Kapitel 4

Ljudanalys

Det finns många olika sätt att analysera och representera signaler. Liksom man kan syntetisera samma signal med hjälp av ett antal olika syntesmodeller kan man analysera samma ljud genom att välja olika representationer av signalen. Men vilken är den korrekta eller bästa?

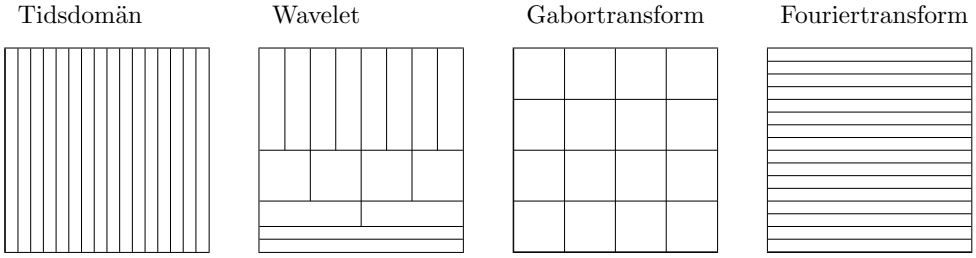
För att analysera perceptuella attribut i en signal behöver man också ta hänsyn till perceptionen av ljud. Analys med hjälp av signaldeskriptorer har flera viktiga användningsområden, bl.a. för sökning i musikdatabaser eller för att matcha två ljudfragment med varann. Det är också grunden för adaptiva effekter, dvs sådana som automatiskt varierar sina egna parametrar beroende på egenskaper i originalljudet och adaptiva syntesmodeller där man använder en ljudsignal för att kontrollera olika syntesparametrar.

4.1 Signalrepresentationer

Vid det här laget finns det många transformer att välja mellan om man vill analysera eller visualisera ljudsignaler. Fouriertransformen visar ett statiskt spektrum, lämpligt till analys av den stabila delen av en ton. Sonogrammet används när man vill visa hur spektrumet varierar över tid. De flesta vanliga transformerna delar så att säga in tid-frekvensplanet i ett raster eller gitter. Har man väl bestämt det rastret så har man bestämt signalrepresentationen och varje specifik signal har en bestämd transform. Men det finns också andra sätt att analysera signaler som adaptivt utgår från signalen själv och där representationen av signalen kan utföras på flera olika sätt. Ytterligare metoder för signalanalys kommer från ickelineär tidsserieanalys.

4.1.1 Tid-frekvensrepresentationer

Vi har redan använt några vanliga signalrepresentationer, nämligen signaler i tidsdomänen och i frekvensdomänen; både i form av amplitudspektrum och korttids-fouriertransformen. Andra möjligheter finns också, som bygger på andra transformer av signalen. Man kan tänka sig ett tid-frekvensplan, som man kan rastrera eller dela in i ett rutnät på flera olika sätt, som i figur 4.1. Tidsdomänen ger bäst upplösning i tid, men ingen upplysning om frekvens, medan fouriertransformen ger bäst upplösning i frekvens men ingen lokalisering i tid. Mellan dessa ytterligheter kan man konstruera flera slags indelningar. Gabortransformen ger till exempel en bästa upplösning i tid och frekvens sammantaget och fås av att använda



Figur 4.1: Tid-frekvensrepresentationer kan ses på som raster med tid på den horisontala axeln och frekvens på den vertikala.

komplexa exponentialfunktioner med ett gaussiskt fönster till att analysera signalen (den är alltså nära besläktad med korttids-fouriertransformen). Wavelets har god tidsupplösning vid höga frekvenser, och god frekvensupplösning men dålig tidsupplösning vid låga frekvenser. Även om man kan konstruera många fler sätt att indela tid-frekvensplanet så gäller en begränsning, nämligen att varje ruta i en sådan rastrering måste ha samma yta. Det har att göra med den fundamentala osäkerhetsrelationen mellan tid och frekvens, som säger att ju bättre precision man vill ha i frekvens, desto sämre precision får man i tid och omvänt. Å andra sidan kan man ha en reduntant representation, på så sätt att rutorna i tid-frekvensplanet överlappar varann.

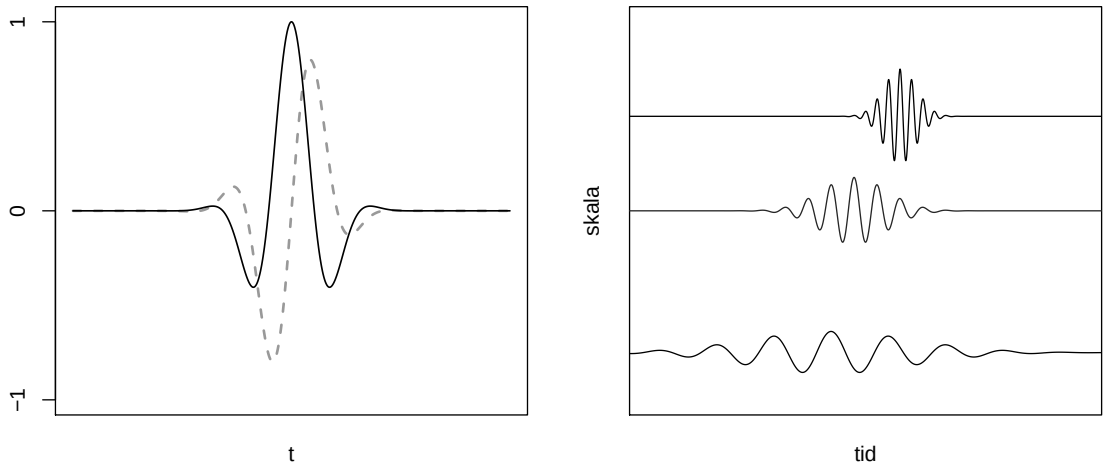
En bra representation av en signal är en som koncist sammanfattar viktiga egenskaper i signalen. Om man exempelvis har en signal som består av glest utspridda impulser separerade av tystnad är tidsdomänen antagligen ett gott alternativ. En signal som består av några få konstanta sinustoner kan däremot uttryckas koncist genom att ange deras amplitud, frekvens och fas, vilket man erhåller genom fouriertransformen. Komprimerade filformat som mp3 utnyttjar i princip samma idé om koncisa beskrivningar av signalen, nämligen genom att utföra en perceptuellt motiverad representation av signalen så att man kan reducera information som ändå inte hörs på grund av maskering.

Ett glissando är en signal som inte har en kompakt representation i något av fallen som visas i figur 4.1. Både gabortransformen och fouriertransformen utgår från sinusoider med konstant frekvens, och kan därför inte beskriva glissandot på ett ekonomiskt vis. Men det finns en speciell transform för den situationen också, som vi ska introducera senare.

4.1.2 Vågelement, konstant Q-transform

Eftersom örats frekvensupplösning på ett ungefär är proportionell mot frekvensen kan det vara befogat att använda en transform som analyserar ljudet med motsvarande indelning av spektrumet. Vågelement, även kallat krusning eller wavelet, är en basfunktion som kan användas till sådan analys. Frekvensindelningen följer då något musikaliskt intervall som oktaver eller terser.

I gabortransformen använder man en fönstrad sinusoid med olika position i tid och frekvens till att analysera signalen. Fönstret har i det fallet alltid samma längd. Vågelement



Figur 4.2: Morlets vågelement: till vänster realdelen (heldragen linje) och imaginärdelen (streckad linje) av basfunktionen, till höger visas flera vågelement vid olika tidpunkter och skala.

däremot har de två parametrarna tidpunkt och skala. Istället för att ha samma fönsterlängd så sträcks eller krymps den funktionen man analyserar signalen med. Man kan tänka sig det som en solfjäder eller en bälg som kan fällas ut eller tryckas ihop, men som alltid har lika många veck. En sak som gör analys med vågelement mera komplicerad men också mera anpassningsbar är att man står fritt att välja basfunktionen själv så länge den uppfyller vissa minimala krav.

Om $g(t)$ är en basfunktion för waveletanalys, så krävs det att den har ändlig energi, eller närmare bestämt att $\int |g(t)|\, dt < \infty$ och $\int |g(t)|^2\, dt < \infty$, samt att medelvärdet är noll, dvs $\int g(t)dt = 0$. Transformdomänen för vågelement är ett plan med koordinaterna tidsskift (d) och skala (s). Basfunktionerna fås ur moderfunktionen $g(t)$ genom tidsskift och ändring av skala:

$$g_{d,s}(t) = \frac{1}{\sqrt{s}} g\left(\frac{t-d}{s}\right)$$

En basfunktion som har använts i analys av ljud är Morlet-funktionen

$$g(t) = Ce^{-t^2/2}e^{i\omega_0 t} \tag{4.1}$$

uppkallad efter Jean Morlet, som är nära besläktad med basfunktion som används i gabortransformen (se figur 4.2). Eftersom det är en komplex sinusoid kan man dela upp transformdomänen i magnitud och fas precis som i fallet med fourieranalys.

Några användningsområden av waveletrepresentationen har föreslagits (Kronland Martinet, 1988): De kan användas för att transformera ljudet och modifiera waveletrepresentationen för att sedan göra en invers transform tillbaka till tidsdomänen. Den inversa transformen kan utföras som additiv syntes. Med hjälp av skräddarsydda vågelement kan man analysera ljud och finna förekomst av olika intervall mellan toner, som oktav eller tritonus. Basfunk-

tionen består då av summan av två sinusoider med det intervallet i frekvens som man vill framhäva i analysen.

