(Vorläufig! Obacht: einige Ungenauigkeiten...)

1. Der analoge Ton

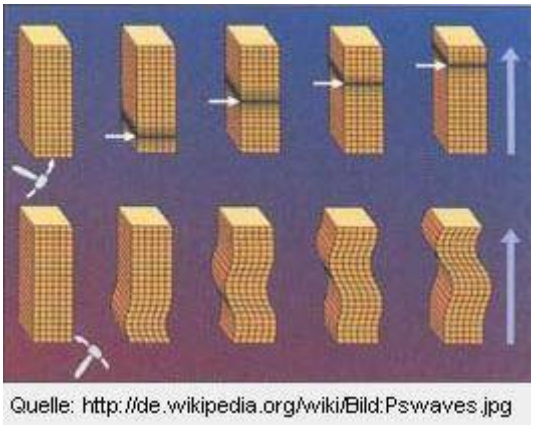
Um zu verstehen, wie Musik *digitalisiert* wird, muss man erst einmal wissen, was *analoge* Musik eigentlich ist.
Genau genommen dürfen wir hier nicht von Musik sprechen, sondern müssen uns allgemein auf Schall- und Schallwellen konzentrieren. Denn auch Stimme oder Geräusche können selbstverständlich digitalisiert werden.

Schall ist in erster Linie eine Welle und unterteilt sich in drei (respektive vier) Bereiche:

- Infraschall: Frequenzbereich unter 16 - 20 Hz. Dieser Schall ist für den Menschen nicht hörbar, dazu ist er zu tieffrequent
- (hörbarer) Schall: Frequenzbereich von 20 Hz – 20 kHz. Der für den Mensch hörbare Schallanteil
- Ultraschall: Frequenzbereich 20 kHz - 1 GHz. Nicht hörbar, da zu hochfrequent
- Hyperschall: Frequenzbereich über 1 GHz. Hier gelten die physikalischen Gesetzte für Schall nur bedingt

Exkurs: Longitudinal- und Transversalwellen

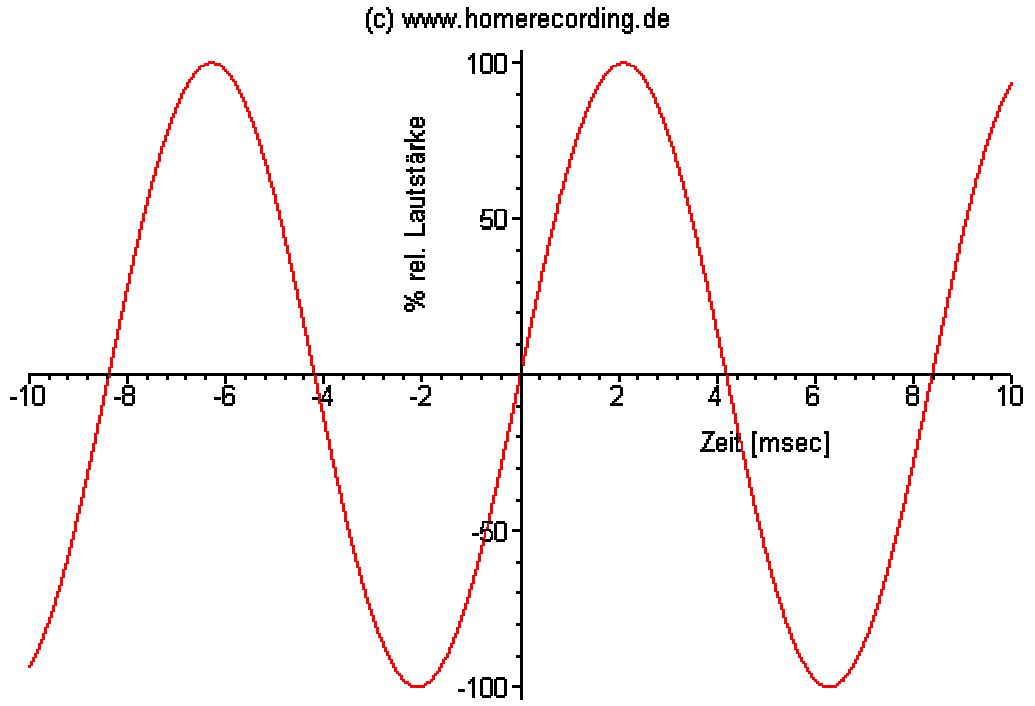
In Luft und allen anderen (ruhenden) Gasen ist Schall eine Druckschwankung, also eine Longitudinalwelle (Längswelle), bei der die schwingenden Teilchen (z.B. Moleküle) längs der Ausbreitung laufen. Das Gegenteil ist übrigens eine Transversalwelle (Querwelle), bei der die Bewegung quer zur Ausbreitung erfolgt (z.B. Wasserwellen heben und senken sich - laufen aber zum Ufer hin). Am besten lassen sich Longitudinal- und Transversalwellen am Beispiel von seismischen Wellen, also Erdbebenwellen, darstellen (in der u.a. Grafik sind oben Longitudinalwellen und unten Transversalwellen abgebildet):



Quelle: http://de.wikipedia.org/wiki/Bild:Pswaves.jpg

Diese Druckschwankungen werden nun von unserem Ohr aufgenommen und dann an unser Hirn weitergegeben.
Sie werden mechanisch erzeugt, als Erzeuger fungiert also immer ein mechanischer Erreger. Dies können die Stimmbänder, Saiten eines Instrumentes oder aber auch die Membran eines Lautsprechers sein - sie haben alle die Eigenschaft gemeinsam, dass sie mechanisch schwingen und damit die Moleküle der Luft anregen.

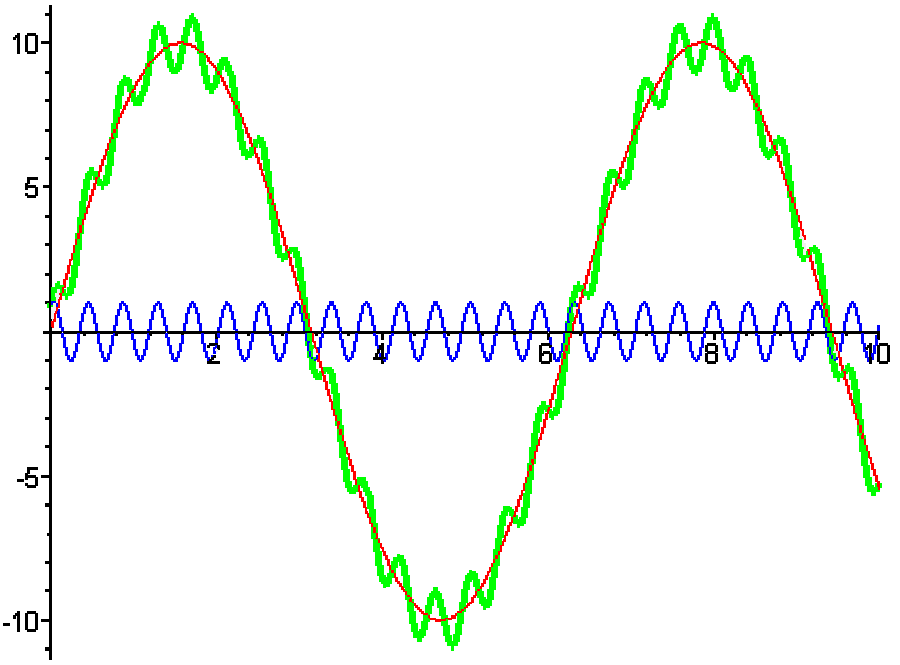
Um diese Druckschwankungen zu verdeutlichen, bedient man sich meistens eines zweidimensionalen Diagramms. An diesem Diagramm kann man genau genommen nur das (transversale) Abbild des Schalldruckes sehen. Die horizontale Achse ist die (absolute) Zeitachse, die vertikale Achse spiegelt die Größe des Druckes zu dem Zeitpunkt wieder. Oft ist die vertikale Achse relativ zu sehen (also quasi von 0 - 100%, manchmal hat man auch absolute Bezeichnungen in dB als Lautstärkeäquivalent oder Pa(scal) als Druckangabe.



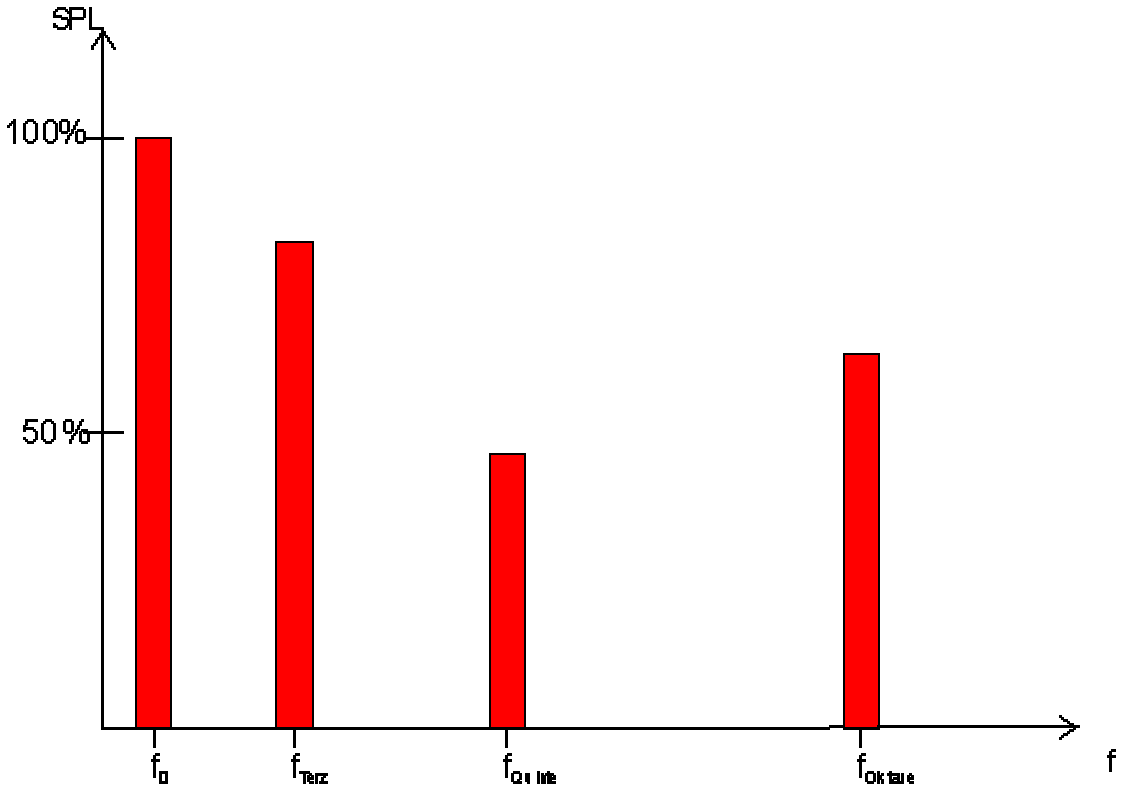
Bei dieser Grafik handelt es sich um einen Sinuston. Dieser heißt Sinuston, weil die Druckschwankungen entlang einer Sinusfunktion der Form $y(t) = \hat{y} * \sin(\omega * t + \varphi$ laufen. Hierbei steht $\hat{y}$ für die maximale Elongation, sprich: Die Amplitude w steht für die Kreisfrequenz und ? für den Phasenwinkel. Die einfachste Form der Sinusfunktion ist y(t) = sin(t).

