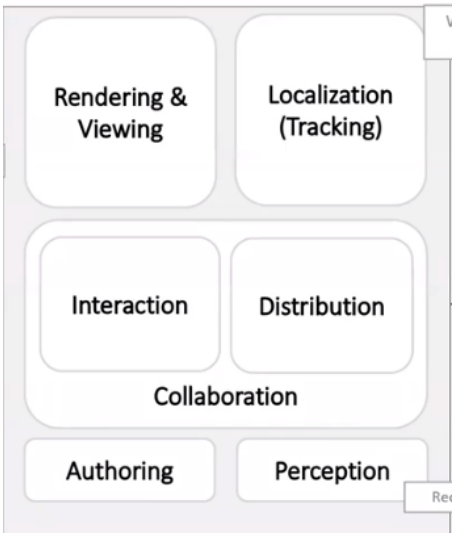
Zusammenfassung Multimedia

VO1 – Einführung

Applications: Medical & Biosensor

Components of mixed reality.



PLAYMANCER

3D-Serious Game Environment with real-time motion capturing and bio-signal feedback for physical rehabilitation

ProFiTex

Support fire fighters with mission-relevant information based on various sensor data

Definition – Was ist Multimedia und was sind Multimediasysteme

Multi – Viele

Medium – Substanz

Im Endeffekt also ein Mittel zur Übertragung von Information.

Es gibt Speichermedien, Übertragungsmedien, Informationsaustauschmedien und Präsentationsmedien

Wichtig – Perception, Representation intern, Repräsentationsräume (Bildschirm, headmounted display…), Dimensionen (3D, 2D), Zeitabhängigkeit oder Zeitunabhängigkeit (Audio und Video sind zeitabhängig und haben eine zeitliche Sequenz, nicht so Text/Grafik) Bei zeitabhängigen ist die Verarbeitungszeit processing time kritisch.

4 Charakteristika

1. Multimedia Systeme sind computer gesteuert
2. Multimedia Systeme sind integriert (Man kann also in dem system Daten verarbeiten)
3. Die Informationen sind digital vorhanden
4. Das Interface der finalen Präsentation ist interaktiv (Also nicht nur zum Anschauen)

Definition Multimediasystem

> *A multimedia system is characterized by computer-controlled, integrated production, manipulation, presentation, storage and communication of independent information, which is enc least through a continuous (time-dependent) and a discre (time- independent) medium.* (Steinmetz/Nahrstedt)

Beispiele Multimediasysteme

Flexibilität

Man kann verschiedene Arten von Medien abspielen, ein Videorecorder ist nach dieser Definition also kein MS

Integration

Es gibt einen unabhängigen Medienspeicher

Ermöglichst computergesteuerte Medienkombination

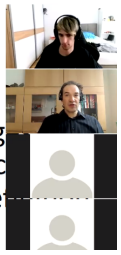
Definition

A multimedia system is characterized by the integrated computer-controlled handling of independent discrete and continuous media.

Heutige Definition Multimedia

Multimedia is often used as an attribute of systems, products, etc., without satisfying the characteristics above. Thus, two notions of MM can be distinguished:

– MM, strictly speaking (by definition above)
– MM, in the general sense
  – In common usage, multimedia refers to an electronically delivered combination of media including video, still imag audio, and text in such a way that can be accessed interac Much of the content on the web today falls within this de as understood by millions.

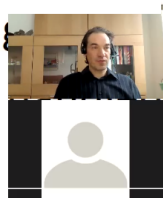


Heutige Definition Multimedia Device

an electronic device, such as a smartphone, a videogame system, a stand-alone VR HMD, or a computer, for example.

these devices have one main, but also other functions beyond their intended purpose, such as reading, writing, recor video, streaming listening to music, and playing video

while previous media was always local, many are now through web based solutions, particularly streaming.



Menschliche Sinne

Morton Heillig baut 1955 Sensorama und schriebt Paper darüber. Er legt fest, dass ca.:

Vision	~ 70%
Hearing	~ 20%
Smelling	~ 5%
Tasting	~ 4%
Touch/haptic perception	~ 1%

Diese Werte gelten.

Das Problem für die Medien über Vision und Hearing hinaus, ist, dass wir keine Standards für die digitale Erzeugung und Wahrnehmung haben.

Text

Es gibt verschiedene Repräsentationen wie ASCII / ANSI / ISO / UTF character sets

Marked-up text wie SGML, HTML, XML…

Hypertext wie das WWW

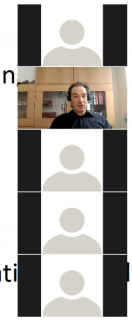
1963 erster Standard (ASCII) 7bit per character = 128 characters

Später dann ANSI, da gab es dann 256 characters, da konnte man dann zusätzlich cursor movements definieren. (Also dass der cursor runter gemoved wird)

Dann UNICODE UTF8 1 byte für standard English, auch abwärts kompatibel.

Textoperationen

- character and string operations
- editing
- formatting — interactive, non-interactive, page description lan
- pattern matching and searching — indexes, signatures
- sorting
- compression — Huffman, Lempel-Ziv
- encryption
- language specific operations — spelling checking, parsing, stati lysis of writing style
- conversion to other media types (text to speech, semantic analysis to generate images or 3D environments)



Audio

Schall ist die wellenförmige Ausbreitung von Druckwechseln in elastischen Medien. Die Medien können Luft, Wasser … sein.

Die einfachste Darstellung sind Sinusschwingungen. Außerdem kann man jede Schwingung durch Kombination von mehreren Sinusschwingungen erzeugen.

Dafür sind Frequenz, Periode und Amplitude wichtig.

rived:

Frequency f: Number of oscillations per second

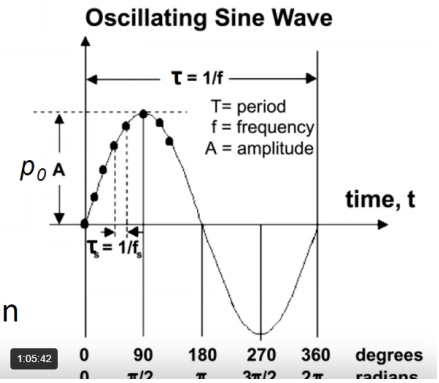
Unit: Hertz [Hz].

Period τ: Duration of one oscillation **τ** =1/f

Unit: second [s]

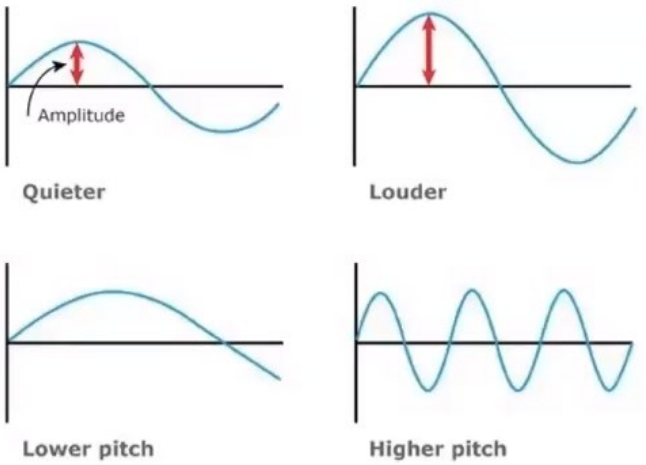
Amplitude p_0: Pressure value at maximum compression

Unit: Pascal [Pa]



Hohe Frequenz = Hoher Ton.

Hohe Amplitude = Lauter Ton



Die Wellenlänge ist die Distanz die in einer Periode zurückgelegt wird (in M)

Wavelength λ: the distance travelled during one period.
Unit: [m]

$$\lambda = s * \tau = s / f$$

Speed of sound s: depends on the propagation medium: Unit: [m/s]

Medium	s [m/s]	λ [m] 16Hz	λ [m] 100Hz	λ [m] 4 kHz	λ [m] 20kHz
Air (0°)	331,0	20,68	3,31	0.08	0,0166
Air (20°)	343,6	21,48	3,44	0,09	0,0172
Water (20°)	1484	92,75	14,84	0,37	0,0742
Brick	3650	228,13	36,50	0,91	0,1825

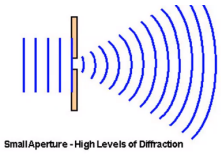
Jede Verdopplung der Frequenz bedeutet für ZuhörerInnen eine Oktav Erhöhung. Das heißt natürlich, dass die Verschiebung nicht linear ist. Von 100 auf 200 Hertz ist eine Oktave, aber auch von 20.000 zu 40.000.

Die Amplitude ist die Lautstärke.

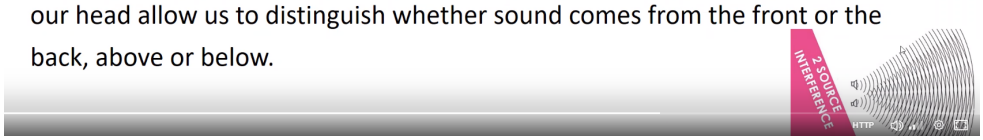
Effekte

Biegung/Reflektion und Interferenz

– Diffraction: Waves deviate from their straight path and "peek" around the corner. Increased noticeable when wavelength is greater than obstacle ("listening around the corner").

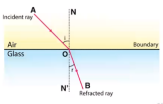


– Interference: Superposition of the direct sound wave and indirect waves created by diffraction and reflection. Interference effects that take place in the area of our head allow us to distinguish whether sound comes from the front or the back, above or below.



Brechung und Dispersion

Refraction: Sound waves change direction when they travel through areas with different sound velocities.

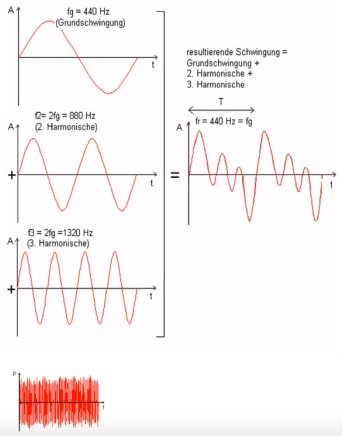


Dispersion: For different pitches, the strength of refraction is different. This effect is called dispersion.

Grundton zu Überlagerung

- Tone:

single sine wave is called a tone.
(t) is sinusoidal function



- Sound:

ound is created from a superposition of undamental tone (heard as pitch) and vertones (f=n*fg) (heard as timbre).
(t) function general periodic function with requency fg.

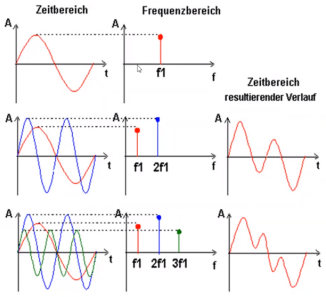
- Noise:

No more fundamental recognizable, aperiodic p(t) function

Ein Grundton ist einfach ein einzelner Ton, eine Welle. Spielt man aber ein Instrument, so hat man einen Sound. Dieser ist zusammengesetzt aus mehreren Grundtönen die sich überlagern. Ein Geräusch hingegen ist ungeordnet und unerkennbar.

- Frequency spectrum:

ery periodic oscillation of any aveform (= sound) can be represented a superposition of fundamental and rmonics. If the amplitudes of all the cillations involved are plotted in a agram as a function of frequency, the so-called frequency spectrum is obtained



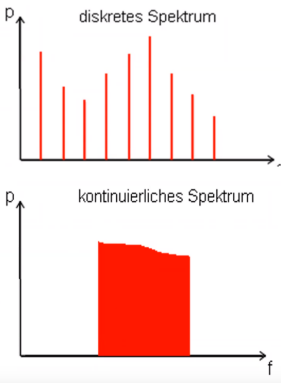
Ein Frequenzspektrum zeigt dann die Historie der höchsten Frequenzen der einzelnen Wellen.

Diskrete Spektren sind dabei nur an einzelnen Punkten abgegriffen. Kontinuierliche Spektren sind hingegen unendlich Werte auf einer Zeitachse.

Discrete spectrum: Sounds contain only ration frequencies that have an integer ratio damental frequency.

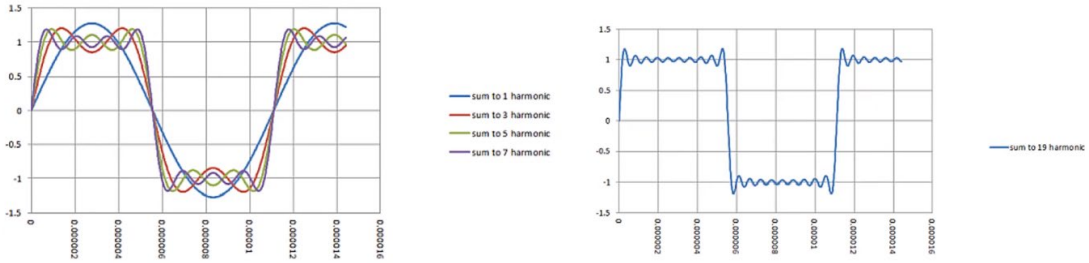
tinuous spectrum:

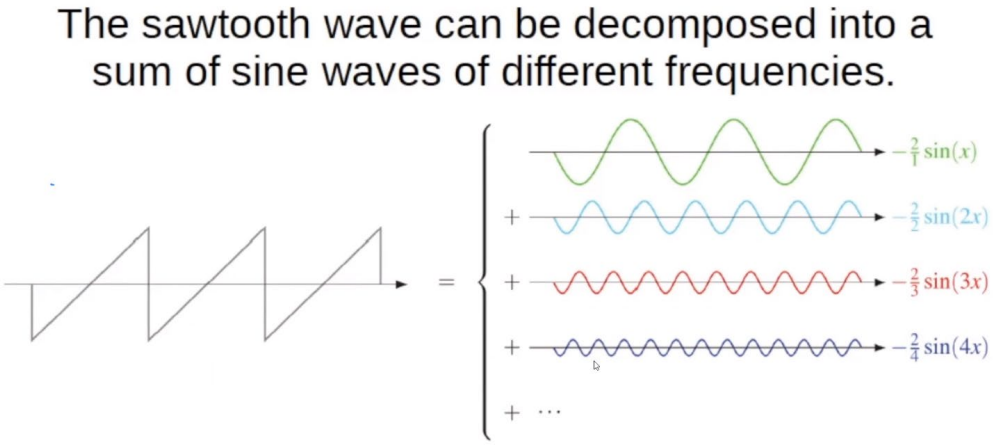
neral sound events (noises) contain an infinite nber of individual oscillations whose quency values extend continuously along the x-axis. The result is a continuous mathematical function: p=p(f)



Sinuswellen

Jede periodische Schwingung lässt sich durch mehrere Sinusschwingungen erzeugen. Das hat Fourier festgestellt, eine Zusammensetzung, also die Summe, aus mehreren Schwingungen heißt deswegen Fourierserie oder Fourierreihe.



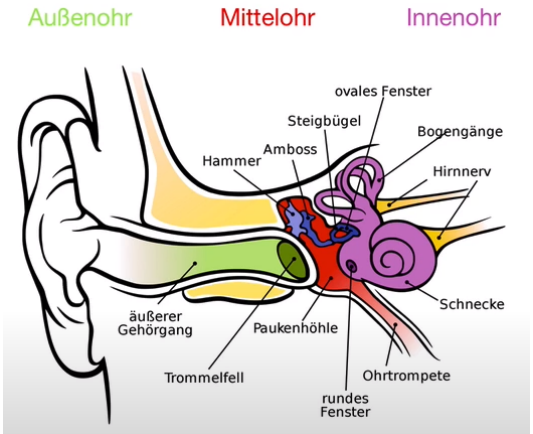


Diskrete Fouriertransformation

Hier wandelt man ein Signal in die Frequenzen um, die beteiligt an der Erzeugung des Signals sind.

VO2 – Psychoakustik

Beschäftigt sich mit der Wahrnehmung von Schallereignissen.



Der Schalldruck versetzt das Trommelfell in Schwingung. Hammer und Steigbügel dämpfen die Intensität notfalls ein bisschen. Dann geht es in die Gehörschnecke, dort befinden sich das Cortiorgan, darin Haare, die Haare werden auch in Schwingung versetzt, diese wandeln die Schwimmung dann zuerst in chemische und dann in elektrische Signale um.

Richtungshören

Beim Richtungshören sind drei Methoden relevant

interaurale Zeitdifferenzen

interaurale Schalldruckdifferenzen

Klangfarbe

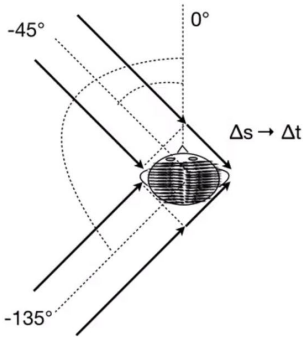
links / rechts:

- Schalldruck- und Zeitunterschiede
- Zeitunterschiede → tiefere Frequenzen
- Schalldruck → höhere Frequenzen

Aufgrund von **Klangfärbung** verursacht durch Kopf und Außenohr kann nicht nur zwischen links und rechts sondern sondern auch zwischen oben und unten und hinten und vorne unterschieden werden.

Interaurale Zeitdifferenzen

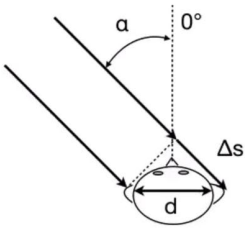
interaural bedeutet „zwischen den Ohren“



Man kann einer bestimmten Zeitdifferenz immer zwei mögliche Winkel zuordnen!

Unterscheidung: oben / unten → nicht möglich

Durch Interaurale Zeitunterschiede kann das Gehirn nur einen horizontalen Winkel errechnen, und dieser ist zweideutig.



$$\Delta s = d * \sin(\alpha)$$

$$\Delta t = \frac{\Delta s}{c}$$

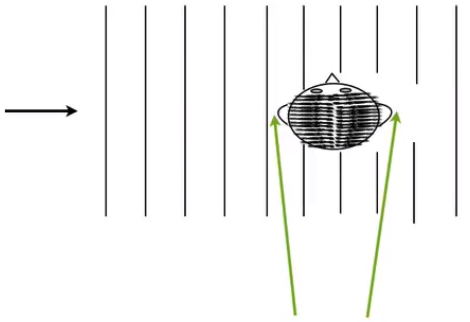
maximale Zeitdifferenzen bei 90° oder 270°
bei d= 0,22 und c= 344 → 0,63 ms

d ~ 0,17 bis 0,22 m

Die größte vorkommende Zeitdifferenz entspricht in etwa 0,63 Millisekunden. Der kleinste noch wahrnehmbare Unterschied beträgt ca. 0,03 ms, was einem Winkel von etwa 3° entspricht.

Interaurale Schalldruckdiffrenzen

Ab einer Frequenz von etwa 1600 Hz kann sich der Schall nicht mehr vollständig um den Kopf herumbeugen.



links / rechts-Unterscheidung möglich

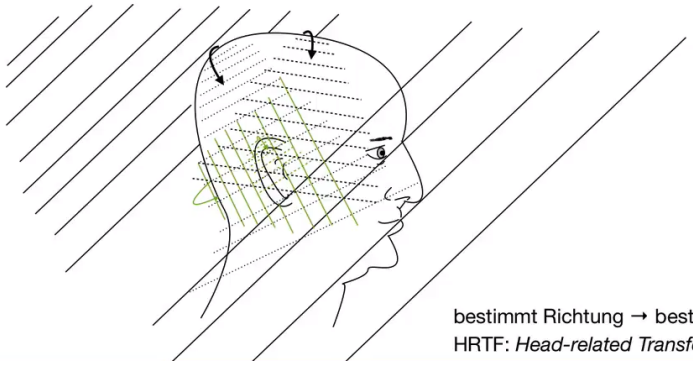
oben / unten bzw. vorne / hinten nicht

unterschiedliche mittlere Schalldrücke

Klangfarbe

Hierbei geht es vor allem um die head related transfer function. Jede Kopfform verändert das Schallereignis ein bisschen, das gilt auch für den Einfallswinkel auf den Kopf.

Durch die Beschaffenheit unseres Kopfes und vor allem der Ohrmuscheln gelangt breitbandiger Schall mit einer ganz bestimmten richtungsabhängigen Klangfärbung an das Trommelfell.



bestimmt Richtung → bestimmten Filterkurve
HRTF: *Head-related Transfer Function*

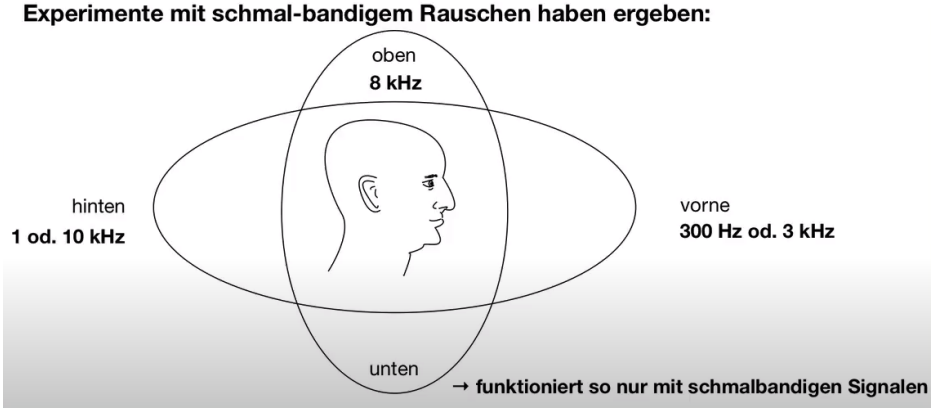
Diese Klangfärbungen nehmen wir nicht bewusst als solche wahr, sondern entnehmen ihnen Richtungsinformationen.

Wir müssen wissen, wie eine Schallquelle aus einer bestimmten Richtung klingt.

→ Lernprozess

Kinder haben ein schlechteres Richtungshören.

Experimente mit schmal-bandigem Rauschen haben ergeben:



→ funktioniert so nur mit schmalbandigen Signalen

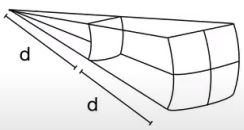
- Richtungsbestimmung ist bis zu 2 Grad genau wenn der Schall von vorne kommt
- anderen Richtungen → ca. 10 Grad

unter 100 Hz	kaum Ortung möglich	
100 Hz bis 1600 Hz	Zeitunterschiede (Phasenunterschiede)	Klangfarbe
ab 1600 Hz	Schalldruckunterschiede Zeitunterschiede (Hüllkurven)	

Entfernung

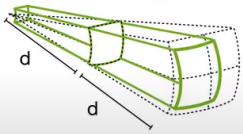
1. **Die Lautstärke:**
   Je näher, desto lauter.
   Voraussetzung ist, dass man die Schallquelle kennt und weiß, wie laut sie in einer bestimmten Entfernung sein kann.
2. **Die Klangfarbe:**

tiefe Frequenzen kugelförmig



Abstandsverdoppelung → -6 dB

hohe Frequenzen gerichtet

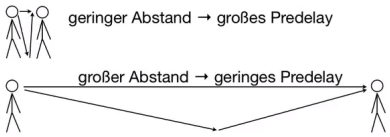


Abstandsverdoppelung → z. B. -4 dB

Je näher man an der Schallquelle ist, desto mehr Bässe hört man.
Zusätzlich: aufgrund der **Luftabsorption** klingt ein sehr weit entferntes Schallereignis dumpfer.

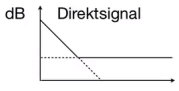
Predelay ist der zeitliche Abstand zwischen Direktsignal und erster Reflektion. Das ist bei nahen Schallquellen allgemein größer als bei Entfernten. Man hört also das ursprüngliche Signal und dann kurz danach, aber deutlich getrennt den Hall. Bei weit entfernten Schallquellen sind Hall und Ursprung schon sehr nah beieinander.

3. Das Predelay:



ca. 1 bis 10 ms

4. Hallanteil:



zunehmender Abstand → mehr Hall

5. Erfahrung:

z. B.: Flüstern immer näher als Schreien

Das Gehörte wird in Bezug auf Klangfarbe und Lautstärke immer mit einem Erfahrungswert verglichen.

Entfernung:

- Lautstärke: laut nah, leise fern
- Klangfarbe: viele Bässe nah, wenig Bässe fern; hell nah, dumpf fern
- Predelay: lang nah, kurz fern
- Hallanteil: trocken nah, hallig fern
- Erfahrung

Tonhöhen

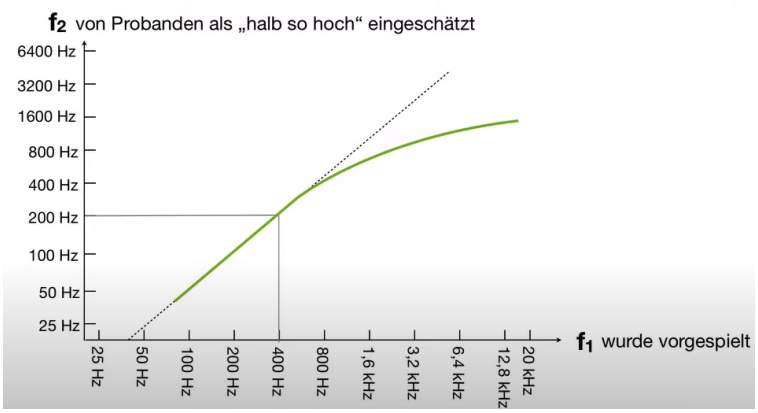
Wahrgenommene Tonhöhe hängt maßgeblich von der Frequenz eines Sinustones bzw. der Grundfrequenz einer komplexen Schwingung ab.

20 Hz bis 20 kHz

Verdoppelung der Frequenz → Oktave höher

Halbierung der Frequenz → Oktave tiefer

Das stimmt aber nur in einem gewissen Frequenzbereich. Für sehr hohe Frequenzen stimmt diese Regel nicht mehr.



Ab 500 Hz stimmt es nicht mehr.

Einheiten – Mel und Bark

Frequenz:

- physikalisch Größe
- Schwingungen pro Sekunde
- [Hz]

Tonhöhe:

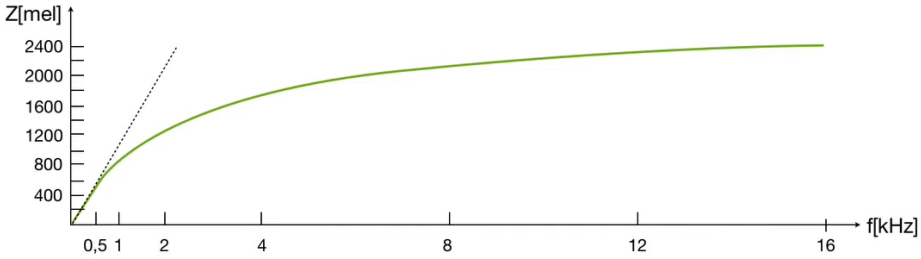
- psychoakustische Größe
- wie hoch wir etwas empfinden
- meist ebenfalls in Hz
  oder als Notenname angegeben

Tonheit:

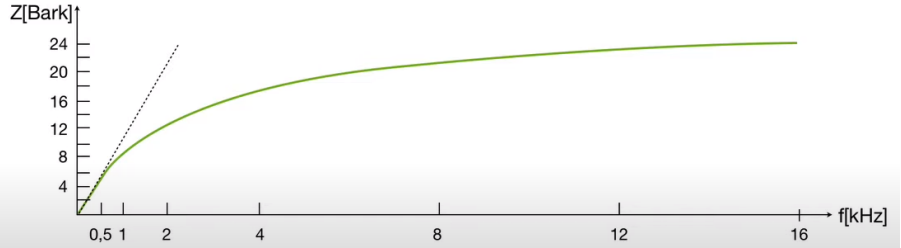
- psychoakustische Größe
- [mel]

Die Tonheit ist eine empfundene Größe

Der gesamte Hörbereich wird auf einer Skala von 0 bis 2400 Mel angegeben.
125 Hz = 125 Mel
Bis 500 Hz sind Mel und Hz identisch.



100 Mel = 1 Bark



Empfindung von Tonhöhenunterschieden

gerade noch wahrnehmbarer Tonhöhenunterschied

JND ... **j**ust **n**oticable **d**ifference

bis 1 kHz: JND = ca. 1 Hz

ab 1 kHz: Fähigkeit Tonhöhen zu unterscheiden nimmt ab

ab 5 kHz: Fähigkeit Tonhöhen zu unterscheiden nimmt dramatisch ab

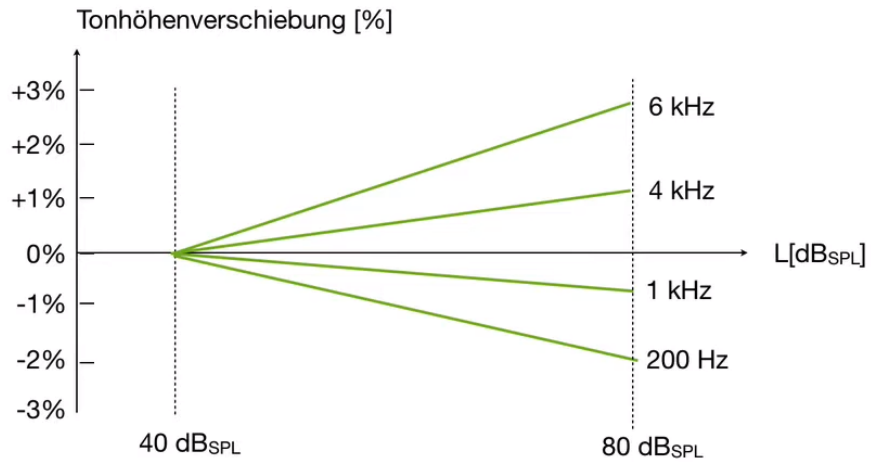
ab 10 kHz: es ist kaum noch möglich Tonhöhen zu unterscheiden

darum endet beispielsweise die Kirchenorgel auf C8 (ca. 4 kHz)

Unschärfeprinzip

Je kürzer der Ton, desto schwieriger seine Tonhöhe wahrzunehmen.

Schalldruck und Tonhöhenwahrnehmung



Je lauter desto klarer sind Tonhöhen unterscheidbar.

Maskierung

Der Maskierungseffekt bedeutet, dass man einen Ton schlechter oder nicht hört, weil ein andere ihn überdeckt.

Maskierung bedeutet, dass man einen Ton nicht oder schlechter hört, weil er von einem zweiten verdeckt wird

→ später mehr über Maskierung

Wird ein Ton von einem anderen maskiert, so bewegt sich die Tonhöhenwahrnehmung etwas von dem Maskierer weg

- wird ein Ton von einem höheren Ton partiell maskiert so wird er etwas tiefer wahrgenommen
- wird ein Ton von einem tieferen Ton maskiert, wird er etwas höher wahrgenommen

Die Verschiebung ist bei 300 Hz am größten und kann bis zu 33 Cent betragen

Tonhöhen in Klängen

Bisher immer nur Bezug zu Sinustönen, was aber wenn man Grundton und Obertöne hat? (Obertöne = ganzzahlige Vielfache)

Generell ist es so, dass, im Vergleich zu einem Sinuston gleicher Frequenz, ein zusammengestellter Klang eher tiefer wahrgenommen wird. Das liegt daran, dass die Obertöne den Grundton maskieren.

