslængde ligger et sted herimellem.

OQSTFT beregnes for frekvenser med en afstand på omkring 5 Hz, således at det for en samplingsfrekvens på 44.100 Hz svarer til en almindelig STFT med en vindueslængde på 4.096 samples og med en zero-padding på 4.096 samples.

Der anvendes en hop-size, som er variabel, forstået således at hop-size kan angives med en parameter. Implementeringen skal afprøves på de 2 datasæt ADC2004 og TEST1, og i de 2 datasæt er referenceresultatet registreret med

⁴For en illustration af aliasing-problemet henvises der til figur 3.1.

⁵Vi angiver vindueslængder i forhold til 44.100 Hz, da det derved bliver nemmere at sammenligne de angivne vindueslængder med vindueslængder angivet i andre skrivelser. En samplingsfrekvens på 44.100 Hz er meget normal at anvende.

forskellige tidsintervaller. I ADC2004 er referenceresultatet registreret med en afstand på 256 samples, hvilket svarer til omkring 0,006 sekunder. Derfor anvendes der i forbindelse med dette datasæt en hopsize på 256 samples ved en samplingsfrekvens på 44.100 Hz. I datasæt TEST1 er referenceresultatet registreret med en afstand på 0,01 sekunder, hvilket giver anledning til en hopsize på 441 samples ved en samplingsfrekvens på 44.100 Hz.

5.3 Estimering af melodiens fundamentalfrekvens

Jeg har implementeret en modificeret SHS, som er baseret på definitionen i formel 4.16. SHS er som udgangspunkt blevet valgt, fordi metoden giver mulighed for at identificere toner, som er meget overtonerige, men det er også muligt at identificere toner, som er meget overtonefattige.

Ved anvendelse af OQSTFT er signalet repræsenteret på en lineær frekvensakse, hvor afstanden mellem hver frekvens er på omkring 5 Hz. SHS beregnes på baggrund af denne repræsentation, men de frekvenser, som SHS beregnes for, er repræsenteret på en lineær frekvensakse med 4 gange så høj opløsning. Man kunne også have valgt at anvende en logaritmisk frekvensakse, men der ser ikke ud til at være nogen problemer ved anvendelsen af den lineære frekvensakse, og i implementeringen er det simplere at anvende den lineære frekvensakse. Det ser ikke ud til at give bedre resultater, hvis der anvendes en 8 gange så høj opløsning.

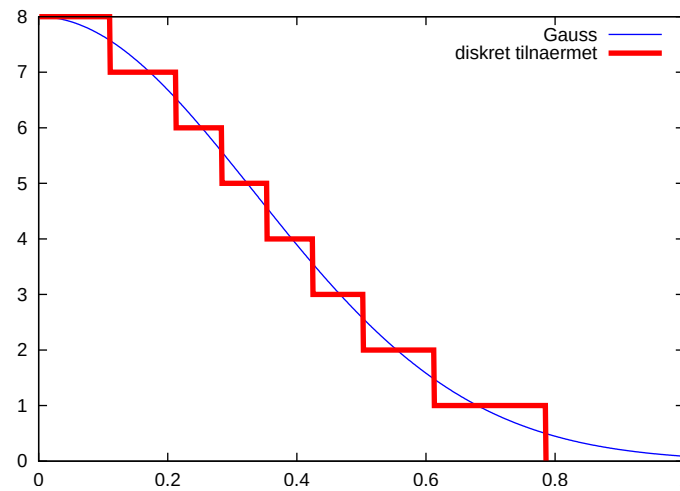
SHS beregnes for frekvenser mellem 100 og 1.200 Hz, hvilket antages at være det område, som melodiens fundamentalfrekvens befinder sig inden for.

Ved beregningen af $\mathbf{P}_{\text{middel}}$ anvendes der i formel 4.13 en vægt $\mathbf{w}$, og det blev foreslået at denne defineres ved den gaussiske funktion i formel 4.14. Det er dog beregningstungt at anvende denne gaussiske funktion, så den gaussiske funktion tilnærmes med en diskret funktion, som antager 8 værdier. Den diskrete funktion og den gaussiske funktion er vist på figur 5.1. Når den tilnærmede funktion anvendes, kan $\mathbf{P}_{\text{middel}}$ udregnes mere effektivt: Først udregnes den akkumulerede sum af styrkerne for hver frekvens. Ved differensen mellem de akkumulerede summer for 2 forskellige frekvenser, kan den gennemsnitlige styrke for frekvenserne herimellem findes. Bredden af hvert niveau i den tilnærmede gaussiske vægt afgør hvilke intervaller, der beregnes gennemsnit over. Der beregnes altså 8 gennemsnit, og disse lægges sammen, for at give et vægtet gennemsnit.

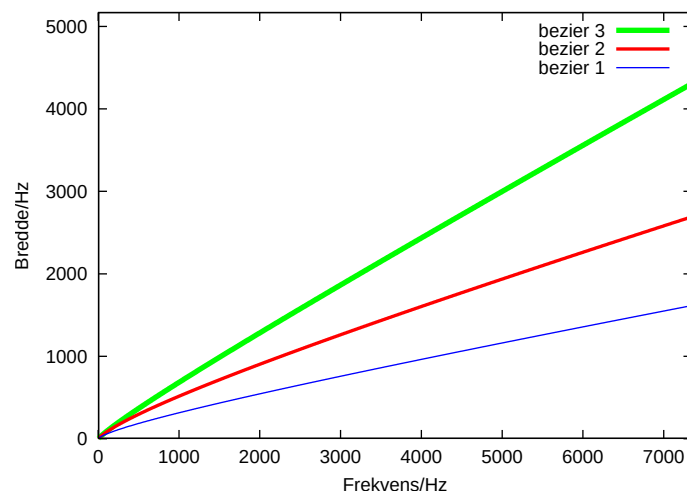
Den anvendte ρ -funktion er beskrevet ved en kvadratisk Bézier-kurve, og der er defineret 3 udseender for denne kurve. Med en parameter (ρ) kan en af de 3 udseender vælges. De 3 kurver er vist på figur 5.2. Der kunne også have været valgt andre udseender for funktionen ρ , men test har vist, at et udseende, som minder om de 3 kurvers udseender, giver gode resultater.

Der er givet mulighed for, at c i formel 4.13 kan sættes ved hjælp af en variabel.

I forhold til den udglattende faktor kan værdierne af u_- og u_+ fastsættes med hver sin parameter. Ved at sætte disse parametre lig 0, kan man altså sige, at den udglattende faktor ikke anvendes. Ledet med u_- medtages ikke i det tilfælde, hvor n i formel 4.15 er lig 1, således at $\mathbf{P}_{\text{vægt}}$ beregnes for en formodet fundamentalfrekvens.



Figur 5.1: Her vises den tilnærmede Gauss-funktion, hvor der anvendes 8 diskrete værdier. 1. akse angiver andelen af ρ .



Figur 5.2: Her vises de 3 forskellige Bézier-kurver. Med en parameter kan en af de 3 kurver vælges som udseendet af den funktion, der bestemmer ρ .

Ved beregningen af SHS medtages op til den 20. overtone. Da den dybeste tone forventes at have en fundamentalfrekvens på omkring 150 Hz, er det nødvendigt med op til 20 overtoner, fordi området omkring 2.000 til 3.000 Hz ofte indeholde kraftige overtoner i tilfældet med menneskelig sang.

Endelig er det muligt at sætte b ved en parameter. b anvendes ved definitionen af funktionen $\mathbf{h}$ i formel 4.16, således at $\mathbf{h}(n) = b^{n-1}$.

Da melodien ofte har en høj fundamentalfrekvens, mens kraftige akkompagnementstoner ofte har lavere fundamentalfrekvenser, er der givet mulighed for at tildele høje fundamentalfrekvenser større vægt. Dette kan betragtes som en form for postprocessing, da den tildelte vægt tilføjes efter udregningen af SHS. Der kan gives flere forslag til funktioner, som kan tildele de høje frekvenser større vægt. Her har vi valgt funktionen ν^h , da denne funktion har vist sig at være brugbar:

$$\mathbf{SHS}_{post}(\nu) = \nu^h \mathbf{SHS}_{norm}(\nu) \quad (5.1)$$

I nedenstående tabel er det oplyst hvilke parametre, der kan sættes:

Parameter	Anbefalede værdier	Forvalgt værdi
b (til beregning af $\mathbf{h}$)	0,9-0,99	0,95
u_- (til beregning af $\mathbf{P}_{udglat}$)	0,0-0,3	0,3
u_+ (til beregning af $\mathbf{P}_{udglat}$)	0,0-0,3	0,1
c (til beregning af $\mathbf{P}_{norm}$)	1,0-1,2	1,0
ρ (til beregning af ρ)	1 2 3	2
h (til beregning af $\mathbf{SHS}_{post}$)	0,0-0,2	0,0

5.3.1 Postprocessing

Fokus i denne opgave har ikke ligget på postprocessing-delen, og derfor er det en meget simpel postprocessing, som jeg har implementeret. Det er også muligt at undlade en postprocessing — i så fald er det til hvert tidspunkt den formodede fundamentalfrekvens, som SHS har tildelt størst vægt, som udvælges som et bud på melodiens fundamentalfrekvens.

Den valgte postprocessing er et forsøg på at fastholde melodiens fundamentalfrekvens, hvis denne til tider bliver lidt svagere. Der tages udgangspunkt i, at det i afsnit 4.1.2.1 blev nævnt, at Grey i [21] fandt, at den mest betydningsfulde klangfarve-dimension er det spektrale hylster. Med postprocessingen ønskes det at følge instrumenter, således at det er muligt at tildele større vægt til toner fra det instrument, som antages at have melodien.