Doch nun stellt sich die Frage: **Was ist ein Ton?** Ein Ton wird dann als Ton bezeichnet, wenn es eine Periodizität gibt, sprich: Wenn sich das Signal wiederholt. Dies passiert x Mal pro Sekunde. Dadurch ergibt sich die Einheit 1 $\frac{1}{s}$ oder auch 1 Hz (Hertz).

Beim Sinuston ist das Anerkennen als Ton noch recht einfach, denn hier sieht man ja ohne Probleme, dass es Wiederholungen gibt. Schwieriger wird es dann bei realen Instrumenten, wenn wir es nicht mit einem reinen Sinuston zu tun haben. Hier hat man neben dem Grundton (rot) noch so genannte Obertöne (blau), die sich hauptsächlich aus Terzen und Quinten des Grundtones zusammensetzen. Diese stehen dann folgendermaßen in Wechselwirkung (grün):



Man hat dann ein Verhältnis von Grund- zu Obertönen, das man in einem Frequenz-Lautstärke-Diagramm (manchmal auch Formanten-Spektrum genannt) darstellen kann:



Schwierig wird es hier aus zwei Gründen:

1. Prügeln sich Physiker und Musiker darum, ob das noch ein Ton ist. Für die Physiker ist nur ein Sinuston ein Ton, die Musiker sehen das (aus gutem Grund anders): Die Obertöne werden vom menschlichen Ohr nicht als separate Töne aufgefasst, sondern als Klangfarbe des Tons. Anhand dieser Klangfarbe beurteilen wir, was dieses Instrument ist - ob nun Gitarre oder Geige, ob Klavier oder Cembalo, alle können sie ein *a* spielen. Trotzdem hört es sich immer

anders an. Der einzige Grund dafür ist die andere Verteilung von Obertönen. Da man bei realen Instrumenten selten in der Lage ist, einen reinen Sinuston zu spielen (außer zum Beispiel bei Dudelsäcken), betrachten die Musiker dies noch als Ton, während die Physiker es schon *Klang* nennen; *Klang* wird von den Musikern aber als jenes Gebilde aus verschiedenen Tönen betrachtet (einerseits Akkorde oder auch das Zusammenspielen verschiedener Instrumentgruppen).

2. In den meisten Fällen hat man es bei realen Instrumenten nicht nur mit einem Oberton zu tun, sondern direkt mit einer Anzahl im zweistelligen Bereich - jedoch ist das Ohr nur begrenzt in der Lage, zu viele Obertöne als periodisch zu erfassen. Und was nicht als Periode erfasst wird, ist ein *Geräusch* und kein Ton mehr! Das heißt: Streng genommen ist der Übergang zwischen Ton und Geräusch fließend.

Das zeigt auch: Hören ist zum größten Teil interpretieren!

2. Umwandlung des Tons in Spannung

Nachdem man sich Jahrhunderte lang damit zufrieden geben musste, den Ton so mechanisch, wie er ist, zu akzeptieren, ging man Ende des 19. bis Anfang des 20. Jahrhunderts dazu über, eine Abbildungsmöglichkeit des Tons zu suchen, bei der man eine Verhältnisgröße zum Druck hat. Diese fand man unter anderem in der Spannung und im Magnetismus, wobei erstere die hauptsächliche "Transportgröße" ist – denn Magnetisierung geschieht auch über den Umweg der Spannung.

Man musste nun einen Druckempfänger konstruieren, der relativ linear die Druckänderung als Spannungsänderung herausgibt. Das daraus resultierende Produkt ist jedem bekannt und nennt sich Mikrofon. Diese erzeugen über verschiedene Methoden Spannungen, die dann weiterverwertet werden können. Die Empfänger sind entweder Lautsprecher (zur direkten Übertragung) oder Medien zur Speicherung (beispielsweise Magnetbänder).
Lautsprecher "drehen den Spieß einfach um" und lenken Membranen aus, die wiederum die Luft anregen.

Jedoch gibt es auch elektrische Tonübertragung ohne den Umweg über Schallerzeugung. Bestes Beispiel hierfür ist die E-Gitarre, nicht umsonst im deutschen Sprachgebrauch auch als *Stromgitarre* bezeichnet. Hierbei schwingen die Stahlsaiten über einen Permanentmagneten und induzieren in die Wicklungen um den Magneten Spannung. (Induktion ist das Phänomen, dass Magneten Elektronen in Metallen systematisch in eine Richtung bewegen.). Diese Spannung kann dann genauso, wie die Spannung des Mikrofons weiterverwendet werden.

Wir sehen also: **Die elektrische Spannung ist die wichtigste Grundlage für das Digitalisieren von Tönen.**

Im Audio-Bereich gibt es verschiedene Bezugsspannungen. Die Bezugsspannung $\mathbf{U_0}$ ist die Spannung, nach der der elektrische Pegel $\mathbf{U_p}$ (auch in dB angegeben) mit der Formel

$$U_p = 20 * \log_{10}\left(\frac{U}{U_0}\right)$$

berechnet wird. Während in den meisten Ländern in professionellen Studios und Rundfunkanstalten die Bezugsspannung 0,775 V (= 0 dB**m** oder auch dB**u**) beträgt, ist er in Großbritannien 1 V (= 0 dB**v**).

Bei vielen semiprofessionellen Geräten ist der Bezugswert 0,32 V, was nach folgender Rechung etwa –10 dBv entspricht:

$$U_p = 20 * lg\left(\frac{0{,}32\,V}{1\,V}\right) \approx -9{,}897\,dBv$$

In Deutschland trifft man manchmal auch auf die Bezeichnung dB**r**. Dies ist insofern eine Besonderheit, als dass 0 dBr = 6 dBm = 1,55V nicht den leisesten Wert, sondern die **Voll**aussteuerung angibt.
Diese ganzen Spannungen mögen verwirrend sein, haben sich aber über die Jahre bei fehlender übergreifender Standarisierung etabliert und existieren nun nebeneinander.

3. Das Digitalisieren

Die erste Frage, die nun beantwortet werden muss, ist: Was heißt eigentlich *digital*? Das Wort *digital* leitet sich aus dem englischen Wort *digit* ab, was nichts anderes als *Ziffer* heißt. (*Digit* kommt übrigens vom lateinischen *digitus*, der Finger, und indiziert, dass die Menschen früher mit den Fingern gezählt haben.) Damit gibt es schon einen Hinweis darauf, dass Digitalisieren eine Informationsumwandlung ins binäre System ist.

Exkurs: binäres System

Doch was ist das binäre System? Das binäre System ist das so genannte "Zweier-System", das im Gegensatz zum meist beim Rechnen benutzten "Zehner-System" steht.
Das binäre System kennt nur Einsen und Nullen (beim PC: Strom EIN und Strom AUS) und nicht, wie das Dezimalsystem die Ziffern 0 - 9.
Das Dezimalsystem ist wie folgt strukturiert:

10000er	1000er	100er	10er	1er	=
7	1	2	5	6	71256

Man berechnet 7 x 10000 + 1 x 1000 + 2 x 100 + 5 x 10 + 6 x 1 = 71256.
Oder mit Potenzen: $7 \times 10^4 + 1 \times 10^3 + 2 \times 10^2 + 5 \times 10^1 + 6 \times 10^0 = 71256$

Das binäre System hingegen ist wie folgt strukturiert:

64er	32er	16er	8er	4er	2er	1er	=
1	0	0	0	1	1	1	71
1	1	0	1	0	0	0	104
1	1	1	1	1	1	1	127

Man berechnet: 1 x 64 + 0 x 32 + 0 x 16 + 0 x 8 + 1 x 4 + 1 x 2 + 1 x 1 = 71
Oder mit Potenzen: $1 \times 2^6 + 0 \times 2^5 + 0 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 71$

Zählt man die Spalten (ohne das Ergebnis), so stellt man in diesem Fall fest, dass es sich um 7 Spalten handelt. Damit haben wir eine 7-Bit-Zahl.

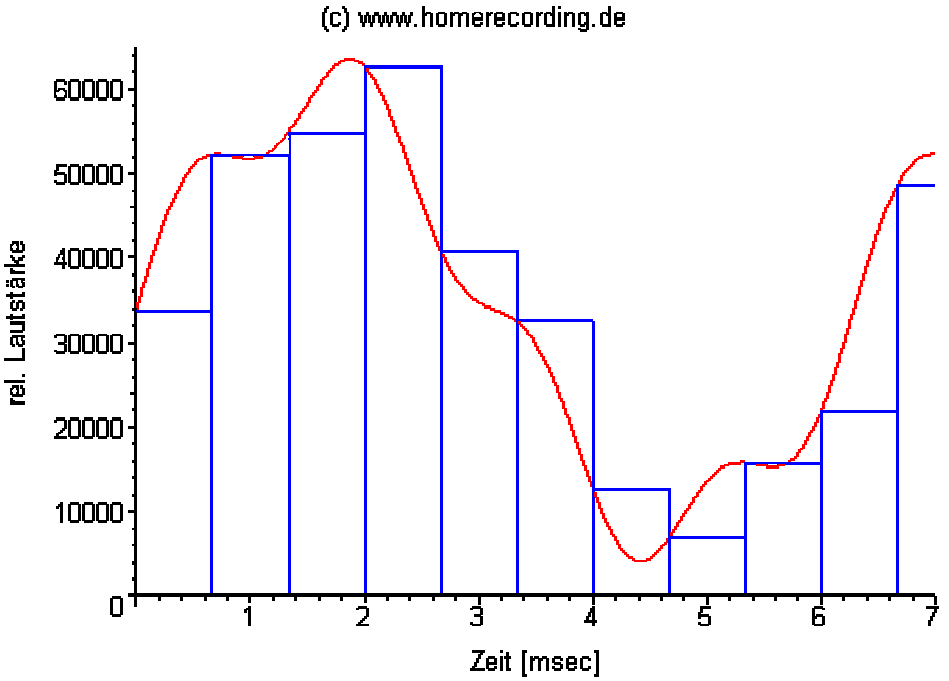
Allgemein kann man sagen, dass die höchstmöglich darstellbare Zahl bei *n* Bit 2^n-1 ist, in diesem Fall also 2^7 - 1 = 127. Man muss aber in einem Punkt etwas aufpassen: Die **Auflösung** oder **Wortbreite** bei 7-Bit beträgt 128 Zahlen - denn die Null gibt es auch im binären System. Damit wird klar, das der Unterschied zwischen 16 und 24 Bit (zwei vielgenutzte Wortbreiten) nicht um den Faktor 1,5 sondern um den Faktor **256** unterschiedlich ist.