Residual-Effekt

wenn der Grundton eines Klanges komplett fehlt, wird die Tonhöhe nachgebildet

selbst wenn ein paar der ersten Obertöne fehlen

„Effekt der Grundton-Nachbildung“

Grundton muss nicht unbedingt besonders stark vorhanden sein, um eine saubere Tonhöhe zu empfinden

ein Klang kann allerdings „dünn“ werden, wenn man den Grundton zu stark absenkt

Kritische Frequenzbänder

Man hat bei verschiedenen psychoakustischen Experimenten festgestellt, dass es einen Unterschied zu machen scheint, ob die Frequenz zweier Testsignale nahe nebeneinander liegt oder weiter auseinander.

Schallereignis	Pegel pro Ton	Gesamtpegel	Ergebnis
1 Sinuston: 920 Hz	3 dB_{SPL}	3 dB_{SPL}	gerade noch hörbar
2 Sinustöne: 920 Hz +940 Hz	0 dB_{SPL}	3 dB_{SPL}	gerade noch hörbar
4 Sinustöne: 920 Hz +940 Hz +960 Hz +980 Hz	- 3 dB_{SPL}	3 dB_{SPL}	gerade noch hörbar
8 Sinustöne: 920 Hz +940 Hz +960 Hz +980 Hz +1000 Hz +1020 Hz +1040 Hz +1060 Hz	-6 dB_{SPL}	3 dB_{SPL}	gerade noch hörbar
9 Sinustöne: 920 Hz +940 Hz +960 Hz +980 Hz +1000 Hz +1020 Hz +1040 Hz +1060 Hz	-6,5 dB_{SPL}	3 dB_{SPL}	nicht hörbar

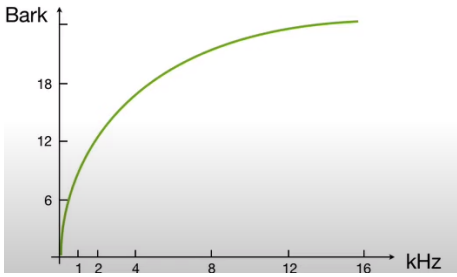
Man hat mehrere Töne zusammen abgespielt und dachte zuerst, dass nur der gesamte, aufaddierte Pegel relevant für die Hörbarkeit ist. Das stimmt aber nicht, die ersten 8 Sinustöne dieses Beispiels lagen einfach in kritischen Frequenzbändern.

Man kann sagen, dass bezüglich der Wahrnehmung zweier Signale innerhalb eines kritischen Frequenzbandes andere Gesetze gelten als wenn eines innerhalb liegt und eines außerhalb.

Beim Überschreiten des kritischen Frequenzbandes kommt es meist zu einer sprunghaften Änderung der Wahrnehmung.

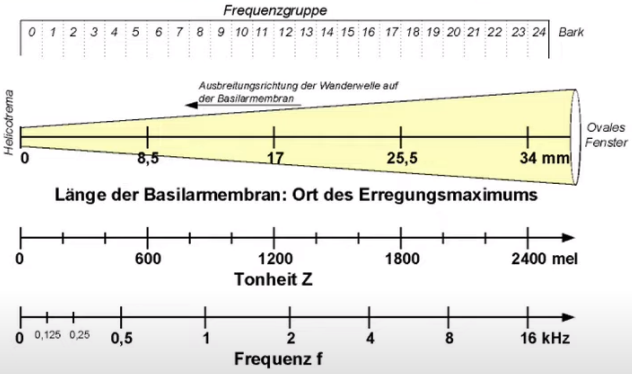
Breite

- bis zu 500 Hz → ca. 100 Hz
- ab 500 Hz → ca. 100 Hz bis 4 kHz
- **1 Bark = Breite eines kritischen Frequenzbandes**



1 Bark = 100 Mel
absolute Angabe möglich
z. B. 10 Bark = 1270 Hz

Ein kritisches Frequenzband reicht beispielsweise nicht fest von 100 Hz bis 200 Hz, sondern jeder Ton ist ein Auslöser für ein kritisches Frequenzband um sich herum!



1 Bark = 100 Mel = 1,3 mm auf der Basilarmembran = 150 Haarzellen = 27 JND

Bezüglich der Wahrnehmung zweier Signale gelten innerhalb eines kritischen Frequenzbandes andere Gesetze als wenn eines innerhalb liegt und eines außerhalb.

bis zu 500 Hz → ca. 100 Hz

ab 500 Hz → ca. 100 Hz bis 4 kHz

1 Bark = Breite eines kritischen Frequenzbandes

1 Bark = 100 Mel

Kritische Frequenzbänder sind keine fest definierten Bereiche mit fixen Grenzfrequenzen sondern variable Bereiche.

Wahrnehmung von Lautstärke

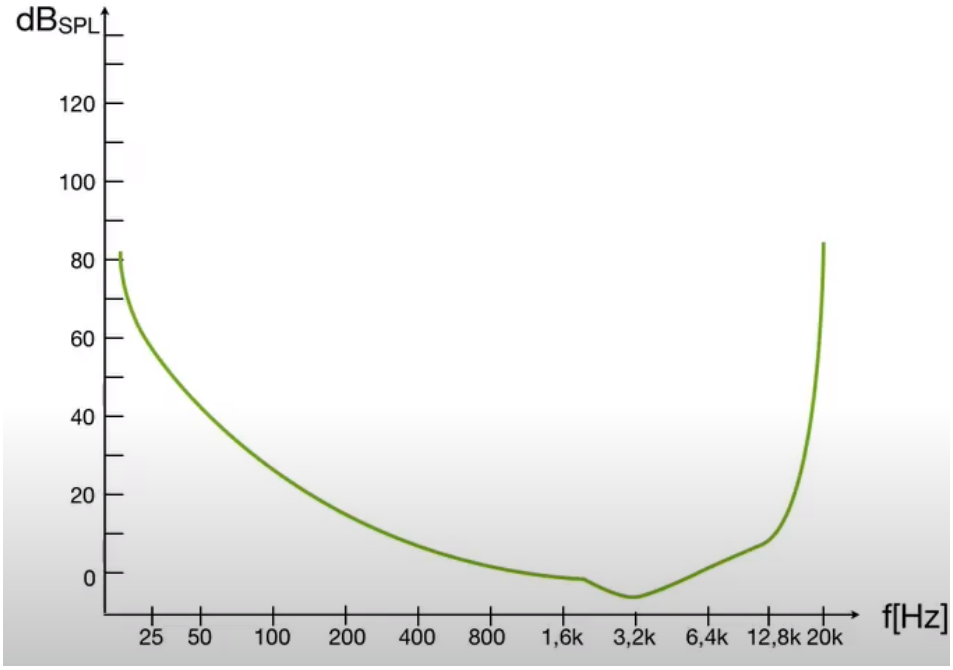
Amplitude → Lautstärke

Hörschwelle: 0,00002 Pa = 0 dB_{SPL}

Schmerzschwelle: 200 Pa = 140 dB_{SPL}

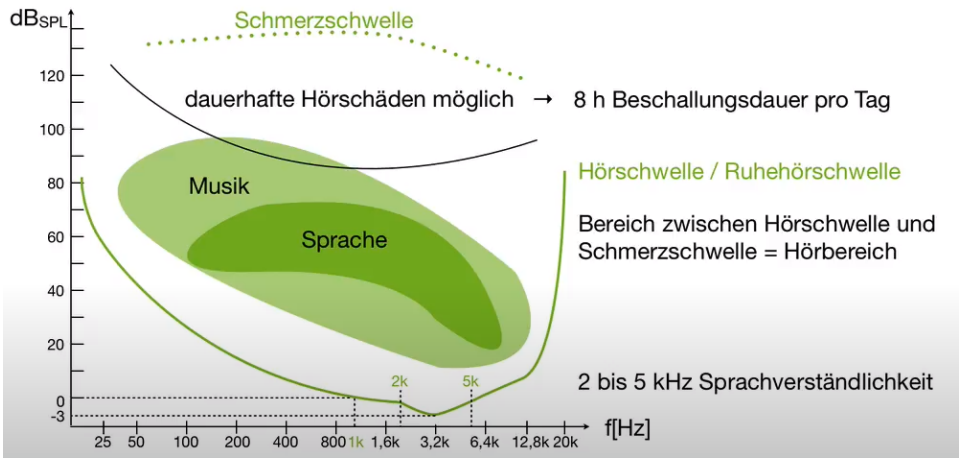
+10 dB = ~ „Verdoppelung der empfundenen Lautstärke“

Lautstärke hängt nicht nur vom Schalldruck sondern auch von der Frequenz ab



Hier Schalldruckpegel zu Frequenz

0dBSPL gilt für 1 kHz



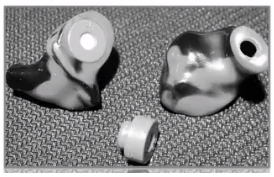
Bei jeweils 10 dB mehr sollte ein Zehntel dieser Dauer nicht überschritten werden!

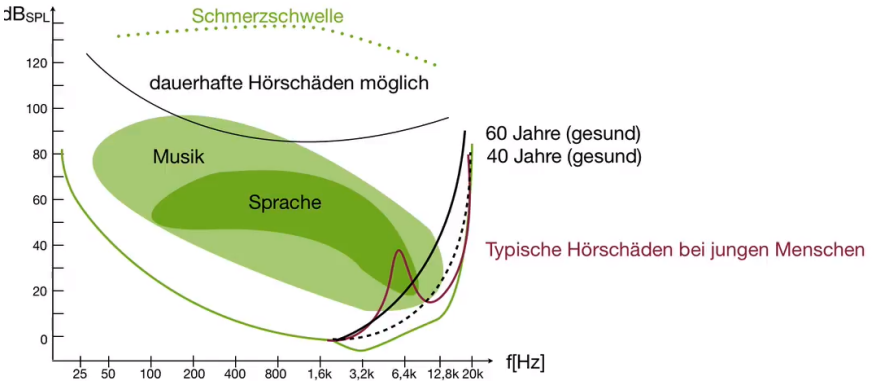
Schalldruckpegel	Beschallungsdauer pro Tag, ab der dauerhafte Hörschäden auftreten können
90 dB	8 h
100 dB	50 min
110 dB	5 min

Vor allem als Toningenieur:

Niemals ohne Gehörschutz in die Disko!!!

Nicht direkt vor dem Lautsprecher aufhalten!!!

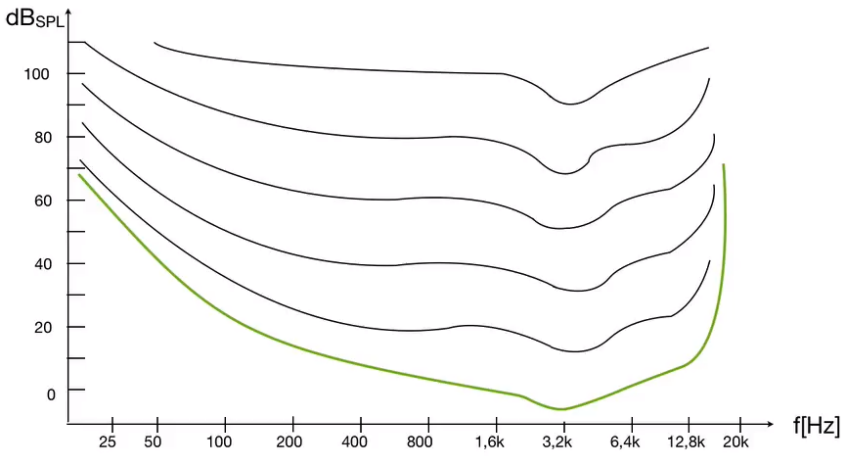




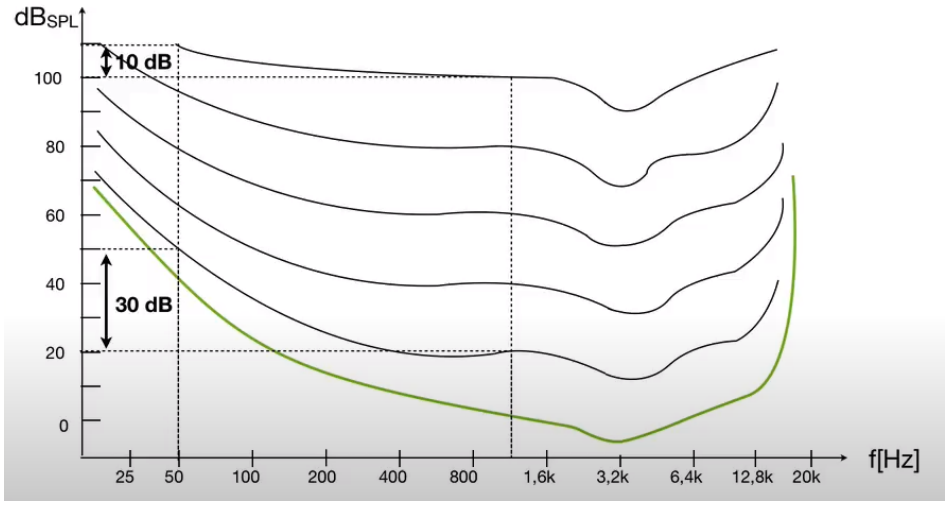
Schallereignis	dB_{SPL}
Knallkörper in Ohrnähe	180
Düsenflugzeug in 30 m Entfernung	150
Gewehrschuss in 1 m Entfernung	140
Sehr laute Diskothek, sehr lautes Rockkonzert, Maximalpegel am Walkman	120

Schallereignis	dB_{SPL}
Kreissäge, Presslufthammer, leise Diskothek oder leises Konzert	100
Straßenlärm, Stadt	80
Normales Gespräch, Musik in Zimmerlautstärke	60
Ruhiges Zimmer bei Tag	40
Ruhiges Zimmer bei Nacht	20
Hörschwelle	0

Kurven gleicher Lautstärke

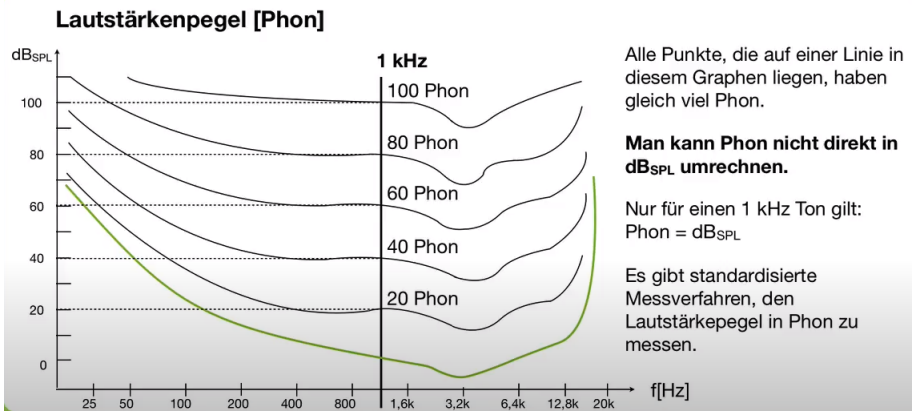


Alle Punkte der grünen Linie kann man gerade noch hören, alle anderen Linien verbinden Punkte die gleich laut wahrgenommen werden. Man ist besonders in 2 – 5 kHz sehr empfindlich, weil das Stimmen sind. Musik mit viel Bass wirkt leiser, weil der Bass eben tiefer ist. Allerdings passt sich das bei zunehmendem Schalldruckpegel an. Je höher, desto geringer wird der Unterschied in der Lautstärkenwahrnehmung.



Lautstärkenpegel

Wird in Phon gerechnet. Zeigt, wie Werte im Vergleich zu einem 1kHz Ton wirken.

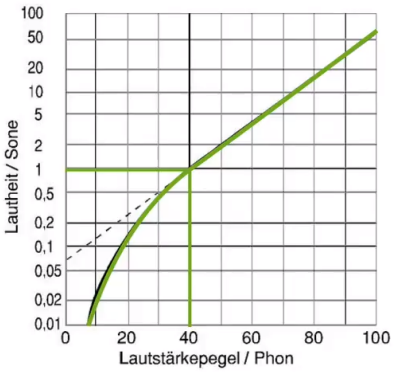


Alle Töne die gleichlaut empfunden werden wie ein 1kHz Ton bei 20dBSPL haben genau 20 Phon.

+10 dB → ~ Verdoppelung der empfundenen Lautstärke
+10 Phon → ~ Verdoppelung der empfundenen Lautstärke
Dies ist aber nicht die ganze Wahrheit!

Ab etwa 40 Phon bedeuten jeweils +10 Phon eine Verdoppelung der empfundenen Lautstärke.
Darunter reichen geringere Änderungen
→ **Lautheit [Sone]**

Deswegen gibt es auch die Einheit Sone… Die einfach diese weirde 40er Grenze anpasst.



40 Phon = 1 Sone

Eine Verdoppelung der Sone-Zahl entspricht immer einer Verdoppelung der Lautheit.

Lautheit / Sone	Pegel /Phon
64	100
32	90
16	80
8	70
4	60
2	50
1	40
1/2	32
1/4	25
1/8	19
1/16	14
1/32	11
1/64	9

Einheiten Lautstärke

Größe	**Schalldruck**	**Schalldruckpegel**	**Lautstärkenpegel**	**Lautheit**
Typ	physikalische Größe	physikalische Größe	psychoakustische Größe	psychoakustische Größe
Einheit	Pascal	dB_{SPL}	Phon	Sone
Bemerkung		$20\log\frac{p}{2*10^{-5}}$	berücksichtigt die Frequenzabhängigkeit unseres Gehörs	1 Sone = 40 Phon
Bedeutung	je mehr desto lauter	+10 dB = Verdoppelung der empfundenen Lautstärke	+10 Phon = Verdoppelung der empfundenen Lautstärke	doppelt so viel Sone = Verdoppelung der empfundenen Lautstärke
Manko	hat sehr wenig mit dem zu tun, wie unser Gehör funktioniert	Frequenzabhängigkeit wird nicht berücksichtigt Verhalten unter 40 dB_{SPL} werden nicht berücksichtigt	Verhalten unter 40 dB_{SPL} werden nicht berücksichtigt	

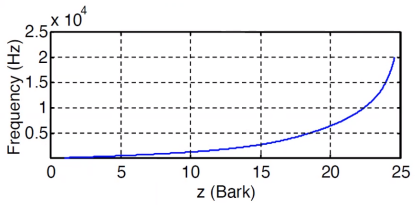
1 Bark = 100 Mel = 1,3mm on the basilar membrane = 150 hair cells = 27 JND
Verdopplung in Mel entspricht Verdopplung der wahrgenommenen Tonhöhe.

Berechnung Bark

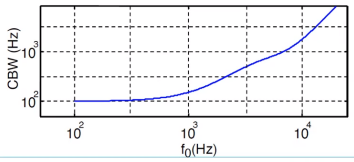
A scale that converts frequency (Hz) into units of critical bandwidth

$$z(f) = 13\arctan(0.00076\,f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right] \qquad \text{(Bark)}$$

named after Heinrich Barkhausen (who proposed the first subjective measurements of loudness)



Bark scale



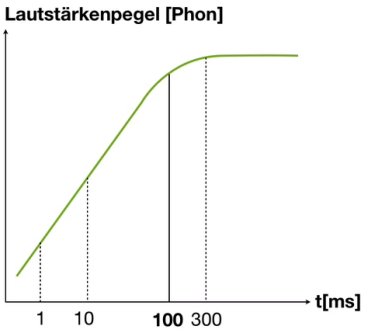
Critical bandwidth

Depending on the base frequency f_0 the wi
frequency group/critical band can be seen
About 100Hz up to 500Hz,
for higher frequencies (20kHz) up to 4kHz bandwidth

TU Delft
Technical University Delft

urtesy of Prof. Dr. Richard Heusdens
& Prof. Dr. Steven van de Par

Lautstärke und Tondauer



Bei einem komplexen Signal tragen kurze Signalspitzen unter 100 ms kaum zur wahrgenommenen Lautstärke bei!

Ausschlaggebend für die wahrgenommene Lautstärke eines Schallereignisses ist der Effektivwert des Schalldrucks integriert über eine Zeit von ca 100 ms!

Maskierung

All die Kurven von vorher gelten nur für Sinustöne. Man kann bei komplexen Schallereignissen nicht einfach alles zusammensummieren und dann die summierte Lautstärke nehmen. Manche Töne maskieren sich nämlich gegenseitig.

innerhalb eines kritischen Frequenzbandes addieren sich die Intensitäten der einzelnen Töne

bei Tönen, die nicht im selben kritischen Frequenzband liegen addieren sich die Lautheiten in Sone

Wenn also zwei Töne in gleichen Frequenzbändern liegen, gilt:

Beispiel:

- 2 Sinustöne mit 60 Phon
- 1000 Hz, 1100 Hz → innerhalb eines kritischen Frequenzbandes

- 60 Phon @ 1kHz ≈ 60 dB_{SPL} ≈ 60 dB_I
- 60 dBI + 60 dB_I = 63 dB_I
- 63 dBI @ 1kHz ≈ 63 Phon
- 63 Phon = ca. **5 Sone**

60+60 = 63 gilt deswegen, weil eine Vergrößerung um 3dBI gleich eine Verdopplung ist.

Andersrum wenn nicht im selben Band:

2 Sinustöne mit 60 Phon

1000 Hz, 2000 Hz → außerhalb eines kritischen Frequenzbandes

60 Phon = 4 Sone

4 Sone + 4 Sone = **8 Sone**

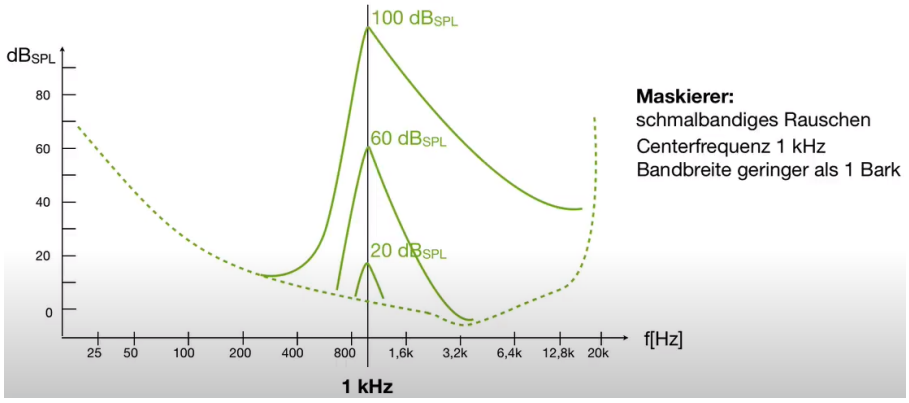
Just **N**oticable **D**ifferences

0,5 bis 1 dB

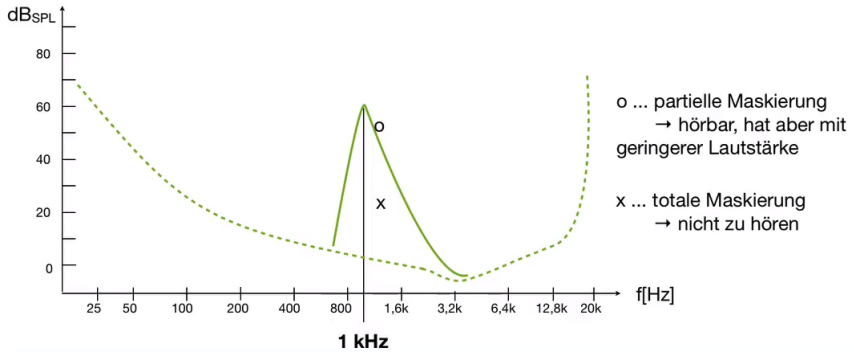
je nach Frequenz und Abhörlautstärke

Maskierungseffekt

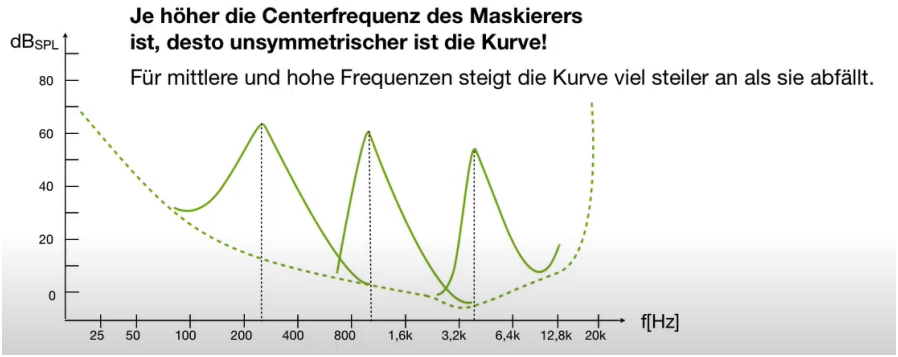
Mithörschwelle ist ähnlich wie Ruhehörschwelle. Ein Schallereignis setzt eine neue Schwelle für weitere Schallereignisse.



Gestrichelt sieht man die Ruhehörschwelle, in grün die Mithörschwelle für verschiedene Schalldruckpegel des Maskierers. Hat der Maskierer 60dBSPL, so passiert Folgendes:



Je höher der Pegel des Maskierers ist, desto breiter und unsymmetrischer wird die Mithörschwelle.



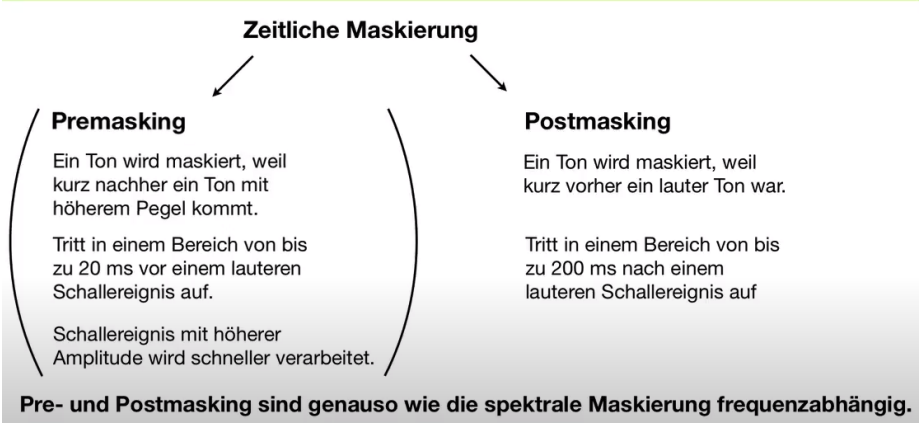
Zeitliche Maskierung

spektrale Maskierung:

- Wir hören einen Ton nicht, weil zeitgleich ein lauterer Ton im gleichen Frequenzbereich vorhanden ist

zeitliche Maskierung:

- Hierbei hören wir einen Ton nicht, weil vorher oder nachher ein lauter Ton vorhanden ist.

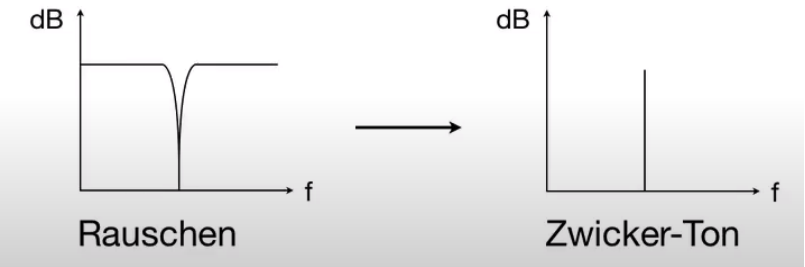


Effekte

Cocktailpartyeffekt – Man hört nur das Gesagte einer Person, man hört selektiv durch Konzentration – Das liegt an der Ortung der Schallquelle. Nach Aufnahme mit Mikro in Mono, geht das nicht mehr.

Adaption – Vorübergehende Anhebung der Hörschwelle. Setzt nach einer Minute ein und bleibt bis nach 3 bis 5 Minuten nach dem lauten Ton.

Zwicker-Ton – Wenn man Rauschen mit einem spektralen Loch abspielt, dann hören ProbandInnen nach dem Vorspielen noch den Ton aus dem Loch nach.



Shepard Skala – Tonleiter die immer höher oder tiefer wird. Illusion bei der die Wahrnehmung ausgetrickst wird.

Jeder einzelne sogenannte Shepard-Ton besteht aus mehreren übereinander „gelayerten“ Sinustönen mit Oktaveabstand.

Der Ton, der in der Mitte des Hörbereichs zu liegen kommt, erhält immer die höchste Amplitude.

Risset Rhythmus – Tempo das immer schneller wird.

Tritonus Paradoxon – Illusion bei der manche Menschen immer zuerst einen tiefen dann hohen Ton hören und manche andersrum.

Oktave – Illusion – Linkes Ohr bekommt umgekehrte Töne als das rechte. Rechtshänder hören hohen Ton rechts, tiefen links.

Deutsch – Skala – Wieder ein rechts links Ding, wo eine Melodie rechts, eine links ist.

McGurk – Effekt – „Nanana / Dadada / Bababa / Gagaga” – Das Auge hört mit

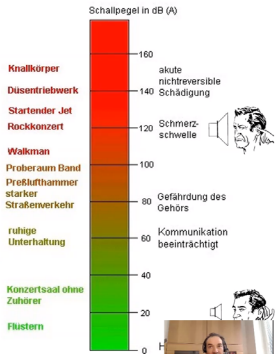
Kombinationstöne – Man hört, wenn man zwei Töne vorgespielt bekommt manchmal noch mehr Töne, nämlich Summentöne und Differenztöne

Sound Pressure Level - Schalldrucklevel

Definition of the sound pressure level:

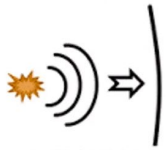
$$L = 20 * \log(p/p_0) \quad [dB]$$

- L… sound pressure level
- p… sound pressure of the considered sound wave
- p_0…reference sound pressure:
  by definition the sound pressure of a 1000Hz tone, which is just still audible. $p_0 = 2 * 10^{-5}$ [Pa]
- dB... decibel, unit of the sound level.



Das braucht es deswegen, weil unsere Wahrnehmung von Lautstärke nach logarithmischen Gesetzen funktioniert. Um also zu verhindern, dass man eine Skala von 1 bis 2 000 000 machen muss, hat man den Schalldrucklevel.