Til enhver tid udtages de 5 højest vægtede formodede fundamentalfrekvenser. SHS tildeler vægte til de formodede fundamentalfrekvenser, og ud fra disse vægte udtrækkes først lokale maksima, og dernæst udtrækkes de 5 højeste maksima. Antagelsen er, at melodiens fundamentalfrekvens kan findes blandt de 5 kandidater. Det antages desuden, at melodiens fundamentalfrekvens som regel har fået tildelt højest vægt.

Det ønskes nu at spore hver af de 5 udtrukne toner/fundamentalfrekvenser gennem tiden. For at spore et harmonisk mønster (underforstået det spektrale hylster for en tone) kan der gives flere bud på metoder, som i større eller mindre grad er robuste overfor frekvensmæssige og tonestyrkemæssige ændringer. Jeg har udført forsøg med at repræsentere styrken af en tones overtoner på en

normaliseret logaritmisk skala, for på denne måde at skabe et klang-begreb. De-similariteten kan eksempelvis findes som en sum af de absolute afstande mellem hver af 2 toners klange. Så når de-similariteten er lav, stammer de 2 toner sandsynligvis fra det samme instrument.

Jeg har dog erfaret at det er lige så effektivt og meget simplere blot at summere differencen mellem 2 toners overtonestyrker. På denne måde kan der på simpel vis gives et mål for similariteten mellem 2 toner.

Til tidspunkt t og $t + 1$ udtrækkes de 5 toner, og similariteten mellem tonerne fra de 2 tidspunkter findes. Først undersøges det, hvilken tone den højest vægtede tone til tidspunkt t er mest similær med til tidspunkt $t + 1$. Denne tone markeres som en efterfølger til den højest vægtede tone til tidspunktet t . For hver af de øvrige toner til tidspunkt t findes på samme måde efterfølgere til tidspunktet $t + 1$.

Det antages, at en tones forløb på denne måde kan følges. Det antages desuden, at en melodi-tone ofte vil starte med at være kraftig og dernæst klinger ud. Så til tidspunkt t udvælges den tone, som har fået størst vægt af SHS. Til tidspunkt $t + 1$ findes efterfølgeren til den tidligere udvalgte tone. Hvis denne efterfølger også har fået tildelt størst vægt af SHS, fortsættes der blot til tidspunkt $t + 2$. Ellers kompenseres efterfølgeren i en vis grad for en eventuelt lavere tildelt vægt af SHS. Efter kompensationen udvælges blandt de 5 toner til tidspunkt $t + 1$ den tone, som nu har fået tildelt størst vægt, og denne tone opfattes som den ny melodi-tone, hvorefter der fortsættes til tidspunkt $t + 2$.

Ved hjælp af den omtalte kompensation er det håbet, at en melodi-tone kan fastholdes, selv om den til tider bliver lidt svagere end andre toner i akkompagnementet.

Kapitel 6

Testresultater

Her vil jeg først præsentere testresultater ved anvendelse af autocorrelation, cepstrum og den generalisering, som blev fremvist i afsnit 4.2.3.

Autocorrelation var på forhånd fremsat som en simpel metode, som det var forventet kunne estimere melodiens fundamentalfrekvens i en vis grad. Derfor kan resultaterne for autocorrelation anses som mindste-mål, som mere avancerede metoder helst skal overgå, før det bliver interessant at anvende de mere avancerede metoder. Dermed ikke sagt at resultaterne for samtlige testfiler skal være bedre for en mere avanceret metode, men de generelle resultater, skal helst være bedre.

Vi vil dernæst præsentere testresultater for denne opgaves endelige bud på en metode for estimering af melodiens fundamentalfrekvens. Metodens implementering blev beskrevet i kapitel 5.

Resultaterne præsenteres ved de 2 scorer Raw Pitch og Raw Chroma, som blev defineret i afsnit 2.2. Det primære datasæt er ADC2004, da dette datasæt blev anvendt ved MIREX-konkurrencerne.

6.1 Autocorrelation, cepstrum og generalisering

Her fremvises resultater for test udført på denne opgaves primære datasæt, nemlig ADC2004. Der bliver kun fremvist gennemsnitlige resultater for hele datasættet. Til beregning af gennemsnitlige resultater tæller hver test-fil lige meget, selv om nogle test-filer er længere end andre. Dette er i tråd med evalueringsmetoderne ved MIREX-konkurrencerne.

Vindueslængder De afprøvede metoder er blevet afprøvet med forskellige vindueslængder. Vindueslængderne er valgt således, at der både har været anvendt vindueslængder, som tydeligt vist var for korte, og vindueslængder, som tydeligt vist var for lange. Forskellige vindueslængder herimellem har været afprøvet, og det generelle mønster er, at de vindueslængder, for hvilke de højeste scorer blev opnået, ikke gav væsentligt bedre resultater end vinduelængder, som var omkring 20% kortere eller længere.

For hver af de afprøvede metoder vil vi i det følgende kun fremvise resultater opnået med de vindueslængder, som viste sig at give de højeste scorer. Som sagt har det generelt kun ringe betydning for resultaterne, hvis vindueslængden ændres med op til omkring 20 %. Ved hver tabel noteres de anvendte vindueslængder.

ADC2004 Autocorrelation $f(x) = x^2$					
		+Hann	+phon	+Hann +phon	+OQSTFT +phon
Pitch i %	58,6	59,8	70,0	70,4	71,7
Chroma i %	65,7	66,8	77,8	78,2	80,9

Tabel 6.1: Her vises resultater for autocorrelation, dvs. $f(x) = x^2$. Vindueslængden med rektangulært vindue er på 800 samples, mens vindueslængden med Hann vindue er på 1.200 samples. For OQSTFT er vindueslængden på 1.200 samples ved 1.000 Hz. De højeste scorer er opnået ved anvendelsen af både OQSTFT og phon-skalaen.

Hann-vindue Ud over at anvende det rektangulære vindue, vil vi også forsøge os med anvendelsen af et Hann-vindue, da Hann-vinduet ikke har de samme store problemer med side lobes som det rektangulære vindue.

OQSTFT Som tid/frekvens-repræsentation vil vi både anvende en almindelig STFT, men vi vil også forsøge os med en OQSTFT, da det forventes at OQSTFT giver en tid/frekvens-repræsentation, som bedre beskriver de væsentlige indhold i signalet.

Phon-skala Da det menneskelige øre især undertrykker meget dybe frekvenser, vil vi desuden forsøge at foretage en preprocessing i form af en filtrering af signalet ved hjælp af en repræsentation på phon-skalaen.¹

6.1.1 Autocorrelation

Jeg har implementeret autocorrelation i forhold til formel 4.5. I tabel 6.1 fremvises resultaterne for denne implementering. Det ses, at det er en stor fordel at filtrere signalet ved hjælp af phon-skalaen. Desuden ser det ud til at være en fordel at anvende Hann-vinduet i stedet for det rektangulære vindue. Ved både at filtrere med phon-skalaen og samtidig anvende OQSTFT opnåedes de bedste resultater blandt de afprøvede kombinationer. Dette er ikke nogen stor overraskelse, fordi phon-skalaen undertrykker de kraftige dybe frekvenser, og fordi det antages, at OQSTFT set i forhold til STFT giver en bedre tid/frekvens-repræsentation af et signal, som indeholder polyfon musik.

6.1.2 Cepstrum

Jeg har implementeret Cepstrum i forhold til formel 4.7. I tabel 6.2 fremvises resultaterne for implementeringen. Det ses, at der opnås gode resultater ved brug af Hann-vinduet. Overraskende nok gav anvendelsen af OQSTFT og phon-skalaen dårlige resultater.

Set i forhold til autocorrelation, så opnår cepstrum væsentligt bedre resultater, når metoderne anvendes uden brug af diverse modifikationer.

¹Phon-enheden er beskrevet i [66, afsnit 2.4.1.2].

	ADC2004 Cepstrum		$f(x) = \log(x)$		
		+Hann	+phon	+Hann +phon	+OQSTFT +phon
Pitch i %	69,5	76,2	70,0	71,2	64,7
Chroma i %	75,5	81,0	77,3	76,9	70,5

Tabel 6.2: Her vises resultaterne for cepstrum, dvs. $f(x) = \log(x)$. Vindueslængden med rektangulært vindue er på 800 samples, mens vindueslængden med Hann vindue er på 1.200 samples. For OQSTFT er vindueslængden på 1.500 samples ved 1.000 Hz. De højeste scorer blev opnået ved brug af Hann-vinduet.

	ADC2004 Generalisering		$f(x) = x$		
		+Hann	+phon	+Hann +phon	+OQSTFT +phon
Pitch i %	67,6	71,1	73,8	74,8	73,9
Chroma i %	74,1	77,2	81,0	82,5	82,6

Tabel 6.3: Her vises resultaterne for $f(x) = x$. Vindueslængden med rektangulært vindue er på 800 samples, mens vindueslængden med Hann vindue er på 1.200 samples. For OQSTFT er vindueslængden på 1.200 samples ved 1.000 Hz. De sidste 3 kombinationer opnår alle høje scorer.

6.1.3 Generalisering

Jeg har implementeret generaliseringen i forhold til formel 4.8. Som et eksempel på brugen af generaliseringen har jeg udført forsøg med $f(x) = x$. Man kunne have valgt andre funktioner, men denne funktion er meget simpel.² Resultaterne er oplistet i tabel 6.3, og her ses det, at Raw Chroma er højest ved brugen af OQSTFT og phon-skalaen. Raw Pitch er højest ved anvendelsen af Hann-vinduet og phon-skalaen. Men der er ikke de store forskelle på resultaterne for de 3 sidste kombinationer i tabellen.

