

# Audiokompression am Beispiel AAC

Medientechnologie IL

Andreas Unterweger

Vertiefung Medieninformatik  
Studiengang ITS  
FH Salzburg

Sommersemester 2015

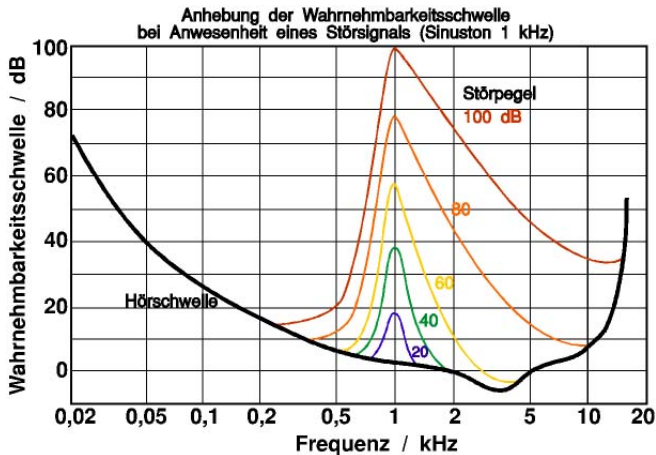
- Schall (englisch *sound*)
  - Zeitliche und örtliche Druckschwankungen
  - Breitet sich als mechanische Welle aus (Trägermedium: zumeist Luft)
  - Schalldruck  $p$ , wobei  $[p] = \frac{N}{m^2} = Pa$  (Pascal)
- Audiosignal
  - Repräsentation einer Schallwelle als Signal
  - Mikrofon: Wandelt Schall in Audiosignale um
  - Lautsprecher: Wandelt Audiosignale in Schall um
- Akustik
  - Beschäftigt sich mit Schall und dessen Ausbreitung
  - Wichtiges Teilgebiet: Psychoakustik (Schallwahrnehmung)

- Physiologie
  - Ohren registrieren Schall, Gehirn verarbeitet Signale (ohne Details)
  - Amplituden- und Phasendifferenzen ermöglichen „Richtungshören“
- Empfindlichkeit
  - Wahrnehmbarkeitsschwelle: ca. 20  $\mu$ Pa
  - Schmerzgrenze: ca. 200 Pa (frequenzabhängig)
  - Hörbares Frequenzspektrum: ca. 20 Hz bis 20 kHz (altersabhängig)
  - Besonders empfindlich bei 1-5 kHz (Stimme: 300 Hz bis 3,4 kHz)
- Lautstärke
  - Subjektiv wahrgenommener Schalldruck
  - Frequenzabhängig (bei gleich bleibendem Schalldruck)

- Weber-Fechnersches Gesetz:  $f' = C_f \cdot \ln\left(\frac{f}{f_0}\right)$ 
    - Wahrgenommene Frequenz  $f'$  ist proportional zum Logarithmus der tatsächlichen Frequenz  $f$  (darum: eine Oktave  $\rightarrow$  doppelte Frequenz)
    - $f_0$ : Wahrnehmungsschwelle (ca. 20 Hz)
    - $C_f$ : Proportionalitätskonstante
    - Auflösungsvermögen ca. 3,6 Hz zwischen 1 und 2 kHz
  - Stevensches Potenzgesetz:  $p' = C_p \cdot (p - p_0)^{n_p}$ 
    - $\log(\text{Lautstärke } p')$  ist proportional zu  $\log(\text{Schalldruck } p)$
    - $p_0$ : Wahrnehmungsschwelle (ca. 20  $\mu\text{Pa}$ )
    - $C_p$ : Proportionalitätskonstante
    - $n_p \approx 0,6$  ( $\sqrt{10}p$  entspricht  $2p'$ )
    - Gilt nur für mittlere bis hohe Lautstärken
- $\rightarrow$  Verwendung logarithmischer Größen (Schalldruck- und Lautstärkepegel)
- $$L_p = 20 \log\left(\frac{p}{p_0}\right), \text{ wobei } [L_p] = \text{dB (engl. } \mathbf{p} \mathbf{r} \mathbf{e} \mathbf{s} \mathbf{s} \mathbf{u} \mathbf{r} \mathbf{e} \mathbf{ } \mathbf{l} \mathbf{e} \mathbf{v} \mathbf{e} \mathbf{l} \mathbf{)}$$