Kleine Anmerkung noch: In digitalen Mischpulten wird oft mit 32 Bit gearbeitet (was einer Auslösung von 4,3 Milliarden Zuständen entspricht). Dies erscheint auf den ersten Blick extrem viel. Berechnet man jedoch das Datenaufkommen, wenn man 24 Spuren mit 16 Bit Signalen belegt und diese abmischt (mathematisch gesprochen: addiert), so kommt man schon auf 21 Bit. Die nächste meist verwendete Größe (man geht meist in 8 Bit-Schritten) ist 24 Bit. Um noch genügend Headroom zu bieten, geht man auf Seiten der Mischpulthersteller auf Nummer sicher und arbeitet mit 32 Bit.

Der Computer kennt, wie wir eben festgestellt haben, nur die Zustände 0 und 1. Es muss also eine Möglichkeit gefunden werden, die anliegende Spannung (die ja nicht konstant ist) zu messen und in eine Zahlenfolge mit n Bit Wortlänge zu übersetzen.
Dies wird mit einem AD-Wandler (**A**nalog-**D**igital-Wandler) vollzogen, der in regelmäßigen Abständen die Spannung abliest.

Doch wie arbeitet ein AD-Wandler? Wie man auf den oberen Abbildungen erkennt, ändert sich die Spannung eines Audiosignals ständig – sie ist nicht gequantelt (in einzelne Abschnitte unterteilt). Innerhalb des Wandlers wird deshalb das Signal in festen, zeitlich diskreten Abständen *entnommen.* Dazu wird ein Kondensator mit dem Augenblickswert der Signalspannung geladen. Der Kondensator wird nun durch einen 'Schalter' vom Signal auf den Eingang des Wandlers umgeschaltet. Die in der Kapazität gespeicherte Ladung wird anschließend dazu verwendet, den so 'eingefrorenen' Wert des analogen Signals in ein digitales, maschinenverarbeitbares umzusetzen, also zu wandeln. Die 'Blitzlichtaufnahme' des Eingangssignals und das halten dieses Wertes bezeichnet man als Sample-and-Hold.

Im Diagramm sähe das dann so aus:

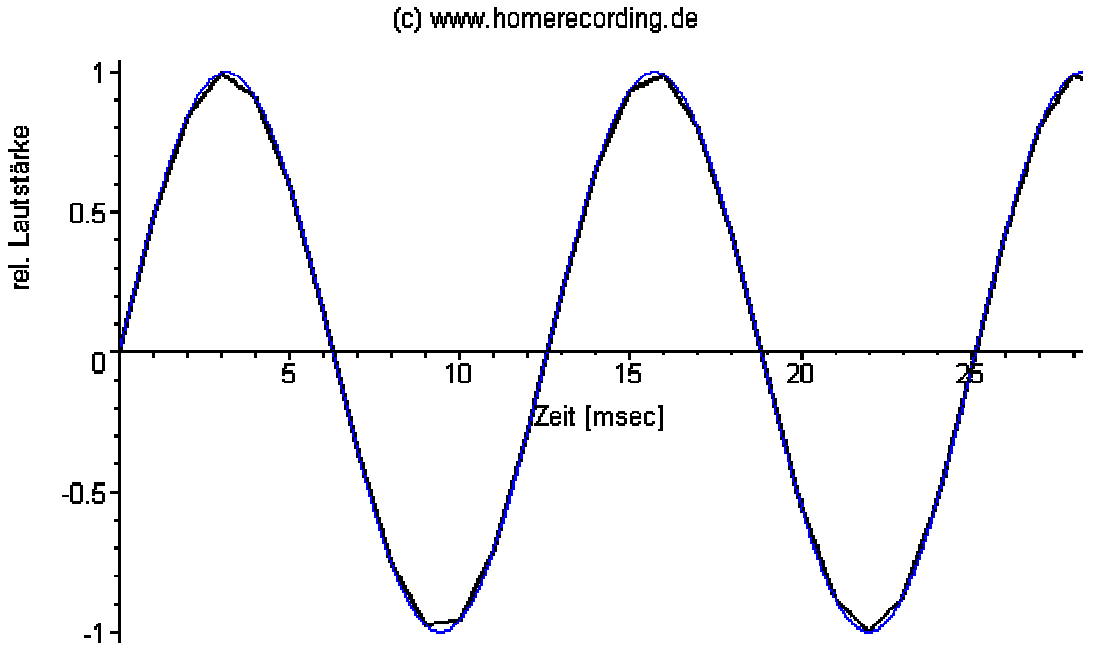


Hier würde in 1 msec 1,5 mal gemessen, das entspräche einer Sampling-Frequenz von 1,5 kHz.

Die regelmäßigen Abstände, mit denen gemessen wird, befinden sich in der Größenordnung von Mikrosekunden oder als Frequenz ausgedrückt im Bereich von Kilohertz. Warum das nötig ist, zeigt folgende Überlegung, die sich hauptsächlich auf Überlegungen von Harry Nyquist (1889 - 1976) und Claude Shannon (1916 - 2001) stützen.

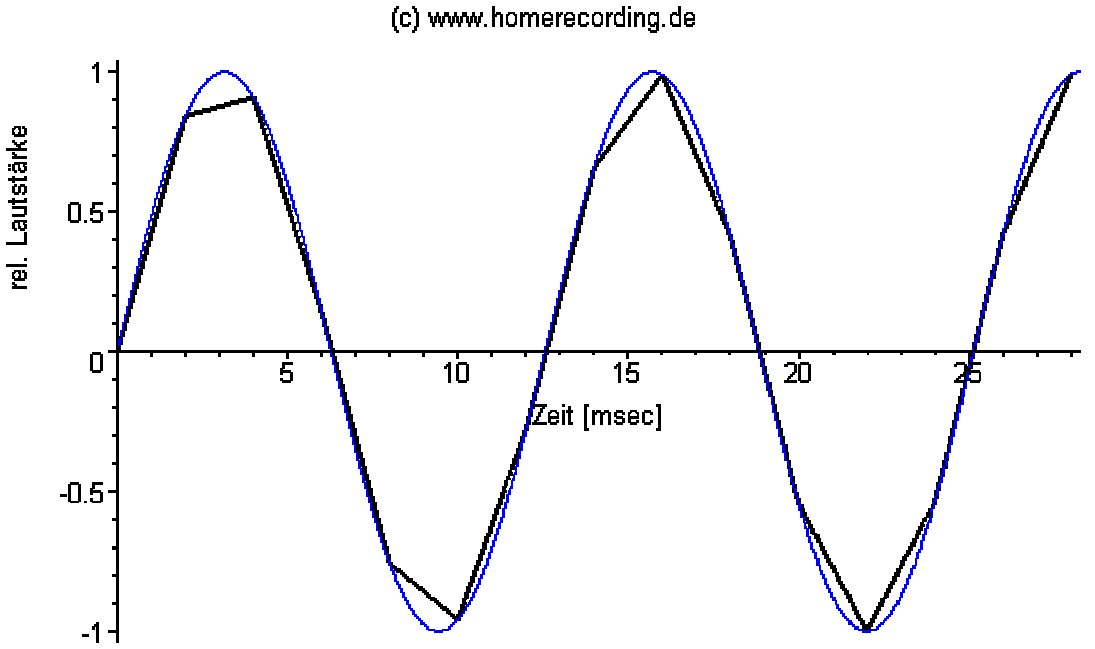
Exkurs: Nyquist-Shannon-Abtasttheorem

Ein AD-Wandler taste ein Sinussignal mit der Frequenz 80 Hz (blaue Linie) ab. Er speichert diese Daten als Werte, die nachher verbunden werden können und dann wieder eine Kurve ergeben (schwarze Linie). Diese Kurve sieht anfangs noch recht eckig aus - dazu aber später mehr. Tastet man nun mit einer Samplerate von 1000 Hz ab, so ergibt sich folgendes Bild:

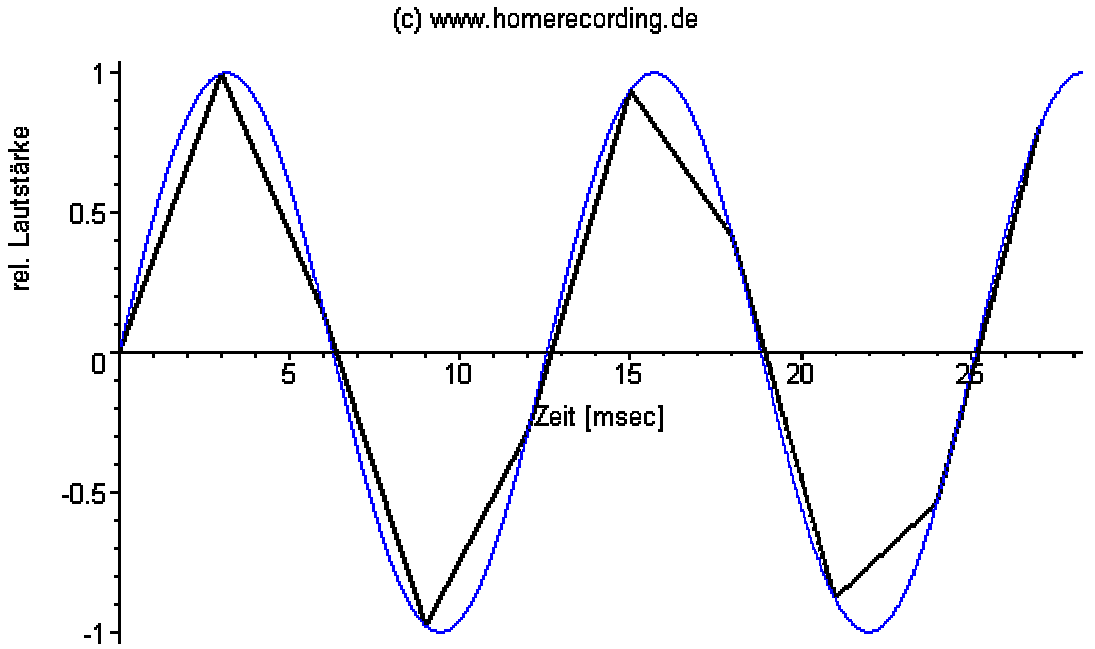


Diese Kurve ist noch sehr nach am Original dran.