Som det blev nævnt i 4.2.3, så benytter autocorrelation og cepstrum sig af 2 forskellige metoder, hvorpå spektrummet modificeres, og som vi så i resultaterne for de 2 metoder, så kan begge metoder i en vis grad anvendes til bestemmelse af melodiens fundamentalfrekvens. Ved at vælge en f , som man kan sige "ligger imellem" de 2 f -funktioner, som anvendes i autocorrelation og cepstrum, var det håbet, at det var muligt at opnå højere scorer. Det har vist sig at alle 3 metoder opnåede omtrent samme scorer, men med forskellige modifikationer.

6.1.4 Resultatanalyse

Det mest interessante ved tabel 6.1, 6.2 og 6.3 er, at de højeste scorer, specielt hvad angår Raw Chroma, ligger særdeles højt set i forhold til de scorer, som de deltagende metoder ved MIREX-konkurrencerne opnåede. De højeste Raw Chroma scorer, som blev opnået ved anvendelse af de 3 afprøvede metoder, er oplistet i tabel 6.4 sammen med Raw Chroma scorerne for de deltagende metoder ved MIREX-2006 og MIREX-2008. Tabellen er ordnet således, at den højeste score ligger øverst.

²Jeg har faktisk også udført forsøg med $f(x) = x^{1,5}$, sådan som det indirekte blev gjort i [34], men resultaterne herved var ikke lige så gode som for $f(x) = x$.

ADC2004 Raw Chroma i %	
MIREX-2008 Durrieu, Richard, David	86,2
MIREX-2008 Cancela	85,9
MIREX-2008 Rao, Rao	85,2
MIREX-2006 Dressler	84,0
MIREX-2008 Ryynanen, Klapuri	83,5
Generalisering +OQSTFT +phon	82,5
MIREX-2006 Ryynanen, Klapuri	82,3
Cepstrum +Hann	81,0
Autocorrelation +OQSTFT +phon	80,9
MIREX-2008 Cao, Li, Liu, Yan	76,7
MIREX-2006 Poliner, Ellis	76,4
MIREX-2006 Brossier	68,7
MIREX-2006 Sutton, Vincent, Plumbley, Bello	65,4

Tabel 6.4: Her vises Raw Chroma scorer for de metoder, som deltog ved MIREX-2006 og MIREX-2008. Desuden vises Raw Chroma for de 3 metoder, som blev afprøvet i de foregående afsnit, og for disse metoder er resultatet markeret med fed. Det ses, at de 3 metoder er konkurrencedygtige i forhold til flere af metoderne, som deltog ved MIREX-konkurrencerne, på trods af at alle disse metoder er langt mere komplicerede.

Alle metoderne, som deltog ved de nævnte MIREX-konkurrencer, er langt mere komplicerede end de 3 metoder, som blev afprøvet her. Alligevel har de 3 metoder opnået scorer, som gør dem konkurrencedygtige i forhold til flere af de komplicerede metoder.

Ved at anvende en anden f-funktion i generaliseringen, kunne man måske opnå endnu højere scorer, men det er et generelt problem ved de 3 metoder, at de egentlig ikke er designet til at finde multiple fundamentalfrekvenser i et signal. Og dette vil muligvis være nødvendigt for at opnå endnu højere scorer, fordi det ellers kan være svært at give et bud på, hvilke toner et signal består af.

6.1.5 Yin

Jeg har implementeret det omtalte 2. skridt i Yin (se afsnit 4.2.4 formel 4.10), og implementeringen har været afprøvet på ADC2004 databasen, men resultaterne er ikke overbevisende. Der opnås resultater, som minder om dem, der blev opnået med autocorrelation, så på den anvendte database fremstår Yin ikke som en bedre metode end autocorrelation. Yin fremstår snarere som en dårligere metode, fordi autocorrelation sammen med Hann-vindue og phon-værdier opnår langt bedre resultater.

En årsag til, at der i [10] rapporteres om så store forbedringer set i forhold til autocorrelation, kan være, at autocorrelation i [10] udregnes ved hjælp af formel 4.4. Forsøg har vist, at hvis autocorrelation udregnes ved hjælp af denne formel, så opnås der et resultat, som ligger omkring 5%-point under det resultat, som blev rapporteret i tabel 6.1.

6.2 Modifieret SHS

Her præsenterer jeg resultater for den modificerede SHS, hvis implementering blev beskrevet i kapitel 5. Metoden bliver afprøvet med en række forskellige parametre, og der bliver testet på begge denne opgaves datasæt, nemlig ADC2004 og TEST1. Dette giver mulighed for at undersøge, om der kræves forskellige parametre for de 2 datasæt.

Jeg vil præsentere gennemsnitlige resultater for hvert datasæt, således at hver fil tæller lige meget ved udregningen af gennemsnittet — dette er i tråd med evalueringsmetoderne ved MIREX-konkurrencerne. I nogle tabeller, hvor der gives resultater for både TEST1 og ADC2004, noteres der også et gennemsnit for begge datasæt. Til beregning af dette gennemsnit vægter hvert datasæt lige meget, således at det ikke har betydning, at det ene datasæt indeholder flere testfiler end det andet.

Der er blevet udført test med adskillige forskellige kombinationer af parametre. På grund af begrænsede ressourcer er nogle kombinationer testet mere intenst end andre.

Jeg vil i det følgende gennemgå, hvorledes den modificerede SHS er blevet testet, og fremvise de væsentlige resultater undervejs.

Hver overordnet test-scenarie er blevet testet med en almindelig STFT med en vindueslængde på 1.700 samples (ved 44.100 Hz samplingsfrekvens), og med OQSTFT med en vindueslængde som for en frekvens på 1.000 Hz er på 1.700 samples. Jeg har eksperimenteret med forskellige vindueslængder, og de nævnte vindueslængder ser generelt ud til at være gode at anvende. Antagelsen er, at OQSTFT generelt giver et bedre billede af indholdet i et signal, og derfor forventes det, at det vil være en fordel at anvende OQSTFT.

Først i de sidste test-scenarier er der blevet afprøvet preprocessing og postprocessing, så generelt er det kun den centrale del der afprøves. Man kan argumentere for, at især preprocessing burde blive afprøvet sammen med forskellige parametre for den centrale del, men forsøg har vist, at den tendens, som beskrives under afprøvningen af preprocessing, er generel for forskellige valg af parametre for den centrale del.

6.2.1 Almindelig SHS

Først har jeg afprøvet den almindelige SHS. Dette indebærer at P_{udglat} i formel 4.16 erstattes med amplituden A . Dermed er der kun én parameter tilbage, nemlig b ved beregningen af h i formel 4.16. Nedenstående tabel viser, hvilke værdier b har været sat til:

Almindelig SHS — parametre					
b	0,7	0,75	0,8	0,85	0,9

Resultaterne for de 2 datasæt, og for STFT og OQSTFT er oplistet i nedenstående tabeller:

Almindelig SHS — TEST1					
STFT					
b	0,7	0,75	0,8	0,85	0,9
Pitch i %	59,2	58,8	57,3	54,2	48,9
Chroma i %	70,1	70,4	70,3	69,9	68,6
OQSTFT					
b	0,7	0,75	0,8	0,85	0,9
Pitch i %	65,3	65,8	65,4	63,2	58,7
Chroma i %	73,0	73,6	73,8	73,4	72,3

Almindelig SHS — ADC2004					
STFT					
b	0,7	0,75	0,8	0,85	0,9
Pitch i %	68,1	69,0	68,2	64,7	56,8
Chroma i %	76,1	76,7	76,3	74,4	70,2
OQSTFT					
b	0,7	0,75	0,8	0,85	0,9
Pitch i %	70,9	72,5	72,8	70,3	63,8
Chroma i %	79,0	79,9	80,1	78,7	75,2

Det ses af tabellerne, at der er en svag tendens til, at b helst skal være lidt lavere for TEST1 end for ADC2004, men generelt set ser det ud til, at $b = 0,75$ er et godt valg. Desuden ses det, at der generelt fås en lavere score ved TEST1 end ved ADC2004. Dette er ikke overraskende, da det i afsnit 2.1 blev konstateret, at ADC2004 indeholder noget simplere test-filer end TEST1.

Som forventet ser det ud til at være en fordel at anvende OQSTFT i stedet for STFT.

Det overrasker lidt, at resultaterne opnået med den almindelige SHS er knap så gode som de resultater, der blev opnået med de modificerede udgaver af autocorrelation og cepstrum. SHS kan nemlig give et detaljeret billede af, hvilke toner et signal formentlig indeholder.

6.2.2 Modifieret SHS med amplitude

Da den almindelige SHS arbejder på amplituden, mens den modificerede SHS fra formel 4.16 arbejder på styrken, vil jeg først foretage nogle test af den modificerede SHS med anvendelse af amplituden, inden jeg afprøver den modificerede SHS med anvendelse af styrken.

For at begrænse antallet af test-scenarier har jeg valgt at holde ρ og h fast. I nedenstående tabel angives de forskellige værdier, som er blevet anvendt for de forskellige parametre. De valgte værdier gav anledning til i alt 135 scenarier, som blev kørt med anvendelsen af både STFT og OQSTFT.

Modifieret SHS med amplitude — parametre					
b	0,75	0,80	0,85	0,90	0,95
u_-		0,0	0,1	0,3	
u_+		0,0	0,1	0,3	
c		1,0	1,1	1,2	
ρ			2		
h			0		

Da der er så mange scenarier, kan alle resultaterne ikke oplistes i en tabel. Resultaterne for hvert datasæt, samt for hver tid/frekvens-repræsentation er blevet sorteret efter Raw Pitch og Raw Chroma.

