Audio-Kompression

Felix von Leitner

2. Dezember 1999

Die Kompression von Audio-Signalen ist deutlich schwieriger als die von Video-Signalen. Der Bedarf ist riesig und die Unterhaltungs- und Kommunikationsindustrie finanziert seit Jahren ein Forschungsprojekt nach dem anderen auf diesem Gebiet. Trotzdem erreicht man für Broadcast-Qualität bei Audio nur etwa eine Reduktion um den Faktor 10 (bei Video ist es eine Größenordnung mehr).

Einfach gesprochen funktioniert Audio-Kompression so, daß unwichtige Informationen nicht abgespeichert werden. Basierend auf Studien über das menschliche Gehör entscheidet der Encoder, welche Informationen wichtig sind (und welche nicht). Bevor der Mensch etwas hören kann, filtern Ohr und Gehirn unwichtige Daten weg. Audio-Kompression schiebt diesen Schritt lediglich weiter nach vorne.

Das native Sound-Format in digitalem Equipment (wie z.B. Soundkarten) ist PCM (*Pulse Code Modulation*). Hierbei wird n Mal pro Sekunde der Eingangspegel gemessen und in ein Datenwort quantisiert (wobei sich als Wortlängen 16 Bit durchgesetzt haben, in professionellem Equipment auch mehr). Diese Meßwerte nennt man Samples.

Dieses Vorgehen wurde gewählt, weil die Hardware extrem billig zu bauen ist. Sie besteht im Grunde nur aus einem A/D-Wandler (*Analog-Digital-Wandler*), der für ein paar Pfennige zu haben ist. Für qualitativ hochwertige A/D-Wandler kann man auch schon mal 10 Mark ausgeben, aber mehr kosten diese Bausteine nicht. In der Gegenrichtung (d.h. zum Abspielen) braucht man das Gegenstück, nämlich einen D/A-Wandler (liegt in der gleichen Preisklasse).

Die Funktionsweise des Ohres basiert aber nicht auf Samples, sondern auf Frequenzen. Der Gehörgang zerlegt das Eingangssignal biologisch in Frequenz-Bereiche (sogenannte *Frequenz-Bänder* oder *Bänder*) und leitet die Intensität des Eingangssignals auf Nervenbahnen als Erregungen an das Gehirn weiter. Der digitalen Soundverarbeitung kommt entgegen, daß der Mensch Töne nur in einem bestimmten Frequenz-Bereich wahrnimmt. Der Durchschnittsmensch hört keine Höhen über 20 kHz, bei den meisten liegt die Schranke sogar schon bei 16 kHz oder noch tiefer. Mit zunehmendem Alter sinkt diese Schranke weiter.

Grundsätzlich gehen beim Messen natürlich Informationen verloren. Mit 30 Samples pro Sekunde kann man kein 1 kHz Signal abbilden. Ein wichtiges Theorem der Informationstheorie (von Shannon) besagt, daß die Sampling-Rate mindestens doppelt so hoch wie die höchste abzubildende Frequenz sein muß. Die Audio-CD benutzt eine Samplingrate von 44,1 kHz und kann damit theore-

tisch Signale bis 22,05 kHz erfassen. Auf DAT-Bändern wird sogar 48 kHz als Sampling-Rate benutzt. Auf Soundkarten hat sich als obere Schranke 44.1 kHz eingebürgert, wobei viele inzwischen auch 48 kHz unterstützen. Eine wichtige Tatsache bei der Soundverarbeitung ist, daß man manche Töne unbewußt hören kann, die eigentlich außerhalb des hörbaren Spektrums sind. U.a. deshalb benutzen CD und DAT höhere Samplingraten als 32 kHz.

Die normale (verlustfreie) Kompression beseitigt *Redundanzen*, d.h. schon implizit gegebene Informationen.

Bei der Sprachkompression versucht man, die Tonquelle (d.h. die Stimmbänder und den Rachenraum) zu modellieren. Die bisher erfolgreichste Methode dafür ist CELP (siehe Kasten). Für generelle Audiokompression möchte man aber keine Annahmen über die Tonquelle treffen, sondern man setzt am anderen Ende an – dem Gehör.