En närbesläktad transform med konstant Q-faktor (CQT, *Constant-Q Transform*) har föreslagits (Brown & Puckette, 1992). Den utgår från en DFT av signalen, men fönsterlängden är omvänt proportionell mot analysfrekvensen. I princip är det samma sak som waveletanalys med Morlets basfunktion. Trots fördelarna i analys av musik har transformer med konstant Q-faktor inte blivit lika popluära som STFT. Det kan bero på att det är en mera beräkningskrävande operation än vanlig analys med DFT, att det inte finns en invers transform som gör det möjligt att transformera tillbaka till tidsdomänen utan förluster i ljudkvalitet, samt att datastrukturen är mera komplicerad att hantera eftersom man i praktiken har olika steglängd mellan analysfönstrena beroende på frekvens. Men förbättringar görs ständigt, som en nyligen föreslagen implementering av Schörkhuber & Klapuri (2010). De visar att det går att göra en redundant analys av signalen och på så sätt uppnå bättre kvalitet i rekonstruktionen av signalen genom en invers transform. En DFT innehåller precis lika mycket information som motsvarande segment i tidsdomänen, och är därför inte redundant. Genom att avsätta analysfönstrena tätare i tid och/eller frekvens i en CQT än vad som ser ut att vara nödvändigt på ett rutnät av liknande slag som i figur 4.1 får man en redundant representation. Kvaliteten av rekonstruktionen beror också på vilken fönsterfunktion som används. Det kommer sig av att när man summerar överlappande fönster vill man att summan ska vara konstant över tid.

4.1.3 Autokorrelation och wignerfördelning

Ett mått på hur lik en signal är sig själv vid en viss fördröjning är autokorrelationen. Den beräknas på ett sätt som påminner om att falta en signal med sig själv. Vid en viss fördröjning τ, är autokorrelationen integralen av signalen multiplicerad med sig själv förskjuten med tiden τ:

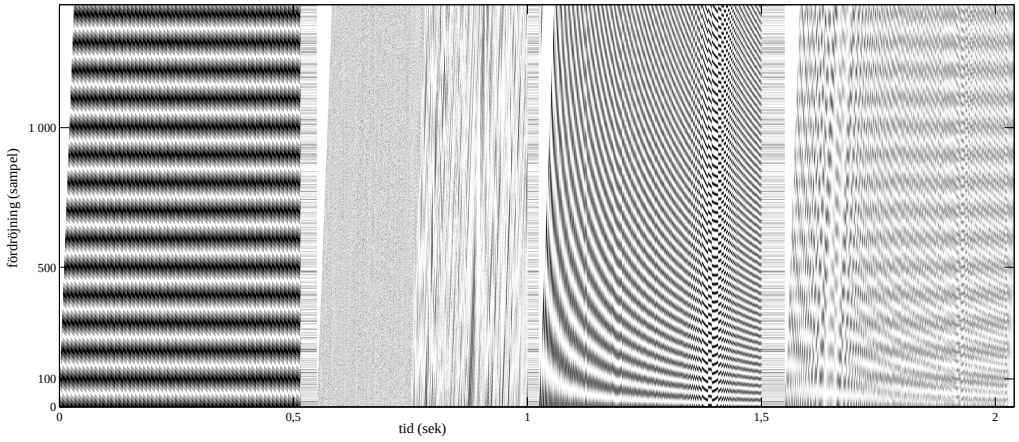
$$r_{xx}(\tau) = \int_{-\infty}^{\infty} x(t)x(t+\tau)dt$$

Autokorrelationen antar alltid sitt maximala värde vid $\tau = 0$, och den är en jämn funktion (symmetrisk runt $\tau = 0$), men man brukar bara se på den positiva halvan av τ-axeln. Om signalen är en sinusoid med godtycklig fas, $x(t) = \sin(\omega t + \phi)$, är dess autokorrelation $\cos(\omega t)$, dvs fasinformationen går förlorad, men frekvensinnehållet blir det samma. För diskreta signaler av längd N kan autokorrelationen beräknas antingen direkt enligt

$$r_{xx}(d) = \frac{1}{N-d} \sum_{n=0}^{N-1-d} x[n]x[n+d]$$

eller med hjälp av en fouriertransform. Signalens effekttäthetsspektrum (eng. *power spectrum*), definierat av $P(\omega) = |X(\omega)|^2$, kan erhållas genom en fouriertransform av autokorrelationsfunktionen. Omvänt kan man beräkna autokorrelationen ur spektrumet genom en invers fouriertransform,

$$r_{xx}(d) = \frac{1}{N} FFT^{-1}\{X(k)X^*(k)\}$$



Figur 4.3: Lokal autokorrelation av en testsignal. Graden av svärta motsvarar amplituden av $R_x(t,\tau)$. En periodicitet på 100 sampel kan ses under den första halva sekunden. Strax efter en sekund ses ett uppåtgående glissando.

där X^* står för det komplexkonjugerade spektrumet. Autokorrelationen kan användas till tonhöjdsanalys, vilket vi ska studera närmare senare.

Lokal autokorrelation, $R_{xx}(t,\tau)$, är ett sätt att mäta hur autokorrelationen förändras över tid:

$$R_{xx}(t,\tau) = x(t+\tau/2)x(t-\tau/2) \tag{4.2}$$

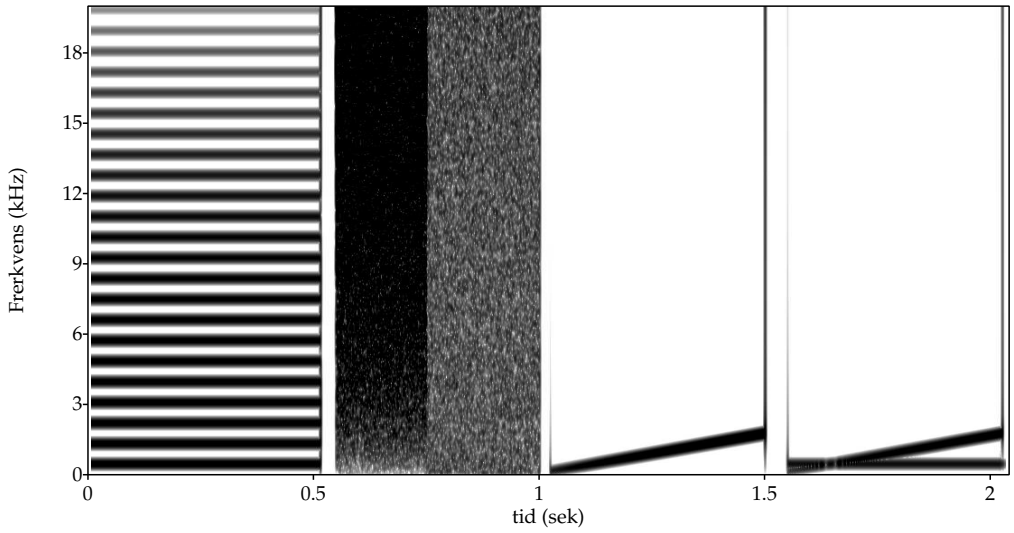
Även om den sällan används, kan man visa den lokala autokorrelationen ungefär som ett sonogram (se figur 4.3). Men den vertikala axeln har enheten av antal sampels fördröjning, så ett uppåtgående glissando ser här ut som ett nedåtgående svep. Sonogrammet av samma signal visas i figur 4.4.

Man kan göra en fouriertransform av den lokala autokorrelationen. I så fall får man *Wigner-Ville-fördelningen*, som är ännu en tid-frekvensrepresentation av signalen:

$$W(t,\omega) = \int_{-\infty}^{\infty} R_{xx}(t,\tau)e^{-i\omega\tau}d\tau$$

Autokorrelationen och den lokala autokorrelationen samt wignerfördelningen är tre exempel på kvadratiska funktioner av signalen, eftersom man i dessa fall multiplicerar signalen med sig själv. Signalen i tidsdomänen och fouriertransformen är däremot lineära funktioner av signalen. Man kan analysera signaler med wignerfördelningen på samma sätt som med korttids-fouriertransformen, men just detta faktum att det är en kvadratisk transform gör den olineär. Det innebär att man exempelvis kan få interferenser (intermodulation) mellan två sinustoner separerade både i tid och frekvens. Många försök att motverka dessa interferenser har gjorts (Hlawatsch & Boudreaux-Bartels, 1992).

Både fouriertransformen och vågelement är lineära transformer, i den meningen att för en signal $x(t)$ och dess transform $X(f,t)$ för frekvens och tid, eller $X(s,t)$ för skala och



Figur 4.4: Sonogram av en testsignal som består av: en fyrkantvåg med grundton 440 Hz, vitt brus, brownskt brus, en stigande sinuston, samma stigande ton mixad med en sinuston på 440 Hz.

tid uppfyller kriterierna för lineära system. Alltså gäller det att signalen multiplicerad med en konstant motsvarar transformen multiplicerad med samma konstant, $kx(t) \leftrightarrow kX(\cdot, t)$, och summan av två signaler motsvaras av summan av deras transformer, $x(t) + y(t) \leftrightarrow X(\cdot, t) + Y(\cdot, t)$. Det finns en mängd av kvadratiska transformer, men det faktum att de inte uppfyller dessa linearitetsegenskaper gör dem svårhanterliga.

Interferenserna kan vara en anledning till att wignerfördelningen och andra ickelineära transformer till nyligen har använts sparsamt för signalbehandling i musikaliska tillämpningar. Kling & Roads (2004) föreslog att använda wignerfördelningen på de separata elementen i en på annat vis dekomponerad signal (m.h.a. *matching pursuit*; se nedan), och sedan summera dessa wignertransformer, i syfte att visualisera signaler med god upplösning i tids- och frekvensdomänen på samma gång.

4.1.4 Spektral modellering

Fasvocodern är ett sätt att överkomma begränsningarna i korttidsfouriertransformen. Den passar bra för ljud med harmoniska spektrum. Idén är att modellera signalen som en summa av sinusoider med tidsvarierande amplitud och frekvens. Men istället för att ha ett fixerat antal deltoner i varje tidsfönster sätter man ett tröskelvärde för hur låg amplitud en delton ska kunna ha för att fortfarande räknas med. Varje delton som har upptäckts i ett tidsfönster försöks paras ihop med en närliggande frekvens i påföljande fönster. En sådan följd av aktiva sinustoner kallas ett spår (*track* på engelska; denna variant kallas ofta *tracking phase vocoder*). Om ett spår inte har en naturlig kandidat att fortsätta till i nästa tidsfönster avbryts spåret, och om en ny delton skulle uppstå på en frekvens som inte kan förbindas bakåt med ett aktivt spår föds ett nytt spår.