En undersøgelse af de 20 bedst placerede scenarier viser, at der ikke er de store forskelle mellem de mest optimale parametre for hvert af de 2 datasæt. Derfor behandles de generelle tendenser samlet herunder:

- Parametren c har kun ringe betydning. Der opnås omtrent samme resultater ligegyldig med om $c = 1,0$ eller $c = 1,1$, men hvis $c = 1,2$ opnås en lidt dårligere score.
- Høje værdier for u_- og u_+ følger med høje værdier for b .
- $b = 0.95$ og $b < 0,85$ giver ikke gode resultater.

Der opnås dårlige resultater, hvis $b = 0.95$ og værdierne for u_- og u_+ samtidig er lave. De dårligste resultater opnås, når b er lav samtidig med at værdierne for u_- og u_+ er høje. Dette er også forventeligt, da høje værdier af u_- og u_+ har lidt samme effekt som en lav værdi for b .

I tabellerne herunder oplistes resultaterne for de 20 bedst placerede scenarier. Det noteres altså hvilket spænd, de 2 scorer findes i.

Modifieret SHS med amplitude — TEST1		
	STFT	OQSTFT
Pitch i %	77,7 - 80,1	81,7 - 82,7
Chroma i %	80,6 - 82,2	84,6 - 85,0

Modifieret SHS med amplitude — ADC2004		
	STFT	OQSTFT
Pitch i %	81,7 - 83,0	81,8 - 82,7
Chroma i %	86,5 - 87,6	86,3 - 86,8

Af de ovenstående tabeller ses det, at OQSTFT generelt giver et markant bedre resultat end STFT i tilfældet med datasæt TEST1. Men overraskende nok, så er dette ikke tilfældet med datasæt ADC2004, hvor STFT generelt opnår lidt bedre resultater. Hvis man ser på resultaterne for begge datasæt, fremstår OQSTFT dog som mest fordelagtig.

Set i forhold til resultaterne for den almindelige SHS, så giver den modificerede SHS markant bedre resultater. Dette skyldes primært anvendelsen af det normaliserede spektrum.

Generelt gælder det, at hvis Raw Pitch er høj, så er Raw Chroma også høj, men hvis Raw Chroma er høj, kan Raw Pitch godt være relativt lav. Som eksempel herpå kan nævnes, at der for datasæt ADC2004 med STFT blandt de

20 bedste resultater sorteret efter Raw Chroma findes Raw Pitch scorer helt ned til 77,0.

Et eksempel på en god sammensætning af parametre er noteret i følgende tabel:

Modifieret SHS med amplitude — gode parametre $b = 0,9 \quad u_- = 0,3 \quad u_+ = 0,1 \quad c = 1 \quad \rho = 2 \quad h = 0 \quad \text{OQSTFT}$			
	TEST1	ADC2004	gennemsnit
Pitch i %	82,5	82,7	82,6
Chroma i %	84,8	86,4	85,6

Med det valg af parametre, som blev angivet i ovenstående tabel, opnåedes generelt gode resultater for begge datasæt. Men der er ikke stor forskel på, hvilke parametre der er mest optimale for de 2 datasæt. De parametre, som gav de 6 højeste Raw Pitch scorer for datasæt TEST1, er også at finde blandt de 20 højeste Raw Pitch scorer for datasæt ADC2004. Omvendt gælder det, at de parametre, som gav de 4 højeste Raw Pitch scorer for datasæt ADC2004, også er at finde blandt de 20 højeste Raw Pitch scorer for datasæt TEST1.

For at undersøge effekten af den udglattende faktor er de højeste Raw Pitch scorer for $u_- = 0$ og $u_+ = 0$ fundet. De ligger ikke overraskende lidt lavere end de scorer, som blev fremvist i ovenstående tabel. Den udglattende faktor ser altså ud til at have en positiv effekt på resultaterne. Det er heller ikke overraskende, at $b = 0,85$ for den højeste Raw Pitch score opnået uden den udglattende faktor. Med anvendelse af den udglattende faktor er det en fordel at b ligger lidt højere end ellers, og dette er forventeligt, da den udglattende faktor netop skulle undertrykke de lave oktav-fejl. Den bedste score opnået uden den udglattende faktor er noteret i tabellen herunder:

Modifieret SHS med amplitude — uden udglat $b = 0,8 \quad u_- = 0 \quad u_+ = 0 \quad c = 1 \quad \rho = 2 \quad h = 0 \quad \text{OQSTFT}$			
	TEST1	ADC2004	gennemsnit
Pitch i %	81,5	82,3	81,9
Chroma i %	84,5	86,3	85,4

6.2.3 Modifieret SHS med styrke

Her præsenteres testresultater for den modificerede SHS, hvor det er styrken, der anvendes, således at det nu er den endelige implementering, som afprøves.

For at begrænse antallet af tests er det ikke alle parametre, som bliver sat til flere værdier. Vi erfarede i forbindelse med den modificerede SHS, som anvender amplituden, at parametren c helst skulle være lig 1 eller være meget tæt på 1. Dermed ikke sagt at dette valg også vil være godt, når det er styrken, som anvendes, men det antages at være fornuftigt at fastholde $c = 1$. Til gengæld prøver vi nu at sætte parametren h til flere værdier. Med de værdier, som er oplistet herunder, er der i alt 135 scenarier.

Modificeret SHS med styrke — parametre					
b	0,85	0,90	0,95	0,97	0,98
u_-		0,0	0,1	0,3	
u_+		0,0	0,1	0,3	
c			1,0		
ρ			2		
h		0,0	0,1	0,2	

Ligesom tidligere er resultaterne fra de forskellige scenarier blevet sorteret efter Raw Pitch og Raw Chroma, og de 20 bedste scenarier er blevet undersøgt nærmere.

Fælles for de bedste resultater er, at b er høj og det er u_- og u_+ også. h ligger lavt og skal helst være lig 0.

I tabellerne herunder oplistes resultaterne for de 20 bedst placerede scenarier. Det noteres altså hvilket spænd, de 2 scorer findes i.

Modificeret SHS med styrke — TEST1		
	STFT	OQSTFT
Pitch i %	76,4 - 77,9	81,4 - 82,3
Chroma i %	79,1 - 80,3	83,8 - 84,6

Modificeret SHS med styrke — ADC2004		
	STFT	OQSTFT
Pitch i %	80,4 - 81,9	82,7 - 83,4
Chroma i %	85,7 - 86,8	87,2 - 87,7

Af ovenstående tabeller ses det, at OQSTFT generelt giver bedre resultater end STFT. Dette står altså i kontrast til afprøvningen på datasæt ADC2004, hvor amplituden blev anvendt. Resultaterne for OQSTFT minder om de resultater, der blev opnået med OQSTFT i tilfældet, hvor amplituden blev anvendt, men resultaterne for STFT er noget lavere ved brug af styrken, set i forhold til når amplituden anvendes.

Et eksempel på en god kombination af parameterværdier og resultaterne for denne kombination er noteret i følgende tabel:

Modificeret SHS med styrke — gode parametre $b = 0,95$ $u_- = 0,3$ $u_+ = 0,1$ $c = 1$ $\rho = 2$ $h = 0$ OQSTFT			
	TEST1	ADC2004	gennemsnit
Pitch i %	82,2	83,2	82,7
Chroma i %	84,5	87,3	85,9

Med det valg af parametre, som blev angivet i ovenstående tabel, opnåedes generelt gode resultater for begge datasæt. De parametre, som gav de 6 højeste Raw Pitch scorer for datasæt TEST1, er også at finde blandt de 20 højeste Raw Pitch scorer for datasæt ADC2004. Omvendt gælder der, at de parametre, som gav de 10 højeste Raw Pitch scorer for datasæt ADC2004 også er at finde blandt de 20 højeste Raw Pitch scorer for datasæt TEST1. Dette vidner om, at der ikke er stor forskel på, hvilke parametre der er mest optimale for de 2 datasæt.

For at undersøge effekten af den udglattende faktor er de højeste Raw Pitch scorer for $u_- = 0$ og $u_+ = 0$ fundet. Konklusionerne følger de konklusioner, der

blev gjort i forbindelse med anvendelsen af amplituden i stedet for styrken. Den bedste score opnået uden den udglattende faktor er noteret i tabellen herunder:

Modifieret SHS med styrke — uden udglat			
$b = 0,9 \quad u_- = 0 \quad u_+ = 0 \quad c = 1 \quad \rho = 2 \quad h = 0 \quad \text{OQSTFT}$			
	TEST1	ADC2004	gennemsnit
Pitch i %	80,4	81,5	81,0
Chroma i %	83,6	86,3	85,0

Yderligere tests Af de foretagne afprøvninger ser det ud som om, der ikke er stor forskel på, om det er styrken eller amplituden, der anvendes. Men da den modificerede SHS blev defineret ved anvendelsen af styrken, er der foretaget yderligere test med forskellige parameterværdier, hvor det er styrken, som anvendes. Ved anvendelsen af OQSTFT og datasæt ADC2004 har jeg udført tests, hvor c sættes til værdierne 1,0, 1,1 og 1,2, og ρ sættes til værdierne 1, 2 og 3. De bedste resultater ved disse yderligere test kunne ikke forbedre Raw Pitch scoren, men Raw Chroma kunne forbedres med 0,2 %-point. Man kan altså ikke sige, at der er opnået væsentligt bedre resultater, og derfor vil jeg foreslå, at parametrene ρ og c ignoreres og blot sættes til de forvalgte værdier.