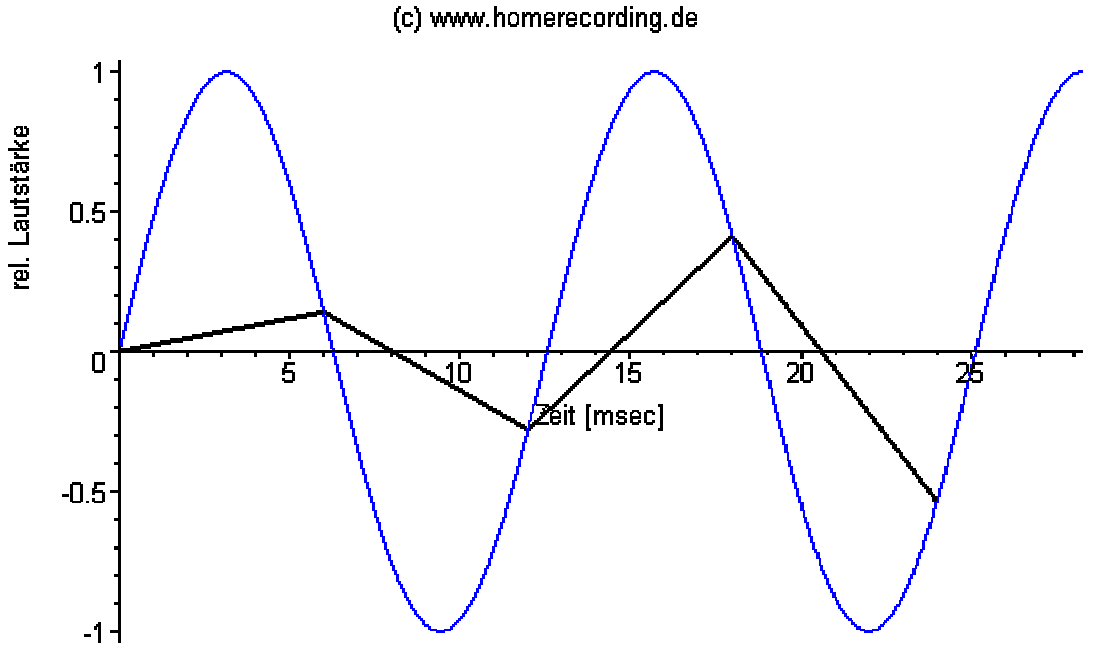
Geht man aber anstatt 1000 Hz nun auf eine Samplingrate von 500 Hz runter, so stellt man sehr schnell eine gewisse "Kantigkeit" fest:



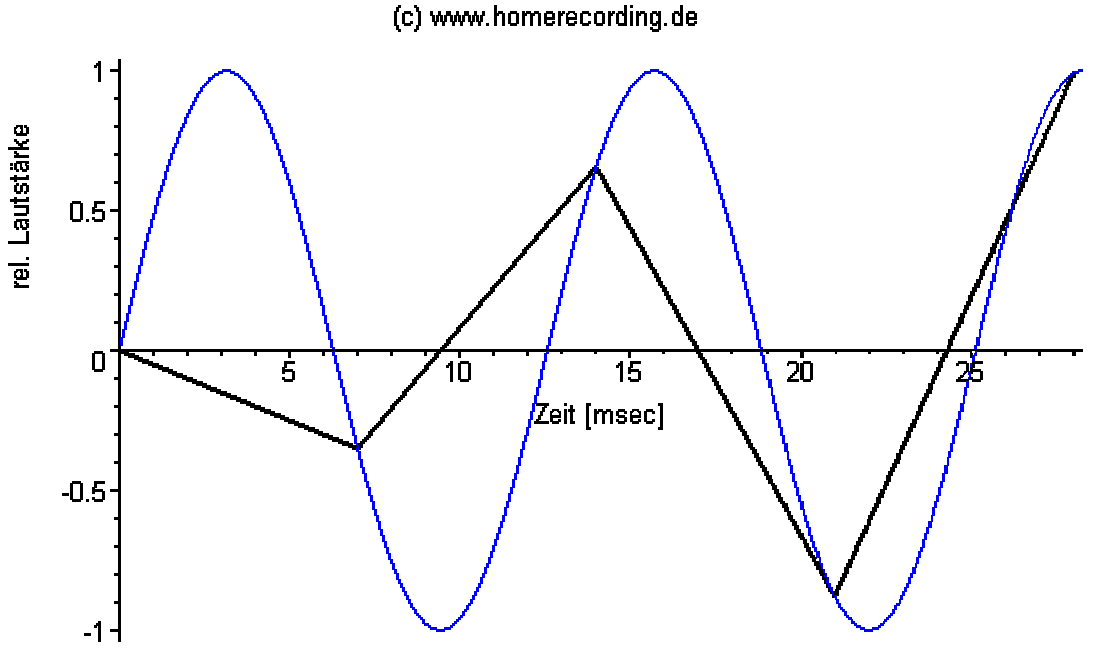
Bei ungefähr 330 Hz wird es noch kantiger:



Nun bei 170 Hz. Man sieht, es wir deutlich eckiger, aber das Original-Signal lässt sich noch erahnen (und vor allem: mittels geeigneter Polynom-Interpolation - dazu später mehr - rekonstruieren). Dass das Signal phasenverschoben scheint, ist hier nicht wichtig, denn a. ist dann das **Gesamt**signal phasenverschoben und b. kann diese Verschiebung wieder herausgerechnet werden. Auch dass die Amplitude so eklatant geringer ist, ist nur Zufall!

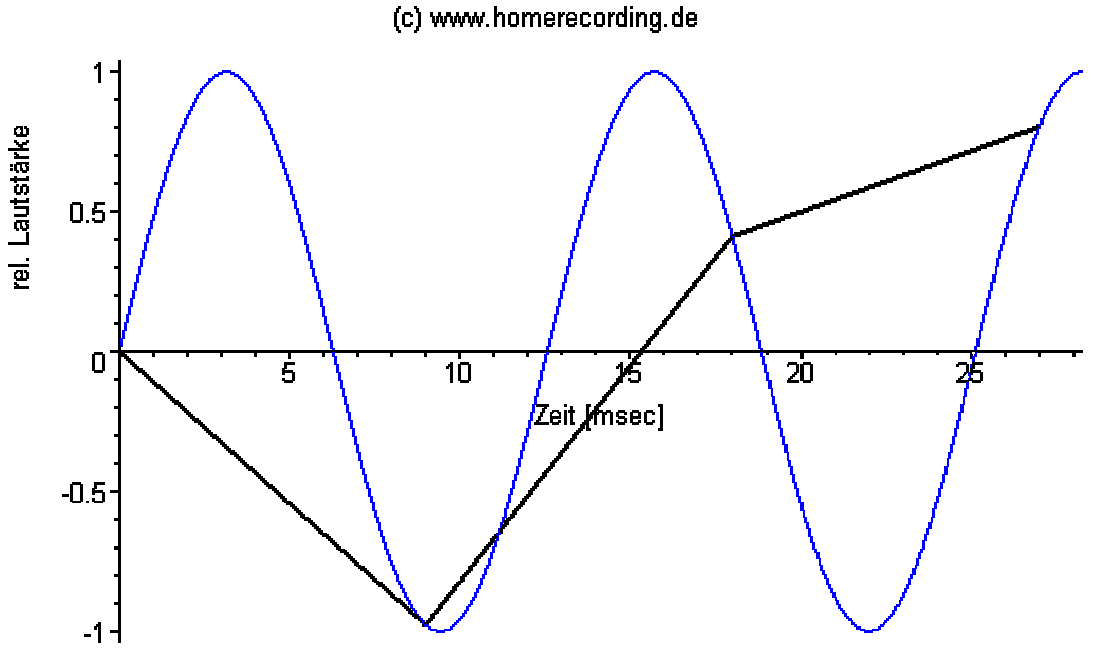


Nun aber bei ca. 145 Hz Abtast-Frequenz:



Hier ist es das erste Mal so, dass sich das Originalsignal nicht wieder rekonstruieren lässt. So sehr man auch die Augen zukneift, man erkennt **keine** drei Berge und zwei Täler mehr.

Bei ca. 110 Hz Abtast-Frequenz wird das ganze noch offensichtlicher:



Harry Nyquist legte die mathematischen Grundlangen für die Bandbreite zur Informationsübertragung (allgemein - nicht auf Musik bezogen) Ende der 1920er Jahre, Claude Shannon formulierte daraus die *Theorie der maximalen Kanalkapazität*. Das Produkt dieser Theorie lässt sich kurz und bündig darstellen:

Die Abtastfreququenz f_{Abtast} muss größer als 2 Mal die maximal abzutastende Nutzfrequenz f_{max} sein.