Det finurliga med fasvocodern är att sinustonerna kan variera i både amplitud och frekvens mellan två analysfönster, vilket gör det möjligt att uppfånga ett vibrato, glissando eller andra gradvisa förändringar i tonen. Amplituden interpoleras vanligen lineärt mellan två fönster. Frekvensen estimeras genom att beräkna fasen av sinustonen i de två tidpunkterna som svarar mot de två analysfönstrena. Fasvocodern används ofta till att sträcka ut ljud i tid utan att förändra tonhöjd, eller transponera det till andra tonhöjder utan att ändra hastighet. Resyntesen kan antingen utföras genom en invers fouriertransform eller direkt genom additiv syntes av sinustoner.

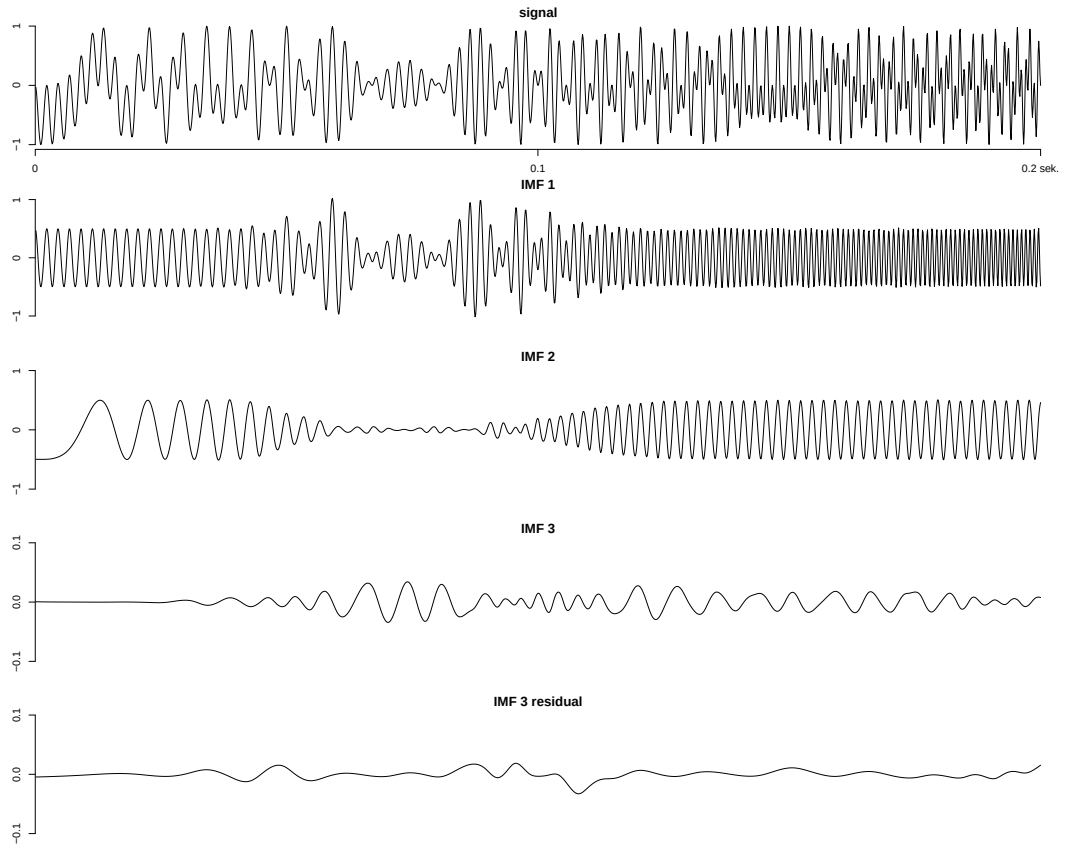
En begränsning med fasvocodern är att den inte är lämplig för brusiga ljud, eftersom det skulle behövas ett stort antal sinustoner för att komma i närheten av en trogen resyntes. Ett sätt att komma förbi den begränsningen är att införa en stokastisk signalkomponent, vilket man gör i SMS (*spectral modelling synthesis*). Analysen utförs först med fasvocodern, sedan resyntetiserar man den analyserade signalen som kallas den deterministiska (eller sinusoida) komponenten. Återstoden eller den stokastiska komponenten av signalen kan man då komma åt genom att subtrahera ut den deterministiska komponenten från originalsignalen. Om fasvocodern nämligen har identifierat en mängd tidsvarierande sinustoner med deras korrekta amplituder, frekvenser och faser, så behöver man bara invertera den deterministiska signalen och summera den med originalsignalen. Vid resyntes modelleras den stokastiska komponenten som en spektral kurva som bestämmer amplituden vid olika frekvenser, medan fasen vid varje frekvensband väljs slumpmässigt (Serra & Smith, 1990).

SMS har i sin tur en begränsning, nämligen att den inte är optimalt ägnad för att representera transienter. En transient är exakt lokaliserad i tid, men analyserad med SMS skulle den återges med färgat brus som varar lika länge som ett helt analysfönster, eller antagligen längre eftersom man bör använda överlappande fönster. Svaret på det problemet är att ta med transienterna i den spektrala modellen (Verma et al., 1997). TMS (*Transient Modelling Synthesis*) bygger alltså på SMS, men lägger till en extra analysfas där man extraherar transienter.

4.1.5 Adaptiva representationer: EMD

En helt annan strategi för signalanalys än de ovan beskrivna är att utgå från signalen och anpassa basfunktionerna efter den. Fördelen är att man kan komma fram till en koncis representation av signalen, vilket är användbart i datakomprimering, bland annat. Det finns många tillvägagångssätt, men det de har gemensamt är att de använder en algoritm som successivt dekomponerar signalen. I första steget finner algoritmen en grov approximation till signalen, som sparas och filtreras ut. Residualen, eller det som är kvar i signalen, analyseras på samma sätt, och man får successivt allt mindre residualer. Processen upprepas tills man når ett stadium där residualen är tillräckligt liten, eller man sätter en övre gräns för antalet iterationer. Vi ska se närmare på två ganska olika varianter av signal-adaptiva algoritmer, nämligen EMD (Empirical Mode Decomposition) och överbestämda representationer.

EMD utgår från hela signalen i tidsdomänen. Idén är att dela upp signalen i flera ”inneboende modalfunktioner” (*intrinsic mode functions*, IMF). Algoritmen isolerar de snabbaste oscillationerna och extraherar dem från signalen, vilket kallas att sålla signalen. Specifikt ser man på extrempunkterna där signalen når ett lokalt maximum eller minimum och drar en kurva som förbinder alla maxima och en annan som förbinder alla minima. Så tar man genomsnittet av de två kurvorna, vilket blir en ny kurva (residualen) med långsammare

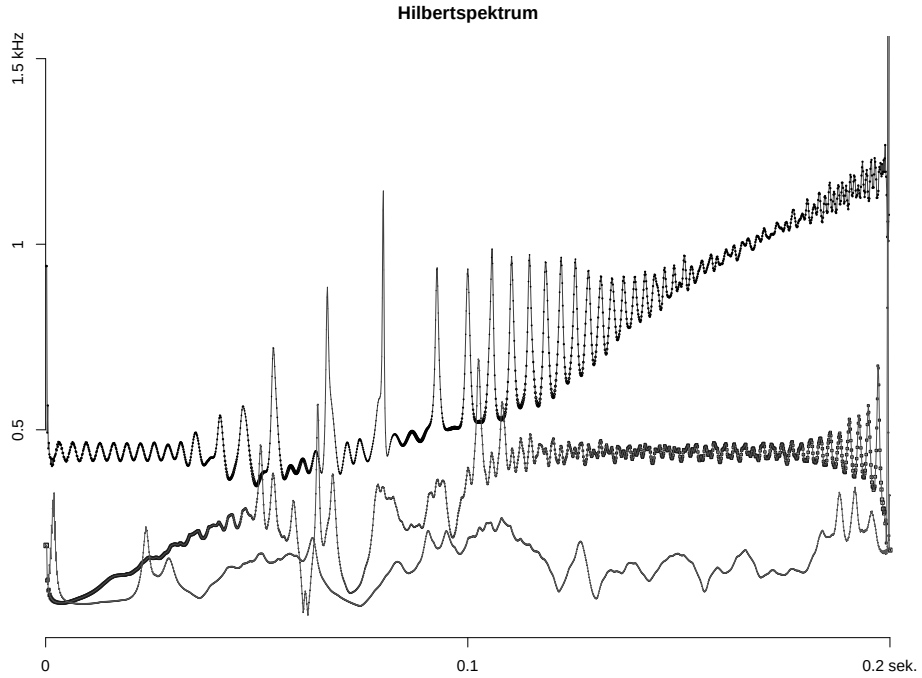


Figur 4.5: Dekomponering i empiriska komponenter av en signal som består av en mixtur av FM och en sinuston som gör ett uppåtgående glissando. Lägg märke till att amplitudskalan är expanderad med en faktor 10 i de två nedersta kurvorna.

oscillationer än den ursprungliga signalen. Därefter subtraherar man den nya kurvan från den ursprungliga signalen och får en kurva där varje oscillation innehåller en nollgenomgång. Processen upprepas på residualen, som gradvis kommer att innehålla allt långsammare oscillationer.

De första stegen av dekomponering av en signal visas i figur 4.5. Signalen är summan av en ton genererad med FM-syntes och en sinusoid som gör ett stigande glissando. Den lägsta delen av glissandot fångas upp av IMF 2, medan den högsta delen ingår i IMF 1. Den tredje komponenten och dess residual är redan mycket svagare än de två första.

EMD fungerar bra på signaler som är uppbyggda av ett antal sinusoider eller andra regelbundna vågformer, men är mindre lämpad på stokastiska signaler. Men i en studie av just stokastiska signaler med gaussfördelning och olika grader av autokorrelation visade det sig att man i grova drag kunde förstå EMD som en slags filterbank med konstant Q-faktor. Något förenklat gäller därför att första steget i sållningen tar ut översta oktaven av signalen, nästa steg oktaven under den, osv (Flandrin et al., 2004). Samtidigt kan det vara missvisande att föreställa sig EMD som en filterbank, eftersom varje steg i processen alltid sållar bort de högsta frekvenserna som är närvarande i signalen vid varje tidpunkt, till skillnad från



Figur 4.6: Hilbertspektrumet av testsignalen visar tydligt den stigande sinustonen. Amplituden markeras av linjernas tjocklek.

vågelement, där frekvensbanden är fixerade.