Number of sound sources	 x1	 x2	 x10	 x100
Sound pressure change	x1	x1,4	x3	x10
Sound level difference	0dB	+3dB	+10dB	+20dB
Subjective hearing sensation	Basic volume	A little louder	twice as loud as a single violin	Four times as loud as a single violin

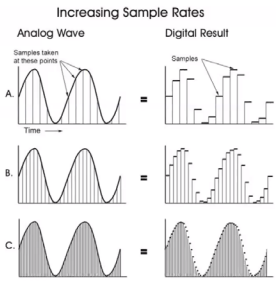
Pegeländerung	Lautstärke Wahrnehmung	Schalldruck Wirkung	Schallintensität Ursache
			
Dezibel	Lautheits-Faktor	Schallfeld-Faktor	Schallenergie-Faktor
+ 20 dB	4,000	10,000	100,000
+ 10 dB	2,000 •	3,160	10,000
+ 6 dB	1,516	2,000 •	4,000
+ 3 dB	1,232	1,414	2,000 •
± 0 dB	1,000	1,000	1,000
– 3 dB	0,812	0,707	0,500 •
– 6 dB	0,660	0,500 •	0,250
– 10 dB	0,500 •	0,316	0,100
– 20 dB	0,250	0,100	0,010

Rechner dazu – www.sengpielaudio.com/calculator-spl.htm

https://www.schweizer-fn.de/akustik/schallpegelaenderung/schallpegel.php

VO3 - Digital Audio

Mikrofon wandelt analoge Signale, also Druckunterschiede in digitale Signale um. Digitale Signale haben eine höhere Bandbreite als analoge.

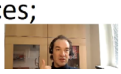


Microphones produce analog signals (continuous-valued voltages)

To get audio into a computer, it must be converted into a stream of numbers;
discrete sampling (both time and voltage)

Sampling: divide the horizontal axis (the time dimension) into discrete pieces; uniform sampling is ubiquitous

Quantization: divide the vertical axis (signal amplitude) into pieces; sometimes, a non-linear function is applied



Das analoge Signal ist eine Spannungsschwankung, diese wird gesampelt, also an gewissen Abgriffspunkten in Zahlen umgewandelt. Anschließend wird quantisiert, also jedem Schritt ein Wert zugewiesen.

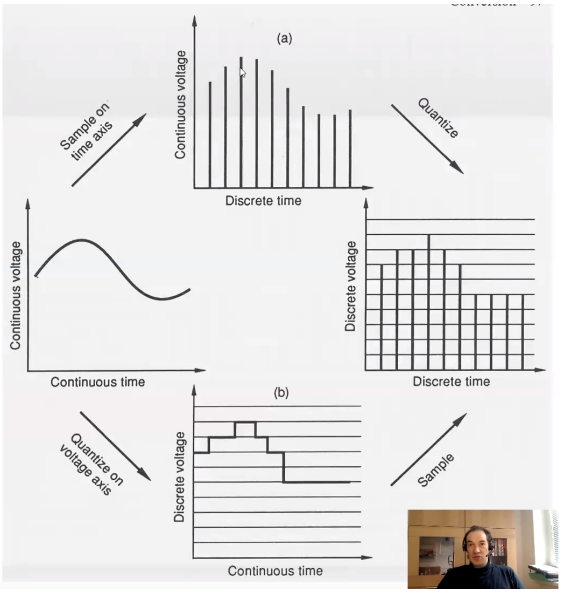
Man kann aber in beide Richtungen vorgehen, also zuerst samplen dann quantisieren oder andersrum.

Digitization Basics

Input: time and voltage continuous waveform

Sampling- Quantizing: are independent of each other temporal order does not matter

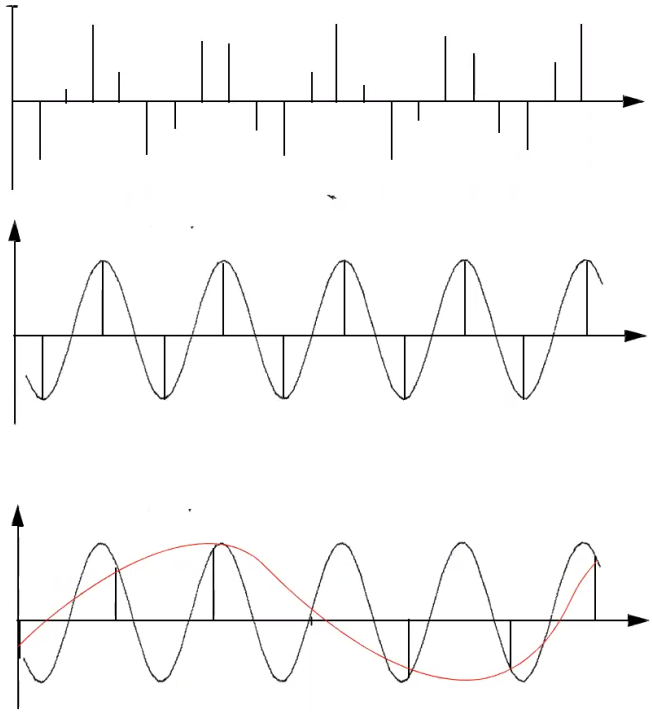
Result: time and voltage discrete format



(a) Input signal is sampled, samples are quantized (common with audio)
(b) Input signal is quantized, then sampled (usual for video)

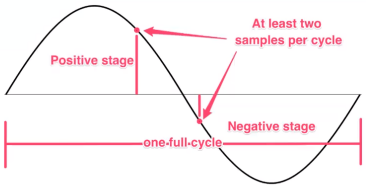
Sampling

Sampelt man mit zu wenig Schritten, bekommt man Aliasing.

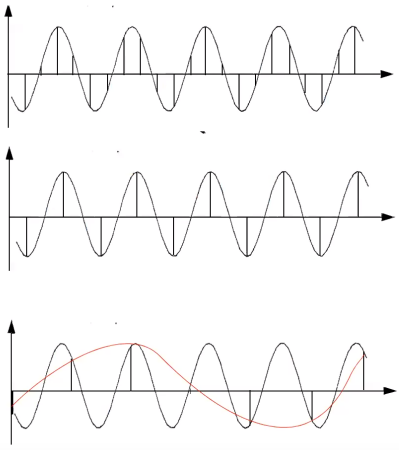


Das Nyquist-Shannon Theorem beasgt, dass man mindestens mit einer Sample frequency von 2x der höchsten Frequenz im Originalsignal samplen muss, damit man loss less rekonstruieren kann.

Alias Component

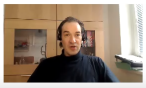


To determine the wavelength and therefore frequency of a sound wave we need at least 2 samples per cycle.



- $f_s > 2 \times f_{max}$
  Reconstructed signal identical with original
- $f_s = 2 \times f_{max}$
  Reconstructed signal same f as original
- $f_s < 2 \times f_{max}$
  Interference frequency
  "alias component"

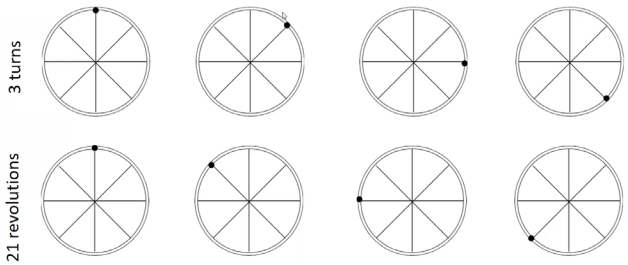
Ex. : f_s = 48kHz, f_{id} = 30kHz, f_{alias} = 18kHz



Alias Component 2/2

Example: turning wheel in movie

- Sampling frequency is fixed (F_S = 24)
- Number of revolutions (u) of the wheel variable

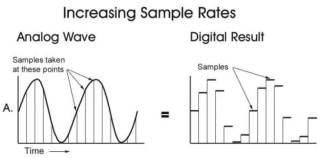


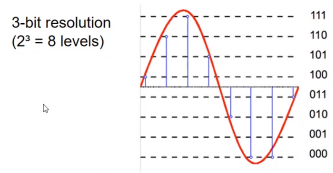
Looks like wheels spins

Quantisierung

Bei der Quantisierung sucht man zu jedem gesampelten Wert diskrete numerische values. Also wie da oben rechts für die Werte 111, 110, 101 …

Quantization



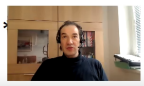


- Sampled voltage values are assigned discrete numerical values:
  Total voltage range divided into quantization intervals *Q*
- ➡ Continuous value assigned to nearest numerical value.
- Quality: *Q* should be as small as possible – is a result of the length of the data word size:

V_{pp} = peak to peak range (max range of signal) or $2 * abs(V_{max})$

8 bit *(256 intervals)*: $Q = \frac{V_{pp}}{2^8} = (4 * 10^{-3}) * V_{pp}$ = 0.004 * V_{pp}

16 bit *(65536 intervals)*: $Q = \frac{V_{pp}}{2^{16}} = (15 * 10^{-6}) * V_{pp}$ = 0.000015 *

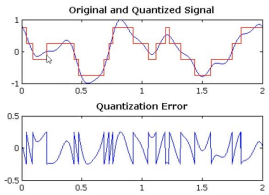


24 bit (*16 777 216 intervals*) is also used

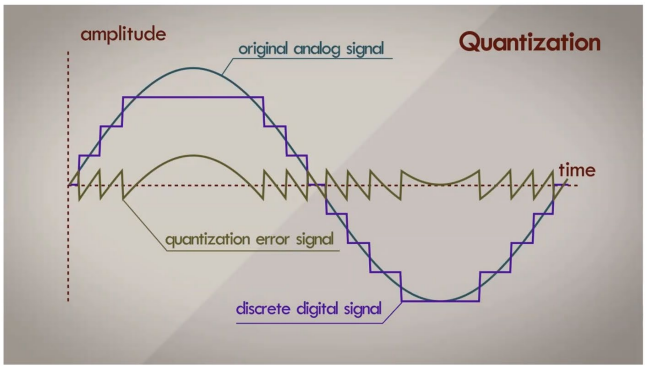
Quantisierungsfehler

Beim Quantisieren trifft man natürlich nicht immer den ursprünglichen Wert in der echten Welle. Um das zu gewährleisten, bräuchte man ja unendlich viele Werte, was den Sinn der Quantisierung zunichtemachen würde. Diese Abweichung vom Originalwert nennt man Quantisierungsfehler.

Quantization Error

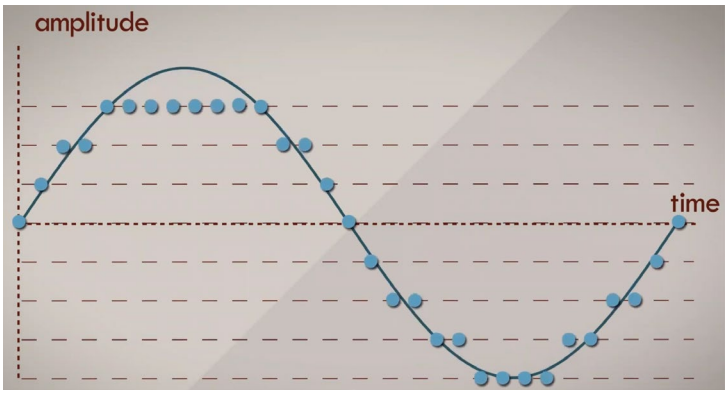


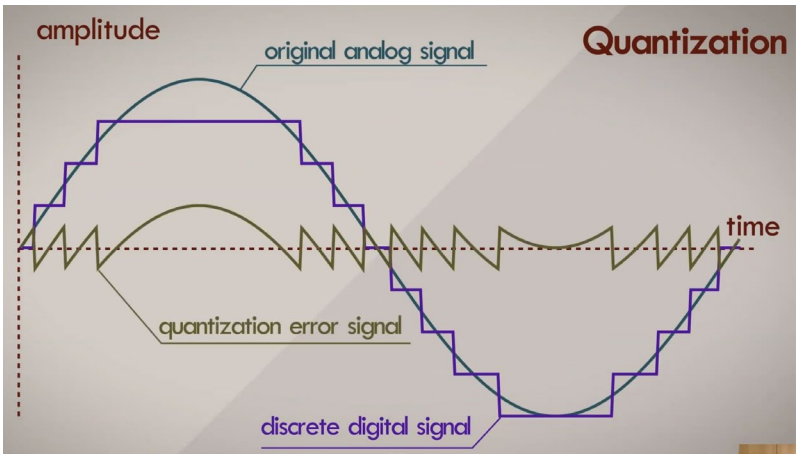
- Numerical value deviates from the sampled value by ≤ 0.5 Q.
  Error is audible as "quantization noise" during reconstruction.



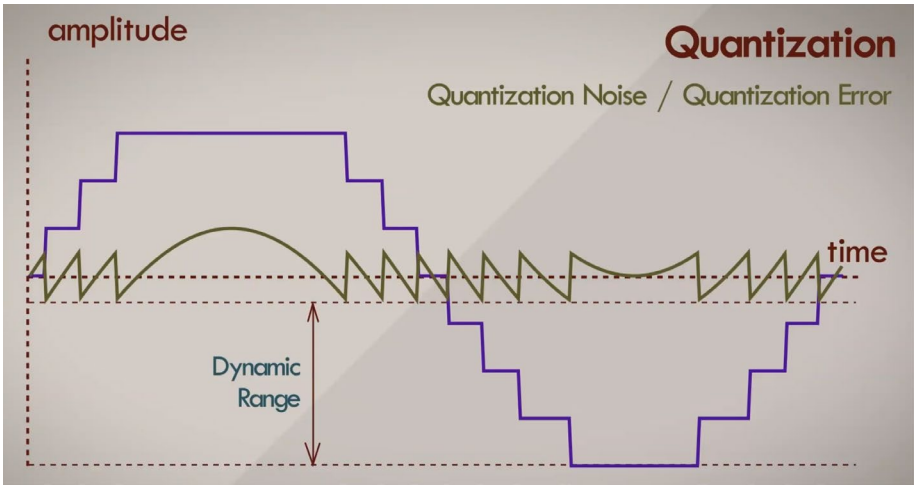
Accuracy of measurement Sample Resolution

Beim Sampling und bei der Quantisierung geht es nur darum, wie fein man einteilt. Es geht nicht um die Qualität eines Sounds (wie bei Bildern), sondern lediglich darum, wie viel noise da ist.





Hier sieht man in ?braun? das noise.

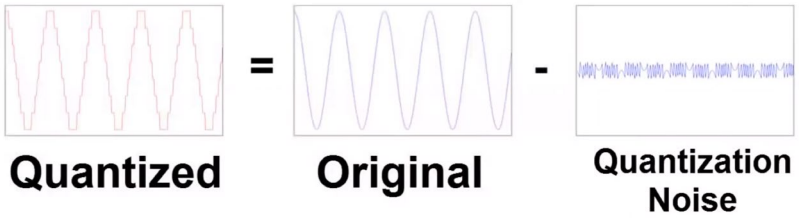


Die Dynamic range wird also dann vom Noise beschränkt, das Noise überdeckt nämlich, maskt also, die gewollten Sounds.

Quantisierungsrauschen ist nie größer als das least significant bit, das liegt daran, dass es ja eine Rundung ist, also maximal der Rundungsfehler der Fehler sein kann.

Dithering

Ist eine Art Quantisierungsrauschen loszuwerden. Was man machen kann ist, dass man aus nicht-random noise, also Quantisierungsnoise, random noise einmischt. Damit wirkt es nicht mehr so störend.



Signal to Noise Ratio / Dynamic range

DR = Bereich vom lautesten zum leisesten Signal

SNR..Signal to noise ratio [dB] – in General:

- Ratio of power of a signal (meaningful input) to power of background noise (unwanted input)
- $SNR = \frac{P_{Signal}}{P_{Noise}}$

In Audio:

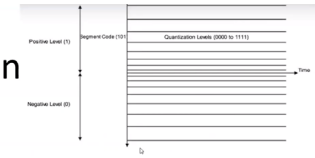
- Quantization noise generates sound pressure level $\mathbf{L_{noise}}$
- Maximum usable signal Sound pressure level $\mathbf{L_{max}}$ (loudest possible signal)
- SNR = $\log\left(\frac{P_{Signal}}{P_{Noise}}\right)$ = L_{max} - L_{noise}

In audio **Dynamic Range** is SNR:

Resolution	8 bit	12 bit	16 bit	24 bit
Dynamic Range	49.8 dB	73.7 dB	96.3 dB	144.5 dB

Non-linear quantisation

.inear; Non-linear quantization



Linear quantization: *Q* constant; common method in audio engineering

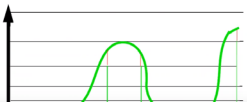
Nonlinear quantization: *Q of different size*

small values -> small quantization error

large values -> large quantization error

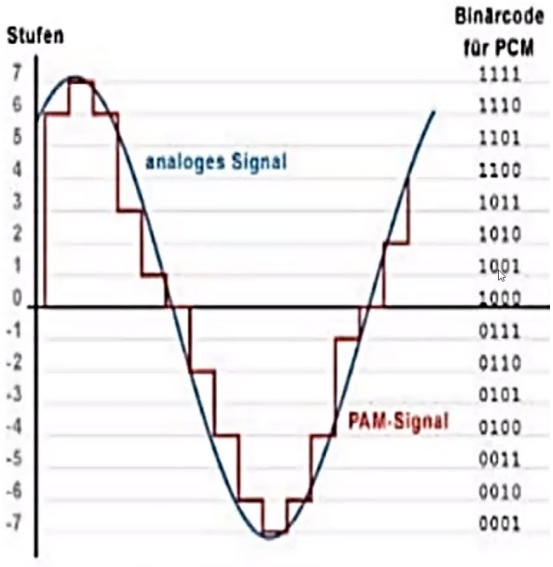
Noise masked if high desired signal -> passable quality even with small resolution

Usage for systems with low transmission bandwidth



Hier teilt man in der Mitte feiner ein um weniger noise zu bekommen.

Encoding – PCM Pulse Code Modulation

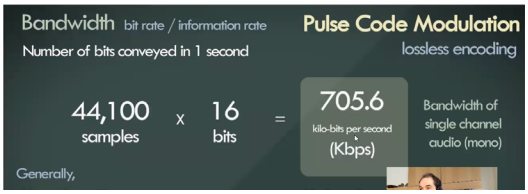


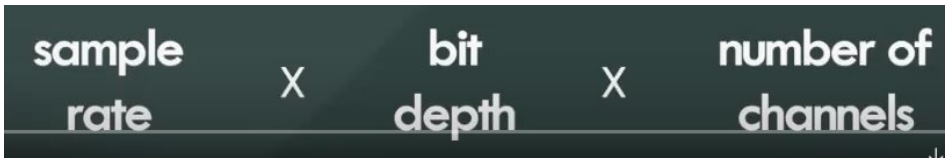
Ist lossless – also beim Abspeichern geht nichts verloren, wird bei wav benutzt.

- Disadvantage: **High data rate**

  Sampling frequency [Hz] * bit depth [bit] = data rate [bit/s].

  Ex. : CD-Audio(mono):
44,1kHz; 16 bit -> 705600 bit/s





Interleaving

Hat man zwei Channels, so speichert man abwechselnd LRLRLRLR ab.

Speichern

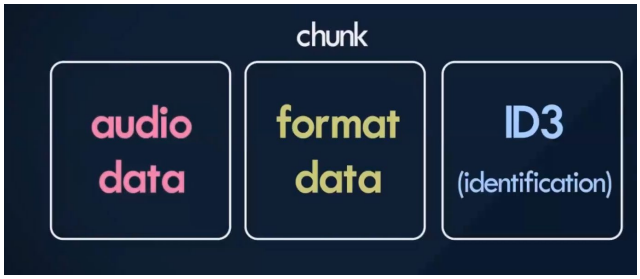
Beim Speichern in files muss der Computer nicht nur wissen, welche bits wo sind, sondern auch, um was es sich eigentlich bei dem Bitstream handelt. Ohne diese Info kann er nicht bestimmen wie er die Daten wiedergibt. Ebenso muss dem Computer bekannt sein wie encoded wurde und mit welcher bit depth, auch die channels und sampel rate müssen bekannt sein, um das Signal dann richtig zu interpretieren.

Man kann zusätzlich auch Infos wie Bandname, Songname usw. speichern.

.wav

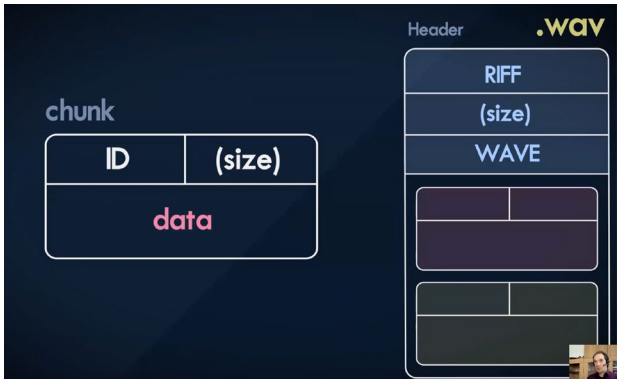
In wav gibt es mehrere header, in einem steht RIFF, was im Endeffekt einfach Resource Interchange File Format heißt. Was ein Containerformat ist, das Daten in getaggten chunks speichert.

Ein chunk kann Audiodaten haben, ein chunk Formatierungsdaten, ein weiterer Identifizierungsdaten.

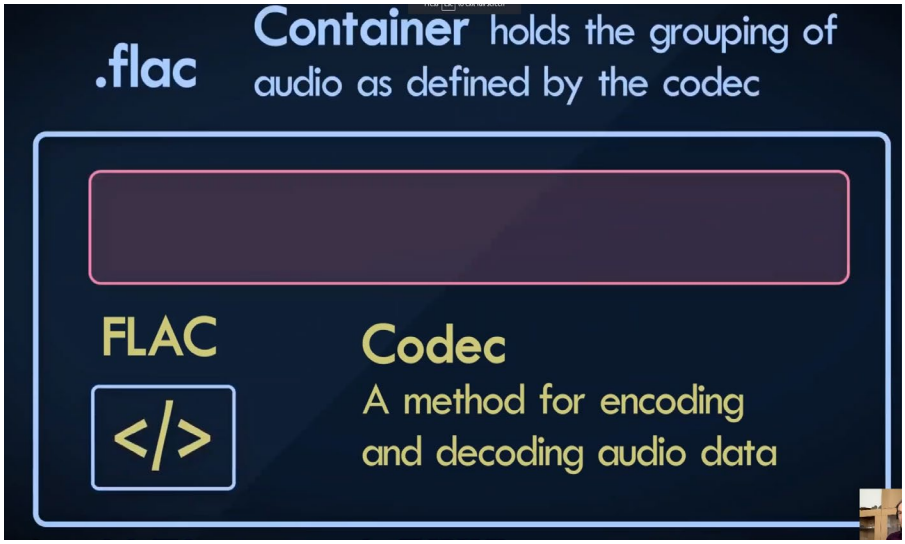


chunks sind so aufgebaut:

Insgesamt:

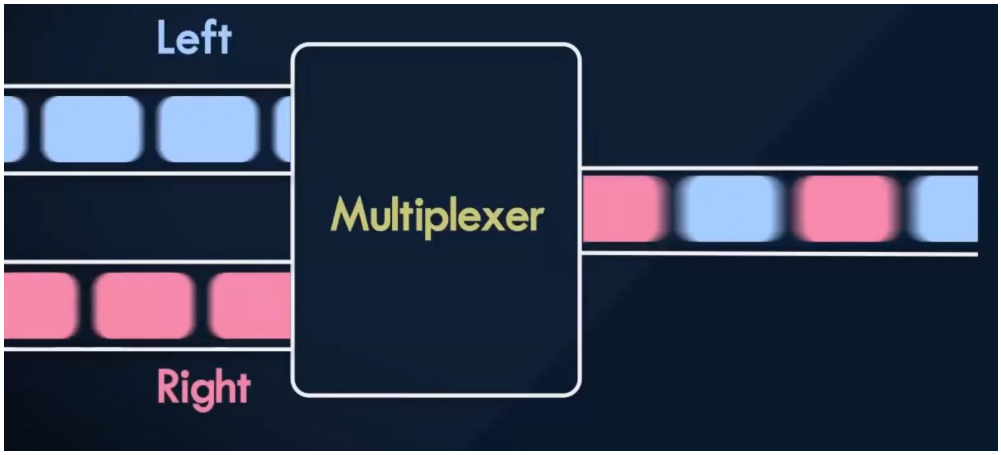


Container vs. Codec



Multiplexer

Macht interleaving

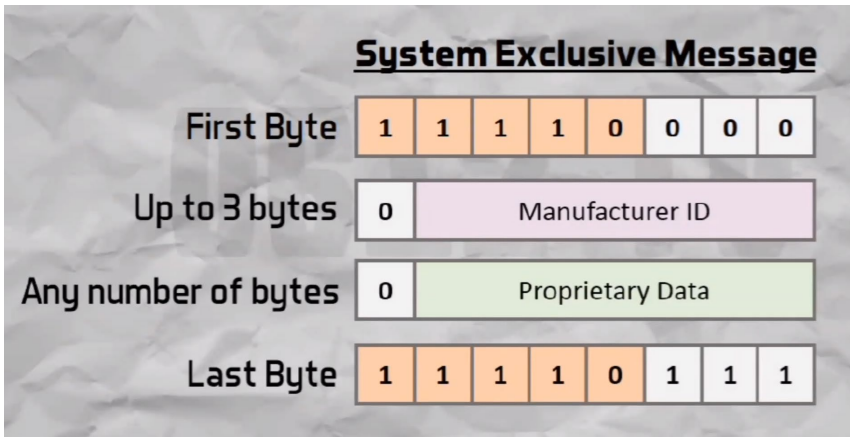


MIDI

Ist ein interface/eine Interfacesprache über das Computer Synthies und Instrumente kommunizieren können.

MIDI Nachrichten beschreiben immer events. Bspw. einen Knopfdruck.

Die ersten 4 bit beschreiben was die Nachricht ist (also welches event – Note on oder off…), die nächsten 4 bit sind der Channel auf dem das Signal kommt. Der zweite und dritte Byte der Nachricht ist abhängig vom Event. Bei key down oder up sind das bspw. die Notennummer und die Geschwindigkeit. Bei Control changes Control ID und der neue Value für den Control. Es gibt auch noch System Exclusive Messages, die sehen so aus:



DAW

Eine digital audio workstation.

VO4 & VO5 – Compression

Codec

Ist ein Paar von Algortihmen zum codieren und encoden von Daten.

Der Codec kann natürlich auch komprimieren, dabei unterscheidet man zwischen lossy und lossless.

Es gibt im Wesentlichen drei Kategorien:

Statistical – Die einfach schauen was wie oft vorkommt

Spatial/Time Domain – Die schauen was wann wo kommt

Universal – Die universal funktionieren

MPEG Group

Motion Picture Experts Group

350 Mitglieder, die sich seit Ewigkeiten zusammensetzen und Standardisierung für diese Formate bestimmen.

Ziel für neue Standards ist, dass man nur die halbe Größe für die gleiche Qualität hat. Außerdem:

Symmetrische und Asymmetrische Kompression

Symmetrisch heißt, dass das Komprimieren gleich lang dauert wie das Dekomprimieren.

Sowas ist bei Echtzeitanwendungen wichtig.

Random-access Playback

Heißt, dass man an einem beliebigen Punkt das File öffnen und abspielen kann.

Synchronisation

Audio und Video muss synchron bleiben.

Außerdem

Datenfehler möglichst geschluckt.

Wenn nötig soll delay in Kompression und Dekompression gesteuert werden können.

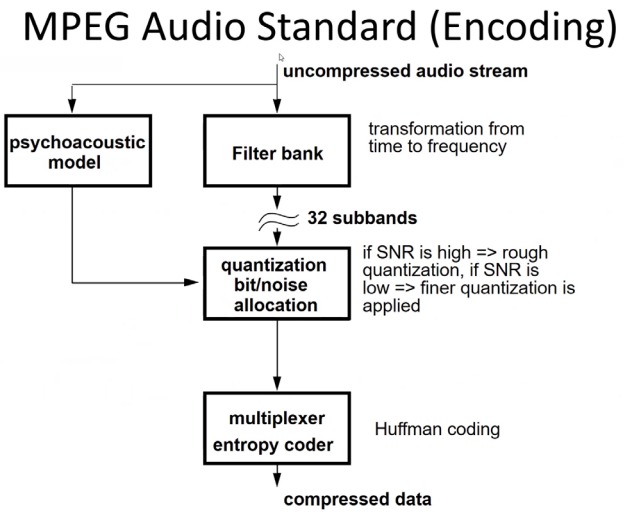
Files sollen bearbeitbar sein.

Die Formate sollen auch auf billigen Endgeräten benutzbar sein.

MPEG.1/2 Layer-3 … MP3

Lizenz schon ausgelaufen, jeder kann einen MP3 Codec machen.

Der Encoder ist hier nicht spezifiziert. Der Standard spezifiziert sehr genau wie der Bitstream aussehen muss der daherkommt, den man dann decoded, aber wie man encoded ist nicht genau drinnen.



Stream kommt rein, man teilt in 32 Frequenzbänder auf. Gleichzeitig macht man ein psychoakustisches Modell. Auch da gibt es eine Transformation in den Frequenzbereich, denn das psychoakustische Modell arbeitet auch mit Frequenzen. Dabei schaut man wo wird was maskiert. Die Sachen fliegen raus. Der Verlust ist also nur im psychoakustischen Modell und in der Regel eh das, was man nicht hören kann. Dann wird mit Huffman coding komprimiert und das kommt dann alles in den Decoder beim Abspielen.

MPEG-1 Layer-3 has been specified 1993

since then, research on perceptual audio coding has progressed and codes with better compression efficiency became available: MPEG-2 Advanced Audio Coding (AAC) and other proprietary compression systems.

Basic task is to compress audio data in a way that

- the compression is as efficient as possible (file size small)
- the reconstructed (decoded) audio sounds exactly (or as close as possible) to the original audio data
- requires low complexity (in software or inexpensive hardware)
- offers flexibility for different application scenarios.

General audio signal consists of many frequency components

- each frequency component influences the hearing threshold curve
- Audio signal varies with time

Hearing threshold different at each point in time

Psychoacoustic model

- Calculates current hearing threshold curve in each case

Variable quantization

- Frequency components quantized so that quantization noise just below current hearing threshold curve
- Where masking is used, coarse quantization, small word leng

Alle Frequenzkomponenten verändern die Hörschwelle. Mittels variabler Quantisierung kann man dann das wegschneiden, was eh nicht gehört wird.