6.2.4 Preprocessing

For at begrænse antallet af testscenarier har jeg valgt at anvende de parametre, som blev foreslået som nogle gode parametre i forbindelse med anvendelsen af styrken. Der har været foretaget afprøvninger med andre valg af parameterværdier, og resultaterne fra disse afprøvninger afviger ikke fra de resultater, som noteres i det følgende.

Phon-skala Med de nævnte parametre blev det undersøgt, om det er en fordel at filtrere signalet med phon-skalaen. Det var ikke forventet at dette nødvendigvis ville give bedre resultater, men det kunne jo tænkes, og derfor er scenariet afprøvet. Som det fremgår af nedenstående tabel, blev der ikke opnået bedre resultater, men de var på den anden side heller ikke væsentligt dårligere.

Modifieret SHS med styrke			
$b = 0,95 \quad u_- = 0,3 \quad u_+ = 0,1 \quad c = 1 \quad \rho = 2 \quad h = 0 \quad \text{OQSTFT} \quad \text{phon}$			
	TEST1	ADC2004	gennemsnit
Pitch i %	80,9	81,9	81,4
Chroma i %	84,1	86,7	85,4

libvorbis Jeg har også foretaget afprøvninger, hvor testfilerne er blevet indkodet ved hjælp af kompressionsalgoritmen libvorbis. På denne måde filtreres signalet indirekte, således at uhørbare frekvenskomponenter ikke tages med. Der blev indkodet med 45 kBit/s, hvilket er en meget hård kompression, men det er alligevel let for et menneske at høre melodien.

Det viste sig ikke at være en fordel at anvende kompressionsalgoritmen. Det var dog heller ikke en stor ulempe at anvende kompressionsalgoritmen, da resultaterne næsten var identiske med dem, som blev opnået uden brug af kompressionen.

Der var egentlig ikke store forventninger til, at kompressionen ville forbedre resultaterne, men det, at resultaterne ikke blev væsentligt dårligere er indirekte en god nyhed, for hvis resultaterne blev væsentligt dårligere, ville det være et tegn på, at den modificerede SHS ikke er robust overfor fjernelse af frekvenskomponenter, som er uvæsentlige for mennesker.

6.2.5 Postprocessing

Som det tidligere blev nævnt, har der ikke været fokuseret på postprocessing-delen i denne opgave. Derfor nøjes jeg med blot at afprøve implementeringens meget simple postprocessing ved anvendelse af de parametre, som også blev anvendt ved afprøvningen af preprocessing-delen. Der er ikke de store forventninger til, at den foreslåede postprocessing vil forbedre resultaterne væsentligt, for idéen med implementeringens postprocessing er blot, at meloditoner bliver fastholdt, selv om de hen mod slutningen blive svagere. Som det ses af resultaterne i tabellen herunder, har postprocessingen generelt set kun en ringe en positiv effekt.

Modifieret SHS med styrke — postprocessing $b = 0,95$ $u_- = 0,3$ $u_+ = 0,1$ $c = 1$ $\rho = 2$ $h = 0$ OQSTFT			
	TEST1	ADC2004	gennemsnit
Pitch i %	82,1	83,8	83,0
Chroma i %	84,4	87,7	86,1

6.2.6 Konklusion

Testresultaterne viste, at inkluderingen af det normaliserede spektrum gav en væsentlig forbedring set i forhold til den almindelige SHS. Derudover viste det sig, at det ikke havde stor betydning, om styrken eller amplituden blev anvendt i forbindelse med den modificerede SHS. Men når man ser på de gennemsnitlige resultater for de 2 datasæt, opnåedes de bedste resultater ved anvendelse af styrken.

Hvad angår valg af tid/frekvens-repræsentation, så viste det sig at være en stor fordel at bruge OQSTFT i stedet for STFT i forbindelse med den almindelige SHS, men i forbindelse med den modificerede SHS havde OQSTFT ikke de samme tydelige fordele. I tilfældet med datasæt ADC2004 og brug af amplituden viste det sig endda, at det var en fordel at anvende STFT. Men når man betragter begge datasæt, var det generelt en fordel at bruge OQSTFT. Hypotesen om, at anvendelsen af OQSTFT generelt ville være fordelagtig i forhold til den modificerede SHS, er altså bekræftet. Men man skal huske på, at der kun er tale om en lille fordel, og at der kun har været testet på et lille antal test-filer.

Hvad angår den udglattende faktor, så viste den sig at være fordelagtig at anvende. I forbindelse med anvendelsen af amplituden opnåedes kun lidt bedre resultater, når den udglattende faktor tages med, men i forbindelse med anvendelsen af styrken opnåedes en del bedre resultater, når den udglattende faktor tages med. Som forventet er det især Raw Pitch, som nyder godt af den udglattende faktor.

De 2 former for preprocessing, som blev afprøvet viste sig ikke at have positive effekter på resultaterne, men der var sådan set heller ikke store forventninger

til, at de 2 former for preprocessing skulle forbedre resultaterne. Forsøgene med de 2 former for preprocessing kan snarere set som eksperimenter.

Den simple postprocessing var der heller ikke store forventninger til. Der var blot tale om et simpelt forsøg på at fastholde melodi-tonen lidt længere, end den ellers ville have været fastholdt. Problemet er, at der også er risiko for, at ikke-melodi-toner bliver fastholdt længere. De gennemsnitlige resultater for de 2 datasæt viste kun en svag positiv effekt ved brugen af den implementerede postprocessingen.

6.3 Overordnet vurdering af resultater

På trods af, at autocorrelation som udgangspunkt blev opfattet som en meget simpel metode og derfor tjente som et godt mindstemål ved vurdering af mere komplicerede metoder, så viste det sig at autocorrelation i forbindelse med ADC2004 opnåede gode resultater. Det blev konkluderet i resultatanalysen for autocorrelation, cepstrum og generaliseringen, at metoderne viste sig at være konkurrencedygtige i forhold til en del af de deltagende metoder ved MIREX-konkurrencerne.

I forbindelse med ADC2004 opnåede den modificerede SHS væsentligt bedre resultater end autocorrelation og formåede dermed at overgå det fastsatte mindstemål. Ved test af den modificerede SHS i forbindelse med de 2 datasæt, blev det erfaret, at de mest optimale parametre for hvert af de 2 datasæt ikke adskilte sig væsentligt fra hinanden, forstået således, at de parametre, som gav gode resultater for det ene datasæt også gav gode resultater for det andet datasæt.

Deltagerne ved MIREX-konkurrencerne kunne anvende begge datasæt under udarbejdelse af deres metoder. Da de forvalgte parameter-værdier for den modificerede SHS er valgt således, at der opnås gode resultater for både ADC2004 og TEST1, er det også disse parametre, som ville have været valgt, hvis den modificerede SHS deltog i en MIREX-konkurrence. Derfor kan man sige, at de deltagende metoder ved MIREX-konkurrencerne og denne opgaves forslag til en metode har lige gode vilkår, og derfor er det rimeligt at foretage en sammenligning af resultaterne.

I tabel 6.5 er resultaterne for de deltagende metoder ved MIREX-2006 og MIREX-2008 oplistet sammen med resultaterne for den modificerede SHS med postprocessing. Det ses, at den modificerede SHS indtager en 3. plads i Raw Pitch score, mens den modificerede SHS indtager 1. pladsen i Raw Chroma. Det ses desuden, at flere af de deltagende metoder ved MIREX-konkurrencerne opnåede Raw Pitch og Raw Chroma scorere, som er næsten ens. Dette er ikke tilfældet med den modificerede SHS. Så noget tyder på, at den modificerede SHS set i forhold til flere af de andre metoder, har store problemer med at bestemme, hvilken oktav en tone befinder sig i.

Det er interessant at bemærke, at den modificerede SHS er inspireret af den spektrale normaliseringsfaktor, som blev anvendt i [5] og [6], og forfatterne til disse artikler deltog ved MIREX-2008 med en metode, som netop anvendte den spektrale normaliseringsfaktor. Den omtalte metode er noteret i tabel 6.5 som "MIREX-2008 Cao, Li, Liu, Yan". Mens denne metode klarede sig dårligst af de deltagende metoder ved MIREX-2008, så klarer den modificerede SHS sig altså væsentligt meget bedre.

ADC2004 Raw Pitch i %	
MIREX-2008 Durrieu, Richard, David	85,7
MIREX-2008 Cancela	85,1
Modificeret SHS	83,8
MIREX-2006 Dressler	82,9
MIREX-2008 Ryynanen, Klapuri	82,4
MIREX-2006 Ryynanen, Klapuri	80,6
MIREX-2008 Rao, Rao	77,1
MIREX-2008 Cao, Li, Liu, Yan	75,3
MIREX-2006 Poliner, Ellis	73,2
MIREX-2006 Sutton, Vincent, Plumbley, Bello	62,6
MIREX-2006 Brossier	57,4

ADC2004 Raw Chroma i %	
Modificeret SHS	87,7
MIREX-2008 Durrieu, Richard, David	86,2
MIREX-2008 Cancela	85,9
MIREX-2008 Rao, Rao	85,2
MIREX-2006 Dressler	84,0
MIREX-2008 Ryynanen, Klapuri	83,5
MIREX-2006 Ryynanen, Klapuri	82,3
MIREX-2008 Cao, Li, Liu, Yan	76,7
MIREX-2006 Poliner, Ellis	76,4
MIREX-2006 Brossier	68,7
MIREX-2006 Sutton, Vincent, Plumbley, Bello	65,4

Tabel 6.5: For datasættet ADC2004 vises her Raw Pitch og Raw Chroma scorer for de metoder, som deltog ved MIREX-2006 og MIREX-2008. Desuden noteres resultaterne for den modificerede SHS. Det ses, at den modificerede SHS indtager en 3. plads, hvad angår Raw Pitch, og hvad angår Raw Chroma, så indtager den modificerede SHS en 1. plads.