Der Schlüssel zur verlustbehafteten Audio-Kompression ist, auch *Irrelevanzen* zu beseitigen, d.h. die Teile des Signals, die der menschliche Zuhörer sowieso nicht wahrnehmen würde. Es gibt drei Basis-Techniken zur Audiokompression:

1. **Predictive Coding**.

   Das Wissen über das vergangene Signal wird für eine Vorhersage des nächsten Samples benutzt. Die Kompression besteht darin, daß man nur noch die Differenz zwischen dem Signal und der Vorhersage abspeichern muß, die weniger Bits beansprucht.

2. **Sub-Band Coding**.

   Man teilt das Audio-Spektrum in Frequenzbänder auf. Dann kann man ausnutzen, daß fast alle Bänder viel weniger (oder weniger wichtiges) Signal beinhalten als das lauteste Band. Die Kompression beruht darauf, daß man für die wichtigen Bänder mehr Platz aufwenden kann als für die unwichtigen, in manchen Fällen kann man sich die Übertragung auch ganz sparen. Die Bits werden dynamisch auf die Bänder verteilt, und dem Encoder fällt die aufwendige Selektionsarbeit zu, für die ein sogenanntes psychoakustisches Modell benutzt wird. Der Encoder muß neben den Audiodaten auch Informationen über die Bitverteilung übertragen. Der Decoder muß das Signal nur synthetisieren.

3. **Spectral oder Transform Coding**.

   Man berechnet über die Wellenform periodisch eine Fourier-Transformation. Weil sich die transformierte Darstellung eines Signals nur langsam ändert, muß man sie nur viel seltener übertragen. Transform Encoder benutzen gewöhnlich eine hohe Nummer an Subbändern und betrachten von der Frequenz her benachbarte Samples zusammen.

   Sub-Band Coding wird oft als Spezialfall von Transform Coding verstanden.

Die Audio-Kompressionsverfahren vollziehen zunächst konzeptionell den gleichen Schritt wie das Gehör, indem das Signal nach der Frequenz aufgeschlüsselt wird. In einem Computer ist das eine relativ aufwendige Angelegenheit, bei der man mit komplexen Zahlen hantiert und viel Fließkomma-Arithmetik vollzieht. Erfreulicherweise ist diese Umwandlung für sehr viele Meßverfahren in der Physik wichtig, so daß sich schon Leute Gedanken gemacht haben, wie man das effizient erledigen kann, als von Audio-Kompression noch keine Rede war. Standard-Verfahren sind die *Fourier-* und die *Diskrete Cosinus-Transformation.*

Die Theorie für diese Verfahren ist, daß man ein Signal annähern kann, indem man

verschiedene Wellen überlagert. Welche Funktionen man so verknüpft, ist im Grunde egal (man benutzt gewöhnlich sinusförmige Wellen). Wenn man das aber standardisiert hat, muß man die Funktionen nicht mitspeichern, sondern nur die Koeffizienten.

Frühe Versuche der Audio-Kompression haben einfach die Sampling-Rate herabgesetzt. ISDN benutzt z.B. eine Sampling-Rate von 8 kHz und legt jedes Sample mit einer einfachen Codierung in einem Byte ab. Analog-Telefonie hat eine Qualität, die mit einer Samplingrate von 5 kHz vergleichbar ist. Das funktioniert für Sprache relativ gut, denn auch der Grundton sehr hoher weiblicher Stimmen liegt bei unter einem Kilohertz. Man hört den Unterschied aber trotzdem, und zwar wegen der sogenannten Obertöne. Natürliche Audiosignale haben neben der Hauptfrequenz auch Erregungen bei Vielfachen dieser Frequenz.