A Basic Perceptual Audio Encoder

Consists of the following building blocks:

1. Filter bank—is used to decompose the input signal into subsampled spectral components (time/frequency domain)
2. Perceptual model—using either the domain input signal and/or the output of the analysis filter bank, an estimate of the actual masking threshold is computed by rules from psychoacoustics
3. Quantization and coding—the spectral components are quantized and coded; noise introduced should be kept below threshold
4. Lossless Huffman Encoding of bitstream

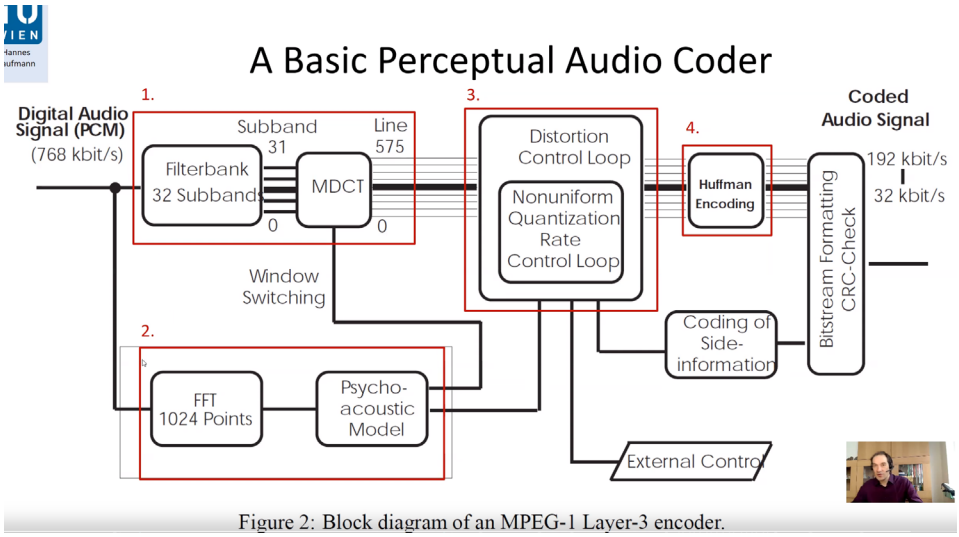


Figure 2: Block diagram of an MPEG-1 Layer-3 encoder.

Als Vorbereitender Schritt werden einzelne Frames in den Encoder geworfen, nicht das ganze Lied. Die Modifizierte Diskrete Cosinus Transformation wandelt dann diese 32 Subbänder in sehr viele Frequenzbereiche. (Die kann man dann nämlich besser quantisieren) Mit dem Psychoakustischen Modell kann ich dann bestimmen, was ich weglassen kann.

Im Detail

Man hat ein Audiosignal, einen Bitstream. Man teilt ihn auf in 576 samples. Die Filterbank schmeißt alles weg, was über oder unter der kHz Hörschwelle liegt. Dann geht man in eine MDCT um in den Frequenzbereich umuwandeln. Ebenso wird mit einer FFT in den Frequenzbereich umgewandelt. Basierned darauf wird dann im psychoakustischen Modell alles wegeschnitten, was nicht hörbar ist. (weil maskiert) Anschließend wird dann alles demnach quantisiert und dann noch huffman encoded. (Auch im Frequenzbereich)

Filterbank

Besteht eigentlich aus zwei Filterbänken. Zunächst einer Polyphase Filterbank, hier werden alle Frequenzen < 16 Hz und > 16kHz weggeschnitten.

Dann werden mit einer DCT (MDCT) nochmal aus jedem Frequenzband 18 feinere Subbänder gemacht.

Perceptual Model

Hier wird meist eine FFT verwendet, um eine Frequenzdarstellung zu erreichen, auf dem gearbeitet werden kann. Hier wird dann all das gelöscht, was sowieso eigentlich maskiert ist.

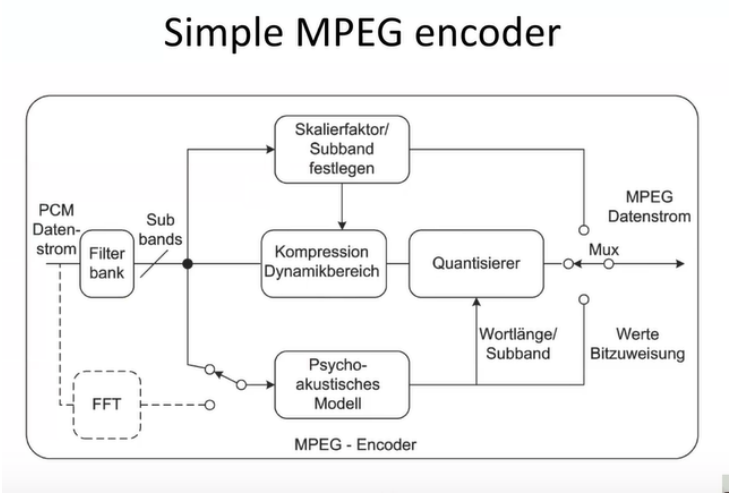
Quantisierung

Hier wird nicht-linear quantisiert. Das durch zwei iteration loops (rate loop und noise control loop)

Man sagt dem Algorithmus mit welcher Bitrate man quantisieren will. Der Algorithmus muss also schauen, wie man einteilen kann.

In der äußeren Loop wird geschaut, wie viel man quantisieren kann, ohne dass das Rauschen über den maskierten Bereich schießt. (Maskierungsschwelle) Hier sucht also der Algorithmus für jedes Band einen optimalen Skalierungsfaktor. Dieser muss dann abgespeichert werden.

In der zweiten Schleife schaut man, wie viele Bits man braucht wenn man mit Huffman codiert. Wenn die resultierende Nummer an Bits die verfügbare Nummer überschreitet, dann verändert man den global gain (über den ganzen Frame hinweg). Dadurch nämlich versucht man die Töne so aneinander anzupassen, dass man möglichst ähnliche Werte in verschiedenen Bändern hat und damit dann bei Huffman weniger Speicher braucht. Man speichert dazu aber, dass man den gain verändert und um wie viel, um später wieder auszugleichen was zwecks Speichern verändert wurde.



Erneut noch mehr im Detail…

Aufteilen in Frames

Split audio file into “frames”

During encoding 576 time domain samples are taken & transformed to 576 frequency domain samples. If there is a transient 192 samples are taken.

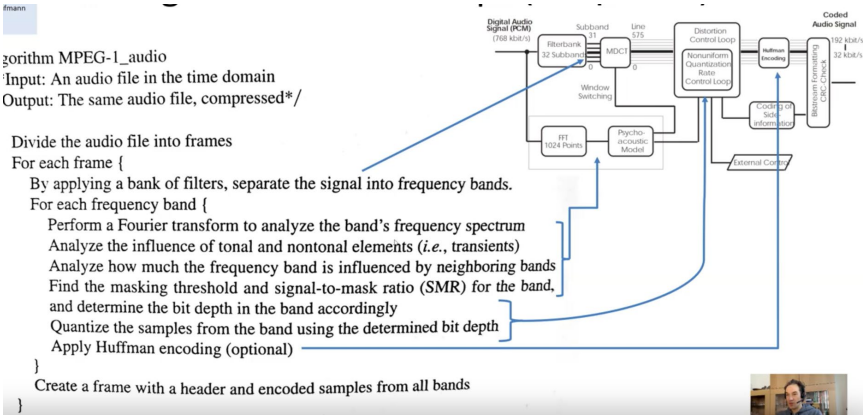
- This is done to limit the temporal spread of quantization noise accompanying the transient.

The decoder expects a frame consisting of 384, 576, or 1152 samples

- Typically in MP3 each "frame" has a length of 1152 samples = 0.026 sec (26,122449 ms)

The following processing happens in each frame individually - with some overlap with previous/next frame to minimize temporal effects

Für jeden Frame



Beispiel Teilen in Bänder

Example filter bank with 4 subbands

Input - 1 x PCM: frequency range 0 - 20kHz, sampling rate 48 kHz; word length 16 bit per sample

Output - 4 x PCM

- Subband 1: 0Hz - 5kHz; sampling rate 12kHz (!); 16 bit
- Subband 2: 5kHz -10 kHz; sampling rate 12kHz (!); 16 bit
- Subband 3: 10kHz - 15 kHz; sampling rate 12kHz (!); 16 bit
- Subband 4: 15 kHz - 20 kHz; sampling rate 12kHz (!); 16 bit

Seems to be a violation of the Nyquist-Shannon theorem BUT (see [1])

- IF a band does not start at zero frequency but at some higher value and can be shifted by a linear translation, smaller sampling rate – at least twice the width of the non-zero frequency interval - is required.
- For example, in order to sample the FM radio signals in the frequency range of 100–102 MHz, it is not necessary to sample at 204 MHz (twice the upper frequency), but rather it is sufficient to sample at 4 MHz (twice the width of the frequency interval).

Critical Sampling

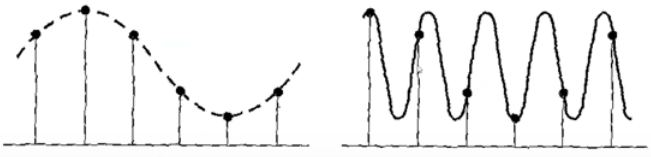
Typically 32 subbands in MP3

Critical sampling of digital filter banks

- Sampling rate in subbands reciprocal to the number of subbands
  - Reduced sampling rate in subbands is called critical sampling rate

Critical sampling of subbands ensures that filter bank has no influence on data volume

- More channels are generated, but their sampling frequencies are correspondingly lower.

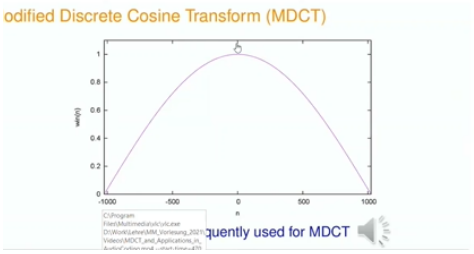


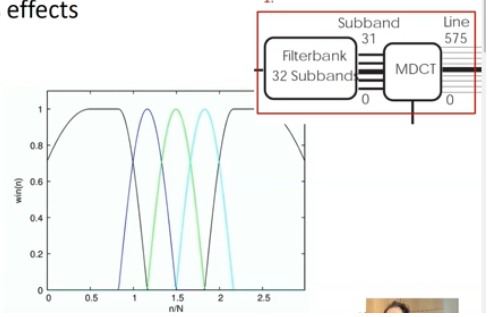
- Left: Samples / signal without specification of frequency range
- Right: Samples / signal with indication of frequency range (right)

Subband Coding

Wenn man am Ende jedes Bandes einfach abschneiden würde, hätte man clipping. Deswegen braucht man einen Übergang, einen overlap. Da benutzt man die MDCT.

Performs transformation of subbands into frequency domain
Overlaps with neighboring frequency bands (but does not increase sample size)
MDCT also used for avoiding temporal alias effects

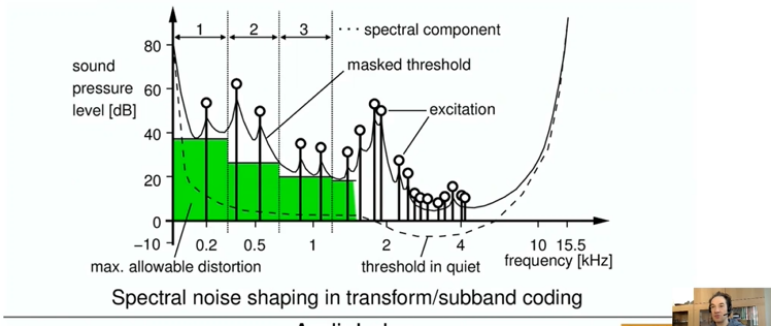


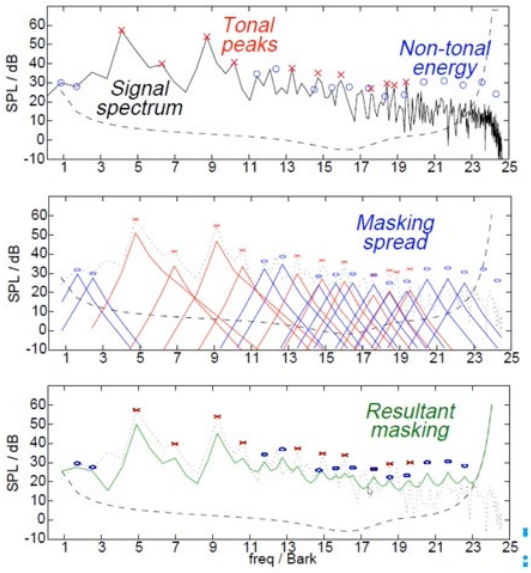


Psychoakustisches Modell

First: Fast Fourier Transform into frequency domain
– Psychoacoustic model operates on frequency spectrum
Analyzes audio signal section by section for masking effects
– Algorithms simulate human hearing
Calculates masking threshold per subband
– calculates maximum allowed quantization noise per subband
Quality of the psychoacoustic model mainly responsible for overall quality of coding

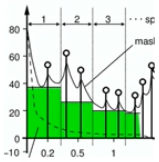
- For each subband: How high can quantization noise be, so that it is still masked?





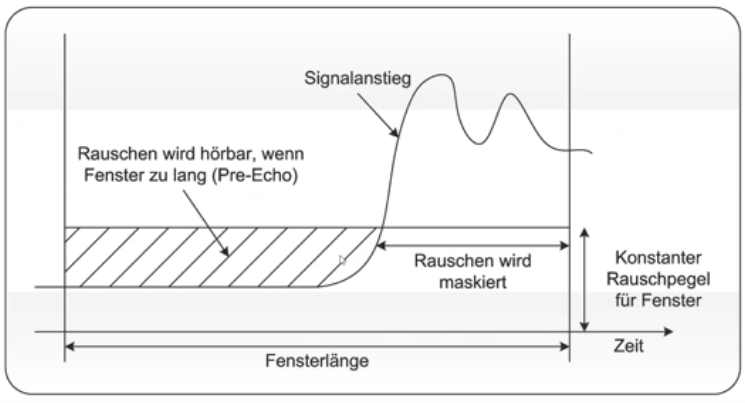
Rot = Tonal, also Gesang, Gitarre… Blau ist Schlagzeug usw. Grün ist dann die finale Masking Kurve

Because of narrow subbands
- high masking threshold for noise possible
- frequency components require only few bits, which corresponds to high compression



In general
- the higher the subband number the narrower the subbands
- the narrower the subbands the higher the compression

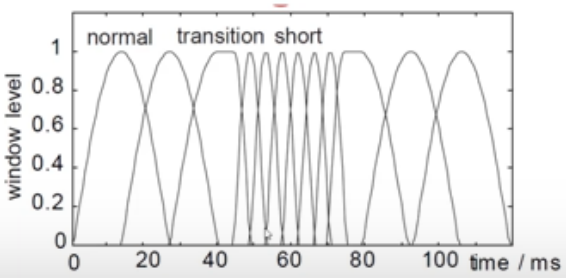
Pre echo



Quantisiert man über diesen Block wo zuerst alles ruhig ist und dann plötzlich sehr laut, dann hat man als Resultat ein Rauschen das halt relativ gesehen zum leisen Bereich sehr laut ist.

Lösung dazu

Adaptive switching of time-window



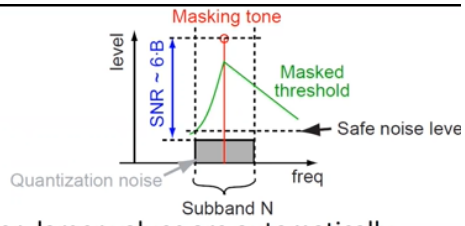
Frequency Resolution / Time Window

High frequency resolution with long time window
+ : many subbands, high compression
  – Subbands well adaptable to hearing threshold curve
- : Quantization noise not changeable for a long time
  – bad with signal increases

Low frequency resolution with short time window
+ : Quantization noise changeable in short time intervals
  – Noise can be well matched to time curve
- : Fewer subbands, moderate compression

Quantisierung

Layer-3 Encoding Algorithm



Quantization and coding

– quantization is done via a power-law quantizer; larger values are automatically coded with less accuracy

– two nested iteration loops (rate loop and noise control loop)

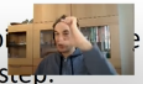
Outer iteration loop (noise control loop)

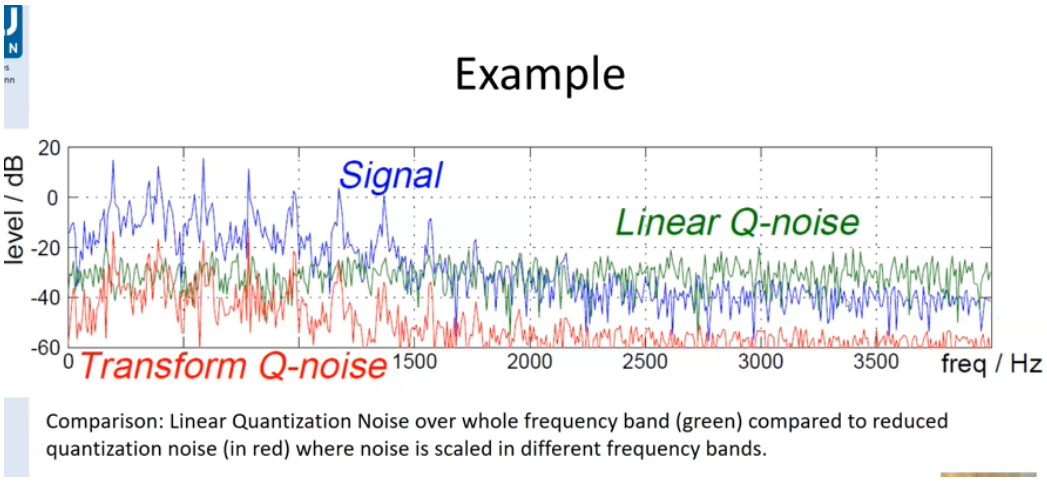
– If the quantization noise in a given band is found to exceed the masking threshold (allowed noise) - supplied by the perceptual model - the scale factor for this band is adjusted to reduce the quantization noise.

Inner iteration loop (rate loop – modifies overall coder rate)

– Check if Huffman coding is small enough

– If number of bits resulting from coding operation exceeds the number of bi… to code a block of data, adjust global gain -> results in larger quantization ste…





Comparison: Linear Quantization Noise over whole frequency band (green) compared to reduced quantization noise (in red) where noise is scaled in different frequency bands.

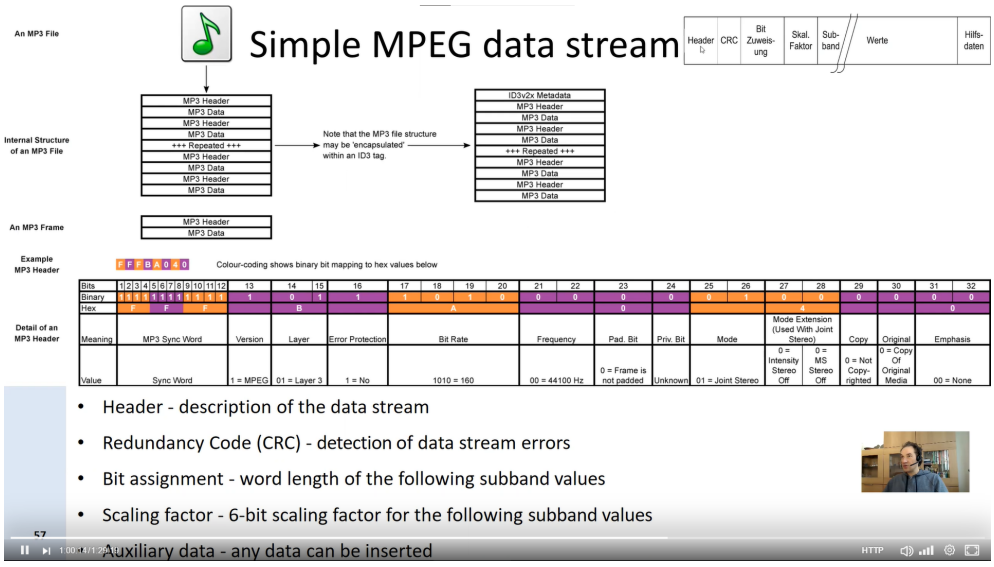
Huffman Encoding

Final quantized bands will be Huffman encoded = lossless encoding (the lossy part happened before)

This is all still done in the frequency domain!

Explanation of Huffman encoding -> later for images

MP3 File

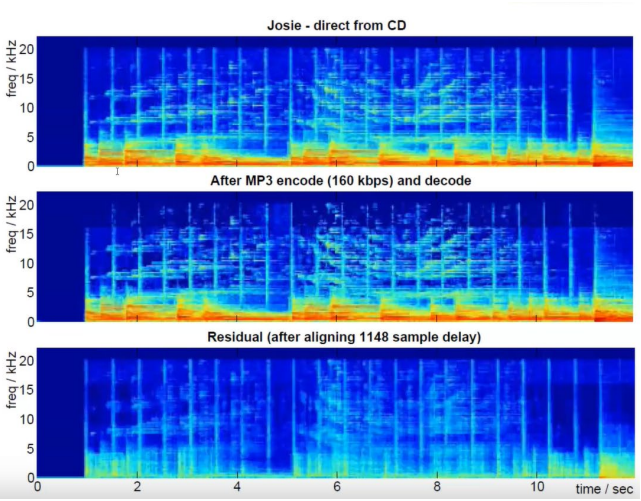


Probleme bei MP3 Kodierung

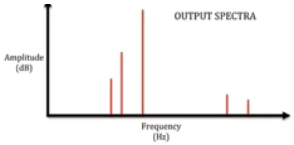
Chop off high frequency (above 16 kHz)

Occasional other time-frequency 'holes'

Quantization noise under signal



Common types of artifacts



- loss of bandwidth (no signal >16kHz)
- pre-echoes
  Name is misleading: the basic coding artifact is noise spread out over some time even before the music event causing the noise
- roughness, double-speak
  at low bit-rates and lower sampling frequencies there is a mismatch between time resolution of the coder and the requirements to follow the time structure of some signals. (e.g. robot like "double-speak")

Qualitätsmessung bei Codecs.

Man macht Hörtests mit vielen Leuten und worst case Szenarios.

AAC

Nachfolger von MP3

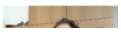
successor of MP3 format

Specified in MPEG-2 (Part 7) (1997) and MPEG-4 (Part 3) (1999)

AAC generally achieves higher sound quality than MP3 at the same bit rate

HE-AAC and HE-AAC v2 (in MPEG-4/Part 3)

- extensions of the original AAC, but do not replace it
- are optimized exclusively for low bit rates and are used, for example, for live broadcasts of digital radio and television.
- At higher bit rates (from 96 kbit/s), they perform worse than AAC and should not be used



AAC hat mehr Samplingraten, auch höhere. Außerdem 48 channels. Erneut variable Framelength und diesmal aber nur eine einfache Filterbank. Nur eine MDCT.

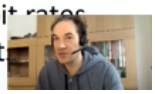
Frequenzen über 16kHz werden auch erfasst. Außerdem gibt es weitere technische Module, beispielsweise das Temporal Noise Shaping = ein weiteres perceptional model.

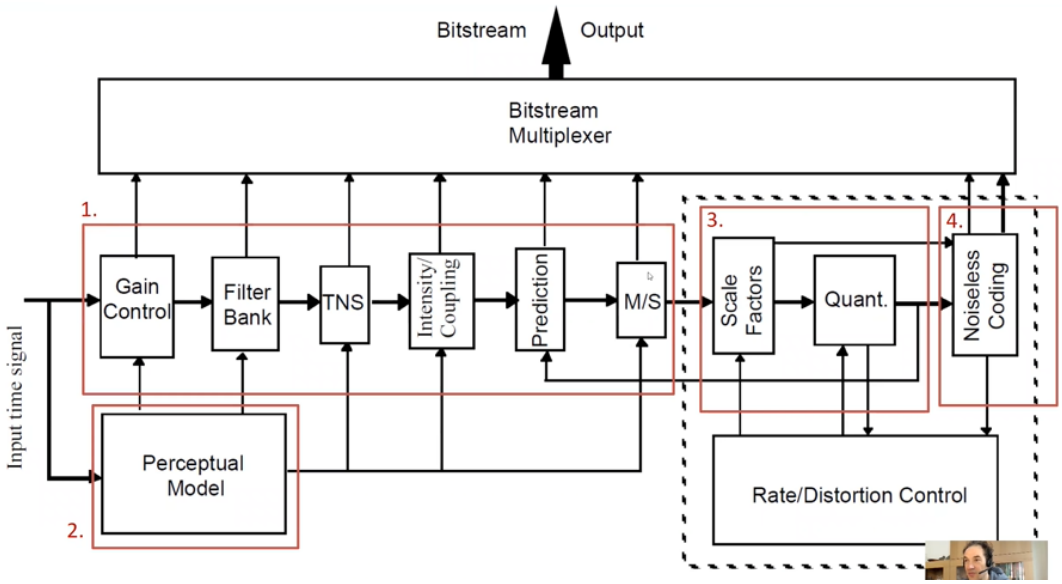
AAC

1. Multi Channel Audio – up to 48 audio channels
2. Sample frequencies from 8KHz - 96KHz
3. Simpler filter bank (pure MDCT used)
4. Better stationary and transient response due to block sizes of 1024 and 128 samples
5. Excellent handling of high frequency signals
6. CD quality audio at 64Kbits/sec
7. Much better quality of audio at lower bit rates (down to 32Kbps)

MP3

1. Stereo signal – maximum of only 2 channels
2. Sampling frequencies from 16KHz ~ 48KHz
3. Hybrid filter bank (requires more computational power)
4. Poorer stationary and transient response due to block sizes of 576 and 192 samples
5. Signal handling up to 15.5/15.8 KHz
6. CD quality audio at 128Kbits/sec
7. Audio quality is poorer at low bit rates and may present coding artifact





Dolby Digital – Eine weitere Standardisierungsbehörde

Dabei gibt es 5.1 mode - Hier gibt es 6 channels, 5 für normal range speakers und einen subwoofer.

Mögliche Kanalkonfigurationen bei Dolby Digital

Name	Modus	Vorne			Hinten		Subwoofer (LFE)	Verwendung
		Links	Mitte (Center)	Rechts	Surround Links	Surround Rechts		
Dolby Digital 1/0	Mono	Nein	Ja	Nein	Nein	Nein	Nein	Vor allem alte restaurierte Filme.
Dolby Digital 2/0	Stereo	Ja	Nein	Ja	Nein	Nein	Nein	Audiokommentar. Kann auch Dolby Surround Informationen enthalten.
Dolby Digital 3/0	3-Stereo	Ja	Ja	Ja	Nein	Nein	Nein	
Dolby Digital 2/1	Stereo mit Mono-Surround	Ja	Nein	Ja	Ja		Nein	
Dolby Digital 2/1.1	Stereo mit Mono-Surround und LFE	Ja	Nein	Ja	Ja		Ja	
Dolby Digital 3/1	3-Stereo mit Mono-Surround	Ja	Ja	Ja	Ja		Nein	
Dolby Digital 3/1.1	3-Stereo mit Mono-Surround und LFE	Ja	Ja	Ja	Ja		Ja	
Dolby Digital 2/2	Quadrophonie	Ja	Nein	Ja	Ja	Ja	Nein	
Dolby Digital 2/2.1	Quadrophonie mit LFE	Ja	Nein	Ja	Ja	Ja	Ja	
Dolby Digital 3/2	5-Kanal-Surround	Ja	Ja	Ja	Ja	Ja	Nein	
Dolby Digital 3/2.1	5-Kanal-Surround mit LFE	Ja	Ja	Ja	Ja	Ja	Ja	Meistens als Haupttonspur bei Spielfilmen oder auch Serien eingesetzt.

VO6 - Spatial Audio

Binaural Audio

Ist Stereo, aber anders als Stereo für speaker explizit für Kopfhörer. Kann 3D Audio in 360 Grad erzeugen. Aufgenommen werden kann sowas mit binaural microphones, also wirklich Kopf/Ohr-Nachbildungen als Mikrofon. Aber auch über Software – Das ist natürlich in Echtzeit wichtig (Computerspiele).

Es gibt aber verschiedene Kopf- und Ohrformen, demnach ist es schwer perfekten Sound für jeden zu erzeugen.

Man nimmt beide Signale für links und rechts, lediglich wird bspw. rechts verzögert und dumpfer.

HRTF – Head Related Transfer Function

Würde man messen wollen, wie genau Ton im Innenohr ankommt, müsste man für jeden Kopf und jedes Ohr die HRTF errechnen.

Dolby Atmos

Gibt verschiedene Standards. Bspw. 7.1.4 = / Lautsprächer für runderhum, einen Subwoofer und 4 an der Decke.

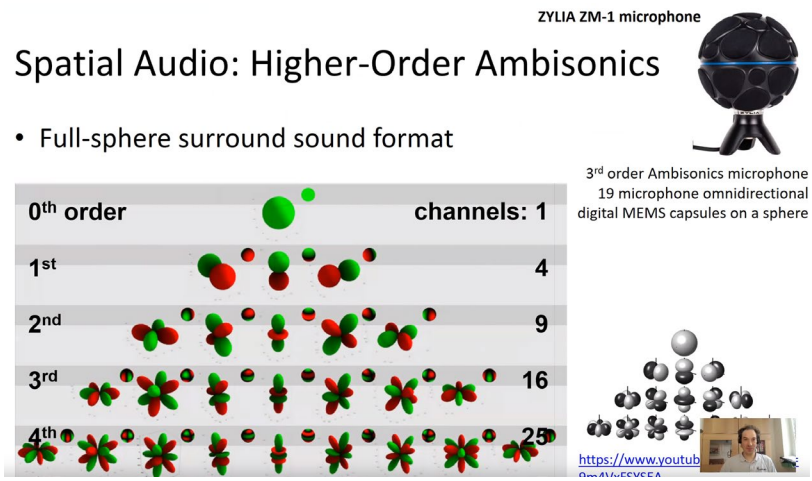
Der Trick dabei ist, dass man nicht Lautsprecher definiert, sondern Audioobjekte die sich iwo im Raum befinden. Die werden dann je nach Position gerendert.

Das wird gemacht indem man in die Audiofiles auch Positionen speichert. Damit wird im Soundfile gespeichert, wie das abgespielt werden muss. Der Computer kann dann errechnen auf welchem Lautsprecher was wie laut gespielt werden muss.

Vorteil ist auch, dass man nicht mit jedem Setup neu mischen muss.

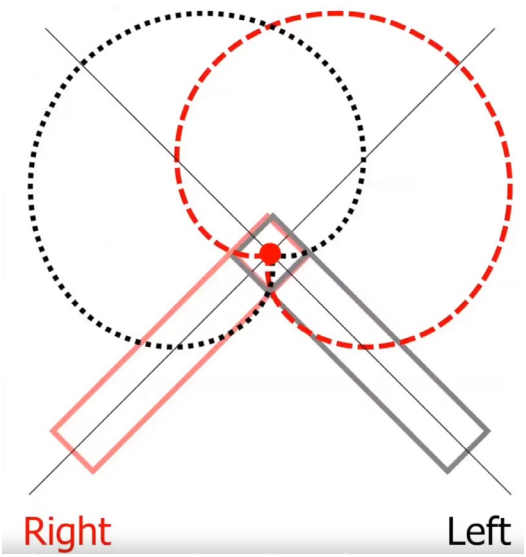
Man braucht aber einen Audiorenderer. Der muss eben wissen wo die Lautsprecher im Raum sind.