- Frequenzmaskierung (englisch *spectral masking*) – **Demo**
  - Signale mit nahe beieinanderliegenden Frequenzen können schwerer (oder gar nicht) unterschieden werden, wenn ihre Amplituden hinreichend verschieden sind (Beispiel: Flüstern bei Rapkonzert)
  - Zahlreiche Studien durchgeführt → Mathematische Modelle verfügbar
- Zeitmaskierung (englisch *temporal masking*)
  - Signale, die hinreichend kurz vor oder nach einem anderen Signal auftreten, werden nicht wahrgenommen (Beispiel: Aufprall einer Bombe auf dem Boden kurz vor deren Explosion)
  - Abhängig vor allem von Amplituden- und Zeitdifferenz (bis zu 200 ms!)
- Weitere Maskierungseffekte (Auswahl)
  - Signal von einem Ohr maskiert das vom jeweils anderen
  - Auswirkung mehrerer gleichzeitig auftretender Maskierungssignale

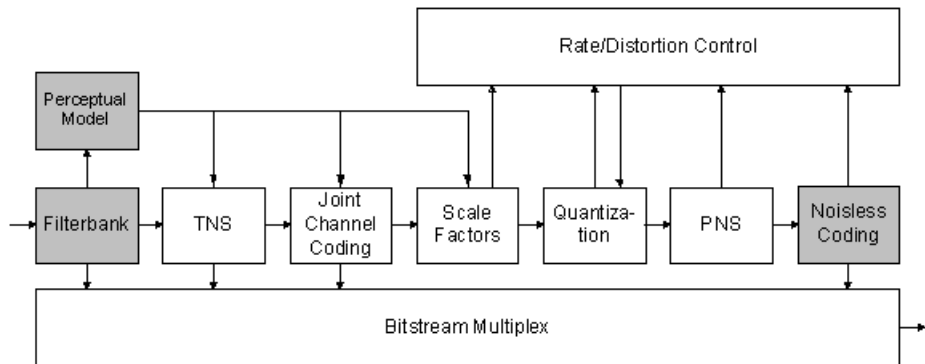
# Menschliche Schallwahrnehmung IV



Quelle: [http://commons.wikimedia.org/wiki/File:Akustik\\_Mithoerschwelle2.JPG](http://commons.wikimedia.org/wiki/File:Akustik_Mithoerschwelle2.JPG)

- Standard für Audio(signal)kompression (Advanced Audio Coding)
- Nutzt Eigenschaften der menschlichen Schallwahrnehmung aus
- Mehrfach standardisiert (verschiedene „Ausbaustufen“)
  - MPEG-2 Part 7 (MPEG-2: Advanced Audio Coding)
  - MPEG-4 Part 3 Subpart 4 (MPEG-4 Audio: General Audio Coding)
- Konfigurierbare Komplexität über Profile
  - AAC definiert verschiedene so genannte Coding Tools
  - Profile legen fest, welche Coding Tools erlaubt sind → Anwendungsfälle
  - Fokus: Low Complexity (LC) Profile (aus MPEG-4)

# Aufbau eines typischen AAC-Encoders I



Quelle: <http://mpeg.chiariglione.org/standards/mpeg-4/audio>

TNS = Temporal Noise Shaping, PNS = Perceptual Noise Substitution



# Aufbau eines typischen AAC-Encoders II

- Wahrnehmung ist frequenzabhängig → **Frequenztransformation**
- TNS: Quantisierungsfehler auf weniger hörbare Frequenzen verteilen
- Ausnutzung von Redundanzen zwischen Kanälen
- **Amplitudenskalierung** durch Maskierungseffekte
- **Quantisierung** (beeinflusst Bitrate und Qualität)
- PNS (optional): Energie von Rauschbändern kodieren, aber ohne Koeffizienten; Decoder ersetzt diese durch pseudozufällige Werte
- Entropiekodierung (Huffman-Kodierung)
- **Datenraten-/Qualitätssteuerung**
- Bitstromgenerierung aus kodierten Daten und Kodierparametern
- Anbringung von Fehlererkennungshilfen (formatabhängig)

- Gesamtes Signal zu groß und uneinheitlich → Stückelung in Blöcke
  - Blockgröße  $N \in \{960, 1024\}$  für lange Blöcke
  - $N \in \{120, 128\}$  für kurze Blöcke (je acht hintereinander)
  - Wechsel zwischen Blockgrößen (mit Übergangsgrößen) möglich
  - Blockgrößenwahl nach Signalcharakteristika
- Frequenztransformation notwendig → DCT?
  - Einzelne Blöcke → Multiplikation mit Rechteckfenster im Zeitbereich
  - Faltung mit sinc-Funktion im Frequenzbereich
  - Zusätzliche (unerwünschte) Frequenzen
  - Andere Fensterfunktion besser (weniger starke Verzerrung)
    - Noch besser: Modifizierte DCT (MDCT) mit teilweiser Überlappung
    - Zusätzlich: Speziell an MDCT angepasste Fensterfunktion(en)