Bei der Kompression von Multimedia gibt es aber auch noch andere Ziele als die bloße Datenreduktion. Bei MPEG waren z.B. wichtige Ziele, daß ein Echtzeit-Encoder mit überschaubaren Kosten in Hardware und ein Echtzeit-Decoder auf verbreiteter Software umsetzbar ist. Weitere Forderungen waren, daß man im Strom vor- und zurückspulen kann, und daß man den Strom auch rückwärts abspielen kann, und man möchte resistent gegen Kaskadierung sein, d.h. wenn jemand einen Song überspielt (und dabei decodiert und wieder codiert), sollen keine neuen Artefakte dazukommen. Außerdem war Design-Ziel, daß die Bitrate fest ist. Die Netzwerk-Technologie ist heute soweit, daß man garantierte Bandbreite reservieren kann, d.h. man muß am Anfang wissen, wieviel Platz man maximal verbrauchen wird. Und wenn man soviel Platz reserviert hat, dann möchte man den auch komplett ausnutzen. Auch möchte man gerne den benötigten Platz in Abhängigkeit der aufgenommenen Zeit berechnen können, damit man kundenfreundliche Angaben wie „auf diese CD passen 4 Stunden Musik“ machen kann.

Deshalb ist bei MPEG nicht der Ansatz, ein gegebenes Signal auf möglichst wenig Platz abzulegen, sondern eine gegebene Bandbreite optimal für das Signal auszunutzen. Das Ziel der Hardware-Realisierung sieht man bei MPEG-Audio z.B. daran, daß die Konfiguration der Filterbänke im Standard fest definiert ist (und nicht vom Signal abhängig ist), und daß kein Wissen über das vergangene Signal benutzt wird (das würde im Encoder viel Speicherplatz benötigen). Eine Filterbank ist das Hardware-Äquivalent von einer Fourier- oder Cosinus-Transformation und besteht aus vielen parallelen Bandpaß-Filtern.

Der MPEG-Standard sieht aber auch vor, daß man ein Signal auf möglichst wenig Platz ablegen kann. Das Stichwort ist hier „VBR“ oder „Variable Bit-Rate“. Der Encoder belegt dann für schwache Abschnitte des Signals weniger Bandbreite. Aber auch ohne VBR hat MPEG Layer 3 ein sogenanntes Bit-Reservoir, das der Encoder in schwächeren Perioden ansammelt und dann bei dynamischeren Perioden benutzt. Dieses Reservoir ist aber nicht sonderlich groß, weil das sonst wieder viel Speicher beim Encoder verlangen würde, was Hardware-Implementationen teurer machen würde. Außerdem darf es nur sehr kurze Zeit verwendet werden, damit man beim vorspulen (wobei man das Reservoir überspringen würde) kein Problem hat.

MPEG steht für „Motion Picture Expert Group“ und hat sich ursprünglich mit der Video-Kompression beschäftigt. Die Ergebnisse dieses Konsortiums werden gewöhnlich bei der ISO standardisiert, alle unter dem Namen MPEG. MPEG-Audio ist in Layers definiert. Layer 1 ist die Basis, die alle En- und Decoder leisten müssen. Decoder für Layer 2 können automatisch auch Layer 1 spielen, aber nicht umgekehrt. Die berühmten MP3-Dateien sind in Wirklichkeit nach MPEG1 Layer 3 codiert (MPEG2 Layer 3 für die sehr kleinen Bitra-

ten). MPEG2 unterscheidet sich von MPEG1 vor allem durch zusätzliche definierte Bitraten, Multichannel-Erweiterungen und der alternativen, inkompatiblen Codierung AAC (Advanced Audio Coding).

Bei perzeptionellem Audio-Coding gibt es das Problem, daß man die Eigenschaften des Gehörs noch nicht vollkommen verstanden hat. Erst in letzter Zeit gibt es fundierte Daten über die verschiedenen Maskierungseffekte. Diese Daten können nur über Hörtests gewonnen werden, an denen viele Testhörer beteiligt sein müssen, damit die Ergebnisse auch verallgemeinerbar sind. Solche Tests können sich nur große Firmen und Regierungen leisten. MPEG Layer 3 ist u.a. in Deutschland bei der Fraunhofer-Gesellschaft in Erlangen entwickelt worden, wobei die Hörtests mit Steuergeldern finanziert wurden.