Ambisonics

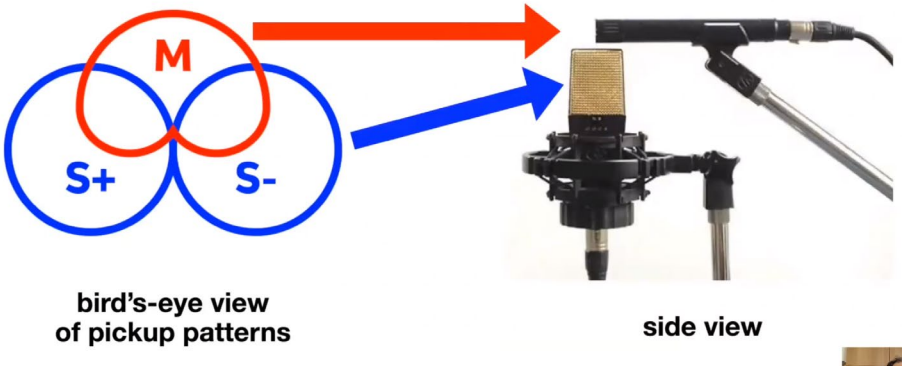


360 Grad Soundaufnahmen.

Stereomikros sind meist so aufgebaut:

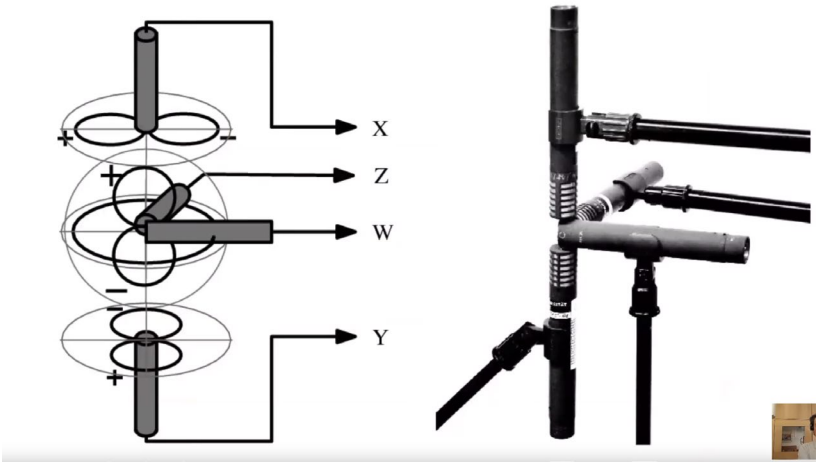


Man kann aber auch so aufnehmen:



Damit deckt man schon mehr ab.

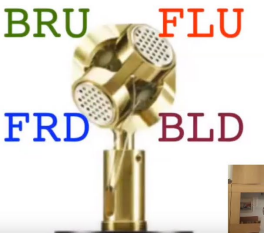
Wenn man jetzt Ambisonics will, braucht man vier Mikros:



Tatsächliche Ambisonicsmikros sehen meist eher so aus, die Aufnahmen werden so errechnet:

W = FLU+FRD+BLD+BRU
X = FLU+FRD−BLD−BRU
Y = FLU−FRD+BLD−BRU
Z = FLU−FRD−BLD+BRU

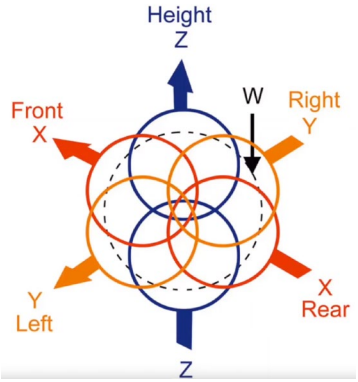
FLU is Front Left Up
FRD is Front Right Down
BLD is Back Left Down
BRU is Back Right Up



Formate

Die raw recordings nennt man A-Format, sie sind in 4 channels.

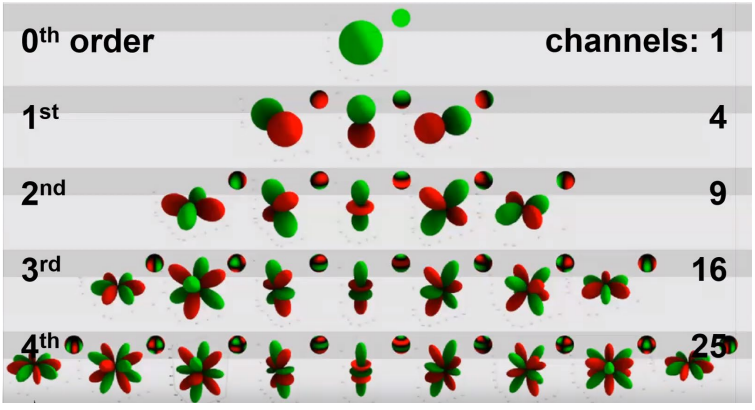
B-Format sind bereits einen Schritt im Umwandeln durchgangen. Die Signale wurden verbunden zu mid-side, up-down, front-back.



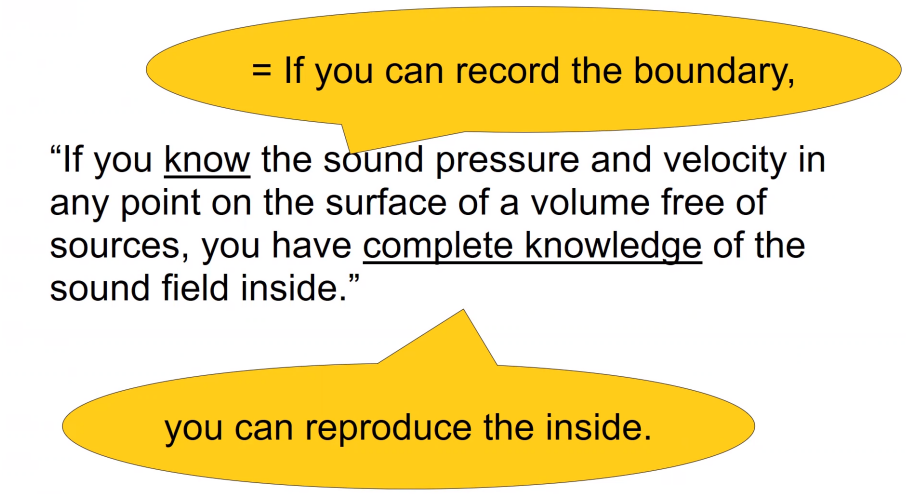
Toll an Ambisonics ist, dass man die Aufnahmen dann runterrechnen kann auf Binaural Audio.

Higher-order ambisonics

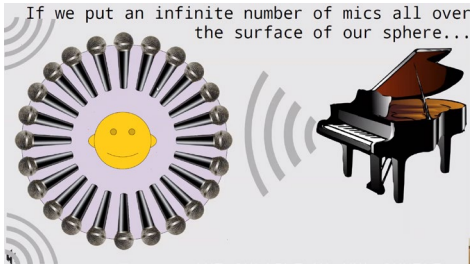
Benutzt man noch mehr Mikros, bekommt man mehr als das von vorher (first-order ambisonics). Damit füllt man die deadspots der vier Mikros.



Kirchhoff-Helmholtz Integral

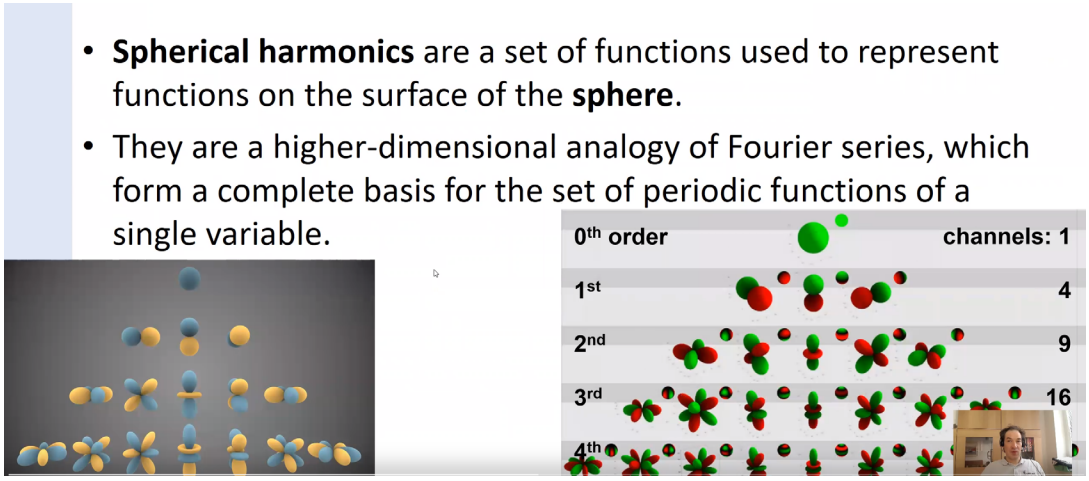


Heißt, wenn man das Äußere, die Hülle komplett aufnehmen kann, kann man das Innere reproduzieren.



Spherical Harmonics

Wie bereits vorher gezeigt, kann man mit X-oder Ambisonics jeweils neue Spheren in den Kreis hinzufügen, was im Endeffekt zu immer dichteren Aufnahmen führt.



5.1 surround vs Ambisonics

5.1 surround

- Localization achieved solely with level
- Poor inter-speaker imaging except at front
- Localization varies with listener position
- Small sweet spot
- Horizontal surround uses six channels
- Speakers in special positions (e.g. ITU)
- One-to-one channel relationship from studio to speakers

Ambisonics

- Localization includes other mechanisms
- Good inter-speaker imaging all round
- Less variation of image with position
- Larger sweet spot
- Even with height, only four channels needed (1st order)
- Decode to 5.1 or other configurations
- Recover B-Format and decode for completely flexible speaker arrays

Wellenfeldsynthese / Virtuelle Schallquellen

- **Echte Schallquellen**: z. B. Gitarre

- **Phantomschallquelle** ist eine Schallquelle, die von einem Hörer an einem bestimmten Ort wahrgenommen wird, aber in Wirklichkeit gar nicht an diesem Ort existiert.
  - wahrgenommene Position der Phantomschallquelle hängt von der Position des Hörers ab
  - Dies ist bei allen Stereo und Surroundverfahren der Fall.

- Eine **virtuelle Schallquelle** ist ebenfalls eine Schallquelle, die nicht existiert. Virtuelle Schallquellen werden aber unabhängig von der Position des Hörers immer am selben Ort lokalisiert.

Wellenfeldsynthese ist eine Möglichkeit solche herzustellen.

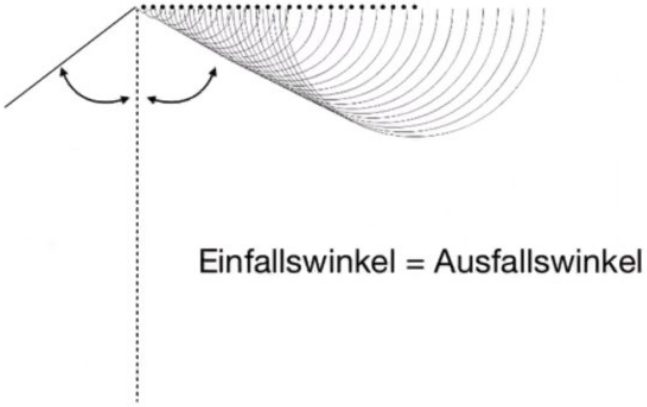
Huygensches Prinzip:

- jeder Punkt einer Wellenfront kann als Ausgangspunkt einer neuen Welle, einer sogenannten Elementarwelle, gesehen werden.

- Die Lage einer Wellenfront ergibt sich aus der Überlagerung ihrer Elementarwellen.

- Huygens konnte mit dieser Theorie die Phänomene Beugung und Brechung erklären.

Huygensches Prinzip:



Kirchhoff-Helmholtz-Integral:

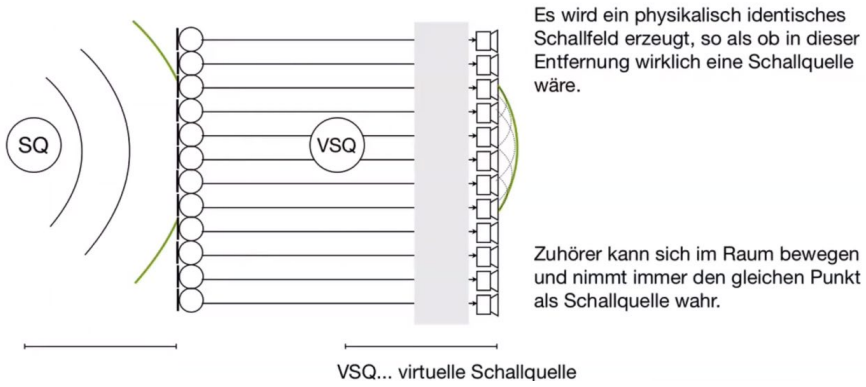
- besagt, dass der Schalldruck an jedem Punkt in einem Raum definiert ist, wenn man den Schalldruck und die Schallschnelle an jedem Punkt der begrenzenden Oberfläche dieses Raumes kennt.

- Man kann jedes beliebige Schallfeld erzeugen, wenn man es schafft auf der Raumoberfläche bestimmte Schallverhältnisse herzustellen.

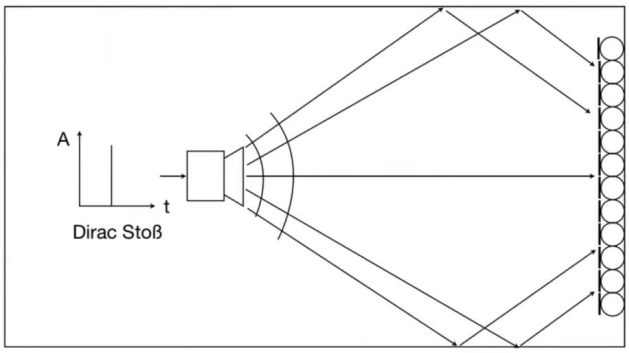
Veranschaulichung – Kirchhoff-Helmholtz-Integral:

- Wenn man mit einem Drumstick auf das Schlagfell einer Trommel schlägt, so bewegt sich das Schlagfell nach einem bestimmten Muster.

- <u>Man könnte genau das gleiche Bewegungsmuster auch hervorrufen, indem man ganz bestimmte Bewegungen am Rand der Membran durchführt.</u>



Aufzeichnen der Impulsantwort eines Raumes:



für unterschiedliche Lautsprecherpositionen

Zwischenpositionen werden bei der Wiedergabe mathematisch aus den Daten extrapoliert.

Vorzüge der Wellenfeldsynthese:

1. Wir können virtuelle Schallquellen hinter der Wand abbilden, die nicht von der Hörerposition abhängen.

2. Wir können durch die Synthetisierung von konkaven Wellenfronten, Schallquellen in den Wiedergaberaum projezieren.

   Mit gewissen Einschränkungen kann der Zuschauer auch um diese virtuelle Schalquelle herumgehen.

3. Es können die akustischen Eigenschaften eines Aufnahmeraumes mit all seinen Reflexionen in den Wiedergaberaum übertragen werden.

Vorzüge der Wellenfeldsynthese:

4. Bewegte virtuelle Schallquellen erzeugen bei dieser Form der Wiedergabe einen Doppler-Effekt.

5. Es könnten komplette Mischungen erstellt werden bei denen jedes Instrument eine eigene virtuelle Position erhält.

   Theoretisch könnte jedes Instrument im gleichen Raum spielen, aber auch jedes Instrument in einem andern.

6. Stereo- und Surroundmischungen könnten auch über dieses System wiedergegeben werden. Dabei könnte man für jeden Lautsprecher eine beliebige Position vor oder hinter der Wand des Raumes wählen.

Nachteile und Probleme:

1. Die Installation einer solchen Anlage ist ein immenser Aufwand.
2. Das Ganze funktioniert nur in einer zweidimensionalen Ebene und nicht dreidimensional.
3. Es kommt zu einer Überlagerung des synthetisierten Wellenfeldes mit der echten Raumakustik des Wiedergaberaumes. Darum müsste der Wiedergabe-Raum stark bedämpft werden.
4. Aufgrund der begrenzten Anzahl an Lautsprechern lässt sich das Wellenfeld nicht 100%ig realistisch herstellen, sondern es kommt zu sogenannten Aliasing Effekten. Es entstehen Punkte im Raum mit schmalbandigen Einbrüchen im Frequenzgang.

Nachteile und Probleme:

5. Wenn der Kreis der Lautsprecher um den Zuhörer nicht ganz geschlossen ist, weil er zum Beispiel durch eine Tür durchbrochen wird, erzeugt dies unerwünschte Effekte – sogenannte Schattenwellen. Diese können vermindert werden, wenn die letzten Lautsprecher in der Reihe etwas im Pegel verringert werden.

6. Bei der Wahrnehmung von in den Raum projizierten virtuellen Schallquellen kann es dann zu Problemen in der Wahrnehmung kommen, wenn sich der Zuhörer zwischen den Lautsprechern und der virtuellen Schallquelle befindet.

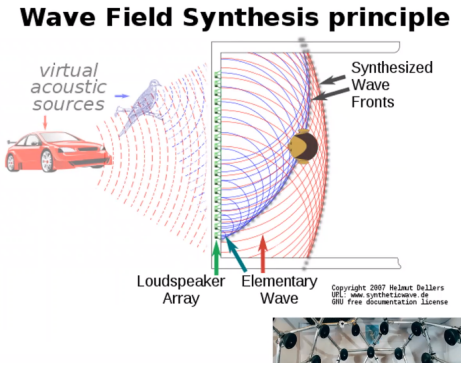
7. Bewegte Schallquellen sind aufgrund der beschränkten Rechnerleistung derzeit nur mit Einschränkungen möglich.

Spatial Audio: Wave Field Synthesis (WFS)

A spatial audio rendering technique, characterized by creation of virtual acoustic environments

Produces artificial wavefronts synthesized by a large number of individually driven loudspeakers.

The localization of virtual sources in WFS does not depend on or change with the listener's position



MPEG-H 3D Audio

MPEG-H 3D Audio (2013 -)

Supports coding audio as audio channels, audio objects, or higher order ambisonics (HOA)

Can support up to 64 loudspeaker channels and 128 codec core channels

Audio objects may be used alone or in combination with channels or HOA components

Use of audio objects allows interactivity or personalization of a program by adjusting gain or position of the objects during rendering in the decoder

Audio is encoded using an improved modified discrete cosine transform (MDCT) algorithm

Transmit immersive sound as well as mono, stereo, or surround sound.

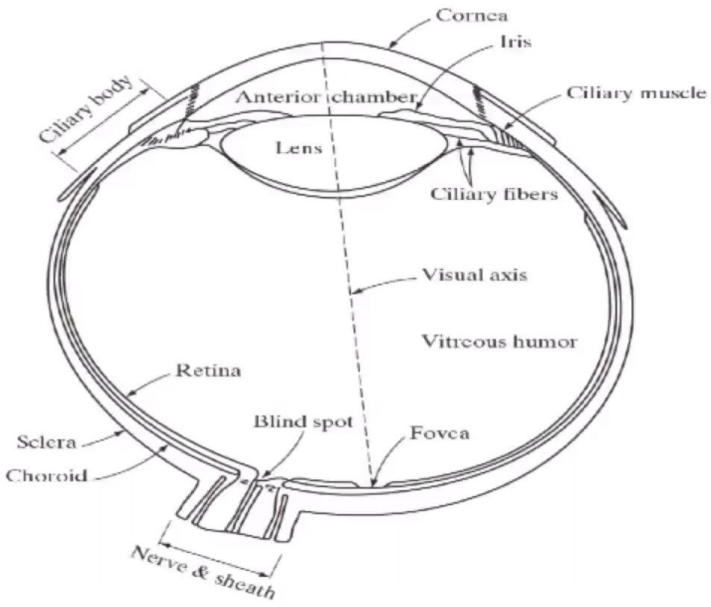
MPEG-H 3D Audio decoder renders bitstream to a number of standard speaker configurations as well as to misplaced sp

Binaural rendering of sound for headphone listen supported

Used since May 2017 in terrestrial South Korean UHD TV broadcast

VO7 & VO8 - Bild

Auge



Licht kommt bei der Cornea rein, durch die Linse auf die Retina. Bei der Retina gibt es Sensoren. Zwei Typen – Cones (Detailsehen, im hellen Licht sehen) und Rods (Für die general idea und gut im dimmen Licht). Es gibt besonders viele Cones bei der Fovea, dort sieht man am besten. Blind spot hat keine Sensoren.

Intensitäten die wir sehen können:

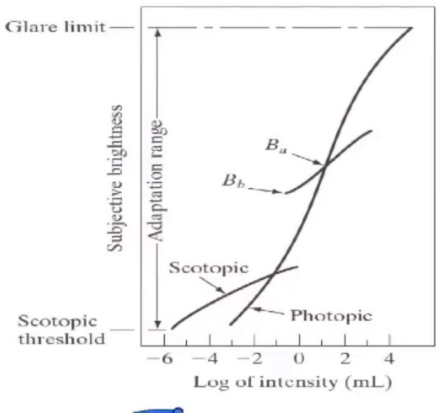


FIGURE 2.4 Range of subjective brightness sensations showing a particular adaptation level.

Wir können aber nicht gleichzeitig low und high intensities sehen. Dazu müssen wir zuerst adapten.

In einem dunklen Raum werden kleine Änderungen nicht wahrgenommen. Wenn der Hintergrund aber sehr hell ist, sieht man den Unterscheid schnell. Das nennt man Weber law bzw. Weber ratio oder Webersches Gesetz.

Licht

Achromatisches Licht = ohne Farbe

Chromatisch = mit Farbe

Farbe hängt ab von surface, Reflektivität, physischen Eigenschaften, Komposition, Lichtquellen, Farbe der Umgebung…

Luminanz / Luminance

Ist die Menge an Licht, die von etwas emittiert wird. Wird in Candela/Quadratmeter = cd/m^2 gemessen.

Brightness

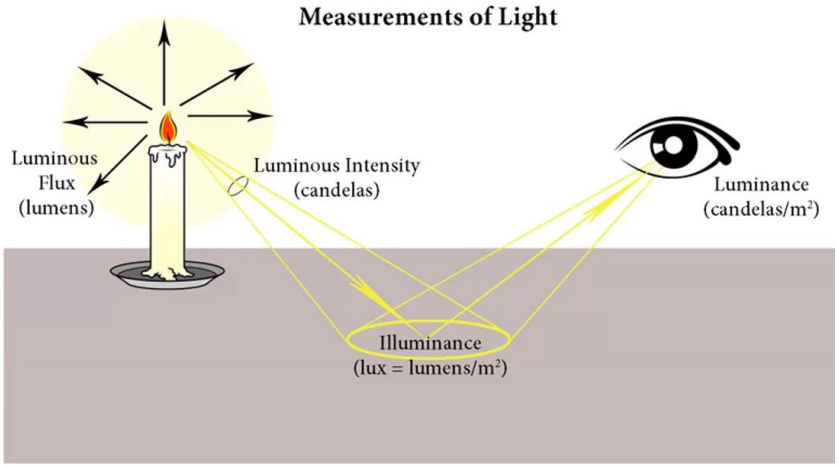
Kein objektives Maß, sondern Wahrnehmungssache. Also wie hell empfindet man etwas.

Dynamic Range

Verhältnis zwischen minimalen und maximalen Intensitäten.

Illuminance

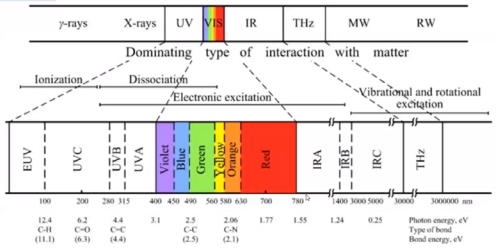
Menge an ausgestoßenen Licht. Gemessen in lx (lux) oder lm/m^2

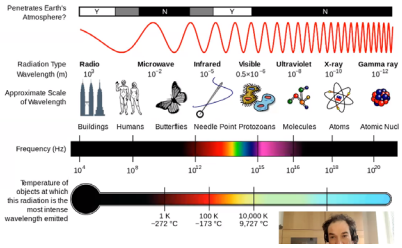


Farbe

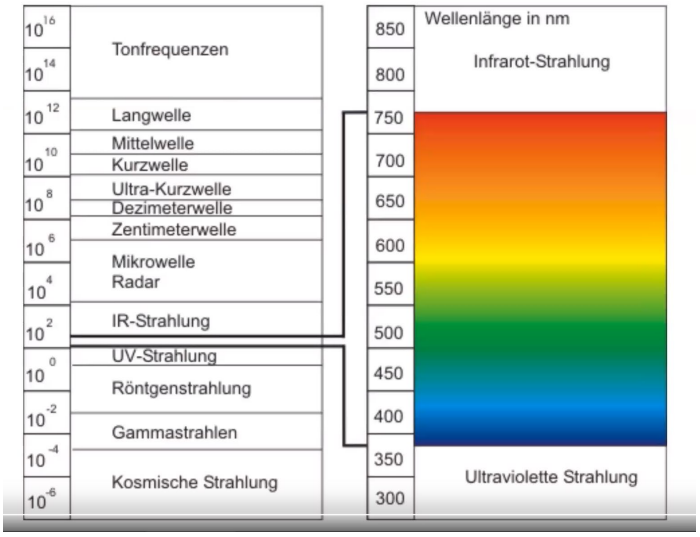
Wir könne Licht zwischen 400 und 700 nm Wellenlänge wahrnehmen.

- Color is how we perceive beams of electromagnetic energy that fall on our retina
- We see light between 400 and 700 nanometers in wavelength (10^{-9} meters)





Light Spectrum – Wavelengths



Für Farbe gibt es verschiedene Darstellungsmodelle.

Printing & graphics industry: color sample books and codes

Artists: tint (adding white to pure pigment), shade (adding black), tone (adding both)

Displays: use hardware capabilities (RGB)

Physicists: use optical models (wavelengths, energy measures)

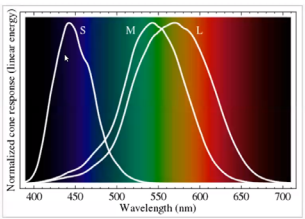
Cones sind für die Farbwahrnehmung verantwortlich.

Man hat drei Arten von Cones die für unterschiedliche Wellenlängen verantwortlich sind.

The retina contains cones that sense either red, green, or blue light

6-7 million cones per eye, concentrated in central area called "fovea"

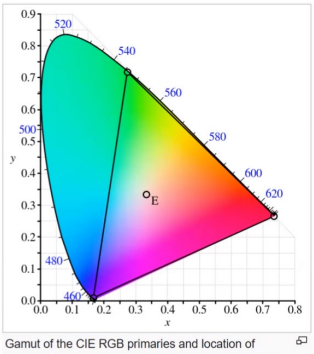
Rods: sensors that surround fovea, and perceive weak "night" images (>100 million per eye)



Normalized response spectra of human cones, to monochromatic spectral stimuli, with wavelength given in nanometers.

Farbmodelle

CIE XYZ Colour Space

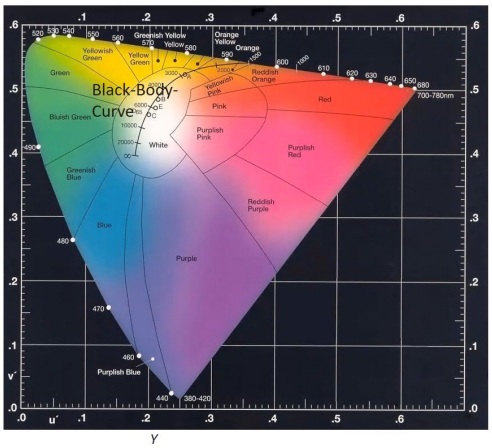


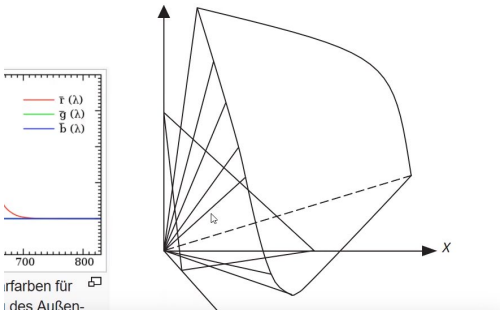
Gamut of the CIE RGB primaries and location of primaries on the CIE 1931 *xy* chromaticity diagram.

Ist die Standardreferenz für colour spaces.

Außen sind die Wellenlängen die wir sehen können, drinnen Mischfarben.

Es handelt sich eigl. um einen Kegel. Ganz unten an der Spitze ist Schwarz, oben in der Mitte der Basis des Kegels Weiß.

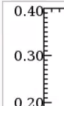


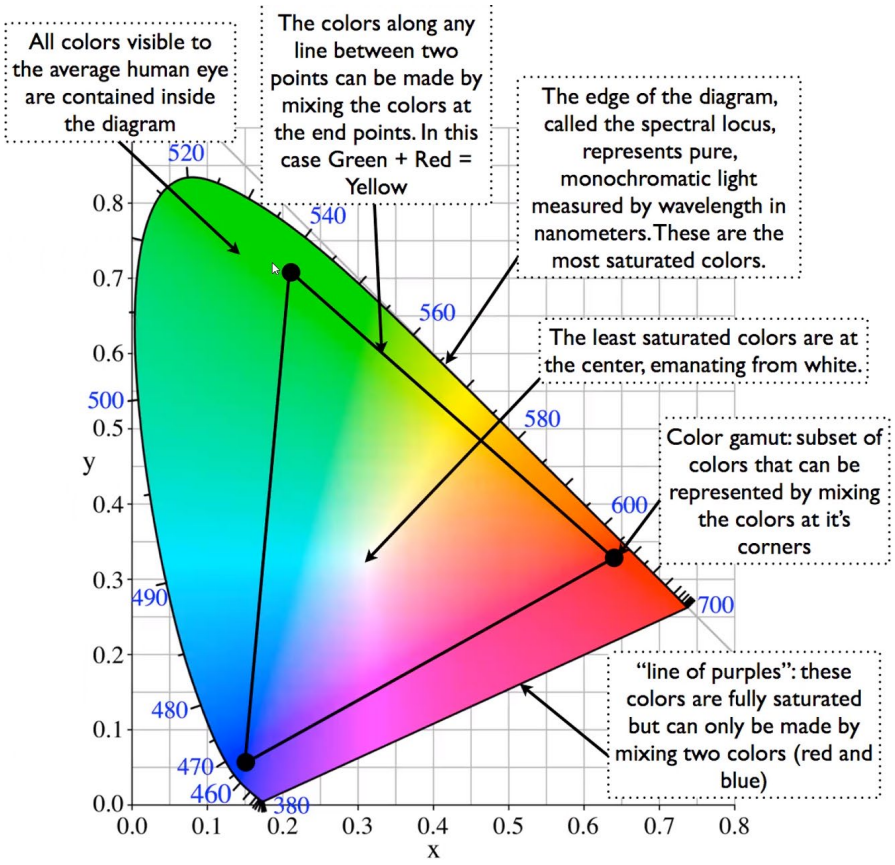


rfarben für
des Außen-

Ein Schnitt durch den Kegel ist was wir in diesen Diagrammen sehen.