- Ausgangsbasis DCT-IV (andere Symmetrie als „herkömmliche“ DCT):

$$Y_k = \sum_{n=0}^{N-1} X_n \cos \left( \frac{\pi \left(k + \frac{1}{2}\right) \left(n + \frac{1}{2}\right)}{N} \right), k \in \{x \in \mathbb{N} \mid 0 \leq x \leq N-1\}$$

- MDCT nimmt doppelt so viele Eingangswerte  $X_n$  ( $2N$ )
- Anzahl der Ausgangswerte  $Y_k$  bleibt  $N$  (Unterabstastung!)
- MDCT verschiebt Eingangsdaten um  $\frac{N}{2}$  ( $\rightarrow N$  muss gerade sein)

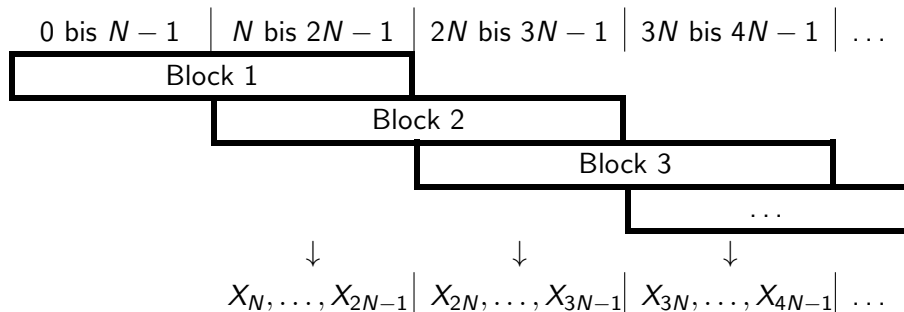
$$Y_k = \sum_{n=0}^{2N-1} X_n \cos \left( \frac{\pi \left(k + \frac{1}{2}\right) \left(n + \frac{N}{2} + \frac{1}{2}\right)}{N} \right), k \in \{x \in \mathbb{N} \mid 0 \leq x \leq N-1\}$$

- MDCT kann auf DCT-IV reduziert werden (vgl. separate Herleitung)
- „Rücktransformation“ (IMDCT):

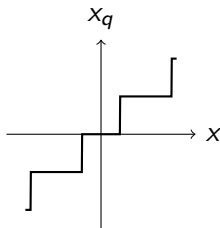
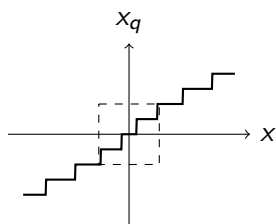
$$X_n^A = \frac{1}{2N} \sum_{k=0}^{N-1} Y_k \cos \left( \frac{\pi \left( k + \frac{1}{2} \right) \left( n + \frac{N}{2} + \frac{1}{2} \right)}{N} \right)$$

$$n \in \{x \in \mathbb{N} \mid 0 \leq x \leq 2N - 1\}$$

- MDCT ist **nicht** invertierbar (Unterabtastung!)
- $IMDCT(MDCT(X_n)) \neq X_n!$
- $IMDCT(MDCT(X_n)) = X_n^a$  (aliased)



- MDCT wird erst durch rücktransformierte Nachbarblöcke umkehrbar
- Ohne Details: Addition der rücktransformierten Signale eliminiert Aliasingfehler im überlappten Bereich (*Time-Domain Aliasing Cancellation* (TDAC)) → Rücktransformation möglich



$$|x_q| = \lfloor |x|^n + M \rfloor$$
$$\text{sgn}(x_q) = \text{sgn}(x)$$
$$n_p = 0,75$$
$$M = 0,4054$$

- Nichtlinear (nach Stevenschem Potenzgesetz, aber mit anderem  $n_p$ )
- „Leise“ Frequenzanteile haben relativ kleine Quantisierungsfehler
- „Laute“ Frequenzanteile haben relativ große Quantisierungsfehler
- Teilweiser Maskierungseffekt für Quantisierungsfehler
- Mit zusätzlichem Offset  $M$  (durch Versuche ermittelt)

- Amplitudenskalierung (vor Quantisierung)
  - Amplituden maskierter Frequenzanteile können verringert werden
    - Amplitudenwertebereich vor Abrundung wird verringert
    - Maskierte Frequenzanteile werden gröber quantisiert
    - Steuerungsmöglichkeit für die Quantisierungsfehlerenergie
- Quantisierung von Koeffizient  $x$  mit Skalierungsfaktor  $s \in \mathbb{N}$ :

$$x_q = \text{sgn}(x) \left\lfloor \left( \frac{|x|}{2^{\frac{s}{4}}} \right)^{\frac{3}{4}} + 0,4054 \right\rfloor$$