Die Firmen und Institute sind in einer Zwickmühle. Auf der einen Seite lassen sich Forschungsausgaben nur rechtfertigen, wenn das Verfahren Verbreitung findet. Auf der anderen Seite findet es nur Verbreitung, wenn es international standardisiert ist. Trotzdem möchte man an Encodern und am besten auch Decodern verdienen. Die ISO schreibt vor, daß Patenthalter an alle zu den gleichen Konditionen Lizenzen vergeben müssen, d.h. der Mann an der Ecke zahlt genau so viel Lizenzgebühren wie Microsoft.

Bei MPEG hat man sich daher dafür entschieden, nur den Decoder zu standardisieren. So ermöglicht man eine weite Verbreitung des Verfahrens, kann aber bei den Encodern Lizenzgebühren kassieren. Für Implementationen des Standards heißt das aber auch, daß der teuerste (und der wichtige und interessante) Teil, nämlich das psychoakustische Modell, nicht offengelegt wird. Encoder kann also nur bauen, wer selbst neben dem nötigen Know-How auch das Geld für Hörtests hat, oder die Technologie von der Fraunhofer Gesellschaft lizensiert.

Die wichtigste Konsequenz ist, daß man relativ weitreichende Verbesserungen am Encoder durchführen kann, ohne daß der Standard (und damit die Abspielgeräte) geändert werden müssen.

Eine andere wichtige Konsequenz ist, daß Aussagen wie „MPEG ist besser als AC-3“ Unsinn sind, denn die Qualität der Kompression hängt praktisch ausschließlich von dem psychoakustischen Modell ab, das nicht mitspezifiziert ist.

Wenn auch der Encoder nicht spezifiziert ist, so sind doch die von ihm ausgenutzten Eigenschaften des menschlichen Gehörs dokumentiert. Am wichtigsten ist die adaptive Hörschwelle (adaptive hearing threshold):

Die leiseste wahrgenommene Lautstärke hängt nicht nur vom Individuum, sondern auch von der Frequenz ab. Das Gehör der meisten Leute ist zwischen 2 und 5 Kilohertz am empfindlichsten. Diese Minimallautstärke hängt aber auch vom Umgebungsgeräusch ab. Das Ticken eines Weckers ist normalerweise gut hörbar, aber nicht während er klingelt (obwohl er natürlich weitertickt). Wenn man das Klingeln abstellt, sinken diese Schwellen wieder. Man spricht davon, daß Geräusche wie das Ticken in diesem Beispiel *maskiert* werden. Dieser Effekt ist bei Musikaufnahmen besonders wichtig, weil ein Schlagzeug z.B. die Bratsche kurzzeitig maskiert. Das Aufnahmegerät nimmt die Bratsche aber mit auf, nur hört man sie eben nicht.

Dieser Effekt tritt interessanterweise nicht nur bei gleichzeitigen Signalen auf. Das Ohr braucht eine kurze Zeit, nachdem der Schlagzeuger Ruhe gibt, bevor sich die Schwelle wieder gesenkt hat, und es gibt sogar einen kurzen Maskierungseffekt *vor* dem ersten Schlag. Je näher das leisere Signal an der Frequenz des maskierenden Geräusches ist, desto lauter kann es sein und trotzdem maskiert werden. Ein Schlagzeug maskiert sowohl eine Bratsche als auch eine Violine, aber die Bratsche kann

dabei lauter sein.

Der Encoder berechnet über dem Audiosignal ständig die Maskierungskurve, die pro Subband sagt, unter welcher Lautstärke ein Signal dort nicht wahrgenommen würde. Dann berechnet er pro Band die Bits, die benötigt würden, damit die Lautstärke des verursachten Quantisierungsrauschens unter der Maskierungskurve liegt. Wenn dabei noch Bits übrig bleiben, tut der Encoder sie in das Bitreservoir. Wenn der Platz inklusive Bit-Reservoir nicht reicht, muß der Encoder Quantisierungsrauschen über der Maskierungsschwelle einfügen.