Color described in x, y (with x+y+z=1)
Luminance = Y
Full color: (x, y, Y)





Anatomy of a CIE Chromaticity Diagram

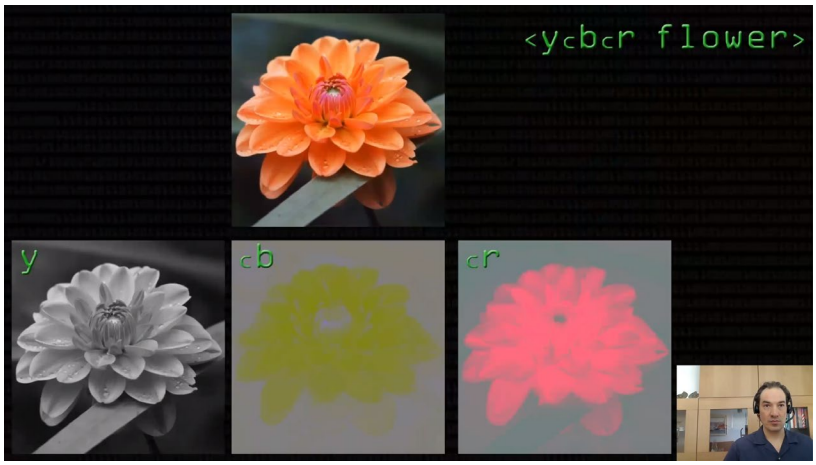
Helligkeitssehen:
~120 Mill. "Stäbchen"/Rods
Farbsehen:
~7 Mill. "Zapfen"/Cones

Wir können Helligkeit sehr viel besser unterscheiden als Farben.

YCbCr

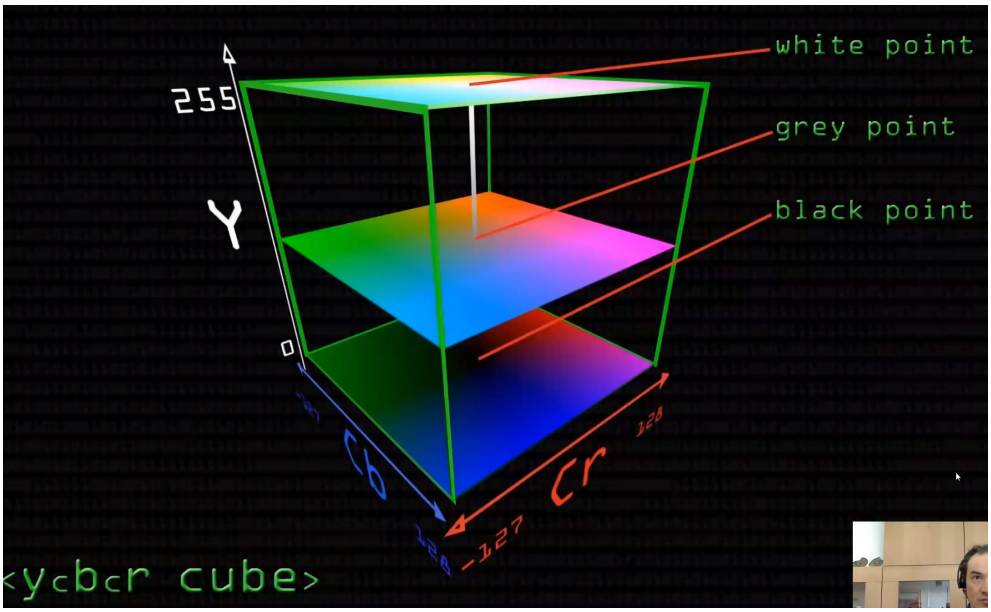
Hier wird die Intensität und Helligkeit von der Farbe getrennt.

Y ist einfach nur Greyscale, also die Helligkeit



Y ist zwischen 0 – 255

Cb und Cr zwischen -127 und 128



Farbmodelle

CIE color space (Commission Internationale de l'Eclairage) used to calibrate other color models 1931 CIE XYZ, tristimulus theory

RGB — for video display drivers

CMYK — cyan, magenta, yellow, black, (subtractive primaries)

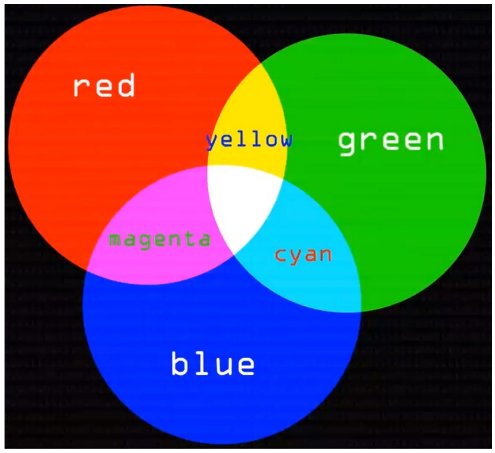
HSB — hue (dominant color), saturation (amount of gray), brightness (intensity)

YUV — used in television industry, Y (luminance), UV (color difference signals) = analog

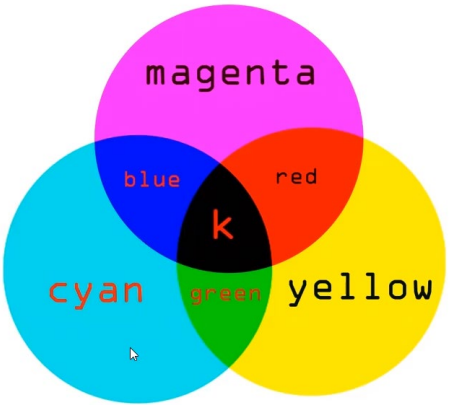
YCbCr = digital; Y (luminance), C_B and C_R are the blue-difference and red-difference chroma components

RGB und CMY

RGB

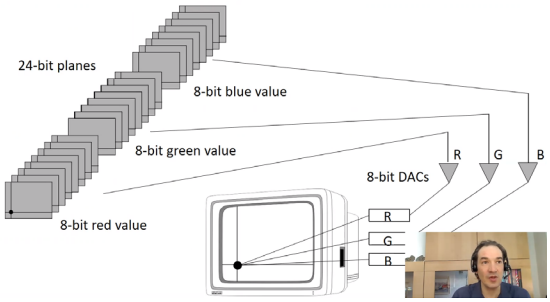


CMYK

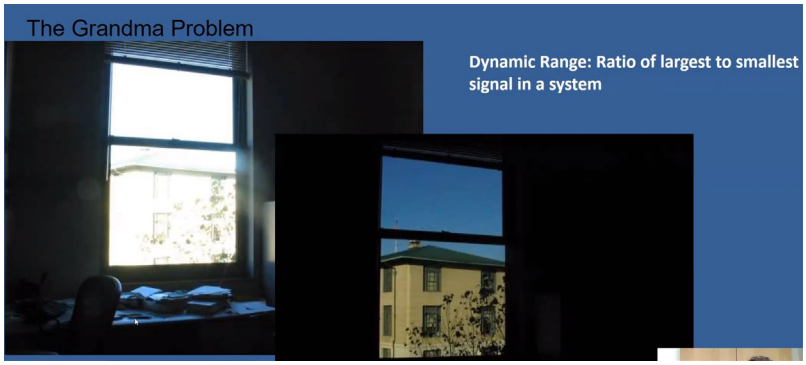


Bit Depth

image descriptor = {
image width = 640
image height = 480
image depth = 24 bit
color model = RGB
encoding = YUV 8:2:2, JPEG}

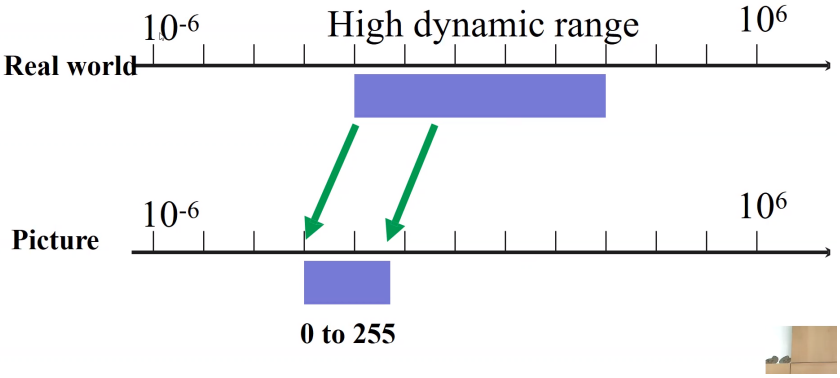


High Dynamic Range



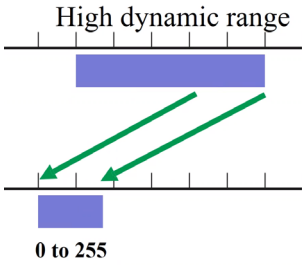
Bei HDR will man eine große Bandbreite an Helligkeiten (und Farben) darstellen.

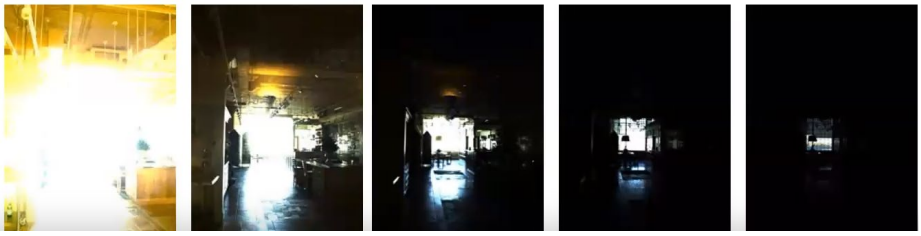
In der realen Welt kann man große Bereiche wahrnehmen.



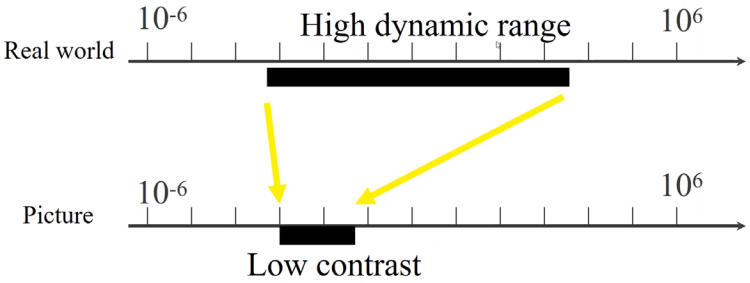
Wenn man in einem Bild lang beleuchtet, also long exposure hat, kann man sehr dunkle Bereiche darstellen.

Belichtet man nur kurz, kann man die hellen Bereiche abbilden.

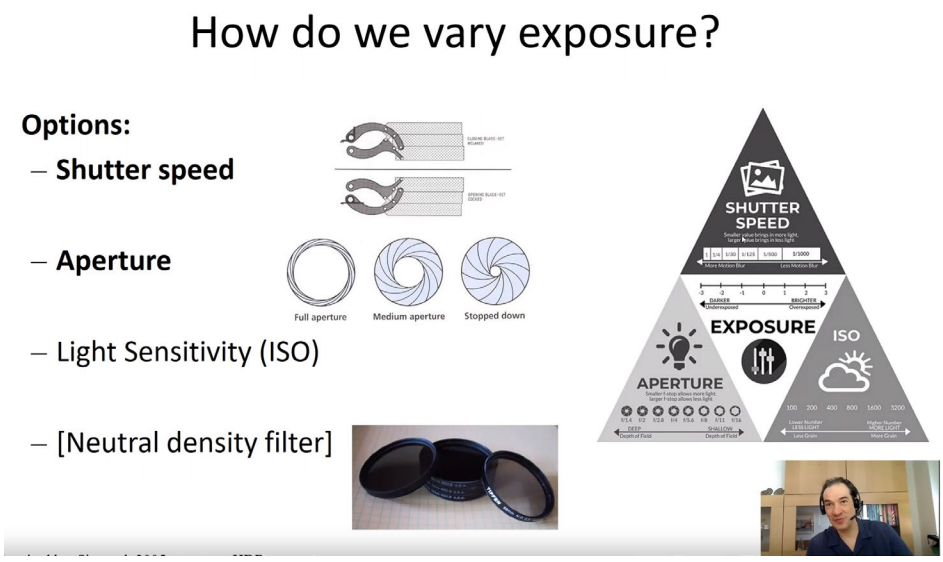




Mit HDR will man alle Bereiche abbilden können. Dazu merged man alle Bilder darüber.



Möglichkeiten zum Variieren der exposure



Shutter speed
- Range: ~30 sec to 1/4000sec (6 orders of magnitude)
- Pros: reliable, linear
- Cons: sometimes noise for long exposure

Aperture
- Range: ~f/1.4 to f/22 (2.5 orders of magnitude)
- Cons: changes depth of field

Light Sensitivity (ISO)
- Range: ~100 to 1600 (1.5 orders of magnitude)
- Cons: noise

Neutral density filter
- Range: up to 4 densities (4 orders of magnitude) & can be stacked
- Cons: not perfectly neutral (color shift), not very precise, need to touch camera (shake)
- Pros: works with strobe/flash, good complement when desperate

e after Siggraph 2005 course on HDR

Beim Mergen muss man aber wissen wie die response curve des Sensors aussieht, damit man weiß, wie man was gewichten muss.

F-stop

Ein stop bedeutet, dass die Lichtmenge die in die Kamera kommt doppelt so hoch ist wie davor. 12 Stops dynamic range würde heißen, dass die hellsten parts 2^12 mal so hell sind wie die dunkelsten.

Dynamic ranges of common devices

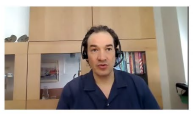
Device	Stops	Contrast ratio
Glossy photograph paper	7 (7 - 7 2/3) [58]	128:1
LCD	9.5 (8 – 10.8)[citation needed]	700:1 (250:1 – 1750:1)
Negative film (Kodak VISION3)	13[59]	8000:1
Human eye	10–14[54]	1000:1 – 16000:1
High-end DSLR camera (Nikon D850)	14.8[60]	28500:1
Digital cinema camera (Red Weapon 8k)	16.5+[61]	92000:1

Das menschliche Auge kann 10-14 Stops sehen. (Ohne Anpassung)

Bit per pixel

How many unique shades are available in grayscale image

Bit Precision of Analog/Digital Converter	Contrast Ratio	Dynamic Range	
		f-stops	Density
8	256:1	8	2.4
10	1024:1	10	3.0
12	4096:1	12	3.6
14	16384:1	14	4.2
16	65536:1	16	4.8



Most digital cameras us a 10 – 14 bit A/D so their theoretical maximum dynamic range is 10-14 f-stops.

Higher precision A/D converter does not necessarily mean greater dynamic range. Total dynamic range is usually limited by noise levels.

Nit

Candela/m^2 = Das Licht das eine Kerze auf einen Quadratmeter abgibt.

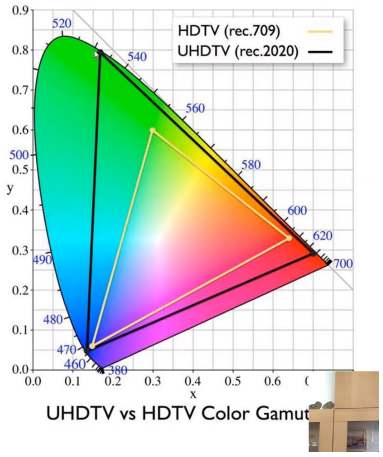
Colour Gamut

Wider Color Gamut

- Current color gamut is limited
  - Current HDTV (BT.709) – 1990
  - Created for CRT, and HDTV
  - Around 100 nits.
  - 8 bpp (bits per pixel)

Today we have better display technology: LCD, OLED …

More Colors: REC. 2020 for UHDTV 10, 12 bpp.



Quantisierung – Linear vs non-Linear

Unsere Wahrnehmung von Helligkeit ist nicht linear. Deswegen muss man mit einer Gamma-Kurve quantisieren, damit man für das menschliche Auge eine gleichmäßige Helligkeitskurve hat.

Wenn man aber sehr weit rauf geht bis 10.000 nit, hat man bei hellen Farben Bandfehler, man braucht also nochmal eine andere Kurve. Die Kurve heißt SMPTE 2084 : HDR Electro – Optical Transfer Function. (EOTF) Diese übersetzt Helligkeit auf verschiedene Displaytechnologien. Je nach Displayart ist die unterschiedlich.

HDR vs SDR

- Content with a wider range of brightness and color
- Requires new monitors to experience HDR content
  - Legacy content (SDR): 8-bit color, 100 nit peak brightness, 709 gamut
    - Dynamic range of about 6 stops
  - HDR content: 10-bit color, 10,000 nit peak brightness, 2020 gamut
    - Dynamic range of about 17,6 stops

HDR is an ambiguous term

- There are several industry specs defining various flavors of HDR

Formate zum Speichern von HDR

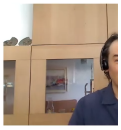
Meistens lossless

Adobe DNG (digital negative)

- Specific for RAW files, avoid proprietary formats

RGBE

- 24 bits/pixels as usual, plus 8 bit of common exponent
- Introduced by Greg Ward for Radiance (light simulation)
- Enormous dynamic range



OpenEXR

- By Industrial Light + Magic, also standard in graphics hardware
- 16bit per channel (48 bits per pixel) 10 mantissa, sign, 5 exponent
- Fine quantization (because 10 bit mantissa), only 9.6 orders of magnitude

JPEG 2000

- Has a 16 bit mode, lossy

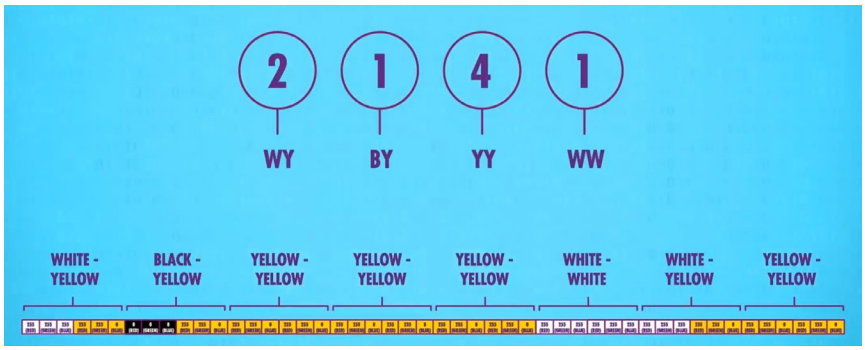
Komprimierung

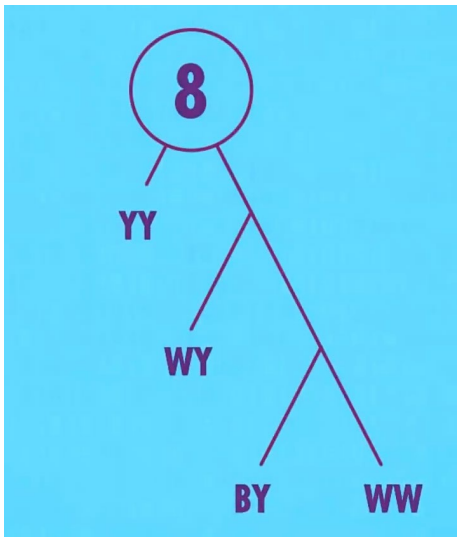
Run-Length Encoding

Redundante Daten werden einfach gezählt und statt sieben gelbe Pixel zu speichern, speichert man einen gelben Pixel und einen Siebener. Es ist außerdem lossless.

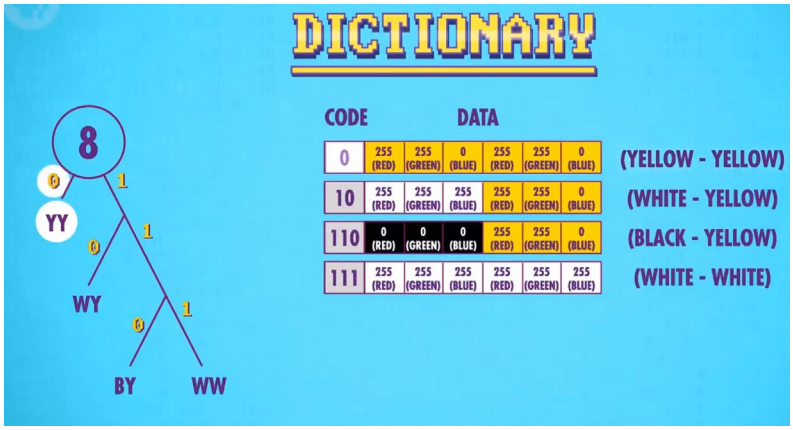
Huffman Encoding

Man baut einen Tree auf, der die kleinsten Frequencies unten anordent, die häufigsten oben. Das ist auch lossless.

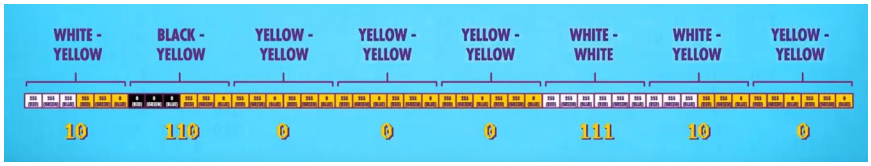




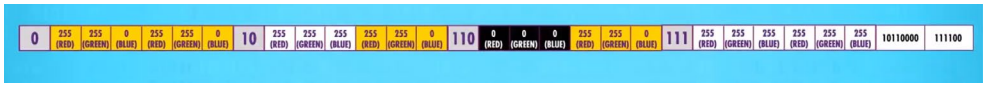
Die Kanten werden dann mit Zahlen beschriftet und damit ein Code generiert.



Die kürzesten Codes sind die, die am häufigsten gebraucht werden.



Man muss aber natürlich auch das Dictionary hinzufügen.



Perceptional Coding

Wie bei MP3 Sachen rausnehmen, die man eh nicht hören kann.

Temporal Redundancy

Bei Videos muss man nur speichern was sich von Frame zu Frame ändert.

Klassifizierung von Compression Techniques

Entropy coding	run-length coding	
	Huffman coding	
	arithmetic coding	
Source coding	prediction	DPCM
		DM
	transformation	FFT
		DCT
	vector quantization	
hybrid coding	JPEG	
	MPEG	
	px64	
	DVI (RTV, PLV)	

Entropy Encoding = Informationsgehalt(Entropie) wird komprimiert. Außerdem ist Entropy Encoding meist lossless und immer unabhängig vom Inhalt der Nachricht.

Source Coding = Informationsinhalt wird gecheckt und danach wird entschieden wie komprimiert werden kann. Ist lossless und abhängig vom Inhalt der Information. DCT ist ein Beispiel.

Hybrid mischt beide Formen.

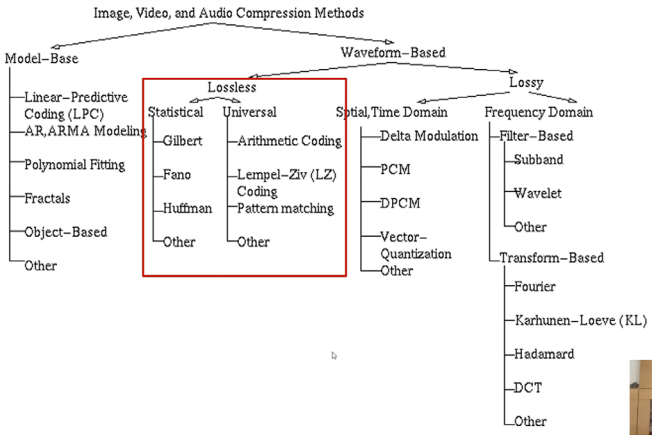
Decompression Requirements

Es gibt symmetrische und asymmetrische compression. Also gleiche Zeit für Komprimieren und Dekomprimieren = symmetrisch. Und vice versa.

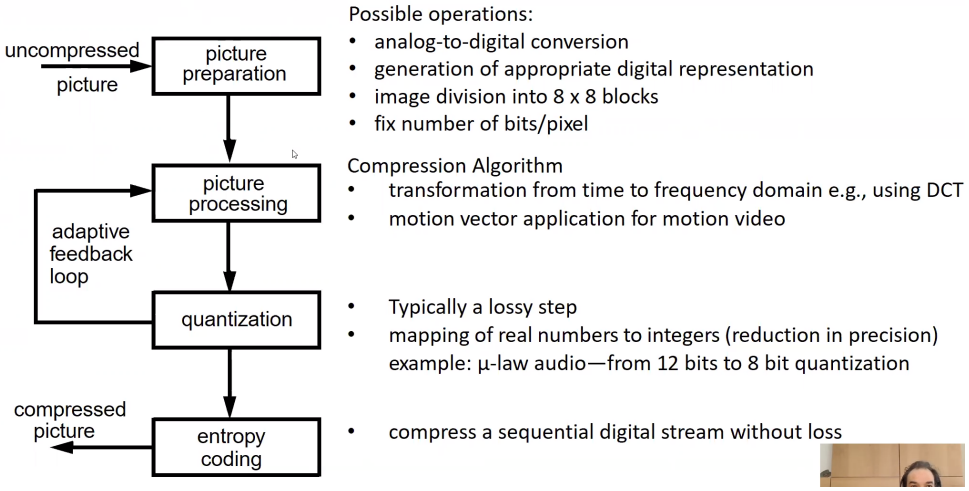
Ersteres braucht man für Dialog mode applications (editing).

Zweiteres für retrieval mode applications (video playback) – Hier wird die compression einmal performed, die decompression hingegen frequently und fast.

Methoden



Schritte für Image Compression



Run-Length Encoding

Man schaut wie häufig ein Wert in einem Signal vorkommt. Wenn der mindestens 4x vorkommt, zählt man die Vorkommnisse, schreibt eine Flag davor, also bspw. ! und dann die Anzahl.

uncompressed sequence (20 Bytes):

ABCCCCCCCCCDEFFFFGGG

compressed sequence (13 Bytes): ABC!9DEF!4GGG

Statistical Encoding

Man schaut wie oft ein Symbol insgesamt in einer Datenmenge vorkommt. Je nachdem was am meisten vorkommt, teilt man dafür Codes zu.

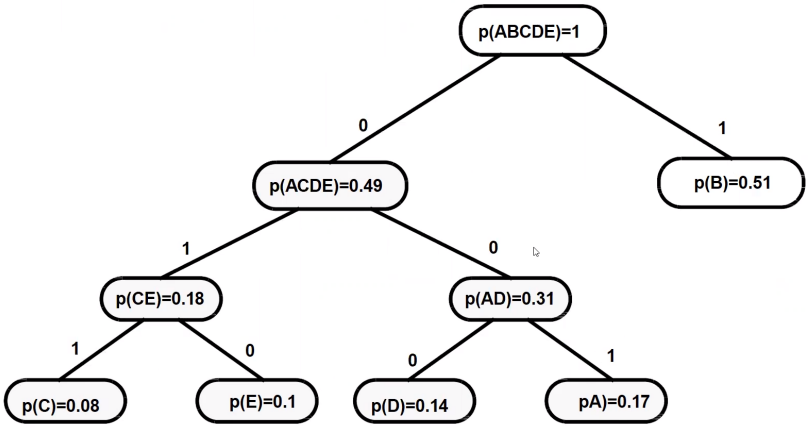
Ex.: p(A)=0.16, p(B)=0.51, p(C)=0.09, p(D)=0.13, p(E)=0.11

first choice: encode A,B,C,D,E as 000,001,010,011,100

Huffman Encoding

Ist eine Entropie Encoding Methode.

```
build ordered list of symbols (increasing probability)
do while list contains at least 2 elements
       construct tree using the first two elements in list
       add parent node for the union of these elements and compute probability
       mark edges by '0' and '1'
       delete the first 2 elements in list; insert parent into list
end
```



Mit diesem Tree kann man die Codes festlegen. C = 011 bspw. P hat einfach 1 und ist der häufigste character.

Der Tree ist immer mit abgespeichert.

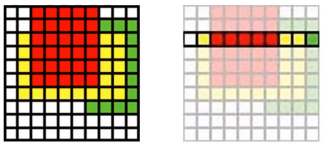
Huffman Coding kann für Musik, Bild und Videos benutzt werden.

Lempel Ziv

Funktioniert nur mit kleinen Datenmengen. Man legt wieder ein Wörterbuch an, das haben Encoder und Decoder.

GIF-Format (Graphics Interchange Format)

- LZW-Compression (Lempel, Ziv, Welch): works line-wise



- 1 st line, 2nd line
- 3rd line is compressed as:
  - 1 white, 1 yellow, 5 red, 2 yellow, 1 green
- Row 4 to 6: „as row 3“

Indexed (1-8 bit Color Lookup Table)

Differential Encoding

Ist source coding. Bei Video bspw. einfach die Unterschiede.

Lossy Compression

Wie bei Audio – was man nicht hören kann, hier was man nicht sehen kann.

The eye is more sensitive to brightness changes than to color changes

The eye is not able to perceive brightness above or below certain threshold values

The eye does not perceive little brightness or color changes. The strength of this phenomenon is dependent on the color

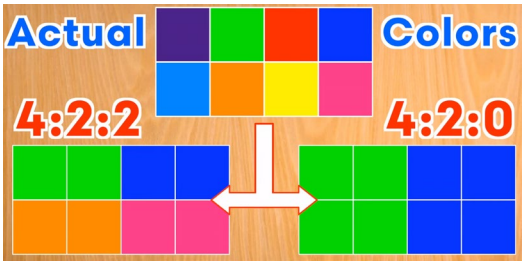
Jpeg

Ist lossy aber soll kaum wahrnehmbare Artefakte liefern. Jpeg unterstützt vier verschiedene compression methods, die wichtigste ist für uns sequential encoding.

Jpeg ist eine compression method ursprünglich.

Es baut darauf auf, dass wir Farben schlechter sehen als Helligkeitsunterschiede. Außerdem dass man Unterschiede in hoher Intensität nicht gut sieht.

1. In YCbCr umwandeln um Farbe und Helligkeit zu trennen
2. Farben downsamplen meist um einen Faktor 2



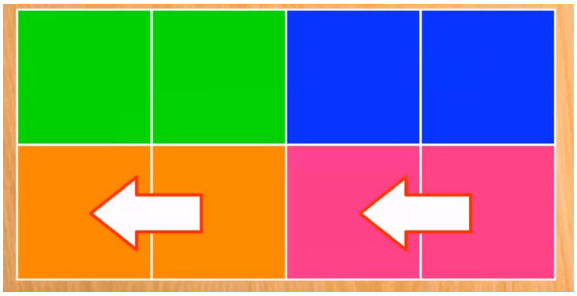
Jeder Pixel hat Luma und Chroma Information. 4:4:4 steht für:

Erster Vierer = Wie viele Pixel pro Reihe haben Luma Information?

Zweiter Vierer = Wie viele Pixel der ersten Reihe haben Chroma Information?

Dritter = Wie viele Pixel der zweiten Reihe haben Chroma Information?