Två egenskaper gör EMD till en ovanlig metod, nämligen att den inte utgår från fasta basfunktioner, och att den i princip opererar på hela signalen på en gång, utan att dela upp den i tidslokaliserade fönster. Det gör att man kan analysera storskaliga tidsförlopp lika väl som de snabbaste variationerna i signalen (Heydarian & Reiss, 2005).

Ytterligare processering av de inneboende modalfunktionerna är möjlig. Om man ser på den momentana frekvensen av varje IMF-komponent får man den så kallade Hilbert-Huang-transformen (Kim & Oh, 2009). Ungefär som med spektrogrammet visar den amplituden av de funna komponenterna som en funktion av tid och frekvens, men med hög upplösning, se figur 4.6.

4.1.6 Överbestämda representationer

I gabortransformen finns det ett unikt sätt att representera signalen med amplitud och fas av basfunktioner centrerade vid olika tidpunkter och frekvenser. Då finns det också en unik invers transform som tar informationen i tid-frekvensplanet tillbaka till tidsdomänen. En helt annan tankegång är att adaptivt leta efter lämpliga basfunktioner som lokalt i tid ägnar sig väl till att representera signalen, och där man har ett helt "lexikon" som består av flera basfunktioner (eller atomer som de ofta kallas i det sammanhanget) än vad som krävs för att kunna representera en godtycklig signal (Goodwin, 1998). Man kan till exempel utgå från gabortransformens basfunktioner och komplettera dem med vågelement. Om lexikonet innehåller både långa sinustoner och korta impulser blir det möjligt att komma över den

inneboende osäkerhetsrelationen i vanlig tids-frekvens-analys. Då kan man åstadkomma sonogram som klart och tydligt visar både korta klick och långa stabila deltoner (Sturm et al., 2009; Kling & Roads, 2004). Sådana sonogram kallas Wivigram, eftersom de består av Wigner-Ville-fördelningen av de atomer som representerar signalen. De funna atomerna analyseras en i taget, och resultatet adderas till en bild.

När man har ett sådant redundant lexikon finns det flera olika sätt att välja atomer för att bygga upp en och samma signal. Exakt vilka atomer som väljs beror på algoritmen, nämligen Matching Pursuit (MP). Det finns olika varianter av MP, men de går ut på att steg för steg dekomponera signalen i atomer. På samma sätt som med EMD tar man successivt ut atomer ur signalen och får en residual, som man fortsätter dekomponera. En strategi för att välja vilken atom man ska välja är att ta den som representerar mest energi i signalen vid varje stadium av dekomponeringen.

Anledningen till att välja en överbestämd transform, är att man då kan hitta ett fåtal koefficienter som tillsammans approximerar signalen väl. Om signalen består av några statiska sinustoner, beskrivs den väl av några få fourierkoefficienter. Men om den istället innehåller många transienter, som perkussiva ljud gör, så kan någon form av wavelets ge en mera relevant beskrivning. En överbestämd analys skulle kunna gottgöra sig av både fourieranalys och waveletanalys, och representera en mångfald av signaler väl med några få koefficienter.

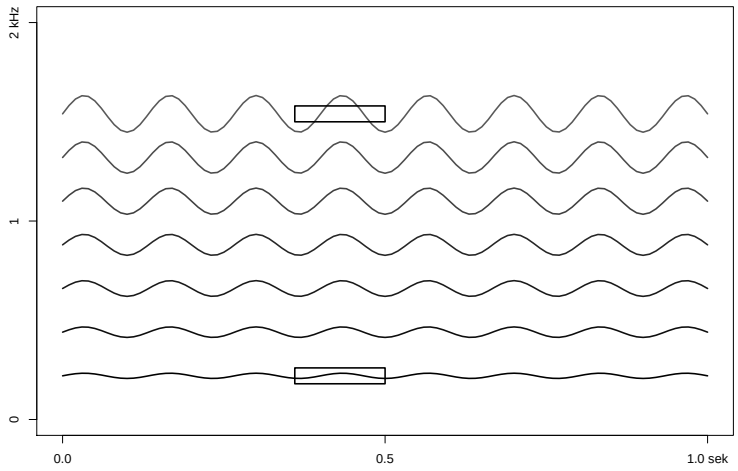
Periodtransformen är ett annat exempel på en överbestämd representation, som används för att söka efter periodiska mönster i signaler (Sethares & Staley, 1999). Den dekomponerar en signal i summan av flera periodiska funktioner, som var och en upprepar sig med olika längd. Också här finns det flera olika algoritmer som leder till olika sätt att dekomponera signalen. Man kan till exempel analysera perioder från korta till längre, eller sortera dem efter bästa korrelation efter periodlängd, eller man kan göra en fouriertransform av signalen och välja periodlängden som motsvarar inversen till frekvensen av den starkaste deltonen. Men i likhet med Matching Pursuit och EMD extraherar man den funna periodiciteten i varje steg av processen och fortsätter analysera residualen.

Periodtransformen har fått användning när det gäller att analysera rytmiska förlopp i musik. Fördelen mot vissa andra metoder är att den kan användas direkt på en ljudsignal, eller åtminstone på lågnivådeskriptorer, medan en del andra metoder kräver att man utgår från en notbaserad representation.

Gemensamt för alla dessa överbestämda representationer, är att det finns fler basfunktioner än det minimum som är nödvändigt för att representera en godtycklig signal, och som en konsekvens av det finns det flera olika representationer av samma signal. Signaler kan representeras med ett fåtal basfunktioner, antingen exakt eller approximerat, vilket kan användas bl.a. till datareduktion.

4.1.7 Analys av variabel tonhöjd

Basfunktionerna i fouriertransformen är sinusoider med konstant frekvens. Det samma gäller för Morlets vågelement och liknande former av analys med hjälp av en filterbank. I korttidsfouriertransformen gör man antagandet att det går bra att representera ljudet som om det vore stationärt under analysfönstrets tidslängd. I praktiken stämmer det inte så bra ifall signalen till exempel kommer från en sångare med kraftigt vibrato eller ett snabbt glissando spelat på violin. Storleken av ett vibrato kan anges som utslaget uppåt och nedåt i frekvens



Figur 4.7: Vibratoproblemet (skematisk illustration): låga deltoner varierar mindre i frekvens och ger en skarp analys med STFT, medan höga deltoner varierar mer. De två rektanglarna ska föreställa lokaliseringen i tid och frekvens vid två olika frekvensband.

i förhållande till en central frekvens, som är grundtonen. Precis som i FM-syntes kan vi tänka på vibrato som en ton som varierar periodiskt i frekvens,

$$f_v(t) = f_0(1 + \Delta \sin(2\pi f_m t))$$

där f_0 är grundtonen, Δ är vibratots utslag och f_m är vibratofrekvensen. Med $\Delta = 0.05$ får man ett vibrato på cirka en halvton. I ett harmoniskt spektrum med $f_v(t)$ som grundton ökar utslaget av vibratot ju längre upp i övertonsserien man kommer (se figur 4.7); för delton nummer k är utslaget nämligen $k\Delta$ Hz. Det gör att om man har valt ett långt nog analysfönster för att fånga upp grundtonen i spektrumet kommer de högre deltonerna att variera i frekvens under den tiden, vilket smetar ut spektrumet för höga frekvenser. En lösning kan vara att använda en CQT, där de högre deltonerna analyseras med kortare fönster än de lägre.

En annan intressant möjlighet i det här sammanhanget är att göra en så kallad ”Fan Chirp Transform” (FChT). Med ”chirp” menas en sinusoid som gör ett glissando. Istället för att bara använda basfunktioner med stabil frekvens, används sinusoider som ökar eller minskar i frekvens i olika takt som basfunktionerna i FChT (Képesi & Weruaga, 2006; Cancela et al., 2010). Det visar sig emellertid att man lika gärna kan variera avspelningshastigheten i den signalen man ska analysera och göra en vanlig fouriertransform av dessa ”tidsvrängda” (eng. time warped) signaler. Om $x(t)$ är en sinuston som gör ett glissando upp en kvint under en sekund, så finner man en funktion $\phi(t)$ som gör att $x(\phi(t))$ bli en sinuston med stabil frekvens. Det betyder i diskret tid att man måste sampla om signalen. För monofoniska signaler (enstämmiga instrument) passar den här metoden bra. Då kan man söka efter den glissandoriktningen som stämmer bäst överens med signalen och plotta den i ett sonogram. Resultatet blir klarare i signaler med snabba frekvensförändringar än med STFT. Dessutom kan metoden användbar vid analys av grundton. En annan tänkbar tillämpning vore att ta

en inspelning med scratching, där en kort snutt från en vinylskiva spelas framlänges och baklänges med varierande hastighet, och genom att finna en invers tidsvrängning skulle man kunna komma tillbaka till den ursprungliga inspelningen på skivan.

Flera av de tidigare nämnda metoderna kan också användas på signaler med varierande tonhöjd. Den spårande fasvocodern följer deltonernas frekvens och interpolerar den mellan analysfönstrena; Hilbert-Huang-transformen kan också visa hur deltoner varierar i frekvens över tid. Än så länge är det bara fasvocodern som är någorlunda standard i programvara för ljudanalys, men det finns alltså intressanta möjligheter för signalrepresentationer som alla är skräddarsydda för olika sätt att se på signalen.

4.2 Signaldeskriptorer

Även om perceptionen av ljud går igenom ett slags frekvensanalys i basilarmembranet är vi sällan medvetna om de enskilda deltonerna. Istället urskiljer vi många andra attribut på högre nivå i ljudet som har med klangfärg, tonhöjd och intensitet att göra. På liknande sätt kan fouriertransformen fungera som en frekvensanalys på låg nivå. Ett första steg för att göra dess data meningsfulla är att göra om det komplexa spektrumet till ett amplitudspektrum. För att utvinna information om diverse perceptuellt meningsfulla attribut ur ett amplitudspektrum måste man bearbeta denna information ytterligare. Fouriertransformen är därför en nyttig utgångspunkt för många olika typer av analys. Utifrån amplitudspektrumet med lineär amplitud och frekvens kan man gå över till logaritmiska enheter, dB och logaritmisk frekvens. Ännu närmare en perceptuellt grundad representation kommer man genom att representera frekvens på en mel- eller barkskala (en Bark motsvarar ett kritiskt band).