Theoretisch könnte ein Encoder Teile des Signals synthetisieren lassen, weil sie z.B. als Obertöne eines anderen Signals erkannt werden. Der Encoder müßte dann nicht die Obertöne mitspeichern, sondern dem Decoder nur sagen, daß sie wieder eingerechnet werden sollen. Das macht aber im Moment kein Encoder.

Bei Mehrkanalsignalen kann man zusätzlich auch Redundanzen zwischen den Kanälen ausnutzen. Normalerweise würde man jeden Kanal einzeln speichern und so den benötigten Platz verdoppeln. Bei MPEG Layer 2 hat man deshalb den „Intensity Stereo“ Modus eingeführt, der für hohe Frequenzen tatsächlich ein Mono-Signal encoded, aber pro Subband die Lautstärken für links und rechts separat dazuspeichern. Das funktioniert deswegen so gut, weil die Signale links und rechts meistens gleich oder sehr ähnlich sind (Sprache und Schlagzeug sind z.B. praktisch immer auf beiden Kanälen gleich) und Stereo-Signale meistens auf einen Frequenzbereich beschränkt sind. Das Ohr nimmt Effekte besonders deutlich als Stereo wahr, wenn auf dem anderen Kanal gar kein Signal oder zumindest keines auf dem gleichen Frequenzbereich ist. Man kann also für das eigentliche Signal doppelt so viel Platz beim Codieren benutzen als wenn man „richtiges“ Stereo benutzen würde.

Geschulte Ohren können Intensity Stereo aber heraushören und so mußte ein besseres Verfahren her. MPEG Layer 3 hat dafür den „Joint Stereo“ Modus eingeführt. Die Idee dabei ist, daß man wie im Radio für manche Frequenzen auf dem linken Kanal die Summe der beiden Eingabekanäle codiert und auf dem rechten Kanal die Differenz. Das hat u.a. wie im Radio den Vorteil, daß ein Mono-Decoder nur den linken Kanal decoden muß und man trotzdem das Signal aus beiden Kanälen hören kann. Für den Encoder hat es den Vorteil, daß er für beide Kanäle nicht gleich viel Platz braucht. Das Differenzsignal beinhaltet ja die Signale nicht mehr, die auf beiden Kanälen sind. Also kann der Encoder für das Summensignal 70% der Bandbreite benutzen und für das Differenzsignal nur 30% (hier benutzt der Encoder spezielle handselektierte Schwellenwerte). Joint Stereo nutzt also die Korrelation und/oder Irrelevanz der Phasendifferenz zwischen den Kanälen aus.

Neu bei MPEG2 sind Mehrkanal-Signale. Zielgruppe sind natürlich Kinofilme und das Heimkino, aber es ist historisch gesehen interessant, daß Stereo nicht prinzipiell für *zwei Kanäle* steht. In den 30er Jahren haben die Bell Labs frühe Experimente mit Stereo-Sound gemacht, allerdings mit drei Kanälen. In den Kinos gab es in den 50er Jahren erstmals Stereo, wobei 4 bis 7 Kanäle gemeint waren. Stereo wurde erst zum Synonym für zwei Kanäle, als es in Form des Plattenspielers die Privathaushalte erreichte. Der Plattenspieler bildete Stereo über die beiden Seiten der Rille ab, und war deshalb hardwaremäßig auf zwei Kanäle limitiert.