Als Formel ausgedrückt sieht das folgendermaßen aus:

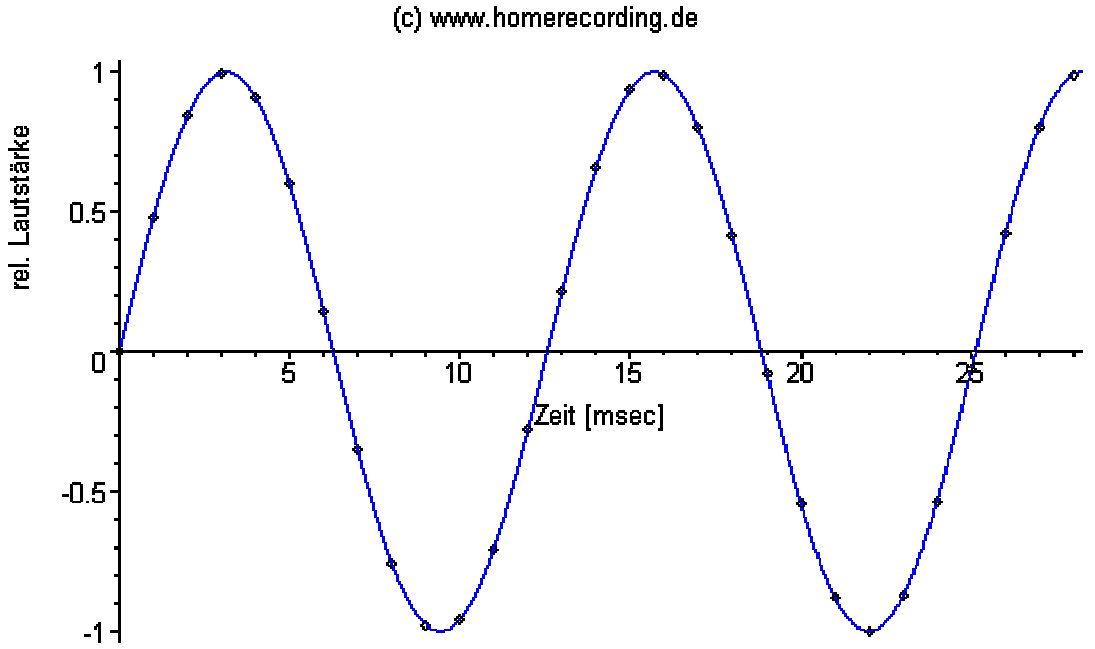
$$f_{Abtast} > 2 * f_{max}$$

Dies nennt man in der Musik die *Nyquist-Frequenz*. Oder auch wissenschaftlicher: Das *Nyquist-Shannon-Abtasttheorem*.

Unterhalb dieser Frequenz kommt es zu sogenannten Aliasing-Fehlern. Wie die sich anhören, kann man hier und hier nachvollziehen.
Es wurde bei beiden Aufnahmen ein gleitendes Sinussignal von 0 - 10 kHz aufgenommen (so genanntes *Sweeping* - von englisch *to sweep*: durchlaufen).
Beim ersten wurde mit 20 kHz abgetastet (wie es laut Nyquist sein muss), beim zweiten mit 10 kHz. Der aufgrund von Aliasing-Fehlern wiederabfallende Ton ab 5 kHz ist genau hörbar.

Doch was ist eigentlich (Polynom-)Interpolation?

Eigentlich sind im PC ja keine Linien gespeichert, sondern einzelne Punkte. Diese sind hier nur zur Veranschaulichung verbunden. Real würde es eher so aussehen:



Die meisten kennen es aus dem Physik-Unterricht, eine so genannte *Ausgleichsgerade* zu zeichnen.
Hierbei hat man es mit einer mit Fehlern behaftete Messung zu tun, bei der man versucht, die (stochastischen) Fehler zurückzuführen auf die klaren physikalischen Zusammenhänge. Die Mathe-LKler unter den Abiturienten werden Interpolation sicher auch komplexer kennen, denn hier bekommt man diverse Punkte vorgesetzt und muss eine Funktion x.ten Grades anhand dieser Punkte bestimmen. Diese Funktionen heißen Polynomfunktionen, denn sie sind aufgebaut nach dem Schema:

$$f(x)=a*x^n+b*x^{n-1}+c*x^{n-2}+...+k*x^2+l*x+m \quad \text{mit}(a,b,c....,k,l,m \in \mathbb{R})$$

Die einfachste Polynomfunktion ist die 2. Grades: $f(x)=x^2$. Diese Polynomisierung wird auch beim Signal des Tones gemacht, doch wer sich mal mit Fourieranalyse auseinandersetzt, wird feststellen, dass diese einerseits endlich ist, und zweitens es viel zu hohen Rechenaufwand benötigen würde, ein Musikstück mit drei Minuten Länge (also knapp 8 Millionen Samples) in eine Funktion zu packen. Man geht viel eher her und zerstückelt dieses Signal in Abschnitte von meist 5 bis 7 Samples, die sich als optimaler Kompromiss zwischen Länge und Berechenbarkeit herausgestellt haben und berechnet diese dann. Zur Berechnung gibt es zwei Verfahren, die sich bewährt haben. Dies ist einerseits das *Newton-Basis-Verfahren* und andererseits das *Lagrange-Verfahren*, welches auch am meisten genutzt wird. Hierauf will ich aber nicht weiter eingehen. Es sei auf tiefergehende Literatur verwiesen.

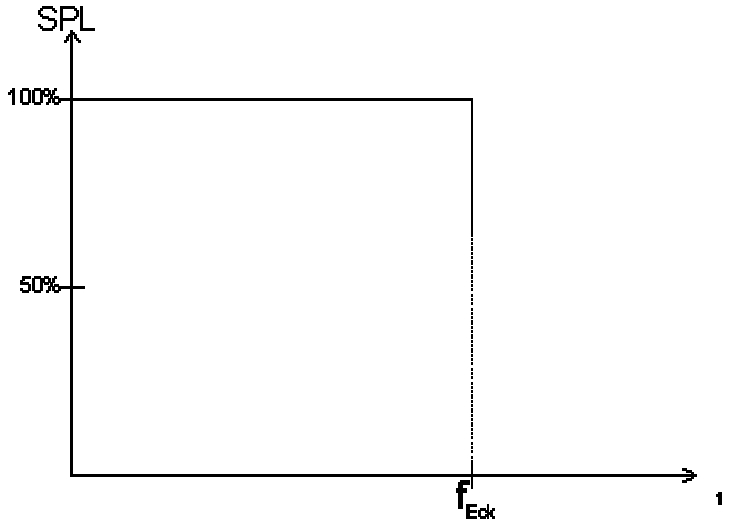
Eine Frage wird sicherlich noch offen stehen: Wenn doch gilt, dass die Abtastfrequenz zwei Mal so hoch, wie die maximal abzutastende Frequenz sein muss, warum wird dann nicht mit 40 kHz gesampelt, sondern mit 44,1 kHz? Denn wir hören schließlich nur bis 20 kHz.

Exkurs: Das 44,1 kHz-Phänomen

Dass wir nur bis maximal 20 kHz hören ist richtig - die meisten (vor allem ältere) Menschen kommen meist nur bis 17 oder 18 kHz, dann versagen die Ohren.
Jedoch werden faktisch auf der CD bis 20 kHz aufgenommen. Das Problem, warum man nun

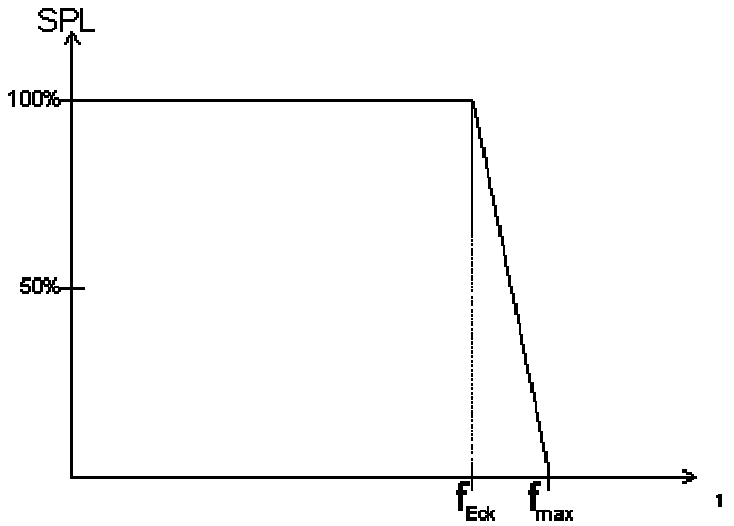
aber nicht mit 40 kHz sampeln kann liegt aber gar nicht auf digitaler Seite, vielmehr ist es ein "Problem" der Hardware. Und auch ein "Problem" der Zeit.

Um alle Frequenzen oberhalb der 20-kHz-Grenze abzuhalten, auf die CD zu kommen, braucht man einen Tiefpass. Ein Tiefpass ist ein Konstrukt aus Spulen und Kondensatoren, dass nur tiefe Töne durchlässt. Tiefe Töne sind hierbei Töne, die tiefer sind, als die Eckfrequenz (also der Frequenz, ab der abgeschnitten wird). Das sieht auf einem Graphen so aus:



Dies wäre ein idealer Tiefpass. Er lässt alle Frequenzen unterhalb f_{Eck} ungehindert durch, jedoch ab 0,1 Hz über der Eckfrequenz wird das Signal **gar nicht** mehr durchgelassen. f_{Eck} wäre gleich f_{max}.

Ein realer Tiefpass lässt dagegen nicht nur die Frequenzen unter f_{Eck} durch, sondern fängt erstens an, die Frequenzen schon unterhalb f_{Eck} in der Amplitude minimal zu beschneiden, und zweitens lässt er auch noch Frequenzen oberhalb von f_{Eck} durch. Ein realer Tiefpass sieht in etwa so aus:



f_{Eck} und f_{max} sind nun eben nicht mehr gleich, sondern f_{max} liegt etwas höher. Je nach Bauweise des Tiefpasses (und Dämpfung) ist der Faktor etwas anders; bei der Audio-CD setzt man einen Faktor von 2,2 ein. Das heißt, es gilt nun nicht mehr:

$$f_{Abtast} > 2 * f_{max}$$

sondern

$$f_{Abtast} \approx 2{,}2 * f_{max}$$

Nun sind wir von 40 kHz Sampling-Frequenz schon bei 44 kHz Sampling-Frequenz gelandet. Aber woher kommen die 100 Hz mehr? Diese Frage kann nicht mit letzter Sicherheit beantwortet werden, denn hierüber streiten sich die Gelehrten. Einige behaupten, dies sei nur ein schlechter Scherz des entwickelnden Ingenieurs gewesen. Andere stellen aber zwei ganz plausible Theorien auf, die ich auch als wahrscheinlicher halte.