Det ses desuden, at flere af de metoder, som har en høj Raw Chroma score også har en næsten lige så høj Raw Pitch score. Dette er ikke tilfældet med den modificerede SHS.

Kapitel 7

Vurdering af den implementerede metode

Her vil vi ganske kort pointere nogle problemer ved den implementerede metode og præsentere idéer til forbedringer.

7.1 Problemer

Selv om den modificerede SHS i forhold til de deltagende metoder ved MIREX-konkurrencerne har opnået et godt resultat, er metoden på ingen måde problemfri.

Et generelt problem ved metoden er, at den ikke er god til at bestemme, hvilken oktav en tone befinder sig i. Ikke nok med at der opstår oktavfejl, således at der bliver gættet på en frekvens, som er den halve eller den dobbelte af fundamentalfrekvensen, men der bliver eksempelvis også gættet på frekvenser, som er 3 gange højere end fundamentalfrekvensen, fordi den 2. overtone måske er meget kraftig.

En form for postprocessing, hvor høje fundamentalfrekvenser vægtes mere favorabelt, er problematisk, fordi der er forøget risiko for, at en overtones frekvens bliver valgt. Dette ville være mindre problematisk, hvis der blev givet et mere sikkert svar på, hvilke fundamentalfrekvenser et signal indeholder.

Adskillige af de mange testscenarier, som blev gennemgået i afsnit 6.2 er blevet undersøgt nærmere. Det viste sig, at resultaterne for visse test-filer berøres i begrænset grad, når metodens parametre ændres, mens resultaterne for andre testfiler i høj grad berøres. Det, som er det store problem, er, at hvis parametrene sættes således, at der opnås gode resultater for én problematisk test-fil, så er der en anden problematisk test-fil, som opnår dårlige resultater.

Generelt handler det om, i hvor høj grad overtonerne vægtes. I nogle test-filer er melodi-tonen meget overtone-fattig, mens akkompagnementet er meget overtone-rigt; og i andre test-filer er det snarere omvendt. For at illustrere problemet med valget af parametre, er der i tabel 7.1 oplistet resultater for ADC2004, hvor der både opgives resultater for de parametre, som endeligt blev valgt, og derudover oplistes resultater for de parametre, der viste sig at være mest optimale for hver enkelt test-fil. Der er kun blevet undersøgt parametervalg i forhold til den modificerede SHS, som anvender styrken, og i alle tilfælde har postprocessing ikke været anvendt.

Af tabel 7.1 ses det, at der specielt i tilfældene med "midi1" og "midi3" opnås dårlige resultater med det endelige parametervalg.

ADC2004 Raw Pitch i %		
test-fil	bedste parametre	valgte parametre
daisy1	92,1	88,1
daisy2	91,8	90,3
daisy3	98,7	91,1
daisy4	98,7	96,8
jazz1	85,2	82,3
jazz2	81,5	80,5
jazz3	91,1	86,7
jazz4	88,5	85,0
midi1	78,7	67,2
midi2	95,9	91,2
midi3	83,7	68,0
midi4	89,4	83,6
opera_fem2	78,9	75,3
opera_fem4	84,4	82,3
opera_male3	73,8	70,1
opera_male5	91,9	89,7
pop1	83,9	80,9
pop2	88,8	85,7
pop3	85,3	80,0
pop4	91,8	89,1
gennemsnit	87,7	83,2

Tabel 7.1: For datasættet ADC2004 vises her Raw Pitch scorer for den modificerede SHS uden postprocessing. Der oplystes både scorer opnået med det parametervalg, som viste sig at være bedst for hver enkelt test-fil, og scorer for det endelige parametervalg. Specielt for 2 testfiler, "midi1" og "midi3", opnås dårlige resultater med det endelige parametervalg.

I forhold til “midi1” ville det være en fordel, hvis høje fundamentalfrekvenser blev tildelt større vægt, fordi melodien ligger højere end akkompagnementet. Dette ville dog være et problem for resultaterne for mange af de andre filer, fordi en højere vægt til de høje fundamentalfrekvenser ville føre til mange fejl, hvor der gættes på en frekvens, som egentlig tilhører melodi-tonens 2. overtone.

I forhold til “midi3” ville det være en fordel at vægte overtonerne så lidt som muligt, da melodi-tonen her primært indeholder en kraftig fundamental-frekvens. Dette ville dog skade resultaterne for mange af de andre filer. Det blev faktisk vist i [66, tabel 6,1, kapitel 6], at hvis man blot gætter på, at den kraftigste frekvens er melodiens fundamentalfrekvens, så kan der i forhold til “midi3” opnås en Raw Chroma score på 87%. Hvad der ikke blev vist i [66] er, at der på denne måde også kan opnås en Raw Pitch score på 87%. Hvis man ser på de scorer, som hver enkelt af de deltagende metoder ved MIREX-konkurrencerne opnåede, så er der faktisk kun én af metoderne nemlig “Durrieu, Richard, David”, som opnåede en score på 87%. Alle øvrige metoder opnåede en lavere score, som i de fleste tilfælde var betydeligt lavere.

7.2 Forslag til forbedringer

I forrige afsnit blev problemerne med oktavfejl nævnt. Ved hjælp af den udglattende faktor er en del af problemerne afhjulpet, således at der er mindre tendens til, at for dybe oktaver bliver valgt. Dette hjælper dog ikke så meget på, at der stadig kan blive valgt en overtones frekvens. Man kunne eventuelt tildele hver frekvens en form for straf, hvis frekvensen ser ud til at være en overtone i stedet for en fundamentalfrekvens. Dette ville eksempelvis være tilfældet, hvis frekvensen, som har den halve frekvens, også får tildelt stor vægt af den modificerede SHS, som anvender den udglattende faktor.

En anden måde, hvorpå oktavfejl muligvis kan elimineres, kunne inkludere anvendelsen af TWM-metoden¹, som burde være god til at bestemme en tones oktav. Det kunne være interessant at undersøge, hvordan denne metode kunne anvendes i forhold til den modificerede SHS. Måske kunne man beregne TWM for de formodede fundamentalfrekvenser, som har fået tildelt højest vægt, og så tildele hver fundamentalfrekvens en straf ved hjælp af TWM. Den fundamentalfrekvens, som derefter har den højeste vægt, kan da udvælges som buddet på melodiens fundamentalfrekvens.

I forrige afsnit blev der også fokuseret på problemerne med enten at favorisere en overtonerig eller overtonefattig tone. Der er brug for bedre viden om, hvorfor mennesker finder nogle toner mere fremtrædende end andre. Eksempelvis kunne man forestille sig, at mennesker mest fokuserer på enkelte kraftige frekvenser i det lave frekvensområde, mens der måske er mere fokus på overtonemønstre i de høje frekvensområder. Hvis dette er tilfældet, kan det måske være en fordel at have 2 mål for vægten af hver fundamentalfrekvens, således at der både anvendes et mål, som er rettet mod overtonefattige toner, og et mål, som er rettet mod overtonerige toner.

Det ville generelt være gavnligt med mere viden om, hvordan mennesker adskiller lyde. Eksempelvis ville det være interessant at undersøge, i hvor høj grad tidsmæssige ændringer af tonehøjden påvirker menneskers evne til at adskille lyde fra forskellige instrumenter. I forhold til den modificerede SHS

¹Der henvises til afsnit 4.4.1.

ville det måske være en fordel at inddrage det tidsmæssige aspekt. Hvis et overtonemønster, registreret til ét tidspunkt, kan genfindes ved et efterfølgende tidspunkt, og mønsteret frekvensmæssigt set er forskubbet lidt som følge af en lille ændring af tonehøjden, så vil det måske være lettere at bestemme, netop hvilke frekvenskomponenter der hører til den aktuelle tone, og derved vil det være lettere at bestemme fundamentalfrekvensen.

Da det ofte er et menneske, som synger melodien, kunne det måske være en fordel at tildele større vægt til frekvenser, som antages at være fundamentalfrekvenser for en menneskelig stemme. Man kunne eventuelt anvende de mål, som blev anvendt i [17]. Her blev en indikator for vibrato blandt andet anvendt.

Da den modificerede SHS ser ud til generelt at give gode resultater, kunne det være interessant at undersøge muligheden for at anvende den modificerede SHS som en egenskabsindikator til brug i metoder, som er baseret på statistisk mønstergenkendelse.

I afsnit 6.2.6 blev det blandt andet konkluderet at det generelt var en fordel at anvende OQSTFT. Det kunne dog stadig være interessant at undersøge muligheden for at anvende en anden tid/frekvens-repræsentation. Specielt kunne det være interessant at undersøge effekten af at bruge Matching Pursuit til generering af en tid/frekvens-repræsentation, da det kunne tænkes, at den derved genererede tid/frekvens-repræsentation bedre kunne gengive det centrale indhold i signalet.

Kapitel 8

Konklusion

Der blev i kapitel 3 præsenteret en lang række metoder for generering af tid/frekvens-repræsentationer — nogle metoder mere komplicerede end andre.

Chirplet-transformationen, som er en kompliceret metode, giver mulighed for at generere en tid/frekvens-repræsentation, som er tilpasset indholdet i et signal, men det kan være svært at generere denne tid/frekvens-repræsentation, fordi der kræves søgninger i et multidimensionelt rum. Det blev nævnt, at en mulig løsning på dette problem er anvendelsen af Matching Pursuit.

Vi så, at constant-Q-transformationen og MWT næsten er ens, og at de anvender en logaritmisk frekvensakse ligesom den musikalske skala. Vi så, at båndbredden af de vinduer, som anvendes i de 2 transformationer, er lineært afhængig af frekvensen, og vi så, at dette er problematisk i forhold til en vibrerende harmonisk tone.