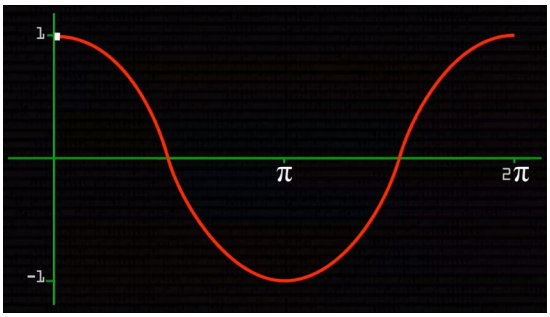
4:2:2 bspw. würde die Farben der jeweils rechten Pixel rüberkopieren.



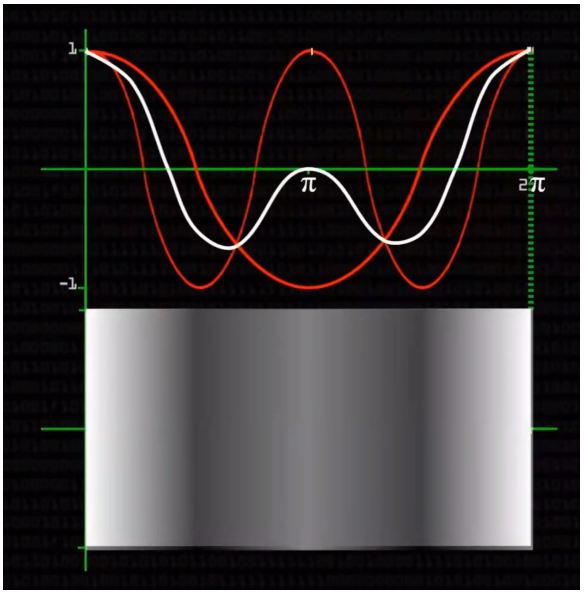
Die Lumavalues machen dann das draus:

3. DCT

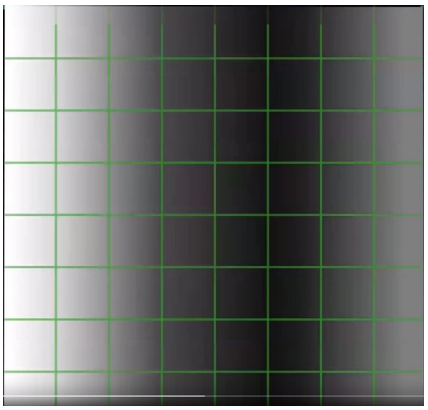
Eine Cosine function schaut so aus:



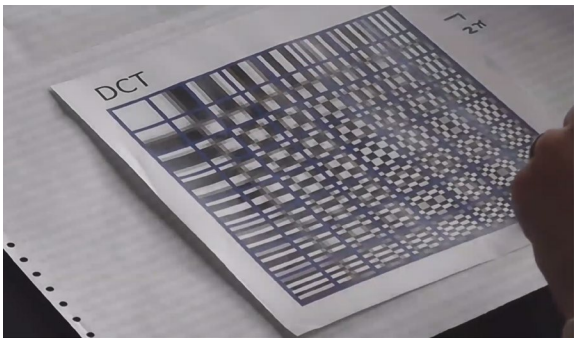
Fügt man eine weitere cosine mit höherer Frequenz bei, bekommt man insgesamt ein neues Mittel:



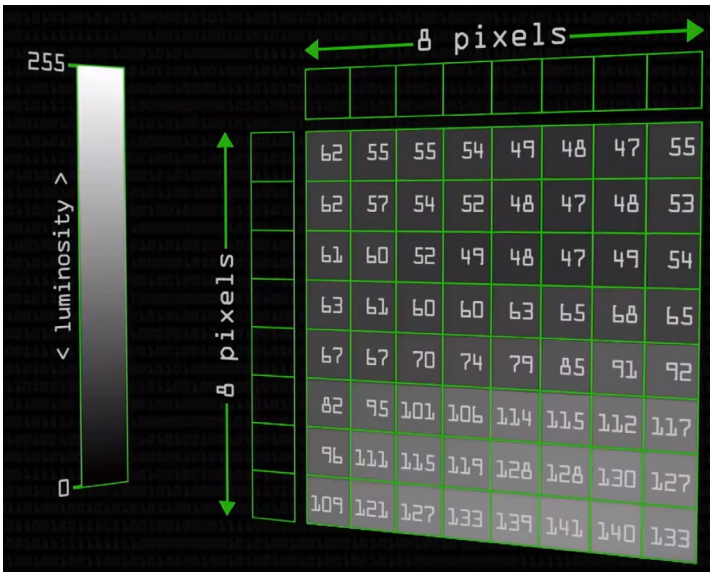
Man kann jedes Bild in 8x8 Pixelgruppen teilen und jede dieser Gruppen durch 8*8 = 64 cosines darstellen.



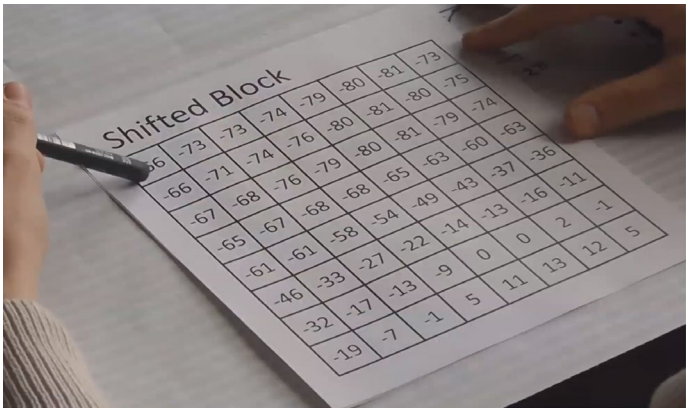
Was man macht, ist, dass man sich alle der 64 cosinewaves anschaut, die contributen könnten und dann gewichtet, welche wie viel beiträgt.

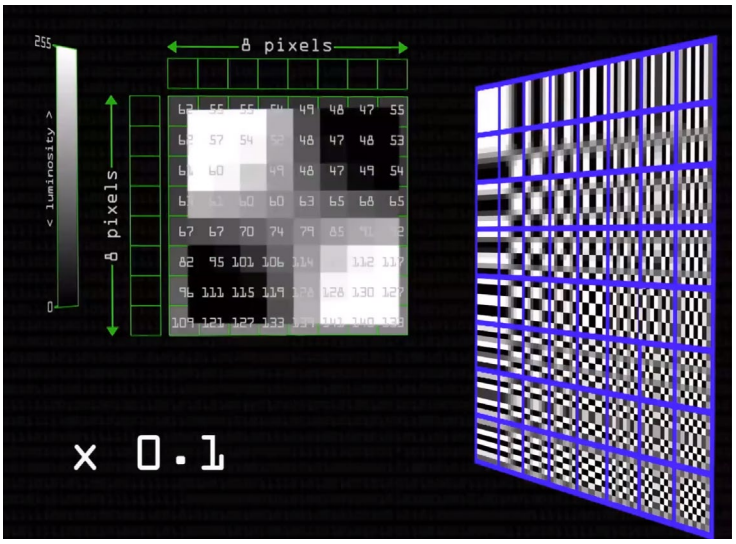


Anhand eines Beispiels: Das ist ein 8x8 Block aus einem Bild

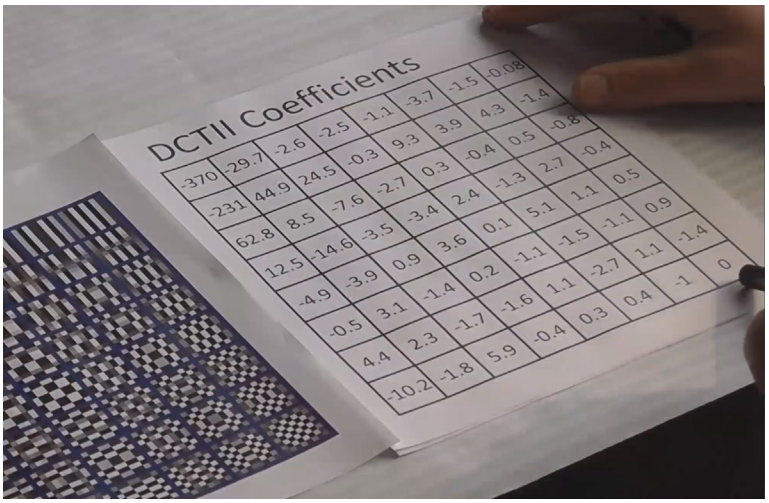


Als ersten Schritt muss man die Werte zwischen -1 und 1 centren. Also 128 abziehen.





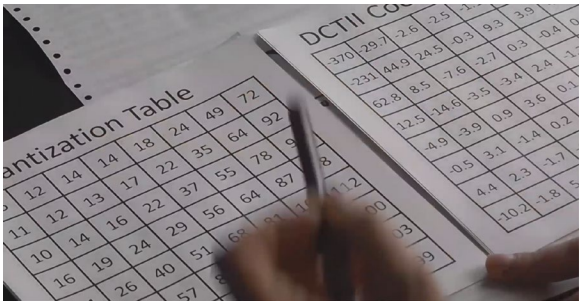
Dann schaut man für alle wie stark man sie gewichten muss, um das zu erzeugen.



Diese Werte können dann bspw. so aussehen. Man sieht, dass der Wert oben links sehr klein ist, das ist der flache Farbwert, der DC coefficient. Alle anderen sind AC coefficient. Auch die Werte rechts unten sind sehr klein, die high frequency cosines tragen nur sehr wenig bei.

4. Quantize

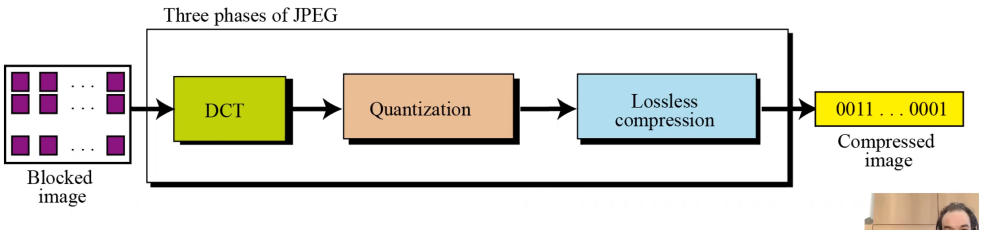
Dann muss man quantisieren. Dazu gibt es einen quantization table. Der hat für jeden coefficient einen Wert durch den man den coefficient dividieren muss.

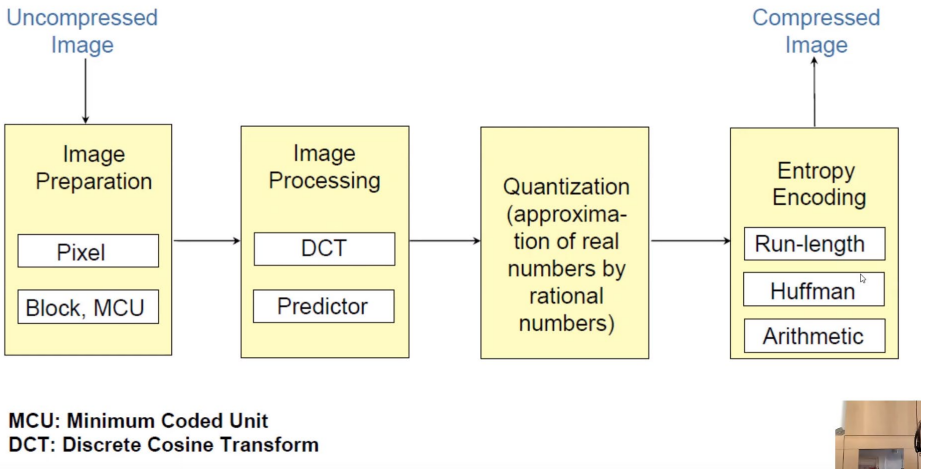


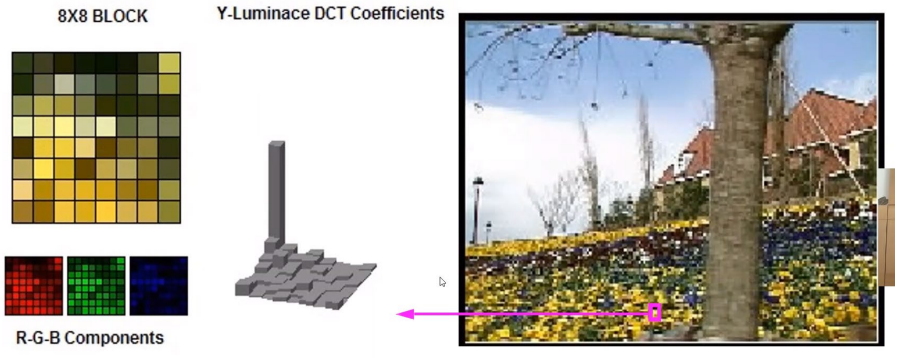
Dann muss man noch runden. (auf den nächsten Integer) und raus kommt das:

Hier sieht man, dass nur neun einen Effekt haben. Anschließend kann man ein Huffman Encoding machen, um das abzuspeichern. (für jedes 8x8 Kastl)

Zusammengefasst

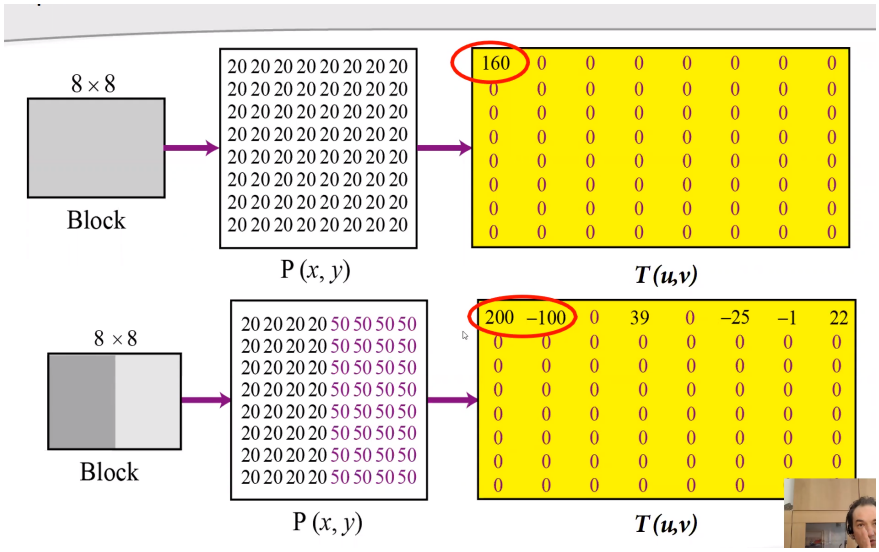




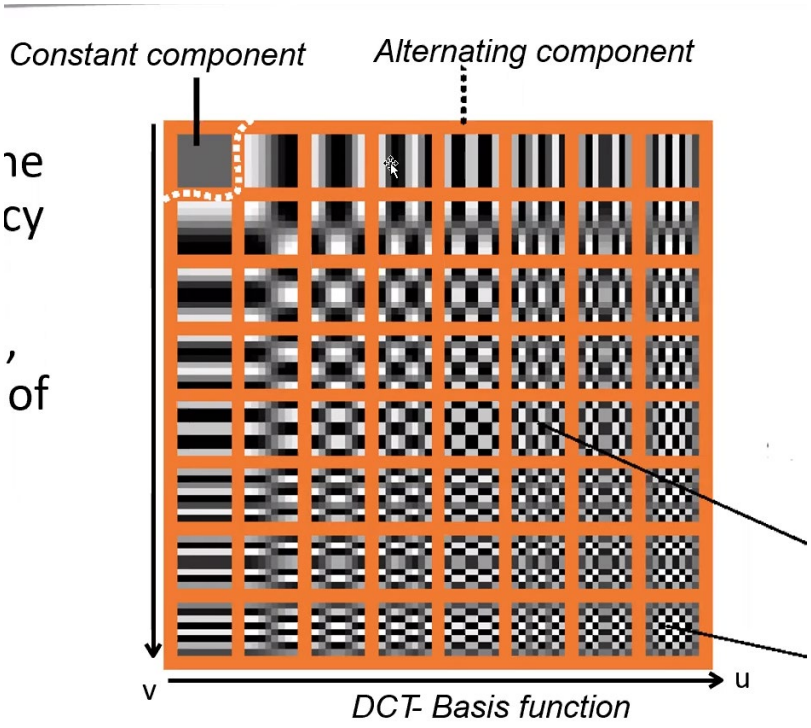


Trivia

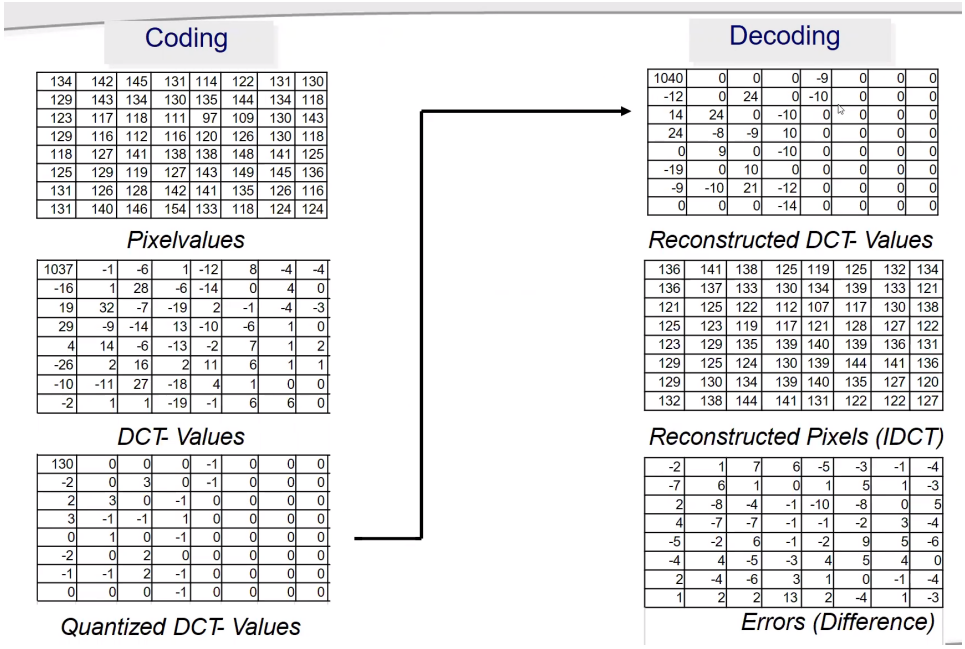
Wenn man einen Block hat wo nur eine Farbe ist, trägt nur der erste Wert bei. Wenn man einen Block hat wo ein Farbverlauf von links nach rechts ist, dann tragen nur die Funktionen der ersten Reihe bei.



Zur Erinnerung



Decoden DCT



134	142	145	131	114	122	131	130
129	143	134	130	135	144	134	118
123	117	118	111	97	109	130	143
129	116	112	116	120	126	130	118
118	127	141	138	138	148	141	125
125	129	119	127	143	149	145	136
131	126	128	142	141	135	126	116
131	140	146	154	133	118	124	124

1040	0	0	0	-9	0	0	0
-12	0	24	0	-10	0	0	0
14	24	0	-10	0	0	0	0
24	-8	-9	10	0	0	0	0
0	9	0	-10	0	0	0	0
-19	0	10	0	0	0	0	0
-9	-10	21	-12	0	0	0	0
0	0	0	-14	0	0	0	0

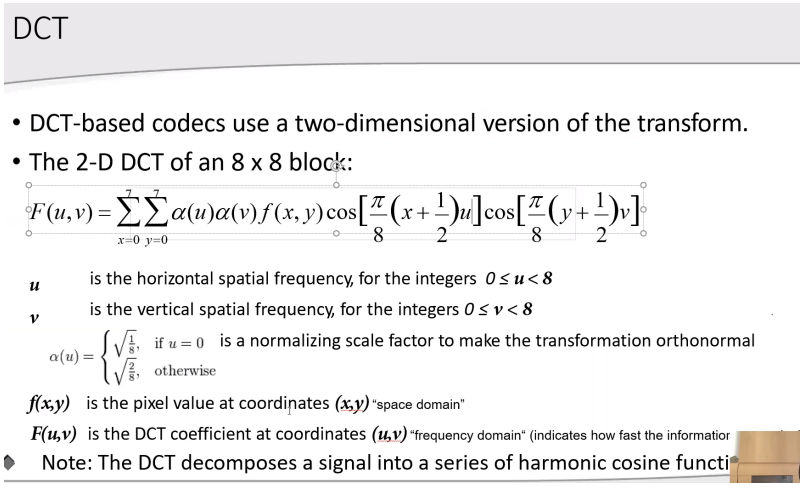
1037	-1	-6	1	-12	8	-4	-4
-16	1	28	-6	-14	0	4	0
19	32	-7	-19	2	-1	-4	-3
29	-9	-14	13	-10	-6	1	0
4	14	-6	-13	-2	7	1	2
-26	2	16	2	11	6	1	1
-10	-11	27	-18	4	1	0	0
-2	1	1	-19	-1	6	6	0

136	141	138	125	119	125	132	134
136	137	133	130	134	139	133	121
121	125	122	112	107	117	130	138
125	123	119	117	121	128	127	122
123	129	135	139	140	139	136	131
129	125	124	130	139	144	141	136
129	130	134	139	140	135	127	120
132	138	144	141	131	122	122	127

130	0	0	0	-1	0	0	0
-2	0	3	0	-1	0	0	0
2	3	0	-1	0	0	0	0
3	-1	-1	1	0	0	0	0
0	1	0	-1	0	0	0	0
-2	0	2	0	0	0	0	0
-1	-1	2	-1	0	0	0	0
0	0	0	-1	0	0	0	0

-2	1	7	6	-5	-3	-1	-4
-7	6	1	0	1	5	1	-3
2	-8	-4	-1	-10	-8	0	5
4	-7	-7	-1	-1	-2	3	-4
-5	-2	6	-1	-2	9	5	-6
-4	4	-5	-3	4	5	4	0
2	-4	-6	3	1	0	-1	-4
1	2	2	13	2	-4	1	-3

DCT Formel



VO9 - Media History

AM and FM radio

AM broadcasts a signal at a constant frequency but adds the actual data as another signal. The first being the carrier wave and the second the data. This results in a modulated signal with the amplitude changes conveying the information. AM stands for Amplitude Modulation. By tuning the radio to the carrier's frequency, the radio's antenna is made to vibrate in that frequency, picking up only signals on that frequency. The antenna creates a current that corresponds to the radio signal, this current is then sent through the radio to the speaker as changes in voltage which, in turn, creates sound.

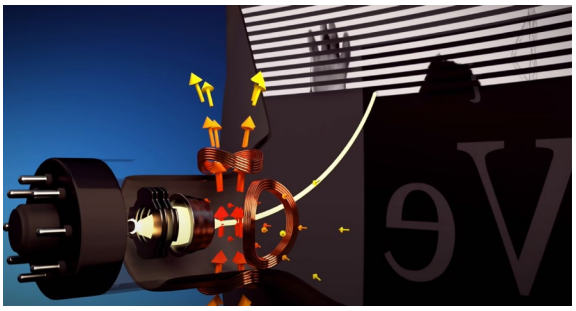
FM changes the frequency of the signal. FM radios have a transformer inside that detects a certain frequency or rather changes in that frequency range. If there is a small change in that frequency that's the signal. This is detected and then sent to the speaker. The advantage here is that you can ignore interferences if they change the amplitude. Additionally, you can use much higher bandwidth due to the mathematical properties of the signal.

AM has more range as AM wavelengths are way longer. High frequency signals are more susceptible to walls and the like. AM waves can reflect from the atmosphere too, making them be amplified along the globe. Also – AM is simple as hell.

Video

Alexander Bane invented scanning with a needle transmitter. Paul Nipkow patented the Nipkow disc. This is basically a disc with holes in it. This light would be picked up by light sensors, creating an electrical signal that could be transmitted. The projector then had to have another disc in front of it that was synchronised with the other disc.

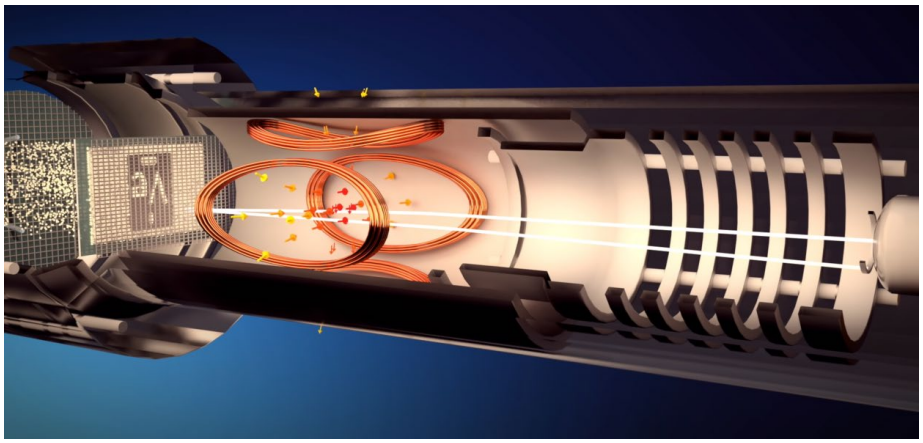
Later, the cathode ray tube came along. It's a glass vacuum tube with an electron gun at the back that shoots electrons to the phosphor on the screen, making the phosphor light up. This electron beam is scanned across the screen, steered by magnets in the gun. The intensity is regulated by changing the voltage of the beam.



Colour was then added with multiple dead ends between this invention and what we ended up with (different pixels). One of them being a colour disc spinning in front of the ray. There was also the triniscope, which was a way of adding three rays with different colours. The final solution was having three colours for every pixel, three colours of phosphor for every small fragment of the screen and three electron guns to determine their relative brightness.

Every other line is scanned each 60^{th} of a second. It takes two scans to scan the whole screen, this is called interlacing. Most of the time the screen is actually empty but this process is so fast that it looks like a moving picture.

There were many vacuum tube designs. One of them would focus an image on the front of the tube, hitting a electron sensitive substance, releasing an electron version of the image in the tube. This would then be sent straight back to a target which is a thin glass plate. From the back of the tube an electron beam would scan the target. The more negative a spot (i.e. the more electrons on the thin glass plate) the more the beam would be reflected.

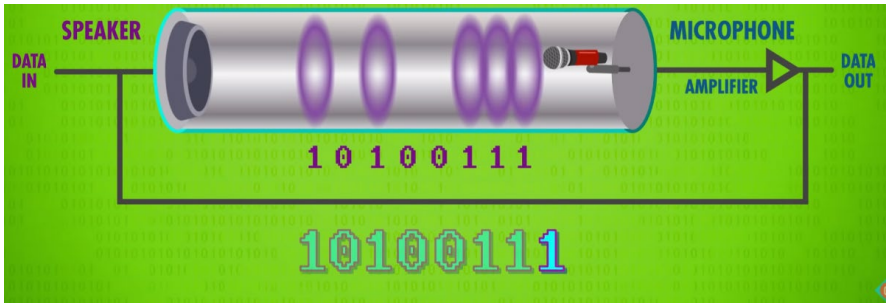


This is how television images were created and displayed but there was no way to record them. This was the era of live broadcasting. What was done was record the transmission and play back the recording at later times. E.g. for live broadcasts that should go live at 8am everywhere in the US. A solution for this came along with the first video tape recorder. This worked with magnetic tape.

Memory and Storage

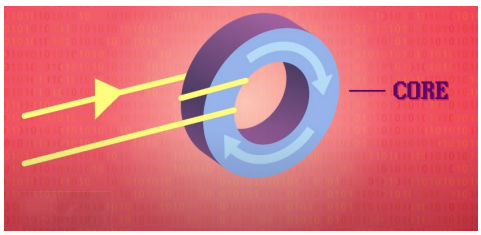
Computer memory in general is non-permanent. This is called volatile storage. The earliest computer storage were punch cards. They were write once though, as you can't unpunch paper.

Then there was delay line memory which is a tube with a speaker and a microphone. If the speaker outputs a pressure wave that's a one and if it doesn't that's a zero. This can be looped as well.

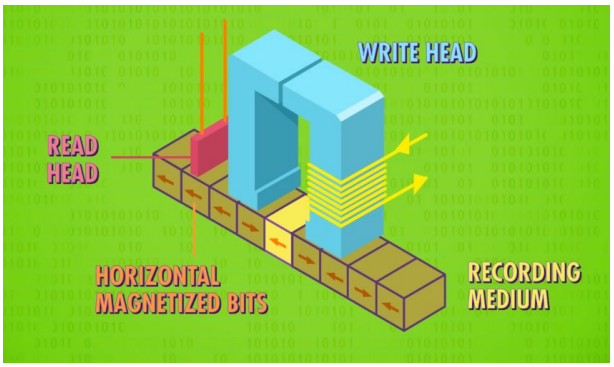


A big drawback with this is that you'd have to wait for bits to come around in the loop. This is called sequential or cyclic access memory but we really want random access memory.

Another newer version was magnetic core memory. This was basically magnetic cores that could be flipped in their polarity to save bits.



Yet another technology was tape drive. Which was a magnetic tape that would be polarised with a read write head.



Tape is sequential, however, as you'd have to unroll it every time you'd want to access certain parts.

The next step after tape were discs. Basically the same idea but with the advantage that memory can be accessed randomly. The seek time – which is the time it takes for the system to access a certain bit of data – was not fast enough for memory but more than fast enough for storage.

CDs and DVDs don't store data magnetically but have little dibs in the discs that reflect light differently.

Today SSDs are taking over.

HDR Standards

HDR is a soup of standards. Generally speaking, HDR is a feature that improves the TV's dynamic range, i.e. the difference between the darkest black and brightest white. HDR is also typically combined with wider colour gamuts.

HDR 10

This is the cheapest / open source HDR standard. It is also the most common format for content. It tells the screen what the lightest and darkest frames in a movie are. This allows then to set the brightness of the whole movie to a certain point.

Dolby Vision

This allows the brightness to be adjusted per scene rather than using one level of brightness for the whole thing.

HDR 10+

Samsung's Dolby Vision rival.

HLG

Another open standard. This can work with non HDR TVs, not giving you a HDR picture then but working nonetheless. It is, however, inferior to both HDR 10 and Dolby Vision.

VO Hannes Kaufmann – Mitschrift Konstantin Lackner 26.06.2021

Formats, Codecs, Containers

A file extension is not a format. The extension is a representation of the container. The container contains the video stream, audio stream and meta data. The meta data contains multiple thing, e.g. the Codec.

There is hundreds of Codecs, here is some important ones:

Video

H.264 / AVC

Most commonly used codec.

H.265 / HEVC

High Efficiency Video Coding – Half the bitrate of H.264 – Very good for 4k streaming but much more complicated to encode.

VP9

Google's royalty free Codec. Good for high resolutions but also hard to encode.

AV1

Currently under development but promising.

Audio

MP3

One of the most famous by MPEG.

AAC

Advanced Audio Codec. Main benefits – widely supported and more efficient than MP3. Does have a limit on audio channels though.

AC-3

Full range of channels, preserves 5.1 surround sound.

File Formats

Standardised rules for storing containers, codecs, meta data and folder structure. There is different formats for online distribution and editing.