1. Ich sprach oben die 70er Jahre an, in der CD "erfunden" wurde. Bauteile (also Spulen und Kondensatoren) waren um einiges ungenauer als heute. Es spricht vieles dafür, dass man eine Reserve von etwa 0,5% (diese nutzt man oft bei Bauteilen) von f_{Eck} hinzufügte, damit man auch wirklich sicher sein konnte, dass es keinen Aliasing-Fehler gibt. 0,5% von 20 kHz sind eben die 100 Hz Differenz.

2. (und das ist die wahrscheinlichste Theorie) Die ersten Digitalaufnahmen wurden mit Videorecordern durchgeführt, wo man die Samples in die Bildzeilen schrieb. Schreibt man nun in eine Bildzeile 3 Samples und nutzt die 588 Zeilen (576 aktive Bildzeilen und 12 für Videotext, VPS, Testsignale, Pay-TV-Decoderinformation und Bildsynchronisation) pro Bild, kommt man bei 25 Bildern pro Sekunde auf eine Abtastung von 44100 Samples pro Sekunde.

Bei DVDs hat man einen Faktor von 2,4, sodass die Abtastfrequenz bei 48 kHz liegt.

Nun hat man sich ja mit der Zeit nicht damit begnügt, bei den 44,1 kHz bzw. bzw. 48 kHz zu bleiben - die nächsten Schritte waren 96 kHz bzw. noch relativ neu die 192 kHz Sampling-Frequenz (was von den 48 kHz aus immer eine Verdopplung ist). Warum macht man diese Schritte?

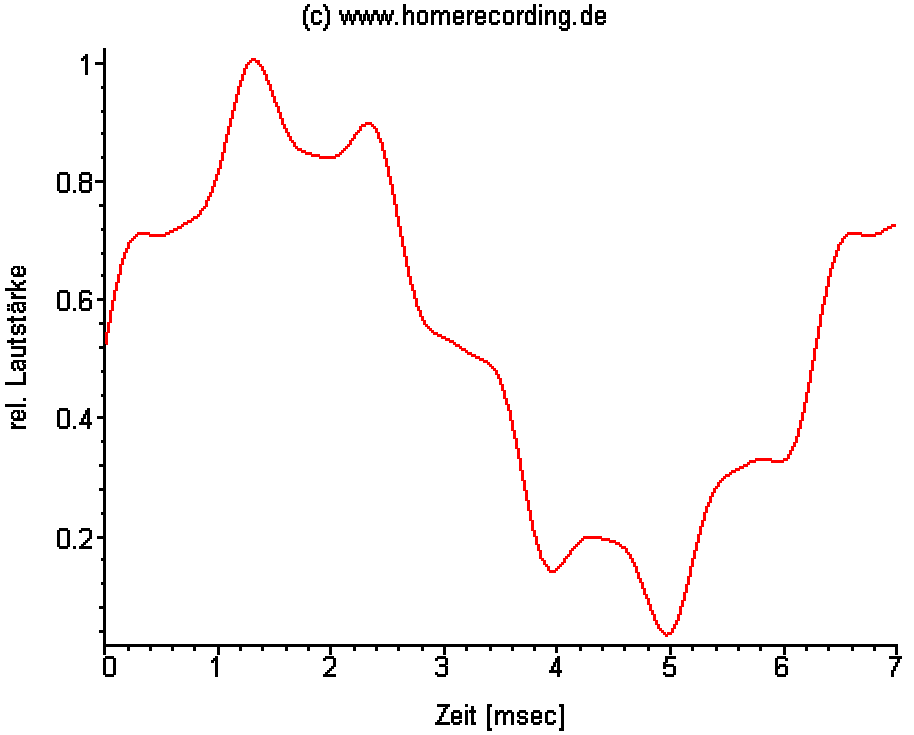
Dies hat mehrere Gründe aus zwei Richtungen: Einerseits tontechnisch-musikalisch und andererseits wirtschaftlich. Letzterer ist relativ einfach beantwortet: Irgendwann hat jeder ein Audiointerface zu Hause, das 48 kHz Sampling-Frequenz bietet - oder auch wirtschaftswissenschaftlich ausgedrückt: Der Markt ist gesättigt! Die verkaufende Industrie kann sich nur dann ein stetiges Einkommen garantieren, wenn sie sich in regelmäßigen Abständen einen neuen Markt schafft. Dieses Phänomen ist bei der ganzen Entwicklung, die in letzter Zeit passiert, absolut nicht zu vernachlässigen!

Der tontechnisch-musikalische Grund rührt aus dem stetigen Vergleich der Digitaltechnik mit der "guten alten" Analogtechnik. Man ist sich heute im Klaren, dass 16 Bit und 44,1 kHz keine Klangverbesserung gegenüber der Analogtechnik gebracht haben (einzig auf Konsumentenseite hat die CD einen Vorteil gegenüber der LP gebracht). Doch warum erhofft man sich durch *Oversampling* Klangverbesserungen?

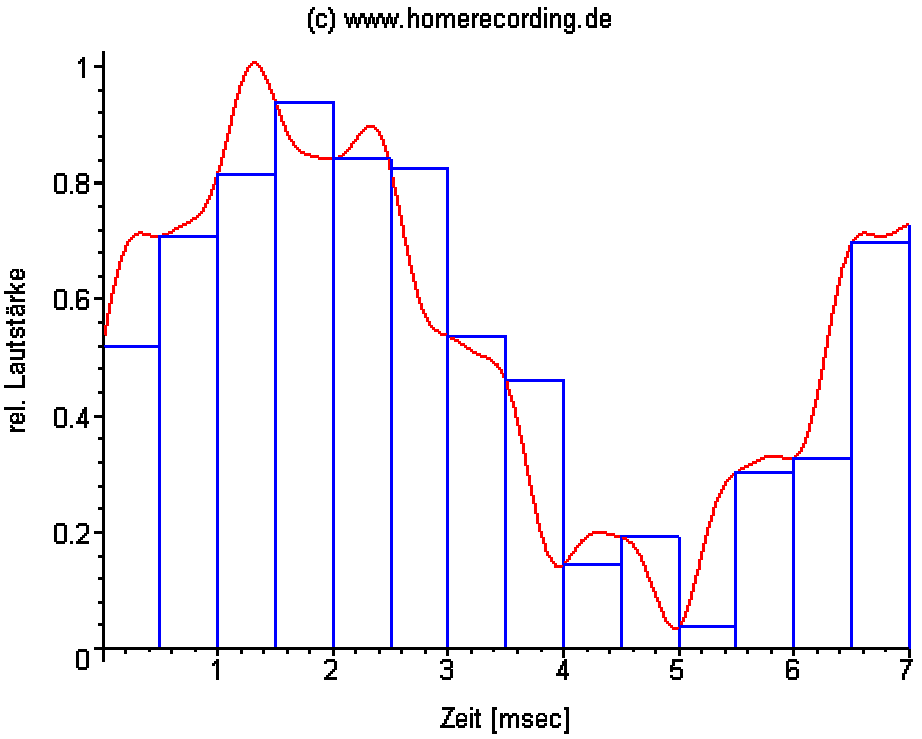
Exkurs: Oversampling

Oversampling heißt zu deutsch *Überabtastung* und beschreibt das Abtasten mit einer Frequenz, die größer als die aus dem Nyquist-Shannon-Therorem resultierende notwendige

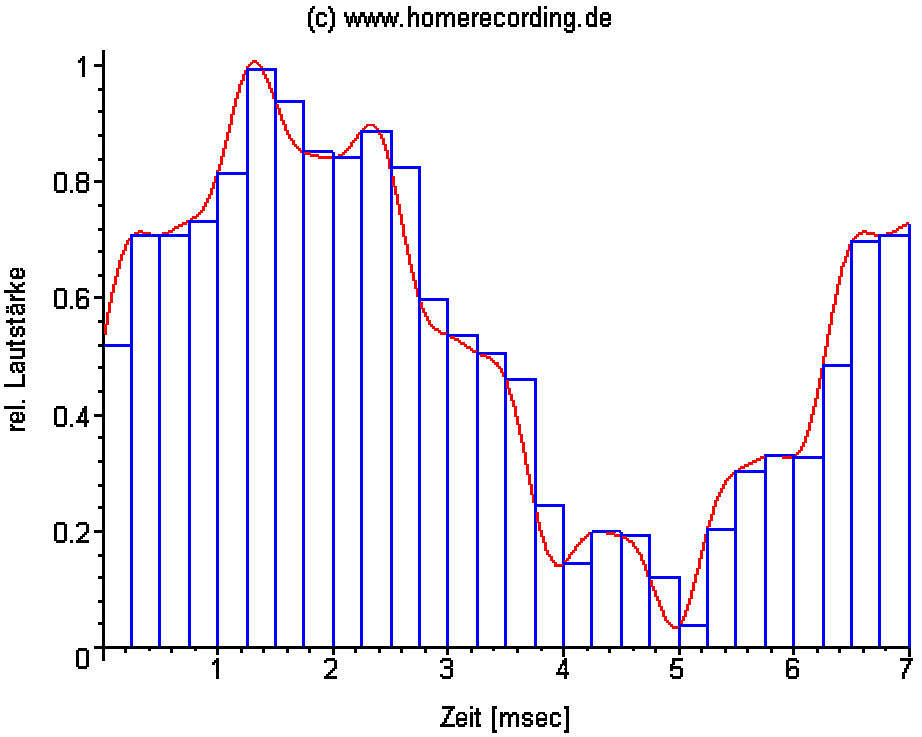
Abstastfrequenz f_{Abtast} ist. Warum das sinnvoll sein kann, lässt sich am besten wieder an Graphen darstellen. Es sei folgendes abzutastendes Signal, dass sich aus drei Einzelsignalen der Frequenzen von ca. 160 Hz, 960 Hz und 1750 Hz zusammensetzt:



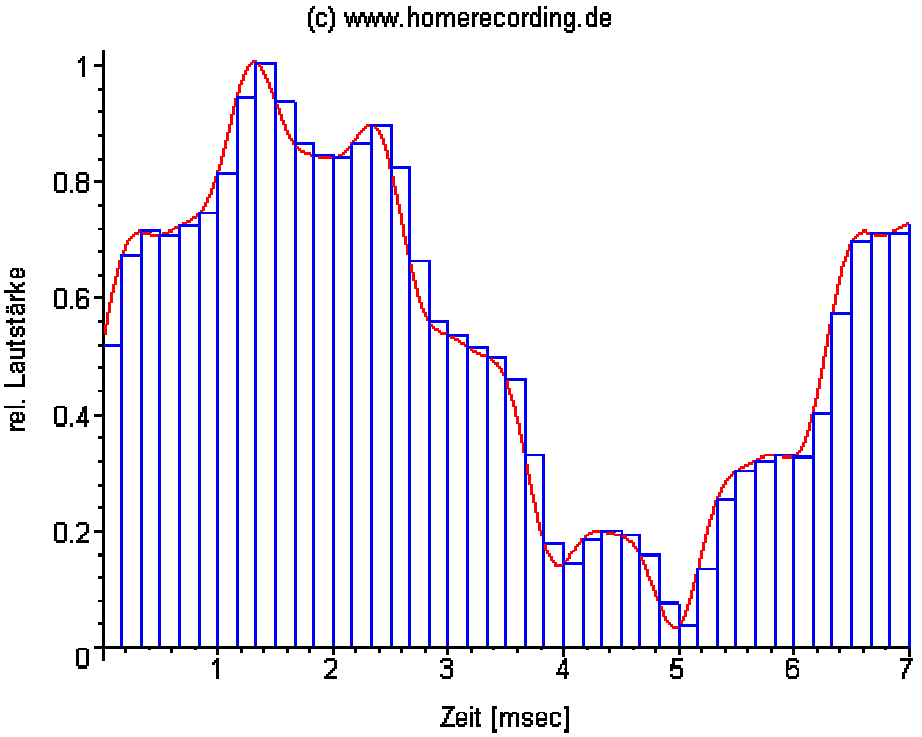
Tasten wir nun mit 2000 Hz ab, so können wir das Signal nur unzureichend erfassen.



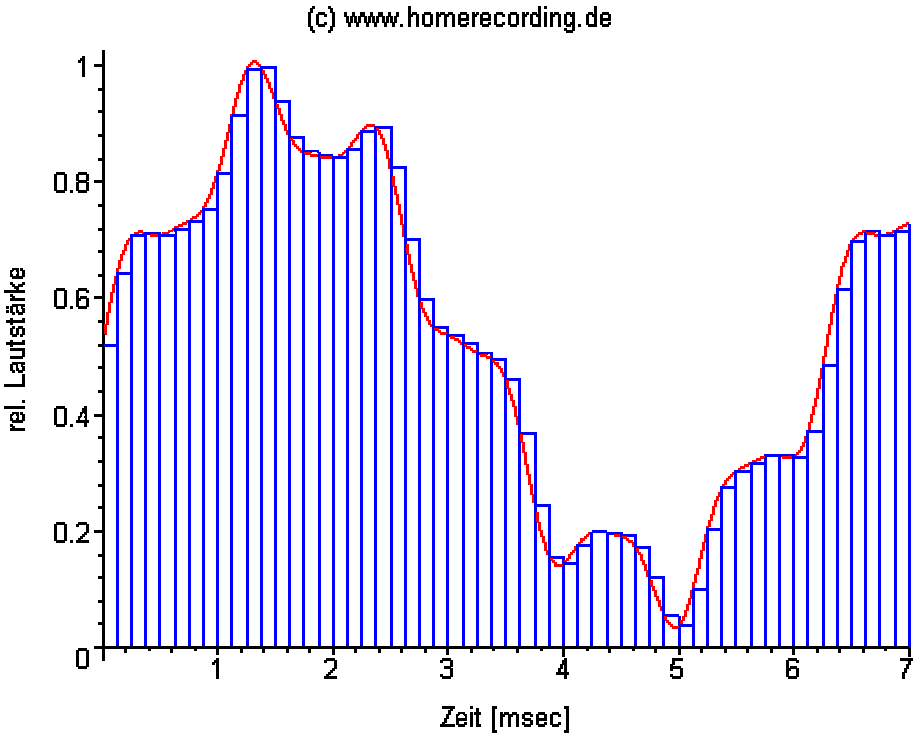
Dies ist aber auch nicht verwunderlich, denn nach Nyquist müssen wir mit mindestens 3,5 kHz abtasten, um auch das höchstfrequente Signal, das vorkommt (also 1,75 kHz) zu erfassen. Wir verdoppeln nun die Sampling-Frequenz auf 4 kHz.



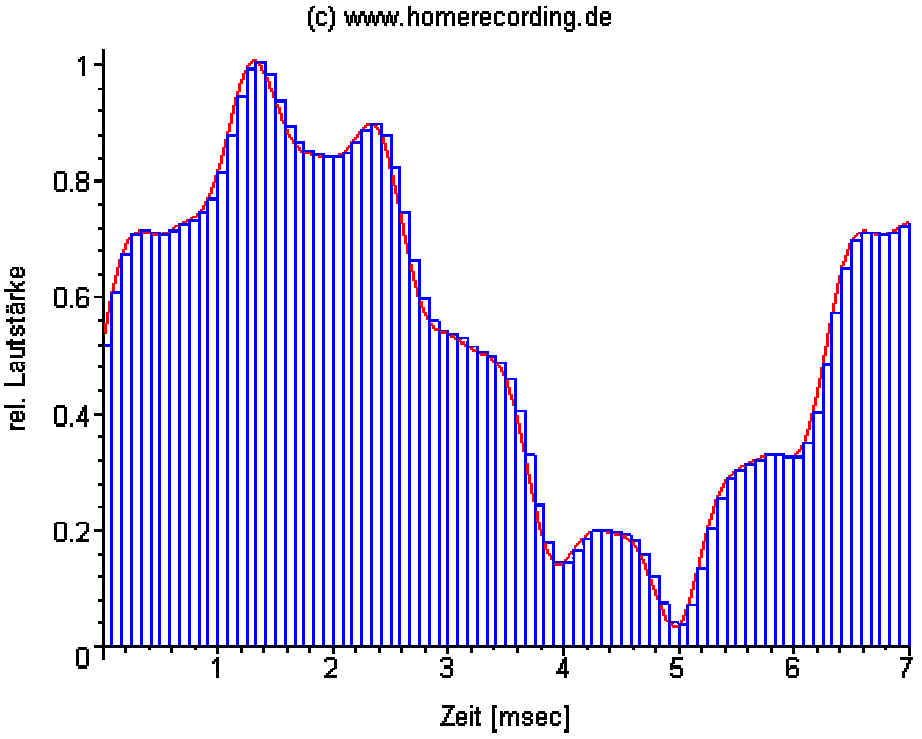
Man sieht, das Signal wird jetzt so erfasst, dass alle Peaks rekonstruiert werden können. Gehen wir nun aber sogar auf 6 kHz Sampling-Frequenz:



Das Signal wird noch genauer abgetastet. Gleiches sieht man auch bei 8 kHz Abtastfrequenz:



Oder gar bei 12 kHz:



Man stellt also fest: **Je höher die Sampling-Frequenz, desto besser die Auflösung des Signals**. Aber auch: **Je höher, die Sampling-Frequenz, desto höher der Speicher-Bedarf**, denn da wo mit 4 kHz ein Sample erfasst und damit gespeichert wurde, sind es bei 12 kHz auf einmal drei Samples!
Man die Abtastung also als Kompromiss zwischen Auflösung und Speicherbedarf sehen.

Doch warum braucht man höhere Sampling-Frequenzen? Sicherlich reichen für das Mischen mehrerer digitaler Signale (also Spuren) 44,1 kHz, wendet man jedoch stark modulierende

Effekte an (z.B. Chorus auf die Gitarre), stellt man fest, dass eben durch diese Modulation die 44,1 kHz nicht mehr wirklich reichen - man hört einen Qualitätsverlust, je nach verwendetem Plug-In und dessen Algorithmus.

Wichtig: Alles bisher geschriebene kann man auch auf die Wortbreite beziehen: Ein Signal mit 24 Bit ist viel besser aufgelöst als ein 16 Bit-Signal (um genau zu sein, 256 Mal so genau), erfordert aber eben auch 1,5x so viel Speicherplatz.

Wir fassen noch einmal zusammen:

Das als Spannung ankommende Signal wird in regelmäßigen Abständen abgetastet. Hierbei sind zwei Größen wichtig: 1. Die Sampling-Frequenz in Hz ("wie oft?") und 2. die Wortbreite in Bit ("wie genau?"). Nun liegen die einzelnen Momentanwerte der Spannungen als Zahlenstring vor.
Aber was passiert mit denen? Die müssen doch noch "verpackt" werden, damit der Computer auch weiß, von wann welche Daten sind und nicht nach der 65.536 Ziffer einfach abschneidet, obwohl das Signal mit 24 Bit noch weitere 16,71 Millionen von den 16,77 Millionen Ziffern enthält.

Und dieses "Verpacken" passiert auch noch. Nur leider nicht einheitlich. Denn die Art der Verpackung hängt zuerst vom Wanderprinzip ab.

Exkurs: Wandlerprinzipien

Es gibt zwei sich grundsätzlich unterscheidende Wandler-Prinzipien: Das erste nennt sich PCM und steht für *Pulse Code Modulation*. Der AD-Wandler erstellt hierbei ein vollständiges Datenwort von *n* Bit. Hiermit arbeiten zum Beispiel das WAVE-Format unter Windows oder das AIFF-Format beim Mac.

Dieses PCM-Verfahren unterteilt sich noch einmal in das *Flash-Verfahren* (jedes Sample wird so lange mit Referenzspannungen verglichen, bis sich die richtige gefunden hat) und das *Wägeverfahren*, das dem Newton-Verfahren zur Bestimmung von Nullstellen sehr nahe kommt. Hier wird das höchste Bit der Referenzspannung ermittelt, die noch vom Signal überschritten wird, wobei eine Differenz zwischen der realen Spannung und der ermittelten bleibt. Diese Differenz wird genauso ermittelt, wie beim 1. Schritt. Da es beim Flash-Verfahren bis zu 2^n Vergleichoperationen braucht, ist dieses für n = 16 Bit nur noch bedingt sinnvoll, für n > 16 Bit ist es absolut unrentabel.