Die Dolby Labs haben mit dem Dolby Surround Verfahren Mehrkanal-Ton im Heim etabliert, der die weiteren Kanäle analog im Zweikanal-Signal codiert. Dolby hat sich schon bei ihrer berühmten Rauschunterdrückung mit der Psychoakustik beschäftigt. Anfangs hat man einfach gemessen, welche Teile des Signals von Kassetten gewöhnlich

verrauscht werden und das Signal einfach lauter aufgespielt und beim Abspielen wieder (mit dem Rauschen) leiser geregelt. Später hat Dolby dann auch verschiedene Kompressionsverfahren für Digital-Sound beim Kinofilm entwickelt, nämlich AC-1 bis AC-3 (jetzt „Dolby Digital“). Auch Dolby spezifiziert wie MPEG nicht den Encoder, sondern nur den Decoder, und ist damit genau so „offen“ wie MPEG. Dolby hat allerdings nach eigenen Aussagen doppelt so viele Anwälte wie Ingenieure, was möglicherweise einer der Gründe dafür ist, daß bisher niemand außer Dolby einen AC-3 Encoder gebaut hat. AC-3 ist heute im Kino und auf DVDs neben MPEG2 (allerdings nur Layer 2, nicht AAC) der verbreitetste Standard zur Audiokompression. Für Kinofilme und DVDs haben sich auch „5.1 Kanäle“ durchgesetzt, wobei der zehntel Kanal für Bass-Effekte gemeint ist und sehr wenig Bandbreite hat.

Im Internet liest man derzeit oft die Frage, ob AC-3 oder MPEG besser sei. Es ist hierbei ganz instruktiv, sich mal die Argumente beider Seiten durchzulesen, denn Dolby vergleicht AC-3 mit MPEG Layer 2 und die MPEG-FAQ vergleicht AC-3 mit AAC (welches viel moderner ist, siehe unten). Aber wie bereits oben gesagt ist diese Frage müßig, weil in beiden Fällen das psychoakustische Modell entscheidet. Beide Standards arbeiten nach den selben Prinzipien. Im Übrigen beantwortet sich diese Frage inzwischen auch dadurch, daß Dolby bei MPEG2 AAC mitgearbeitet hat und dafür sogar MPEG-Lizenz-Ansprechpartner ist.

Neben MPEG und AC-3 ist auch das proprietäre MiniDisc-Verfahren ATRAC ziemlich verbreitet. ATRAC komprimiert immerhin auf ein Fünftel, was zwar im heutigen Vergleich nicht sehr beeindruckend ist, aber bei Erscheinen der MiniDisc beeindruckend war. Ohne ATRAC würden nur 15 Minuten auf eine MiniDisc passen.

Der MPEG-Audio Standard kommt mit zwei Beispielen für ein psychoakustisches Modell. Die Modelle leisten folgendes:

1. Herausrechnen der Verzögerung durch die Filterbänke von MPEG, um die Daten im Zentrum des zu codierenden Frames zu betrachten.

2. Eine Fourier-Transformation der einzelnen Abschnitte, weil man für die Maskierungs-Effekte besser auflösen muß als die Filterbank von MPEG.

3. Eine Hann-Gewichtung mildert Kanten-Effekte des Transformationsfensters vor der Fourier-Transformation.

4. Die Werte werden nach den Subbändern des Gehörs gruppiert.

5. Separation der Werte in tonale und nichttonale Komponenten. Töne maskieren anders als Geräusche. Hier gehen beide Modelle verschiedene Wege: Modell 1 identifiziert Töne anhand von lokalen Intensitäts-Spitzen und summiert den Rest in einer einzigen Rausch-Komponente pro Band auf. Modell 2 berechnet einen Tonalitäts-Index als eine Funktion der Frequenz, der besagt, ob eine Komponente eher wie ein Ton oder wie ein Rauschen ist. Dieser Index wird dann zur Interpolation zwischen den Maskierungs-Eigenschaften benutzt. Der Index basiert auf der Vorhersagbarkeit, denn Töne sind vorhersagbarer als Rauschen. Modell 2 benutzt dafür die Ergebnisse der letzten beiden Fenster.

6. Verteilen der Maskierung. Diese Funktion ist empirisch ermittelt und verteilt die Maskierung auf die benachbarten Subbänder.

7. Die unteren Schranken pro Subband werden auf empirisch ermittelte Werte gesetzt.

8. Die Maskierungsschranken pro Subband berechnen. Das Problem ist hierbei, daß das Modell intern mit den Gehör-Subbändern arbeitet und deshalb teilweise mehrere Werte pro MPEG-Subband hat. Die Modelle verwenden hier verschiedene Heuristiken.