STFT udviste også problemer i forhold til en vibrerende harmonisk tone. Som en simpel løsning på problemet med at generere en god tid/frekvens-repræsentation af en vibrerende harmonisk tone, blev OQSTFT foreslået. OQSTFT kan udregnes effektivt, og det blev antaget, at OQSTFT ville være fordelagtig at anvende i forbindelse med en metode for bestemmelse af melodiens fundamentalfrekvens.

I forhold til de mange vinduesfunktioner, som blev behandlet i forbindelse med STFT, blev der fremvist problemer ved anvendelsen af Gauss-vinduet og Hamming-vinduet. Specielt Hann-vinduet havde i modsætning til de andre nævnte vinduer ikke store problemer med, at energien fra én kraftig frekvens spredtes til frekvenser langt fra den kraftige frekvens.

I kapitel 4 blev autocorrelation og cepstrum samt en generalisering præsenteret. Derudover blev SHS præsenteret, og et par modifikationer til SHS blev foreslået. Den spektrale normaliseringsfaktor skulle give et mere informativt billede af spektrummet ved at beskrive, hvor meget styrken af hver frekvens adskiller sig fra de omgivende frekvensers styrker. Den udglattende faktor skulle give bedre mulighed for at vægte de høje overtoner højt, uden at der samtidig opstår mange oktav-fejl.

I kapitel 6 blev en implementering af autocorrelation, cepstrum og generaliseringen afprøvet på det datasæt, som også blev anvendt ved MIREX-konkurrencerne. I forhold til autocorrelation og de parametre, som blev anvendt i forbindelse med generaliseringen, viste det sig, at der blev opnået gode resultater ved anvendelse af OQSTFT. Det viste sig desuden at generaliseringen opnåede den bedste Raw Chroma score af de 3 afprøvede metoder.

Ved afprøvningen af den almindelige SHS viste det sig at være en stor fordel,

at anvende OQSTFT. Ved afprøvningen af den modificerede SHS viste det sig, at den spektrale normaliseringsfaktor gav store forbedringer i forhold til den almindelige SHS, og det viste sig også, at den udglattende faktor generelt set havde en positiv effekt. Det viste sig desuden, at anvendelsen af OQSTFT i forbindelse med den modificerede SHS generelt set var en fordel.

Test-resultaterne viste, at den implementerede metode set i forhold til de metoder, som deltog ved MIREX-konkurrencerne, gav gode resultater. I forhold til det anvendte datasæt viste det sig, at kun 2 metoder fra MIREX-konkurrencerne opnåede højere Raw Pitch score end den implementerede metode, og i forhold til Raw Chroma score var der ingen metoder ved MIREX-konkurrencerne, som opnåede lige så høj score som den implementerede metode.

I forhold til de metoder, som deltog ved MIREX-konkurrencerne, er den implementerede metode meget succesfuld, fordi den har opnået en høj score, og samtidig er den forholdsvis simpel. Der er dog en del problemer ved metoden, og derfor blev der i kapitel 7 foreslået en del ændringer og tilføjelser til metoden. Det blev nævnt, at det eksempelvis kunne være interessant at undersøge, om tilføjelsen af TWM kan mindske problemerne med oktav-fejl. Det kunne også være interessant at undersøge, om en anden tid/frekvens-repræsentation er mere fordelagtig end den, der blev genereret med OQSTFT. Man kunne eksempelvis undersøge muligheden for at anvende Matching Pursuit.

Den modificerede SHS, som blev præsenteret i denne opgave, er et godt bud på en metode for bestemmelse af melodiens fundamentalfrekvens, men som det også blev nævnt i kapitel 7, kan den formentlig også med fordel anvendes som udgangspunkt for en statistisk baseret metode.

Selv om den implementerede metode viste sig at give gode resultater set i forhold til en del andre nyere metoder, så er problemet med at estimere melodiens fundamentalfrekvens langt fra løst. Forhåbentlig kan flere studier af den menneskelige evne til at opfatte lyde føre til dannelse af nye metoder eller til forbedringer af eksisterende metoder.

Litteratur

- [1] J. W. Beauchamp, "Analysis and synthesis of musical instrument sounds," in *Analysis, Synthesis, and Perception of Musical Sounds*, J. W. Beauchamp, Ed. Springer, 2007, pp. 1–89, <http://books.google.com/books?id=wIZt9rZ27i4C>.
- [2] J. C. Brown, "Calculation of a constant q spectral transform," *Journal of Acoustic of Society of America*, vol. 89(1), pp. 425–434, 1991, <http://www.wellesley.edu/Physics/brown/pubs/cq1stPaper.pdf>.
- [3] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant q transform," *Journal of Acoustic of Society of America*, vol. 92(5), pp. 2698–2701, 1992, <http://www.wellesley.edu/Physics/brown/pubs/effalgV92P2698-P2701.pdf>.
- [4] R. Büssow, "An algorithm for the continuous morlet wavelet transform," *Mechanical Systems and Signal Processing*, vol. 21, no. 8, pp. 2970–2979, 2007.
- [5] C. Cao, M. Li, J. Liu, and Y. Yan, "Multiple f_0 estimation in polyphonic music (mirex 2007)," Thinkit Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, Tech. Rep., 2007, http://www.music-ir.org/mirex/2007/abs/F0_cao.pdf.
- [6] —, "Singing melody extraction in polyphonic music by hamonic tracking," Thinkit Speech Lab., Institute of Acoustics, Chinese Academy of Sciences, Tech. Rep., 2007, http://ismir2007.ismir.net/proceedings/ISMIR2007_p373_cao.pdf.
- [7] E. Chu, *Discrete and Continuous Fourier Transforms - Analysis, Applications and Fast Algorithms*. CRC Press, 2008.
- [8] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
- [9] R. P. P. de Carvalho e Paiva, "Melody detection in polyphonic audio," Ph.D. dissertation, 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=.
- [10] A. de Cheveigné and H. 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.
- [11] K. Dressler, "An auditory streaming approach on melody extraction," Fraunhofer Institute for Digital Media Technology, Tech. Rep., 2006, http://www.music-ir.org/evaluation/MIREX/2006_abstracts/AME_dressler.pdf.
- [12] —, "Sinusoidal extraction using an efficient implementation of a multi-resolution fft," Fraunhofer Institute for Digital Media Technology, Tech. Rep., 2006, http://www.dafx.ca/proceedings/papers/p_247.pdf.
- [13] B. M. E. Gómez, A. Klapuri, "Melody description and extraction in the

- context of music content processing," *Journal of New Music Research*, vol. 32(1), 2003, <http://www.cs.tut.fi/sgn/arg/klap/gomez.pdf>.
- [14] R. H. Ehmer, "Masking by tones vs noise bands," *The Journal of the Acoustical Society of America*, vol. 31, no. 9, pp. 1253–1256, 1959.
- [15] —, "Masking patterns of tones," *The Journal of the Acoustical Society of America*, vol. 31, no. 8, pp. 1115–1120, 1959.
- [16] M. Farge, "Wavelet transforms and their applications to turbulence," *Annual Review of Fluid Mechanics*, vol. 24, pp. 395–458, 1992, http://wavelets.ens.fr/ENSEIGNEMENT/COURS/UCSB/farge_ann_rev_1992.pdf.
- [17] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 253–256, 2006, <http://staff.aist.go.jp/m.goto/PAPER/ICASSP2006fujihara.pdf>.
- [18] D. Gabor, "Theory of communication," *Journ. of Inst. Electrical Engineers*, vol. 93, part III, no. 26, pp. 429–457, 1946, <http://wearcam.org/chirplet/gabor1946.pdf>.
- [19] M. Goto and S. 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*, pp. 31–40, 1999, <http://staff.aist.go.jp/m.goto/PAPER/CASA99.pdf>.
- [20] J. M. Grey, "An exploration of musical timbre," Ph.D. dissertation, Department of Music Stanford University, 1975, <http://ccrma.stanford.edu/STANM/stanms/stanm2/stanm2.pdf>.
- [21] —, "Multidimensional perceptual scaling of musical timbres," *The Journal of the Acoustical Society of America*, vol. 61, no. 5, pp. 1270–1277, 1977.
- [22] R. Gribonval, P. Depalle, X. Rodet, E. Bacry, and S. Mallat, "Sound signals decomposition using a high resolution matching pursuit," 1996, <http://www.irisa.fr/metiss/gribonval/Conf/1996/icmc96.pdf>.
- [23] A. Haar, "Zur theorie der orthogonalen funktionensysteme," *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1910, <http://www.springerlink.com/content/v1n8787g737j5k64/fulltext.pdf>.
- [24] J. K. Hammond and P. R. White, "The analysis of non-stationary signals using time-frequency methods," *Journal of Sound and Vibration*, vol. 190, no. 3, pp. 419–447, 1996.
- [25] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978, <http://www.utdallas.edu/~cpb021000/EE4361/GreatDSPPapers/HarrisonWindows.pdf> and <http://web.mit.edu/xiphmont/Public/windows.pdf>.
- [26] H. Hermansky, N. Morgan, and H. Hirsch, "Recognition of speech in additive and convolutional noise based on rasta spectral processing," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 83–86, 1993.
- [27] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994, <http://lectures.idiap.ch/winter2005-2006/ic-48/courseslide/rasta.pdf>.
- [28] D. J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of Acoustic of Society of America*, vol. 83, pp. 257–264, 1988.