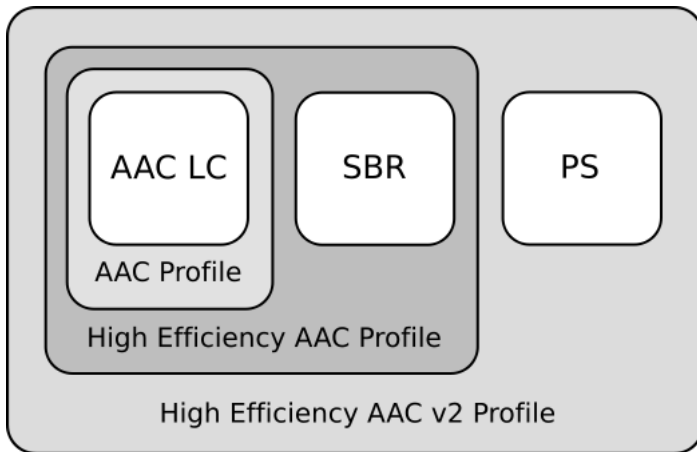
- Koeffizienten werden Skalierungsfaktor „bändern“ zugeordnet
- Nachfolgende Entropiekodierung kodiert Koeffizientengruppen

- Quantisierung verursacht über Skalierungsfaktoren Fehler
  - Kleine Skalierungsfaktoren → Feine Quantisierung
  - Hohe Datenraten und hohe Qualität
  - Große Skalierungsfaktoren → Grobe Quantisierung
  - Niedrige Datenraten und niedrige Qualität
- Skalierungsfaktorsteuerung erlaubt:
  - Datenratensteuerung (Bitrate)
  - Qualitätssteuerung
- Übliche Vorgehensweise im Encoder:
  - Mit „initial guess“ (Skalierungsfaktoren) kodieren
  - Qualität und/oder Datenrate ermitteln (inkl. Entropiekodierung!)
  - Ergebnis mit Benutzervorgaben vergleichen
  - Bei Abweichung Skalierungsfaktoren anpassen und wiederholen



- Übliche Benutzervorgaben für den Encoder
  - Konstante oder beschränkte Datenrate (z.B. bei fixer Kanalbandbreite)
  - Konstante Qualität
- CBR (Constant Bit Rate) Encoding
  - Datenrate vorgegeben und konstant
  - Ziel: Qualität unter Vorgabe maximieren
  - Nachteil: Qualität schwankt über die Zeit hinweg
- VBR (Variable Bit Rate) Encoding
  - Qualität vorgegeben
  - Ziel: Datenrate unter Vorgabe minimieren
  - Nachteil: Datenrate nicht vorhersagbar
- Constrained VBR bzw. ABR (Average Bit Rate) Encoding
  - Min./max. bzw. durchschnittliche Datenrate(n) vorgegeben
  - Ziel: Qualität unter Vorgabe(n) maximieren
  - Nachteil: Qualität durch Einschränkungen nicht optimal

# High Efficiency (HE) AAC I



Quelle: [http://en.wikipedia.org/wiki/File:HE-AAC\\_and\\_HE-AAC\\_v2.svg](http://en.wikipedia.org/wiki/File:HE-AAC_and_HE-AAC_v2.svg)

- SBR (Spectral Band Replication)
  - Verlauf hochfrequenter Amplituden oft ähnlich dem niederfrequenter Amplituden (durch Oberwellen und deren Periodizität)
  - Genaue Unterscheidung hoher Frequenzen schwer/kaum möglich
  - Rauschtoleranz bei hochfrequenten Koeffizienten sehr hoch
  - Hochfrequente Koeffizienten aus niederfrequenten extrapolieren (mit minimaler Zusatzinformation zur „Form“ der hochfrequenten)
  - Nachteil: Extrapolation eventuell falsch → nur für niedrige Datenraten
- PS (Parametric Stereo)
  - Stereosignal (zwei Kanäle) wird als Monosignal (ein Kanal) kodiert
  - Minimale Zusatzinformation zur ursprünglichen Amplitudenverteilung auf die beiden ursprünglichen Kanäle wird mitgeschickt
  - Decoder kann Stereosignal teilweise rekonstruieren
  - Vorteil: Sehr geringer Overhead (wenige kbps)
  - Nachteil: Nur bei sehr niedrigen Datenraten akzeptabel (aber deutlich besser als „echtes“ Stereo mit gleicher Datenrate)

Fragen?