För att verkligen skapa en mera perceptuellt grundad modell, filtrerar man först signalen på ett sätt som motsvarar ytter- och mellanörat, och därefter genom en gammatonfilterbank, som består av bandpassfilter som simulerar innerörats frekvensupplösning. Utgången från gammafiltrena motsvarar svängningarna i basilarmembranet. Ett ytterligare steg är att omvandla gammafiltrenas utgång till nervsignaler i hörselnerven.

Utan att gå den komplicerade vägen att representera alla led i nervsignalen för att komma fram till perceptuella attribut, kan man använda långt enklare metoder och eventuellt anpassa dem så att de kommer närmare en användbar beskrivning. Å andra sidan är många signaldeskriptorer användbara i ljudsyntes och effekter, till exempel för att skapa olika slags adaptiva effekter. Kompressorn är typexemplet på en sådan, där signalens amplitud styr förstärkningen av signalen.

Med signaldeskriptorer menas vanligen signaler med en betydligt lägre samplingsfrekvens än den analyserade signalen. De fyller funktionen att sammanfatta någon egenskap hos signalen över ett visst tidsfönster. Vidare dataanalys av signaldeskriptorer kan vara intressant, som att beräkna medelvärde över tid och standardavvikelse. Om man försöker hitta tonansatser kan det vara nyttigt att ta derivatan av en deskriptor för att se var den förändras fort.

Man kan dela in signaldeskriptorer i olika nivåer beroende på hur nära de relaterar till signalen eller till meningsfulla perceptuella attribut. Lågnivådeskriptorer är de som använder vanliga metoder från signalbehandling och statistik utan någon större ambition om att vara anpassade till perceptionen av ljud, medan deskriptorer på hög nivå brukar modellera hörseln

som fysiologisk process och försöker att beskriva egenskaper av ljud som är meningsfulla för en lyssnare. I det här avsnittet går vi huvudsakligen igenom deskriptorer på låg nivå.

Somliga deskriptorer är lätta att beräkna i tidsdomänen, andra bygger på amplitudspektrumet av en STFT. Dessutom finns det en grupp deskriptorer som tar utgångspunkt i att man har identifierat ett antal deltoner med deras frekvens och amplitud (McDermott et al., 2006).

4.2.1 Amplitudmått

Amplitudkurvor kan utvinnas ur en signal på flera sätt. Hilberttransformen kan lätt användas till att analysera momentan amplitud (se kapitel 3). Denna kurva har en hög tidsupplösning och egentligen för mycket detaljer, så om den ska användas är det en fördel att lågpassfiltrera den. Effektivvärdet (RMS-amplituden, efter *Root Mean Square*)

$$A_{RMS}[n] = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1} x^2[n-k]} \tag{4.3}$$

är ett vanligare mått. Det kan bestämmas över kortare eller längre tidsintervall, vilket gör det möjligt att ställa in dess detaljskärpa. RMS-beräkningen kan betraktas som ett icke-lineärt filter, som mäter genomsnittlig amplitud under de N senaste samplen. Genom Parsevals teorem framgår det att RMS-värdet kan beräknas med i princip samma formel i frekvensdomänen också, nämligen

$$A_{RMS} = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1} |X(k)|^2} \tag{4.4}$$

där X är fouriertransformen av $x[n],\ n = 0, 1, \ldots, N-1$. Ibland kan det vara intressant att se på energin i något begränsat band av frekvenser. Då kan man summera amplituden i (4.4) över detta begränsade band, eller alternativt filtrera signalen i tidsdomänen och beräkna RMS-amplituden enligt (4.3).

Det enda som varierar i hur man implementerar RMS-beräkningen av signalen i tidsdomänen är hur man tar medelvärdet av signalen. Enligt formeln (4.3) skulle man använda det löpande medelvärdet

$$y_n = \frac{1}{N}\sum_{k=0}^{N-1} x_{n-k}^2,$$

men det är också vanligt att använda ett enpolsfilter,

$$y_n = x_n^2 + by_{n-1}$$

där b är aningen mindre än 1, en så kallad läckande integrator.

Två andra sätt att mäta amplitud är att utgå ifrån en s.k. halv- och helvågslikriktare (*half* resp. *full wave rectifier*). Halvvågslikriktare ser bara på den delen av signalen som är positiv, dvs $\max(x_n, 0)$, medan helvågslikriktare innebär att ta absolutbeloppet $|x_n|$ av signalen. I båda fallen följs dessa operationer av lågpassfilter som tar bort så mycket som möjligt av vågformens periodicitet och lämnar kvar ett genomsnittligt amplitudvärde. Ingen av dessa amplitudmätningarna motsvarar exakt den uppfattade intensiteten. Ett viktigt skäl

Algoritm 4.1 Beräkning av RMS i tidsdomänen.

```
// LP_filter är ett valfritt lågpassfilter.

double RMS(double x)
{
    double y = LP_filter(x*x);
    return sqrt(y);
}
```

till det är att de inte tar hänsyn till örats frekvensberoende uppfattning av tonstyrka. För att nå det målet skulle man behöva separera signalen i frekvensområden med en filterbank och därefter göra amplitudanalyser på varje kanal för sig, och slutligen väga och summera ihop resultatet.

Ett annat amplitudmått är toppfaktorn (*crest factor*), som är kvoten mellan det maximala amplitudvärdet och effektivvärdet. Det ger ett mått på hur mycket utstickande toppar vågformen innehåller.

4.2.2 Nollgenomgångar och vågtoppar

Frekvensen av nollgenomgångar (*zero crossing rate*, ZCR), dvs antalet nollgenomgångar per antal sampel, ger en grov indikation på den spektrala balansen mellan höga och låga frekvenser; om signalen består av rena toner tenderar den att vara låg, om den är mera brusig är den hög. Om man på förhand vet att signalen är monofonisk med variabel frekvens, och om dessutom vågformen är oföränderlig över tid, så är frekvensen av nollgenomgångar exakt proportionerlig mot tonens frekvens. Det är naturligtvis orealistiska förutsättningar med de flesta signaler man kan tänkas vilja analysera.

Man kan lätt härleda den väntade frekvensen av nollgenomgångar för vitt brus och en del andra enkla signaler. Vitt brus kan bland annat genereras genom att välja amplitud för varje sampel slumpmässigt i intervallet $[-1, 1]$ med likformig sannolikhetsfördelning. Om x_n är en sådan signal så är sannolikheten att $x_n > 0$ lika stor som att $x_n < 0$. Det kan också inträffa att signalen är exakt 0, även om det är ganska liten sannolikhet för det utfallet. Alltså är $p(x_n > 0) = p(x_n < 0) = 1/2 - p(x_n = 0)$. Antag att $x_n > 0$. Vitt brus kännetecknas av att de successiva amplitudvärdena är helt okorrelerade, och föregående sampelvärde kan därför ha vilket värde som helst oberoende av det nuvarande värdet. Då är sannolikheten att $x_{n-1} > 0$ lika stor som att $x_{n-1} < 0$, nämligen ca $1/2$, och sannolikheten för en nollgenomgång mellan x_n och x_{n-1} är därför $1/2$. Därav följer att vitt brus har väntevärdet ZCR $= 1/2$. En sinuston på nyquistfrekvensen passerar noll för varje nytt sampel, och den har därför den högsta möjliga frekvensen av nollgenomgångar, ZCR $= 1$. Inspelningar av musik har nästan alltid mycket lägre ZCR-värden.

Summeringen av nollgenomgångar sker över ett visst antal sampel, L. Om man väljer ett kort fönster (litet L) får man hög upplösning i tid, men antalet nollgenomgångar är förstås ett heltal från 0 till L. Det innebär att man får en kvantisering av ZCR som är grövre för korta fönster och finare för högre – osäkerhetsrelationen för tid och frekvens gör sig åter påmind. Om man delar antalet nollgenomgångar med L så får man en normaliserad enhet,

ZCR $\in [0, 1]$. Ifall man ska jämföra signaler med olika samplingsfrekvens kan det vara en fördel att konvertera till enheten Hz genom att multiplicera med halva samplingsfrekvensen.

En vanlig situation är att man har en inspelning med störande brus. Antag att den innehåller periodiska toner och svagt vitt bakgrundsbrus. Det betyder att när vågformen är nära noll så kommer bruset att orsaka sporadiska nollgenomgångar, vilket leder till ett högre estimat av ZCR än man hade fått med en ren signal. En tänkbar lösning för att komma förbi det problemet är att använda en robust variant av ZCR, som bara detekterar en uppåtgående nollgenomgång ifall

$$x_n > \epsilon \ \& \ x_{n-1} < -\epsilon$$

för ett passande tröskelvärde $\epsilon > 0$, och motsvarande för nedåtgående nollgenomgångar. Nackdelen med den varianten är att den missar nollgenomgångar både i svaga signaler där $|x_n| < \epsilon$ och ifall signalen passerar nollnivån alltför långsamt.

Det är inte vanligt i ljudanalys att räkna frekvensen av vågtoppar även om det också kan vara en användbar signaldeskriptor. Signalens lokala extrempunkter (maxima och minima) är de punkter där signalens derivata är noll. Det faktumet har utnyttjats i en del ickelineär tidsserieanalys, som vi kommer in på senare. Det är uppenbart att sinustoner har flera toppar per tidsenhet ju högre frekvens de har, och likaså har en mixtur av flera sinustoner, $x(t) = \sum_k a_k \sin(\omega_k t)$, flera toppar än vad dessa sinustoner har för sig. Därför är frekvensen av vågtoppar relaterad till höga frekvenser, och den är också proportionell mot antalet deltoner.