MP4

MPEG-4 Part 14. This can mean… MP4 Container, ISO Base Media File(MPEG-4 Part 12) or H.264 Codec(MPEG-4 Part 10)

The MP4 Format is the container.

Adaptive Bitrate Segmentation

Saves the video in small clips in different resolutions/bitrates. These can then be swapped out if the connection is slowing down or speeding up. This prevents interruption of playback.

HLS

HTTP Live Streaming. One of the files in HLS is the .m3u8 playlist file also called manifest.

MPEG-DASH

Open source, also adaptive Bitrate Segmentation.

Best settings for Youtube:

Container: MP4
Video Codec: H.264
Audio Codec: AAC-LC
Channels: Stereo or Stereo 5.1
Sample Rate: 96khz
Aspect Ratio: 16:9
Frame Rate: 60fps

SDR

Type	Video Bitrate, Standard Frame Rate (24, 25, 30)	Video Bitrate, High Frame Rate (48, 50, 60)
2160p (4k)	35-45 Mbps	53-68 Mbps
1440p (2k)	16 Mbps	24 Mbps
1080p	8 Mbps	12 Mbps
720p	5 Mbps	7.5 Mbps
480p	2.5 Mbps	4 Mbps
360p	1 Mbps	1.5 Mbps

HDR

Type	Video Bitrate, Standard Frame Rate (24, 25, 30)	Video Bitrate, High Frame Rate (48, 50, 60)
2160p (4k)	44-56 Mbps	66-85 Mbps
1440p (2k)	20 Mbps	30 Mbps
1080p	10 Mbps	15 Mbps
720p	6.5 Mbps	9.5 Mbps
480p	Not supported	Not supported
360p	Not supported	Not supported

Stereo 5.1: 512 kbps
Stereo: 384 kbps
Mono: 128 kbps

VO Hannes Kaufmann – Mitschrift Konstantin Lackner 26.06.2021

Video Codecs

A codec encodes and decodes. The intent of a codec can be acquisition, delivery or intermediary.

Containers are boxes in which video data, audio data and meta data are saved.

H.264 Codec

This is the MPEG-4 Part 10 thing now, the codec. MOV is one of the most common containers for this. Levels refers to the specs of a profile. This has to do with the decoding speeds and so on.

The profile of a codec are the broader capabilities. Like Chromasampling and the like, bit depth…

The intent of H.264 often is used for all three intents.

HEVC

Double compression compared to H.264.

ALL-I

In H.264 the typical compression uses IPB or LongGOP. Which basically means that only changing portions of images are saved. This makes editing harder though. There is an I frame with the original picture and many predictive frames in between iframes that have only changing portions of the image.

All-I prevents this by only using I frames, not predictive frames.

Intermediary Intent

When working with footage you can make a proxy of the images for editing. This is not the actual image but a lower res image for editing. Edits are later added to the full res image.

Using these for all steps though comes with the disadvantage of them being very data intensive, they need loads of space.

LED Walls

No spilling (greenscreen) but natural light and correct reflections.

Cameras need to be motion tracked as the angle changes with the camera view.

The unreal engine is used to track camera movement.

PAL Standard

Analog Fernsehstandard

Insgesamt 576 lines

25 fps

VO10 – Video

Component video

RGB alle einzeln über ein Kabel übertragen

Composite video

Rot/Weiß/Gelb – Rot Audio, Weiß Audio, Gelb Video

Die Signale werden gemischt in eine einzige carrier wave.

S-video (separated video)

Kompromiss zwischen component und composite analaog video, zwei lines, eine Chroma eine Luma.

Für ein kontinuierliches Bild braucht man mindestens 12-16 frames.

MJEPG

Einfach JPEG auf jedes File des Films.

Ist aber ziemlich groß, weswegen es bessere braucht.

Komprimierung in Videos

Es gibt zwei gute Tricks dafür:

Background often remains the same for a long time

- Newscast, video telephony applications, soap operas, …
- Very small difference between subsequent frames
- Differential encoding of whole images

Objects remain the same but change position

- Compare blocks of N × N pixels in subsequent frames
- Motion compensation

MPEG Family

Es gibt anscheinend einige MPEG Formate…

MPEG-1
– Coding for VCD (Video CD) quality, bit rate of 0.9 - 2 MBit/s
MPEG-2
– Super-set of MPEG-1: bit rates up to 8 MBit/s, coding for DVD and HDTV
MPEG-4
– Improved coding gain to achieve multimedia for the web and mobility
– Additional new coding schema: coding of objects
MPEG-7
– No compression standard
– Allows multimedia content description (ease of searching)
MPEG-21
– Content identification and (rights) management

Spatial compression (intraframe) vs temporal compression (interframe)

Spatial is applied to individual video frames. Chroma subsampling … Applied to videos this doesn't quite do the trick, hence:
Temporal compression – The trick here is to reduce redundancy. We don't need to save the white background in a video x-times if it doesn't change from frame to frame. All we need to save in that case is the information that the background does in fact stay the same. Similarly, we can also save rotations, translations, yadayada… to save space. (feels like I said save a lot)

This works by once again splitting the image into different sections. (Much like in JPEG) After which the video is split into different frames. Some frames only contain of information like the aforementioned "rotate this part / change colour in that part…" – those frames are called P-Frames. They use about half as much data as I-frames which are frames that contain "all" the picture info. (pretty much just JPEGs) Then there is B-Frames which are predictions / interpolations between P and I Frames and also named rather confusingly. (Why would you call them B and not P frames, wth???)

Generally, no matter what compression method, Bitrate determines quality. Bitrate is basically how many bits are used every second. Compression is always a balancing act between file size and image quality.

Image vs Frame

Single images in a video stream

- **Image** is used as a term to refer to one picture which is a part of the video stream. Image denotes an uncompressed picture which **is encoded with MPEG into a frame or comes out after decoding a frame**
- **Frame** is used as a term to denote a compressed picture within the video stream. It thus describes the format in which a video sequence is stored/transmitted, but a frame first needs to be decoded into an image before it can be played.
- Please keep this difference in mind for the principle of the MPEG compression process!

MPEG-1/2 Encoding

Image preparation phase (very similar to JPEG)

- Each image consists of 3 components Y, Cb, Cr
- Luminance has twice as many samples in horizontal and vertical axes as chrominance (4:2:0)
- 8 bit per pixel per component

Encoding basically as in JPEG

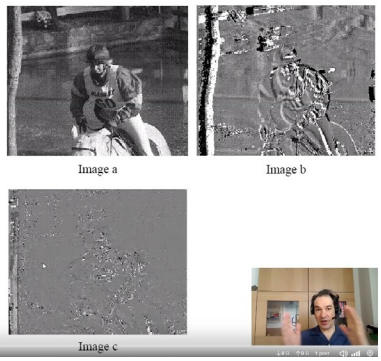
- Separation into blocks – DCT – quantization – entropy encoding
- Enhanced by
  - Differential encoding of similar scenes in subsequent images
  - Considering moving objects

Frames werden als JPEG komprimiert, das sind die I-Frames. Anschließend werden P und B Frames dazu erzeugt.

Die predictions für die nächsten Frames werden mit predictive coding erzeugt. Motion compensation ist der andere Part der vorher beschrieben wurde.

Prediction and motion compensation

- Image a: original image
- Image b: prediction to a previous image
  - Partly homogeneous areas and small values, but not much gain
- Image c: use of motion compensation
  - Very homogeneous values, DCT produces nearly only zero-values!



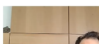
Hier sieht man, dass man, wenn man nur die Änderung anschauen würde, relativ viel neu speichern müsste (von Bild a zu b). Wenn man aber miteinrechnet, dass sich viele Sachen nur bewegen, muss man nur die Bewegung mitspeichern, was natürlich viel weniger Speicheraufwand ist.

Macro Blocks

Bei JPEG war das MCUs (minimum Coded Units), bei MPEG heißt das Macro Block.

Form macro blocks as regions which are more suitable for compression based on motion estimation

- 4 blocks for luminance component (fixed: 4 blocks of 8x8 pixels in MPEG 1&2)
- One block for each chrominance component
- Variable size in MPEG-4 and newer standards
- Similar to MCU concept in JPEG, but different purpose



Predictive Coding and Motion Compensation

Man kodiert ein Bild in seinen Unterschieden zum vorigen Block. Man speichert im Endeffekt den Macro Block zusammen mit einem Motionvektor ab.

Die Suche läuft so ab:

Man schaut zuerst wo der Macro Block zuerst war, dann zieht man einen Bereich darum und schaut wohin er sich bewegt hat.

Example for search algorithm

$$\begin{matrix} b_{x,y} & \cdots & \cdots & b_{x+N-1,y} \\ \cdots & \cdots & & \cdots \\ \cdots & & \cdots & \cdots \\ b_{x,y+N-1} & \cdots & \cdots & b_{x+N-1,y+N-1} \end{matrix}$$

- Search for old position of macro block $B_{x,y}$ =
- Search only within certain window around the current position
  - Speed-up of encoding by limited number of comparisons
  - Window size depends on implementation/configuration
- Consider only the average of all differences, not detailed values:
  - Within the window calculate for each position (u,v)

$$d_{u,v} = \sum_{i=0}^{N-1}\sum_{J=0}^{N-1}\left(b_{x+i,y+j} - p_{u+i,v+j}\right)^2$$

- Set a default threshold: stop searching if a found macro block fi enough“, i.e. $d_{u,v} < d_{threshold}$

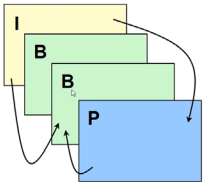
Man schaut also nur wo man ein Pixelmuster findet das sehr ähnlich ist, dann muss man nur noch den Block und Vektor speichern.

Frametypen nochmal

I-frame(Intra-coded frames), P-frame(Predictive-coded frames)[AHA!], B-Frame(Bi-directionally predictive-coded frames)

Efficient coding with fast random access: 3 frame types

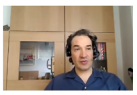
- I-frames: Intra-coded frames
  - Moderate compression but fast random access
- P-frames: Predictive-coded frames
  - With motion compensation, encode current image differentially to the previous image which is encoded as an I- or P-frame
- B-frames: Bi-directionally predictive-coded frames
  - Encode current image differentially to the previous and/or following image which is encoded as and I- or P-frame
  - Can be coded / decoded only after the following I-/P-frame



I-Frames

Self-contained, i.e. represent a full image

- Coded without reference to other images
- Mechanisms like in JPEG
  - DCT on 8 × 8 blocks within macro blocks
- Typical compression gain:
  - I-frames have up to 3 times size compared to P-frames
  - P-frames have 2 – 5 times size compared to B-frames
- Serve as points of random access in MPEG streams
  - E.g.: 3 I-frames per second for reasonably fast random access
  - Or: only one I-frame in several seconds if fast random access is less important than compression gain



P-Frames

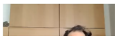
Requires previous I- or P-frame

- Motion estimation for macro blocks
  - Apply DCT, quantization, and entropy encoding to differential macro blocks – very efficient for small differences
  - Add motion vector
- If no prediction possible:
  - Encode macro block as in I-frame

B-Frame

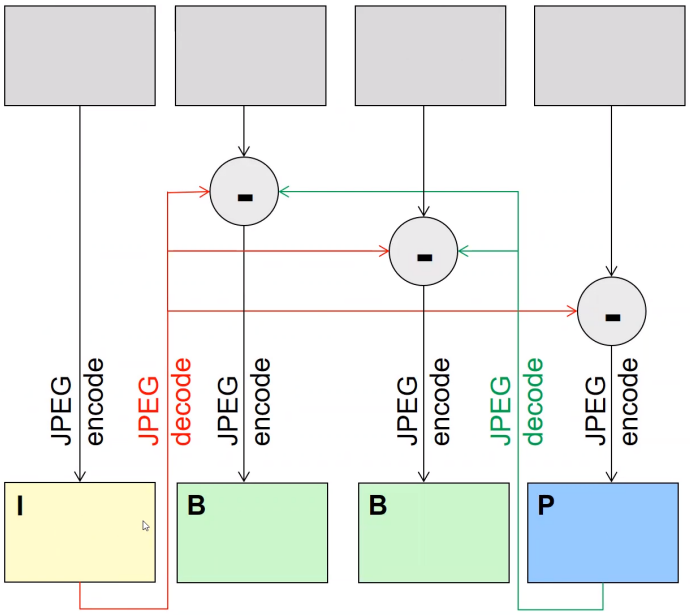
Requires previous and following I- and/or P-frame

- Prediction possible in past and future
  - Some parts of an image will match to a former image, some to a later image, some to an average of both
  - Increased chance to find a match for a macro block
- Quantization and entropy encoding of the macro blocks will be very efficient on such double-predicted images
  - Highest compression ratio will be obtained
- Encoding only is possible after encoding of the following I- resp. P-frame
  - Processing delay, especially for many subsequent B-frames



Der kann vor und zurück schauen, demnach findet er eher einen Block.

Ablauf Kodierung



Die B-Frames brauchen zum Dekodieren sowohl I als auch P.

D-Frame

D-Frame = DC-coded image

- Intra-frame coding, but only of DC coefficients
  - "Very low resolution" image
- Useful for fast forward
- Usually not in use
- Only used for MPEG-1

Speichert nur die DC coefficients. Die ersten coefficients des 8x8 Blocks

Abfolgen

Eine GOP (Group of Pictures, nicht "Grand Old Party") kann jede Anordnung haben. 1 : 2 : 6 war unser Beispiel vorher. Man kann aber wild wechseln wie viele I, P und B man nacheinander haben will.

Für Vorspulen und viel im Video Herumspringen, braucht man I-Frames. (no na ned) Für low bitrate aber eher B-Frames

Bei vielen B- und P-Frames hat man mehr Aufwand beim Dekodieren.

MPEG-2

Hat Profile.

MPEG-2 encoding standard

– Same techniques as in MPEG-1
- Motion compensation, frame types, image encoding (see below)
- But: more applications, other dimensions, other bit rates, …

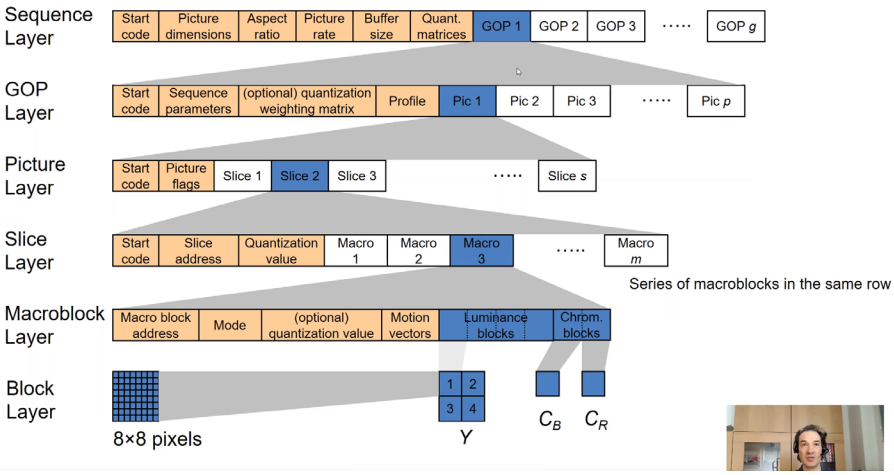
– More a toolkit (profiles) rather than a flat procedure
- Interlaced and non-interlaced frame
- Different color subsampling modes e.g., 4:2:0, 4:2:2 (different resolution of chrominance values in vertical and horizontal direction)
- Flexible quantization schemes – can be changed on picture level
- Scalable bit-streams

Profile:

	Simple Profile No B-frames 4:2:0 Not Scalable	**Main Profile** 4:2:0 Not Scalable	**SNR Scalable Profile** 4:2:0 SNR Scalable	**Spatially Scalable Profile** 4:2:0 SNR or Spatial Scalable	**High Profile** 4:2:0 or 4:2:2 SNR or Spatial Scalable
Low Level 352 pixels/line 288 lines		**≤ 4 Mbits/s**	**≤ 4 Mbits/s**		
Main Level 720 pixels/line 576 lines	**≤ 15 Mbits/s**	**≤ 15 Mbits/s**	**≤ 15 Mbits/s**		**≤ 20 Mbits/s**
High-1440 Level 1440 pixels/line 1152 lines		**≤ 60 Mbits/s**		**≤ 60 Mbits/s**	**≤ 80 Mbits/s**
High Level 1920 pixels/line 1152 lines		**≤ 80 Mbits/s**			**≤ 100 Mbits**

MPEG Data Streams – Layers

MPEG files sind als Layer strukturiert erklärbar.



– Sequence layer: entire video sequence
- Header information like dimensions, aspect ratios, DCT quantization matrices (if not default quantization is used), …
- Max. bitrate, Low-Delay indication, profile indication, …

– Group of Pictures layer: enables random access in a sequence
- One GOP with additional header information like timing parameters for certain recording devices and user data

– Picture layer: primary coding unit
- One frame with header information like frame type (I, B or P), synchronization information, resolution and different picture encoding modes, …

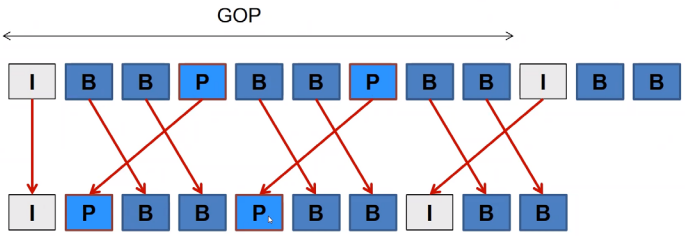
- Slice layer: resynchronization, refresh, error recovery
  - Subdivision of a frame to provide some robustness against data corruption
  - Resynchronization point for playout after data loss
  - Variable number of macro blocks / slice length can be chosen depending on expected bit error rates
- Macro Block layer: basic unit for motion compensation
  - Header information like block position in picture (you can skip macro blocks if they are 0), type (intra or predicted), motion-vector-type, quantizer scale changes, …
  - 4 blocks 8x8 luminance; 2, 4 or 8 blocks 8x8 chrominance
- Block layer: basic coding unit
  - 8×8 pixels: DCT applied at this layer

Frame Decoden

Die Reihenfolge im Stream beim Client, die ankommt, ist eine andere, weil man ja verschiedene Frames für andere zum Decoden braucht.

Different ordering in playout and in streams

- Reordering of frames so that reference frames are available first



Multiplex

Multiplexen heißt, dass im Stream eine GOP ist, dann Audio, dann evtl. Metadaten… dann wieder ein GOP.

MPEG-4

Früher Web und Interaktivitätssachen. Außerdem niedrigere Bitraten.

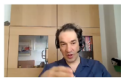
Es gibt wieder Profile.

Baseline Profile
- For applications with limited computing resources
- Videoconferencing and mobile applications
- No B-frames, no CABAC

Main Profile
- Mainstream consumer profile for broadcast and storage applications
- Uses Context-adaptive binary arithmetic coding (**CABAC**) = lossless entropy encoding with high compression rates; decoding requires processing power too

Extended Profile
- Video streaming
- Robustness to data loss and server stream switching

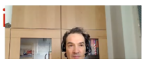
High Profile
- Broadcast and storage applications, particularly for HDTV applications



MPEG-4 vs MPEG-2 (Eher „Was kann MPEG-4 alles“)

- How to achieve an encoding which produces smaller file sizes resp. lower data rates in streaming?
  - Enhanced methods for entropy coding (CABAC)
  - Better prediction accuracy in motion compensation: variable block size
  - Improved search for matches in motion compensation
  - Multi-frame motion compensation (precise segmentation of moving regions)
  - More methods of intra-frame coding
  - New transforms
  - Deblocking filter to smooth edges between blocks
  - Arbitrary shaped frames (video object plane)

→ Significant bit rate saving for equivalent perceptual quali
bit rate or less)

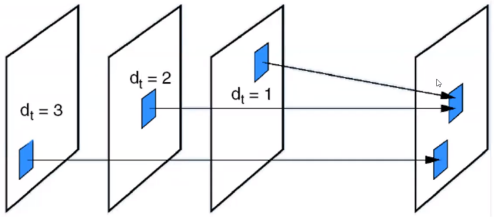


Außerdem kann man auch Macro Blocks mit 16x16 oder 16x8 oder sogar 32x32 machen.

Und es geht motion compensation über mehrere Blocks hinweg.

Prediction from multiple reference frames

- Each block can be coded independent of the other blocks, referencing to any future or past frame
- Also: weighted references to two other blocks (e.g. for scene changes)



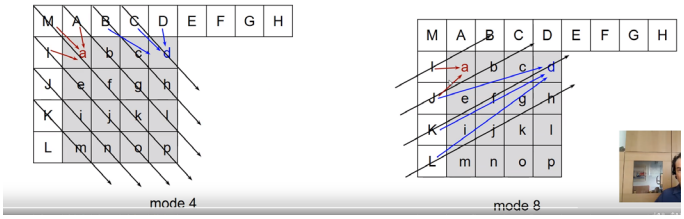
Außerdem Möglichkeit das ganze Bild als Block zu nehmen, das nennt man S-Frame (Sprite Frame) – Ist bspw. bei pans oder zooms praktisch.

Modi für Intra-Prediction

Sind einfach Modi mit denen man Pixelwerte berechnet, Beispiele:

Example for 4x4 luminance block:

- Sample a predicted by round(I/4 + M/2 + A/4) in mode 4
- Sample d: predicted by round(B/4 + C/2 + D/4) in mode 4
- Sample a predicted by round(I/2 + J/2) in mode 8
- Sample d: predicted by round(J/4 + K/2 + L/4) in mode 8



Standards use multiple “Modes”, which are various linear combinations of pixels for prediction of their neighbors w
Macro-Blocks (MBs).

DCT als Integer Matrix und Deblocking Filter

Allow for new transforms

- E.g. 4 x 4 Block Integer DCT Transform: additions and shifts, thus faster processing
- Like a matrix multiplication with $H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix}$
- Multiplication operation combined with quantization

After the transform: use deblocking filter

- Reduce the blocking artifacts in the block boundary and prevent the propagation of accumulated coded noise
- Filtering is applied to horizontal or vertical edges of 4 x 4 blocks block

Redundanz – Redundant slice placement

Um zu verhindern dass Daten verloren gehen, gibt es Möglichkeiten Macro Blocks anzuordnen. Bzw. auch aus anderen slices mit anderen Auflösungen Daten zu holen.

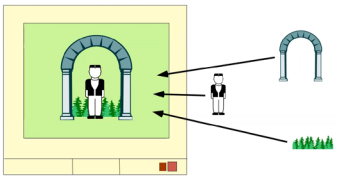
Redundant slice placement in the stream

- Having redundant representations of the same macro blocks
  - E.g. a primary representation can be coded with a low quantization parameter (good quality)
  - A redundant slice can be coded with a high quantization parameter (much coarser quality, but also using fewer bits)

Objektbasierte Videorepräsentation

Different idea (optional mode):

- Representation of a video/audio scene is understood as composition of video objects with respect to their spatial and temporal relationship
- Individual objects in a scene can be coded with different parameters, at different quality levels and with different coding algorithms



Das sollte das Erstellen von Videos vereinfachen. Man könnte damit natürlich einfach verschiedene Presets speichern und neue Szenen zusammenwürfeln.

Video Object Plane

Man konnte auch verschiedene Layers machen mit Planes die sich verschieden zueinander bewegen können.

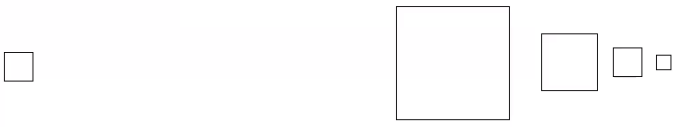
Feature/Standard	MPEG-1	MPEG-2	H.264/MPEG-4 AVC
Macroblock size	16x16	16x16	16x16
Block size	8x8	8x8	16x16, 8x16, 16x8, 8x8, 4x8, 8x4, 4x4
Transform	8x8 DCT	8x8 DCT	4x4, 8x8 Integer DCT
Entropy coding	VLC	VLC	VLC, CAVLC, CABAC
Motion Estimation & Compensation	Yes	Yes	Yes, more flexible: up to 16 motion vectors per macroblock
Reference image	one	one	multiple
Error robustness	Synchronization & concealment	Data partitioning, FEC for important packets	Flexible macroblock ordering, Redundant slice, ..
Bit rate	Up to 1.5 MBit/s	2 – 15 MBit/s	64 kBit/s – 240 MBit

H.265 Verbesserungen

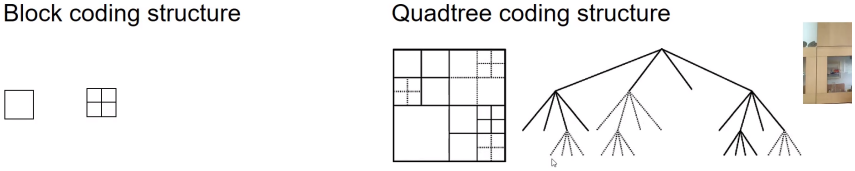
Ungefähr 50% bessere compression als MPEG-4.

Es gibt größere macro blocks, was natürlich gut ist, wenn große Teile gleich sind (Hintergrund), außerdem kann man macro blocks in Baumstrukturen unterteilen.

MPEG-4/AVC	**HEVC**
16x16 macroblock	64x64 to 8x8 coding units



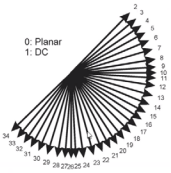
Block coding structure	Quadtree coding structure



HEVC supports partitioning of the macroblock into smaller blocks using a tree structure

Transformationen

MPEG-4/AVC	**HEVC**
Transforms	Transforms
Integer DCT 4x4, 8x8	Integer DCT, Discrete Sine Transform (DST) 64x64 to 4x4
Intra-prediction	Intra-prediction
9 prediction direction modes	35 prediction direction modes



Feature/Standard	**H.264/MPEG-4 AVC**	**H.265/MPEG-H/HEVC**
Partition size	Macroblock16x16	Coding unit 64x64 to 8x8
Partitioning	Down to 4x4	Down to 4x4
Transform	4x4, 8x8 Integer DCT	32x32 to 4x4 Integer DCT 4x4 DST (intra)
Intra prediction	9 predictors	35 predictors
Motion prediction	Vector referencing to several other images	As before, plus Advanced Motion Vector Prediction
Entropy coding	CAVLC, CABAC	CABAC
…	…	…

Deltawert und JND

Es gibt einen Deltawert, der bei verschiedenen Kompressionen zeigt, ob tatsächlich ein merkbarer Unterschied besteht. Ist dieser Deltawert größer 6, ist der Unterschied merkbar.

	H.264		HEVC		
Data Rate	Rez	VMAF	Rez	VMAF	Delta
145	234p	21.50	432p	26.56	5.06
365	270p	52.52	540p	65.12	12.61
730	360p	69.10	720p	78.45	9.34
1100	432p	80.61	720p	87.32	6.72
2000	540p	88.02	1080p	92.94	4.92
3000	720p	92.89	1080p	95.86	2.97
4500	720p	95.06	1080p *	97.53	2.47
6000	1080p	96.99	1080p *	97.53	0.54
7800	1080p	97.71	1080p *	97.53	-0.18

VMAF = quality metric

Delta = difference between VMAF H.264 and VMAF HEVC

Delta >6 = Just noticeable difference for viewers

Vor- und Nachteile von H.265

Great for reach and features
- Clearly best codec for legacy viewers
- Not optimal for HDR

Cost side
- Low quality means:
  - high bandwidth costs
  - Limited access to low-bandwidth markets
- Content royalties an accepted reality

	H.264
Revenue Side	
Reach	
Computers	100%
Mobile with hardware	100%
OTT/Smart TV	100%
Features	
Live	100%
Live transcode	100%
Low latency	100%
HDR	Not optimal (reach of 10-bit AVC unknown)
Cost Side	
Quality	1 - lowest of the bunch
Encoding time	1
Content royalty cost	PPV/Subscripion
FUD Factor	Nokia/Motorola

H.265 vs H.264

Quality
- Assume same quality as H.264 at 60% of the data rate (save 40%)

Encoding time/cost
- Assume 4x H.264, 8x for UHD streams
  - Much less if running your own encoding farm
  - More if you're paying retail by the GB or minute

Storage – assume 60% the cost of H.264 quality

All-Intra

Ein Codec bei dem jedes Bild für sich selbst und wie JPEG bspw. komprimiert wird.

VO11 - Andere Medientypen

Vision	~ 70%
Hearing	~ 20%
Smelling	~ 5%
Tasting	~ 4%
Touch/haptic perception	~ 1%

Geruch

Kann nicht synthetisiert werden. Ein Geruch besteht aus zehntausenden Geruchsmolekülen, meistens spielen nur wenige eine Rolle, aber weil es keinen Standard gibt, ist es sehr schwer zu definieren, was man braucht.

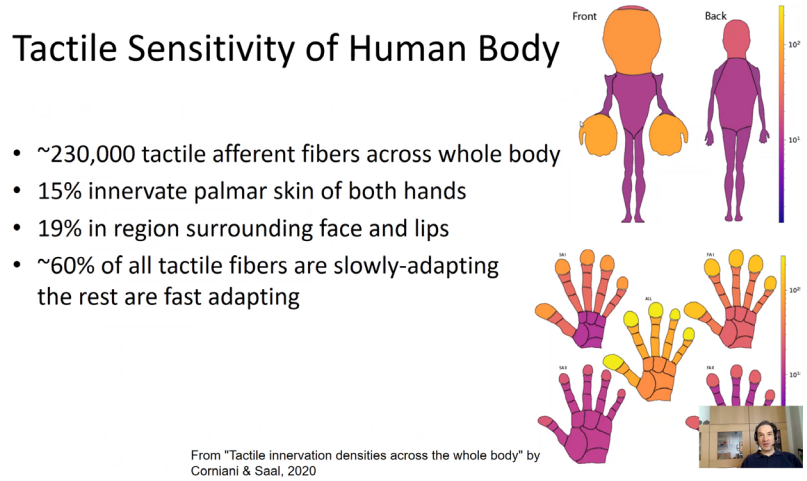
Geschmack

Bessere Situation, weil es mehr Forschung gibt. Man hat 5 Geschmackssensoren auf der Zunge, das hilft natürlich dabei. Mit taste sensor machines kann man virtuell Geschmäcker analysieren und dann auch synthetisieren.

Tastsinn

Kann die Form von Objekten übermitteln, die Oberfläche…

Gewichtung auf den Körper verteilt:



Tactile Brush

Mittels Vibrationen auf der Haut lassen sich Berührungen nachmodellieren.

Je nach Abfolge der Vibration lassen sich damit auch Streichrichtungen simulieren.

Ein-Punkt haptische Geräte

Man hat Kontakt mit einem Punkt, bspw. einem Stift(der Gegendruck gibt), der bewegt sich aber mit der eigenen Hand mit, sodass, wenn diese Bewegung schnell genug ist, der Eindruck einer Oberfläche entsteht. Ab 1000 Hz fühlt sich das ca. kontinuierlich an.