- [29] F. Hlawatsch, "Interference terms in the wigner distribution," *International Conference of Digital Signal Processing, Florence, Italy, September 5-8, 1984*, http://publik.tuwien.ac.at/files/pub-et_12538.pdf.
- [30] F. Hlawatsch and P. Flandrin, *The Wigner Distribution — Theory and Applications in Signal Processing*. Amsterdam: Elsevier, 1997, pp. 59–133, http://publik.tuwien.ac.at/files/pub-et_3288.pdf.
- [31] F. Hlawatsch, T. G. Manickam, R. L. Urbanke, and W. Jones, "Smoothed pseudo-wigner distribution, choi-williams distribution, and cone-kernel representation: Ambiguity-domain analysis and experimental comparison," *Signal Processing*, vol. 43, no. 2, pp. 149–168, 1995.
- [32] B. Jawerth and W. Sweldens, "An overview of wavelet based multiresolution analyses," *SIAM Review*, vol. 36, no. 3, pp. 377–412, 1994.
- [33] K. Jensen, "Timbre models of musical sounds," Ph.D. dissertation, Datalogisk Institut Københavns Universitet, 1999, <http://www.aaue.dk/~krist/TMoMS.pdf>.
- [34] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 1999, pp. 929–932, <http://eprints.kfupm.edu.sa/52349/1/52349.pdf>.
- [35] R. Keren, Y. Y. Zeevi, and D. Chazan, "Multiresolution time-frequency analysis of polyphonic music," in *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, 1998, pp. 565–568.
- [36] Klapuri, Virtanen, Eronen, and Seppänen, "Automatic transcription of musical recordings," in *Proceedings of the Consistent & Reliable Acoustic Cues Workshop*, 2001, <http://www.cs.tut.fi/sgn/arg/klap/crac2001/crac2001.pdf>.
- [37] A. Klapuri, "Wide-band pitch estimation for natural sound sources with inharmonicities," in *106th Audio Engineering Society Convention, Munich, Germany*, 1999, <http://www.cs.tut.fi/sgn/arg/music/munchen99.ps>.
- [38] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," in *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, 2003, pp. 804–816, <http://www.cs.tut.fi/sgn/arg/klap/multiplef0.pdf>.
- [39] —, "Signal processing methods for the automatic transcription of music," Ph.D. dissertation, Tampere University of Technology, Finland, 2004, http://www.cs.tut.fi/sgn/arg/klap/phd/klap_phd.pdf.
- [40] P. J. Kootsookos, B. C. Lovell, and B. Boashash, "A unified approach to the stft, tfds and instantaneous frequency," *IEEE Transactions on Signal Processing*, vol. 40, no. 8, pp. 1971–1982, 1992, <http://espace.library.uq.edu.au/eserv/UQ:11271/n00149998.pdf>.
- [41] R. Kronland-Martinet, "The wavelet transform for analysis, synthesis, and processing of speech and music sounds," *Computer Music Journal*, vol. 12, no. 4, pp. 11–20, 1988.
- [42] N. Kunieda, T. Shimamura, and J. Suzuki, "Robust method of measurement of fundamental frequency by aclos: autocorrelation of log spectrum," in *IEEE International Conference Proceedings of the Acoustics, Speech, and Signal Processing*, vol. 1, 1996, pp. 232 – 235.
- [43] H. Kuttig, M. Niethammer, S. Hurlebaus, and L. J. Jacobs, "Model-based analysis of dispersion curves," *Journal of the Acoustical Society of America*, vol. 119, pp. 2122–2130, 2006, <http://www.cs.unc.edu/~mn/pubs/pdfs/>

- kuttig2006_JASA_model_based_analysis_of_dispersion_curves_using_chirplets.pdf.
- [44] M. K  peri and L. Weruaga, "Adaptive chirp-based time-frequency analysis of speech signals," *Speech communication*, no. 48, pp. 474–492, 2006.
 - [45] M. Lahat, R. Niederjohn, and D. Krubsack, "A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 6, pp. 741–750, 1987.
 - [46] W. C. Lang and K. Forinash, "Time-frequency analysis with the continuous wavelet transform," *American Journal of Physics*, vol. 66, no. 9, pp. 794–797, 1998, <http://physics.ius.edu/~kyle/K/pdf/timefreq.pdf>.
 - [47] A. Lukin and J. Todd, "Adaptive time-frequency resolution for analysis and processing of audio," *Audio Engineering Society Convension Paper, 120th Convention*, 2006, <http://graphics.cs.msu.su/en/publications/text/LukinTodd.pdf>.
 - [48] R. C. Maher, "Evaluation of a method for separating digitized duet signals," *Journal of the Audio Engineering Society*, vol. 38, no. 12, pp. 956–979, 1990, http://www.coe.montana.edu/ee/rmaher/publications/maher_jaes_1290_956-979.pdf.
 - [49] R. C. Maher and J. W. Beauchamp, "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254–2263, 1994, http://www.coe.montana.edu/ee/rmaher/publications/maher_jasa_0494_2254-2263.pdf.
 - [50] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993, <http://www.cmap.polytechnique.fr/~mallat/papiers/MallatPursuit93.pdf>.
 - [51] S. Mann and S. Haykin, "The chirplet transform: A generalization of gabor's logon transform," *Proceedings Vision Interface 1991*, pp. 205–212, June 1991, <http://wearcam.org/chirplet/vi91/index.htm> and <http://eyetap.org/research/wearables/wearcomp/oldchirplet/>.
 - [52] —, "Adaptive chirplet transform: an adaptive generalization of the wavelet transform," *Optical Engineering*, vol. 31, no. 6, pp. 1243–1256, 1992, http://wearcam.org/chirplet/adaptive_chirplet1992/adaptive_chirplet_transform_1992.pdf.
 - [53] R. Meddis and L. O'Mard, "A unitary model of pitch perception," *The Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1811–1820, 1997, [http://www.sx.ac.uk/psychology/psy/PEOPLE/meddis/Meddis_OMard%201997\(Pitch\).pdf](http://www.sx.ac.uk/psychology/psy/PEOPLE/meddis/Meddis_OMard%201997(Pitch).pdf).
 - [54] A. Mesaros, E. Lupu, and C. Rusu, "Singing voice features by time-frequency representations," in *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis*, 2003, vol. 1, 2003, pp. 471–475.
 - [55] B. C. J. Moore, J. I. Alc  ntara, and T. Dau, "Masking patterns for sinusoidal and narrow-band noise maskers," *The Journal of the Acoustical Society of America*, vol. 104, no. 2, pp. 1023–1038, 1998, http://server.oersted.dtu.dk/personal/tda/.downloads/reviewed/moore_et_al_1998.pdf.
 - [56] J. A. Moorer, "The optimum comb method of pitch period analysis of continuous digitized speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 22, no. 5, pp. 330–338, 1974.

- [57] S.-C. Pei and I.-I. Yang, "Computing pseudo-wigner distribution by the fast hartley transform," *IEEE Transactions on Signal Processing*, vol. 40, no. 9, pp. 2346–2349, 1992, <http://ntur.lib.ntu.edu.tw/bitstream/246246/2007041910032270/1/00150139.pdf>.
- [58] M. Piszczalski and B. A. Galler, "Predicting musical pitch from component frequency ratios," *Journal of the Acoustical Society of America*, vol. 66(3), pp. 710–720, 1979.
- [59] W. H. Prosser and M. D. Seale, "Time-frequency analysis of the dispersion of lamb modes," *Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2669–2676, 1999, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.40.1716&rep=rep1&type=pdf>.
- [60] M. Ryyänänen and A. Klapuri, "Note event modeling for audio melody extraction," Tech. Rep., 2005, <http://www.music-ir.org/evaluation/mirex-results/articles/melody/ryynanen.pdf>.
- [61] H.-C. Shyu and Y.-S. Sun, "Construction of a morlet wavelet power spectrum," *Multidimensional Systems and Signal Processing*, vol. 13, no. 1, pp. 101–111, 2002.
- [62] L. M. Smith, "A multiresolution time-frequency analysis and interpretation of musical rhythm," Ph.D. dissertation, Department of Computer Science, University of Western Australia, 2000, <http://staff.science.uva.nl/~lsmith/Papers/MultiresRhythm.pdf>.
- [63] —, "Time-frequency representation of musical rhythm by continuous wavelets," *Journal of Mathematics and Music*, vol. 2, no. 2, p. (in press), 2008, http://emcap.iaa.upf.es/showEmcap/publications/MaM_SmithHoning.pdf.
- [64] K. Toraichi, M. Kamada, S. Itahashi, and R. Mori, "Window functions represented by b-spline functions," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 1, pp. 145–147, 1989.
- [65] M. Unser, "Fast gabor-like windowed fourier and continuous wavelet transforms," *IEEE Signal Processing Letters*, vol. 1, no. 5, pp. 76–79, 1994, <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=07845B9288DF6A2077609D3AAEEE18B1?doi=10.1.1.19.7500&rep=rep1&type=pdf>.
- [66] M. Wendelboe, "Udtræk af melodien fra et musikalsk nummer," Kandidatprojekt, Datalogisk Institut, Københavns Universitet, 2008, http://morwen.dk/university/melody-project/morten_wendelboe_udtraek_af_melodien.pdf.
- [67] E. Wilczok, "New uncertainty principles for the continuous gabor transform and the continuous wavelet transform," *Documenta Mathematica*, vol. 5, pp. 201–226, 2000, <http://www.math.uiuc.edu/documenta/vol-05/08.pdf>.
- [68] R. Zhou, "Feature extraction of musical content for automatic music transcription," Ph.D. dissertation, Chinese Academy of Science, Beijing, 2006, http://biblion.epfl.ch/EPFL/theses/2006/3638/EPFL_TH3638.pdf.