4.2.3 Centroid

Balansen mellan högt och lågt frekvensinnehåll kan beskrivas med den spektrala centroiden, som är känd för att stå i samsvar med en aspekt av klangfärgen, nämligen perceptionen av brilljans. Centroiden kan beräknas som ett vägt genomsnittsvärde eller spektral tyngdpunkt,

$$C = \frac{\sum_{k=0}^{K} k A_k}{\sum_{k=0}^{K} A_k} \tag{4.5}$$

där A är signalens amplitudspektrum. En annan enkel algoritm för att räkna ut centroiden är att parallellt summera amplituden uppifrån nyquistfrekvensen och nedifrån 0 Hz. När man har summerat alla frekvensband så möts de två summorna och är lika vid något band, eller mellan två band. Motsvarande frekvens är centroiden. Det leder till följande formel i kontinuerlig frekvens:

$$\int_0^C |X(\omega)| \, d\omega = \int_C^\pi |X(\omega)| \, d\omega$$

Om man tänker på amplituderna i spektrumet som massa, så kan man också föreställa sig centroiden som den tyngdpunkten spektrumet skulle balansera omkring.

En praktisk fråga är vilken enhet man använder för centroiden. Eftersom den representerar en frekvens vore Hz det mest naturliga valet, men i olika sammanhang kan det vara lämpligt att välja en normaliserad centroid, $C \in [0, 1]$, som fås genom att dividera (4.5) med K, dvs antalet frekvensband.

Algoritm 4.2 Algoritm för centroid.

```
// Input: amplitudspektrum A av längd N, ut: centroiden.
int k=0, m=N-1;
double sum_L=0, sum_H=0, centroid;
while(k != m)
    {
    sum_L += A[k];   // summera från låga frekvenser
    sum_R += A[m];   // och från höga
    if(sum_L > sum_R)
        k += 1;
    else
        m -= 1;
    }
centroid = k/N;
// grovt estimat, kan förbättras med interpolering
```

Centroiden kan för övrigt beräknas i tidsdomänen också. Man utnyttjar då att derivatan av signalen, $\frac{d}{dt}x(t)$, har amplitudspektrumet $|\omega X(\omega)|$. Derivatan kan approximeras genom $\Delta x_n = x_n - x_{n-1}$. Dessutom behövs RMS-amplituden mätas två gånger:

$$C[n] = \frac{A_{RMS}(\Delta x_n)}{A_{RMS}(x_n)}$$

Här får man centroiden uppdaterad med samma samplingsfrekvens som signalen, och genom att välja längd på fönstret för RMS-beräkning kan man reglera hur detaljerat i tid man följer signalen.

4.2.4 Andra spektrala deskriptorer

Många olika signaldeskriptorer har uppfunnits för olika syften (Peeters, 2004; Verfaille, 2003; McDermott et al., 2006). Här och i fortsättningen presenteras en handfull av dem.

Spektral spridning (eller varians) är ett mått på hur utbrett spektrumet är kring sitt medelvärde, centroiden,

$$V = \sum_{k=0}^{N-1} (k/N - C)^2 a_k \tag{4.6}$$

där a_k är amplituden av varje frekvensband delat med summan av alla amplituder och $C \in [0, 1]$ är den normaliserade centroiden. Andra sätt att beräkna spridningen förekommer. Två besläktade mått är skevhet och krökning (*skewness* resp *curtosis*). Det sistnämnda spelar en viktig roll för att skilja mellan spektrum som är koncentrerade eller spridda – skillnaden mellan en sinuston och vitt brus som ytterligheter.

Spektral lutning beräknas genom att anpassa en rät linje (regressionslinje) till amplitudspektrumet. Det är dess lutning som ger detta mått. Enheten är vanligen dB/oktav. Man kan också ange en frekvens, sådan att exempelvis 95 % av spektrumets energi befinner sig

under den (kallat *spectral roll-off*). Det måttet har stora likheter med centroiden, som ju anger den punkt på frekvensaxeln där 50 % av energin befinner sig under den.

Spektral entropi tillämpar Shannons entropibegrepp, som egentligen har att göra med sannolikheter $p(x)$, på amplitudspektrumet. Formeln för entropi är

$$H = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

där man summerar över alla tänkbara utfall av variabeln x. Sannolikheterna måste uppfylla $\sum p(x) = 1$ och $p(x) \geq 0$. Amplituderna i ett spektrum är på inget vis sannolikheter, men om de normaliseras så att $\sum a_k = 1$ så kan man definiera spektral entropi som

$$\hat{H} = -\sum_{k=0}^{K} a_k \log a_k \tag{4.7}$$

där vi också definierar $0 \cdot \log 0 = 0$. Entropi brukar tolkas som ett mått på mängden av information i ett meddelande eller i utfallet av en slumpmässig händelse. Om man slår en rättvis tärning är sannolikheten $1/6$ för vardera av de möjliga utfallen, och entropin är $-\log(1/6) = \log 6$. Om däremot utfallet är säkert, låt säga att en sexa garanterat kommer upp, är entropin lika med 0. Överfört på amplitudspektrum kan entropin tolkas så att en sinuston, med energi vid en enda spektral komponent, har noll spektral entropi, medan flata spektrum (av vitt brus, en impuls eller ett glissando som täcker hela registret) har maximal spektral entropi. Det kan vara praktiskt att normalisera den spektrala entropin så att $\hat{H} \in [0, 1]$ genom att dela värdet i (4.7) med det maximala värdet som är $\log K$. På så sätt blir inte värdet beroende av vilken fönsterlängd man analyserar ljudet med.

En mera subtil detalj är att analysfönstret spelar en viss roll, vilket är speciellt märkbart för sinustoner. För de frekvenser som inte träffar en analysfrekvens exakt uppstår spektralt läckage, med en form som beror på det valda fönstret. Och även om sinustonen sammanfaller med en analysfrekvens resulterar fönstringen i en spektral spridning som gör att den spektrala entropin i praktiken inte kommer ner till 0.

Spektral fluktuation (flux) tar hänsyn till utveckling i tid och anger hur mycket spektrumet förändrar sig från ett analysfönster till nästa.

$$flux_n = \sum_{k=1}^{K} |A[n,k] - A[n,k]| \tag{4.8}$$

Peeters (2004) anger ett annat sätt att beräkna spektral variation eller flux, nämligen genom den normaliserade korskorrelationen mellan spektrumet i det nuvarande och det föregående fönstret. Det leder till formeln:

$$flux_n = 1 - \frac{\sum_{k=1}^{K} A_n[k] A_{n-1}[k]}{\sqrt{\sum_k A_n[k] \sum_k A_{n-1}[k]}} \tag{4.9}$$

De båda formlerna (4.8, 4.9) ser olika ut, men leder till likartade resultat. Om amplituden håller sig oförändrad men frekvensinnehållet plötsligt byts ut, så ger det ett klart utslag i ett högt flux-värde.

4.2.5 Kepstrala koefficienter (MFCC)

Kepstral-koefficienter i Mel-frekvens (MFCC) används flitigt i signalbehandling inom taligenkänning, för att finna likheter inom eller mellan signaler, och diverse andra områden inom MIR. Det engelska ordet, cepstrum, är en förvrängning av ”spectrum”. I själva verket finns det en helt speciell terminologi på detta område, ett slags fikonspråk med ord som ”liftering”, ”rahmonics”, ”quefrency alanysis”, osv (Oppenheim & Schafer, 2004).

De reella kepstralkoefficienterna av en signal x_n med tillhörande spektrum $X(\omega)$ fås av

$$c(n) = \frac{1}{2\pi} \int_{2\pi} \ln |X(\omega)| \, e^{i\omega n} d\omega \tag{4.10}$$

(olika varianter av denna formel förekommer). Man tar alltså logaritmen av amplitudspektrumet, vilket sånär som på en konstant faktor motsvarar att använda dB för spektrumets amplitud. Så tar man fouriertransformen av detta log-amplitudspektrum och får ett kepstrum. Eftersom man tar fouriertransformen av (en ickelineär funktion av) ett spektrum, så kommer man till en tidsdomän av något slag. Det är emellertid inte samma tidsdomän som signalen x_n befinner sig i, utan här representerar de låga koefficienterna långsamma variationer i amplitudspektrumet, sedd på som en signal, medan de höga koefficienterna representerar snabbare variationer. Harmoniska spektrum med flera övertoner har en regelbundenhet som kommer till syne i den kepstralkoefficienten som motsvarar signalens periodicitet (se figur 4.8). Därför kan kepstrumet användas till estimering av tonhöjd.

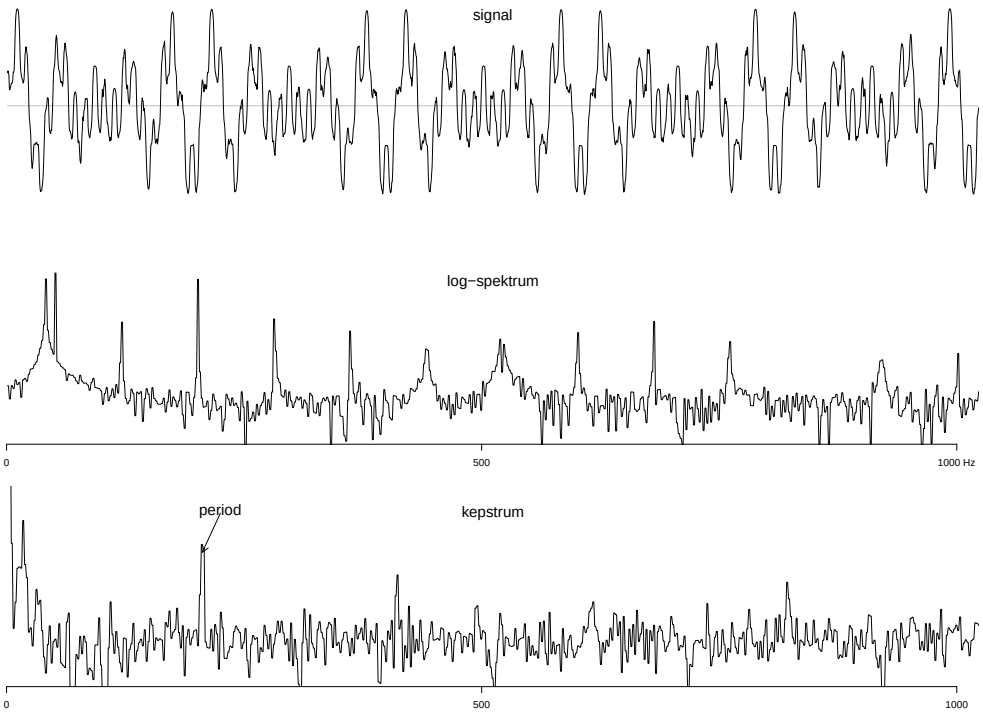
En av de ursprungliga motivationerna till denna formulering av kepstrumet var att man ville ha ett sätt att separera en signal från sitt eko (avfaltning), eller demodulera en ringmodulerad signal. Med tanke på att faltningen av två signaler, $x_n * h_n$, motsvaras av en multiplikation av deras spektrum, $X(\omega)H(\omega)$, så skulle man vilja finna en operation som kan göra om denna multiplikation till en addition. För i så fall kan man använda vanliga filter för att separera de två signalerna. Logaritmen är en operation som konverterar multiplikation till addition, alltså får vi

$$\log(X(\omega)H(\omega)) = \log X(\omega) + \log H(\omega) \tag{4.11}$$

som kan separeras förutsatt att de upptar olika frekvensområden (kvefrensområden vore rätt ord här) i kepstrumet.