Das zweite grundlegende Verfahren ist letzten Endes nur bedingt ein eigenständiges Verfahren - denn auch stehen am Anfang bzw. Ende ein PCM-Signal. Jedoch werden bei dem so genannten *Sigma-Delta-Wandler* keine vollständigen Datenwörter geschrieben, sondern es wird nur mit 1-Bit-Verleichsoperationen gearbeitet. Es wird erfasst, ob das neue Sample höher (1) oder kleiner (0) ist. Dies geschieht wesentlich schneller als das PCM-Verfahren. Eine Weiterentwicklung dessen ist das *Direct-Stream-Digital-Signal*, das Sony und Philips für die SACD erfunden haben. Hier wird die Information auch durch Sigma-Delta-Vergleichsoperationen gewonnen, aber auch als solche gespeichert. Die Samplerate für eine SACD liegt übrigens bei 2822,4 kHz.

Die meisten Soundkarten und auch professionellen AD/DA-Wandler nutzen eines der PCM-Verfahren, wobei das Wägeverfahren den größeren Anteil haben dürfte. Auf einer Audio-CD

werden diese Signale einfach hintereinander gesetzt. Rechnet man das hoch, stellt man fest, dass man bei 16 Bit und 44,1 kHz Sampling-Frequenz einen Datendurchsatz von:

$$\frac{16\text{Bit} * 44100\frac{1}{s}}{8\frac{Bit}{Byte} * 1024\frac{Byte}{kB}} \approx 86{,}13\frac{kB}{s}$$

nur für die Daten der Musik hat. Das sind ziemlich genau 5 Megabyte pro Minute **pro Kanal**. Bei einer Stereo-Datei kommt man damit auf 10,1 MB pro Minute und damit passen auf eine 700 MB fassende CD auch etwa 70 Minuten Musik. Gleiche Berechnung gilt auch in etwas für eine WAV-Datei, jedoch werden hier Daten wie Sampling-Rate, Wortbreite und Länge noch vorne in die Datei geschrieben – so dass diese einige Byte größer ist.
Eine Mono-WAV-Datei der Länge einer Minute, die mit 24 Bit und 96 kHz aufgenommen wurde, hat demnach eine Größe von:

$$\frac{24\text{Bit} * 96000\frac{1}{s}}{8\frac{Bit}{Byte} * 1024\frac{Byte}{kB} * 1024\frac{kB}{MB}} * 60\,s \approx 16{,}48\,MB$$

Als allgemeine Formel gilt für die Größe einer WAV-Datei in Kilobyte: **Wortbreite x Sampling-Frequenz x Länge [in sec] / 1024**

Diese Formel gilt **nicht** für mp3 oder andere verlustbehaftete Formate! Diese kann man nur dann berechnen, wenn sie eine konstante Bitrate haben (z.B. 128 kBit/s). Dann gilt für die Größe in kB: **Bitrate x Länge [in sec] / 1024**.

Nun gibt es aber neben der Speicherung auf dem PC noch andere Datenarten - meist zur Übertragung genutzt. Ich möchte hier kurz die beiden professionellen Übertragungsformate SDIF-2 und AES/EBU vorstellen, und in Anlehnung an das AES/EBU-Format die semi-professionelle Variante S/PDIF.

Bevor ich damit anfange, die einzelnen Formate abzuhandeln, möchte ich aber noch ein paar Worte zum Thema *Wordclock* verlieren.

Exkurs: Wordclock

Das aus dem englischen Sprachgebrauch kommende Wort *Wordclock* setzt sich aus zwei Teilen zusammen: *word* und *clock*. Der erste Teil bezieht sich auf das Datenwort. Wir hatten ja eben festgestellt, dass die Daten eines Samples immer gemeinsam "gepackt" versendet werden. Hinzu kommen, je nach Format, noch weitere Daten. Ein solcher Datenblock nennt sich (Daten-)Wort.
to clock heißt im englischen takten und damit ist auch vom Prinzip klar, was eine Wordclock macht: Sie taktet das Signal.

Aber warum ist das nötig?

Im binären System werden bekanntermaßen nur Einsen und Nullen gesendet - das heißt, das Verarbeitungsgerät muss wissen, wann ein Wort aufhört und das nächste anfängt. Innerhalb eines Computers sind die Daten so kodiert, dass garantiert wird, dass die Wörter immer ganz

ankommen. Extern klappt das aber nicht. Also wird ein externes Kabel genommen, das nur die Wordclock transportiert. Das heißt, bildlich gesprochen: Jedes Gerät, das angeschlossen ist, weiß: "Aha, jetzt ist das Wort fertig!"

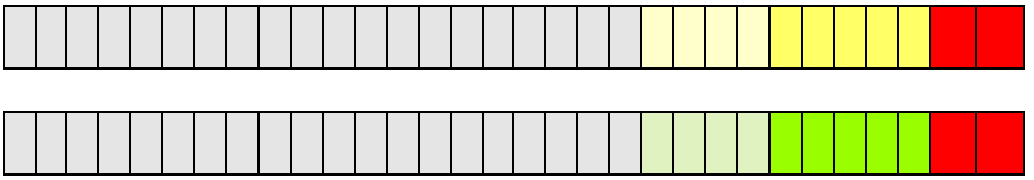
Es gibt auch Signale, die *selbsttaktend* sind. Dies heißt, dass die Information, dass ein Wort zu Ende ist, schon in der Logik des Wortes selber mit transportiert werden. Das ist zum Beispiel beim AES/EBU-Format der Fall.

Im Zusammenhang mit der Wordclock sollte noch der Begriff "Jitter" erklärt werden: Als Jitter (engl. Schwankung) bezeichnet man eine ungleichmäßige Übertragung von digitalen Daten, praktisch eine Gleichlaufschwankung, wie man sie früher von Bandmaschinen oder Plattenspielern her kannte. Dieser Effekt ist natürlich unerwünscht und wirkt sich qualitativ auf das Signal selber aus. Meist entsteht der Jitter bei der direkten Verbindung zweier oder mehrerer Geräte, die digitale Audiodaten austauschen. Grund für einen Jitter können Billigbauteile oder sehr lange sowie minderwertige Kabel sein.

Das SDIF-2 Format

Dies ist wahrscheinlich das älteste digitale Übertragungsformat, das es gibt. SDIF steht für **S**ony **D**igital **I**nterface **F**ormat und wird im Stereobetrieb über drei Kabel übertragen. Im ersten wird die Wordclock gesendet, in den beiden anderen die beiden Kanäle.

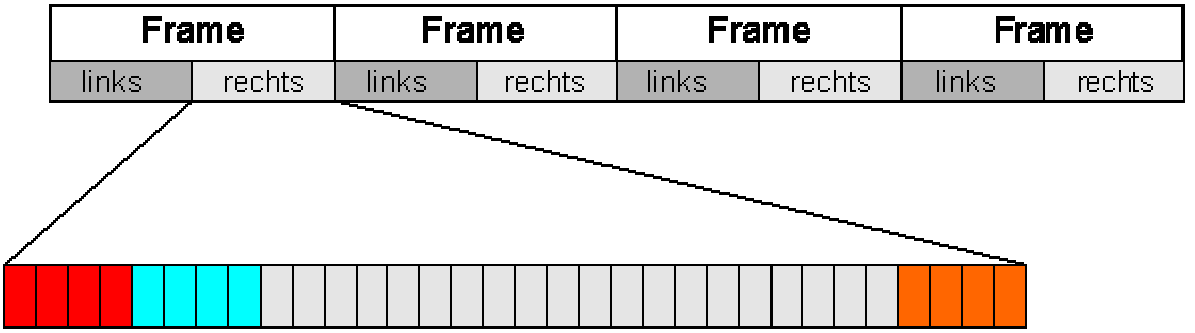
Ein SDIF-2-Wort setzt sich auf 32 Bits zusammen: 20 Bits (grau) sind für die PCM-Audio-Daten, die Bits werden folgendermaßen aufgeteilt:



Im linken Kanal (oben) werden 9 Bits als Userbits verwendet (gelb), im rechten Kanal (unten) werden 9 Bits als Kontrollbits verwendet (grün). Es ist jedoch möglich, die Bits 21-24 (die jeweils hellen) noch als Audio-Daten zu verwenden und somit ein 24-Bit-Signal zu übertragen. Die letzten beiden Bits (rot) sind jeweils anderthalb mal so lang und stellen die Syncbits dar, mit denen das verarbeitende Gerät das Wort an der Wordclock ausrichten kann.

Das AES/EBU-Format

Das AES/EBU-Format ist von der Audio Engineering Society (AES) entwickelt worden. Es ist zweikanalig und wird über XLR-Verbindungen übertragen (manchmal auch über 75-Ω-BNC-Stecker, dann heißt das AES/EBU-ID – das Datenformat ist aber das gleiche). In jedem AED-Datenblock, dieser heißt Frame, gibt es zwei so genannte Subframes. Diese tragen die Daten der Kanäle:



Die ersten vier Bits (rot) sind die Syncbits, die zweiten vier (blau), die Auxbits. Diese können neben den 20 Standard-PCM-Datenbits (grau) zusätzlich als Bits zur PCM-Übertragung genutzt werden. Die letzten vier Bits (orange) enthalten Kontroll- und Userdaten. Im Gegensatz zum SDIF-2-Format hat das AES/EBU-Format eine genormte Belegung für 5.1-Ton. Hierbei werden drei Kabel benutzt, die Belegung der Kanäle ist: L-R, C-Sw, Ls-Rs.

Das S/PDIF-Interface

S/PDIF steht für **S**ony/**P**hilips-**D**igitial-**I**nterface-**F**ormat und ist eigentlich ein semiprofessioneller Ableger des AES/EBU-Formates, wird aber mittlerweile öfter genutzt als das professionelle Vorbild. Mittlerweile hat fast jede Soundkarte und jeder DVD-Player eine S/PDIF-Schnittstelle, sei es nun der gelbe Cinchstecker (der gerne mal mit dem gleich aussehenden Video-Stecker einer Composite-Buchse verwechselt wird) oder der als Toslink bekannte Lichtleiter, beide nutzen sie das S/PDIF, dessen Aufbau keine großartigen Differenzen zum Vorbild birgt; lediglich einige Statusbits (orange bei AES/EBU) sind anders, denn das S/PDIF kann auch Titelkennungen u.ä. übertragen. Oft kann man sogar eine direkte Verbindung zwischen AES/EBU und S/PDIF aufstellen.

Grundlegend kann man daraus ziehen, dass es eben nicht einfach geht, sich zwei Steckverbinder aneinander zu löten (wie das bei der Analogtechnik relativ einfach funktionierte) - denn die Formate werden unter Umständen gar nicht verstanden.

http://homerecording.de/modules/AMS/article.php?storyid=508

http://www.frank-schaetzlein.de/links/software-links-audio.htm