Ett sätt att se på ljudkällor är som kombinationer av excitationer och resonans. Till exempel rösten kan ses på som en serie impulser producerade när stämbanden öppnas och sluts, med en resonans formad av munhålan. Spektrumet av resonansen är gärna en ganska slät kurva med långsam variation, vilket betyder att den delen kan representeras av låga kepstralkoefficienter. Excitationen i harmoniska toner är en övertonsrik periodisk signal, med motsvarande impulståg i amplitudspektrumet, vilket fångas upp i de högre kepstralkoefficienterna. Om de två log-spektrumen i (4.11) betraktade som signaler oscillerar olika fort, så uppfångas dessa oscillationer av olika kepstrala koefficienter. Då kan man filtrera ut resonansen $H(\omega)$ och bara få kvar excitationen. För att göra en invers transform tillbaka till den avfaltade signalen behöver man också fasspektrumet, vilket innebär att man måste använda det komplexa kepstrumet (Proakis & Manolakis, 2007).

Även om det är en elegant idé är det inget som säger att man alltid kan separera excitation och resonans på det sättet. Resonansen av ett instrument som violinen kan uppskattas



Figur 4.8: Överst: signal i tidsdomänen, i mitten: log-amplitudspektrumet av signalen, nederst: kepstrumet. Periodicitet i signalen visar sig som en topp i kepstralkoefficienterna.

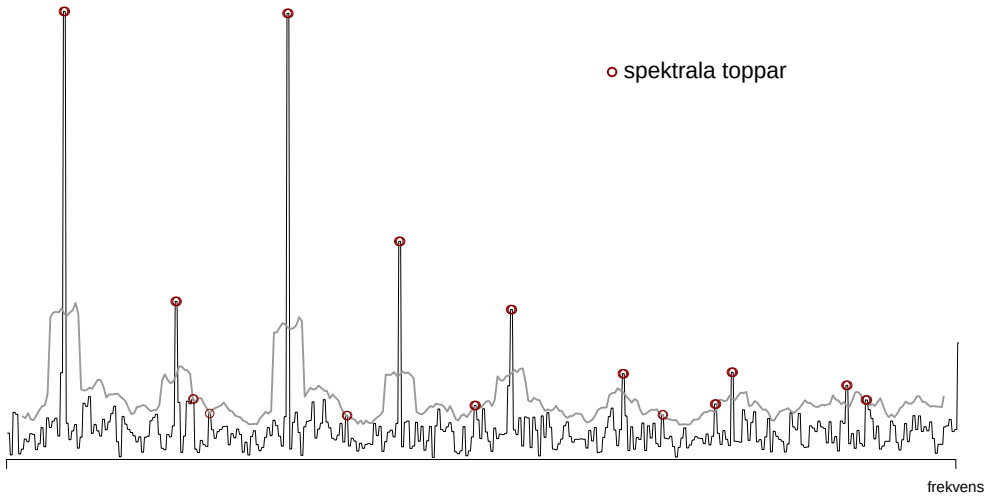
genom att mäta impulssvaret i resonanslådan, och det ger ett spektrum med en ganska taggig kurva, vilken kan vara spridd över samtliga kepstralkoefficienter.

MFCC (Mel Frequency Cepstral Coefficients) är den varianten av kepstrumet som har blivit populär i analys av musik i MIR-sammanhang. Kort sagt går det ut på att indela frekvensaxeln i (4.10) i lika stora steg på en mel-skala. Dessutom används den diskreta cosinustransformen istället för den vanliga fouriertransformen i det sista steget av algoritmen.

4.2.6 Estimering av frekvenser

Så mycket data som ett amplitudspektrum innehåller är det otympligt att handskas med det som sådant. Därför är det vanligt att reducera detta till ett litet antal deltoner, hämtade från topparna i amplitudspektrumet. Det kan delvis försvaras utifrån att starka deltoner har en tendens att maskera de svagare, så att det inte ger någon hörbar skillnad om man avlägsnar de maskerade frekvenserna. Ett sådant reducerat spektrum är användbart för att analysera tonhöjd, harmonicitet, och för att konstruera dissonanskurvor och finna ägnade skalor att spela ljudet i (Sethares, 2005).

Men vad är en spektral topp? Intuitivt sett är det en punkt i spektrumet som är högre än de omgivande punkterna. En naiv algoritm skulle finna de frekvensband P som uppfyller $A(P-1) < A(P)$ och $A(P) > A(P+1)$, och så rangordna dem i storleksordning. Men i många typiska ljudsignaler består inte spektrumet av några få smala toppar, utan de har



Figur 4.9: Detektion av deltoner. Bara de som höjer sig över genomsnittet (här 2 gånger RMS-värdet över 33 angränsande frekvensband, visat med grå kurva) plockas ut.

en viss utbredning över ett större frekvensregister. Dessutom är spektrumet ofta fullt av små lokala toppar och dalar, så att den naiva metoden skulle finna även obetydliga toppar i sluttningarna av större toppar. För ett mera stabilt maximum kan man söka efter kandidater till toppar i en större omgivning,

$$P = \arg\max_{P}\{A(P+m)\},\ m \in \{-2,-1,0,1,2\}$$

med en viss risk för att missa somliga toppar. (Här är arg max(x) en funktion som anger det värdet av x som maximerar funktionen.) Nästa problem är att avgöra vad som är en spektral topp. En naiv algoritm skulle antingen kräva att man letade efter ett fastlagt antal toppar, eller att man satte en fixerad absolut gräns för lägsta möjliga amplitud för en topp. Att fixera antalet toppar vore olämpligt, eftersom ljudet kanske består av en enda klar sinuston i svagt bakgrundsbrus, eller av väldigt många deltoner. I första fallet leder det till falskt alarm, man finner deltoner som inte finns; i andra fallet kan det leda till att man missar deltoner. Att fixera ett konstant tröskelvärde för lägsta amplitud leder till liknande problem. Det får den oönskade konsekvensen att algoritmen upphör att finna några toppar om signalens amplitud blir för svag, oavsett hur framträdande de skulle vara.

En god lösning är att plocka ut de spektrala topparna som höjer sig från genomsnittet i de närmaste omgivande frekvenserna. Det kan man göra genom att jämföra amplituden med RMS-värdet av amplitudspektrumet betraktad som signal, och beräknat utifrån ett antal frekvensband centrerade runt det aktuella (se figur 4.9). Algoritmen innehåller följande moment: Ta ett segment av signalen och fönstra den, beräkna dess amplitudspektrum och finn alla toppar som höjer sig över genomsnittsnivån med någon konstant faktor.

Den enklaste idén till uppskattning av frekvens och amplitud av en spektraltopp i band P är att räkna ut frekvensen utifrån samplingsfrekvens och FFT-storlek, och bara läsa av amplitudvärdet direkt:

Algoritm 4.3 Identifiering av spektrala toppar.

```
// hitta de spektrala toppar
// som höjer sig över tresh*rms(A[k])

void peaks(double *A, int N, int sr, float tresh)
{
// A: Amplitudspektrum av längd N; sr: sampl.frekv.,
// tresh: ratio topp : rms-amp (tresh > 1)
// rms(L): ett rms-objekt som använder löpande medelvärde
    const int L = 32;    // L: rms-filtrets längd
    RMS rms(L);          // initialisera rms-objektet
    float e, frq;

// fyll på halva rms-buffern och ge försprång för symmetri
    for(int i=0; i<L/2; i++)
        e = rms(A[i]);
    for(int i=1; i<N-L/2; i++)
        {
        e = rms(A[i+L/2]);
        // icke-kausalt filter för symmetriskt värde kring A[i]
        if(A[i] > A[i-1] && A[i] > A[i+1] && A[i] > tresh*e)
            {
            frq = (float)i/N * sr;
            printf("%3d\t%7.1f\t%7.2f\n", i, frq, A[i]);
            // skriv ut deltonernas frekvens och amplitud
            }
        }
}
```

$$f(P) = Pf_s/N$$

I praktiken behöver man förbättra estimatet av frekvensen genom att interpolera med hjälp av de närmaste frekvensbanden. Det samma gäller för amplituden. För en ensam sinuston i vitt brus finns det en formel, Cramer-Raos olikhet, som väsentligen säger att felet i estimatet av frekvensen är omvänt proportionellt mot signal-brus-förhållandet, och för konstanta frekvenser förbättras estimatet betydligt ju längre analysfönster man använder. Som en konsekvens kan man uppskatta frekvensen av en stabil sinuston i vitt brus, hur svag tonen än månde vara, genom att analysera tillräckligt lång tid av signalen. Det är visserligen inte till stor hjälp i analys av ickestationära signaler, som praktiskt taget all musik är exempel på.

Estimering av frekvens och amplitud av sinustoner är alltså en process som består av två steg (Hainsworth & Macleod, 2003): Först gäller det att identifiera de korrekta deltonerna, utan att missa deltoner som finns i signalen och utan att hitta deltoner som inte finns där. Eftersom sinustonerna sällan träffar exakt på de frekvenser som bestäms av analysfönstrets

längd måste man på något sätt uppskatta frekvensen. Vanliga metoder utnyttjar antingen fasspektrumet eller interpolerar mellan de två eller tre högsta värdena i närheten av en topp.

4.2.7 Tonhöjdsuppskattning

Tonhöjd kan analyseras på många sätt. Ett av de enklaste, som kan fungera på monofoniska ljud, bygger på autokorrelationen. Att det är möjligt, kommer sig av att autokorrrelationen har ett lokalt maximum vid det värde av fördröjning (d) som motsvarar tonens period. Men det finns flera lokala maxima vid multiplar av periodlängden, så det gäller att hitta det rätta. För alla signaler gäller att autokorrelationsfunktionen antar sitt högsta värde vid $d = 0$, och många ljud har hög korrelation vid korta fördröjningar. Även om den här metoden inte fungerar korrekt för alla slags ljud, så kan den i många fall identifiera korrekt tonhöjd även om grundtonen är frånvarande.

Det är naturligt att vända sig till fouriertransformen för tonhöjdsanalys. Tonhöjden behöver inte representeras av den starkaste frekvenskomponenten i varje ögonblick, då skulle man inte klara av att upptäcka tonhöjden i spektra med frånvarande grundton. En ganska robust algoritm är följande: Multiplicera amplitudspektrumet med en periodisk funktion, till exempel $p(k) = \cos^2(k\omega)$, där k är frekvensband och ω står för perioden. Resultatet integreras från 0 Hz till någon bestämd hög frekvens h. Upprepa för alla realistiska värden på ω. När man finner den period vars integral antar det största värdet, har man en sannolik kandidat till tonhöjd, som är invers till perioden ω. Den högre integrationsgränsen h kan bestämmas utifrån perceptuella överväganden. Över ett visst frekvensområde bidrar övertonerna mindre eller inte alls till tonhöjdsuppfattning. Och även vid låga frekvenser är det primärt de första få deltonerna som styr tonhöjdsuppfattningen.

Mera koncist, den uppskattade tonhöjden är

$$\hat{f_o} \quad = (f_s/N) \quad \arg\max_v \frac{1}{K} \sum_{k=1}^{K} A_k \cos^{2m} \pi k v \tag{4.12}$$

där f_s/N är samplingsfrekvensen delat med längden på analysfönstret, $v \geq 1$ är periodiciteten och m är ett positivt heltal. Ju högre m desto skarpare toppar får funktionen, och därmed mindre tolerans för inharmoniska avvikelser.

Den spektrala metoden för tonhöjdsanalys har den fördelen framför autokorrelationsmetoden att den inte är känslig för små avvikelser från perfekt harmonicitet. Man kan experimentera med olika periodiska funktioner som är mer eller mindre toleranta för inharmonicitet. Om man redan har funnit de spektrala topparna på något sätt som beskrivet i föregående avsnitt, kan man beräkna tonhöjden enbart utifrån dem istället för hela amplitudspektrumet.

Om man utgår från spektrumet $X(\omega)$, så kan man estimera tonhöjder genom att summera energin på de frekvenser där de harmoniska deltonerna är. Då får man en funktion av grundtonen f_0,

$$\rho(f_0) = \frac{1}{K} \sum_{k=1}^{K} \log |X(k f_0)|^2 \tag{4.13}$$

som har en topp vid den frekvensen som förmodligen är grundtonen (Képesi & Weruaga, 2006). Logaritmen av amplituden gör att relativt svaga deltoner blir mera framträdande. Man kan välja godtyckliga värden på frekvensen f_0 förutsatt att man interpolerar det diskreta spektrumet $X(k)$ när de harmoniska övertonerna kf_0 hamnar mellan två frekvensband. Den här metoden kallas "Gathered log-spectrum" eller GlogS.

4.2.8 Ytterligare spektrala attribut

Flera attribut tar utgångspunkt i de spektrala topparna. Tonhöjden är den viktigaste, men det finns andra som har att göra med tonens klangfärg. Tristimuli är tre olika mått, som anger den relativa styrkan av tre spektrala regioner, nämligen grundtonen, de tre första övertonerna och resten av övertonerna (Pollard & Jansson, 1982):

$$T_1 = \frac{a_1}{A}$$

$$T_2 = \frac{a_2 + a_3 + a_4}{A}$$

$$T_3 = \frac{1}{A}\sum_{k=5}^{K} a_k$$

där A är summan av alla amplituder. Tristimuli användes ursprungligen som mått på egenskaper hos färger, men lämpar sig alltså också till att beskriva spektral balans.

Ratio mellan energin av de udda och de jämna deltonerna är en av faktorerna som gör en fyrkantvåg olik en sågtandvåg, en klarinett olik en trumpet. Ett alternativ är att beräkna ratiot mellan udda deltoner och den totala energin. Irregularitet ger ett mått på hur mycket deltonerna varierar i styrka.

$$Irr = \frac{\sum_{k=1}^{K}(a_k - a_{k+1})^2}{\sum_{k=1}^{K} a_k^2}$$

Inharmonicitet mäter hur stor avvikelsen är från ett perfekt harmoniskt spektrum, där alla frekvenser är exakta multipler av grundtonen.

$$Inh = \frac{a_k^2 \sum |f_k - kf_1|}{f_1 \sum_{k=1}^{K} a_k^2}$$

Denna formel fungerar bra så länge spektrumet inte saknar vissa av deltonerna. Ett perfekt harmoniskt spektrum som saknar delton nummer två, framstår till exempel som ganska inharmoniskt i den här modellen.

Förutom de diskuterade attributen, finns det många varianter och utökningar. Exempelvis kan man beräkna differensen av attribut över tid, eller medelvärden och standardavvikelse. Andra varianter fås genom att utgå från lineär amplitud, energi (som är amplituden i kvadrat) eller dB, som är logaritmen av amplituden; eller av att omvandla lineär frekvens till logaritmisk eller någon mera perceptuellt motiverad skala, som barkskalan.

Enligt McDermott et al. (2006) är många av dessa attribut perceptuellt signifikanta, men några av dem är korrelerade med varann. Exempelvis är den spektrala centroiden korrelerad med roll-off. Därför behöver man aldrig använda alla tänkbara attribut för att beskriva ett

ljud, ifall man gör ett vettigt urval av deskriptorer. Samtidigt finns det otaliga andra attribut som vi är bra på att urskilja, men som är mer eller mindre svåra att modellera. Att t.ex. detektera vibrato är aningen mer komplicerat än att detektera de ovan beskrivna attributen. Som nämnt i ett tidigare kapitel, undersökte Schaeffer en mängd morfologiska kriterier för ljud, som vi inte ens har börjat formalisera i termer av signalbehandling. Det är långt ifrån självklart hur man ska ställa upp korrespondenser mellan de schaefferska begreppen och signaldeskriptorer. Här torde det i alla fall finnas utrymme för tolkningar.

4.3

Fortsättning följer ...

Litteraturförteckning

Brown, J. & Puckette, M. (1992). An efficient algorithm for the calculation of a constant Q transform. *J. Acoust. Soc. Am.*, 92(5), 2698–2701.

Cancela, P., López, E., & Rocamora, M. (2010). Fan chirp transform for music representation. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-10).* Graz, Austria.

Flandrin, P., Rilling, G., & Gonçalvés, P. (2004). Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, 11(2), 112–114.

Goodwin, M. (1998). *Adaptive Signal Models. Theory, Algorithms and Audio Applications.* Boston: Kluwer Academic Publishers.

Hainsworth, S. & Macleod, M. (2003). On sinusoidal parameter estimation. In *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx-03).* London.

Heydarian, P. & Reiss, J. (2005). Extraction of long-term structures in musical signals using the empirical mode decomposition. In *DAFX-05 Proceedings.*

Hlawatsch, F. & Boudreaux-Bartels, G. (1992). Linear and quadratic time-frequency signal representations. *IEEE Signal Processing Mag.*, Apr., 21–67.

Képesi, M. & Weruaga, L. (2006). Adaptive chirp-based time-frequency analysis of speech signals. *Speech Communication*, 48, 474–492.

Kim, D. & Oh, H.-S. (2009). EMD: A package for empirical mode decomposition and Hilbert spectrum. *The R Journal*, 1(1), 40–46.

Kling, G. & Roads, C. (2004). Audio analysis, visualization, and transformation with the matching pursuit algorithm. In *DAFx'04 Proceedings.*

Kronland Martinet, R. (1988). The wavelet transform for analysis, synthesis, and processing of speech and music sounds. *Computer Music Journal*, 12(4), 11–20.

McDermott, J., Griffith, N., & O'Neill, M. (2006). Timbral, perceptual, and statistical attributes for synthesized sound. In *Proceedings of the ICMC* (pp. 179–185). Tulane.

Oppenheim, A. & Schafer, R. (2004). From frequency to quefrency: A history of the cepstrum. *IEEE Signal Processing Magazine*, 21(5), 95–99.

Peeters, G. (2004). *A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project.* Technical report, IRCAM, Paris.

Pollard, H. F. & Jansson, E. V. (1982). A tristimulus method for the specification of musical timbre. *Acustica*, 51, 162–171.

Proakis, J. & Manolakis, D. (2007). *Digital Signal Processing. Principles, Algorithms, and Applications. Fourth Edition.* Upper Saddle River: Pearson Prentice Hall.

Schörkhuber, C. & Klapuri, A. (2010). Constant-Q transform toolox for music processing. In *7th Sound and Music Conference (SMC'2010)* Barcelona.

Serra, X. & Smith, J. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4), 12–24.

Sethares, W. (2005). *Tuning, Timbre, Spectrum, Scale.* Springer, second edition.

Sethares, W. & Staley, T. (1999). Periodicity transforms. *IEEE transactions on signal processing*, 47(11), 2953–2964.

Sturm, B., Roads, C., McLeran, A., & Shynk, J. (2009). Analysis, visualization, and transformation of audio signals using dictionary-based methods. *Journal of New Music Research*, 38(4), 325–341.

Verfaille, V. (2003). *Effets audionumériques adaptatifs : Théorie, mise en œuvre et usage en création musicale numérique.* PhD thesis, Université Aix-Marseille II.

Verma, T., Levine, S., & Meng, T. (1997). Transient modelling synthesis: a flexible analysis/synthesis tool for transient signals. In *Proc. of the ICMC, Thessaloniki, Greece.*