9. Berechnen der *Signal-to-Mask Ratios*. Das Verhältnis von Signal-Intensität zur kleinsten Maskierungsschwelle in dem Subband ist das eigentliche Ergebnis des psychoakustischen Modells.

Verfahren	Ziel-Bitrate
MPEG1/2 Layer 1	160+ kbps pro Kanal (DCC: 192 kbps pro Kanal)
MPEG1/2 Layer 2	128 kbps pro Kanal
MPEG1/2 Layer 3	64 kbps pro Kanal
ATRAC	128 kbps pro Kanal
AC-3	64-96 kbps pro Kanal (Kino: 640 kbps total)

Hier ist eine Übersicht der Details der verschiedenen Verfahren:

1. **MPEG Layer 1**.

   Der erste Schritt bei MPEG ist die *Polyphase Filter Bank*, die das Signal in 32 gleich breite Frequenz-Subbänder aufteilt. Die Filter sind verhältnismäßig einfach, haben eine gute Zeitauflösung und die Frequenzauflösung ist auch ganz brauchbar. Bis auf drei Punkte ist das Design ein guter Kompromiß:

   a) Das Gehör teilt das Signal nicht in gleich breite Subbänder auf, sondern die Breite wächst exponentiell zur Frequenz. Bei den Maskierungseffekten kann man die Subbänder des Gehörs beobachten (von 100 Hz für tiefe Frequenzen bis 4 kHz für Höhen), nicht die von MPEG.

   b) Die Filterbank und die Synthese sind verlustbehaftet. Das ist allerdings nicht hörbar.

   c) Aliasing: Benachbarte Bänder haben signifikante Frequenz-Überlappung, weil die Filterbänke nicht scharf abschneiden sondern mit Cosinus approximieren. Ein Ton auf einer Frequenz kann also in zwei Subbändern auftauchen (und so die Kompression verschlechtern).

   Die Filterbank nimmt je 32 Eingabe-Samples und produziert daraus je ein Sample in jedem der 32 Subbänder.

   Ein Layer 1 Frame hat 384 Samples, indem 12 Samples aus jedem der 32 Subbänder gruppiert werden. Der Encoder befragt das psychoakustische Modell und alloziert dann pro Sample-Gruppe Bits. Wenn eine Sample-Gruppe mehr als 0 Bits zugewiesen bekommen hat, wird auch ein 6-Bit-Skalierungsfaktor mitgespeichert (so kann man den Wertebereich des Quantisierers vergrößern).

2. **MPEG Layer 2**

   Layer 2 codiert die Daten in größeren Gruppen und schränkt die Bit-Allokationen in mittleren und hohen Subbändern ein, weil diese für das Gehör nicht so wichtig sind. Die

die Bit-Allokationsdaten, die Skalierungsfaktoren und die quantisierten Samples werden kompakter abgelegt. Durch diese Einsparungen können mehr Bits in die Audio-Daten investiert werden.

Ein Layer 2 Frame hat 1152 Samples pro Kanal, und es werden nicht Gruppen von 12 Samples sondern 3er-Blöcke je 12 Samples pro Subband betrachtet. Die Bits werden pro Dreiergruppe vergeben, aber es gibt bis zu drei Skalierungsfaktoren. Außerdem kann Layer 2 drei quantisierte Werte in einem einzigen, kompakteren Codewort ablegen, wenn er 3, 5 oder 9 Levels für die Subband-Quantisierung vergeben hat.

3. **MPEG Layer 3**

   Layer 3 benutzt eine Modifizierte Diskrete Cosinus-Transformation auf die Ausgabe der Filterbänke, und erhöht damit die Auflösung drastisch. Durch diese Erhöhung kann Layer 3 auch die Aliasing-Effekte rückgängig rechnen (der Decoder muß das wieder draufrechnen). Layer 3 spezifiziert zwei MDCT-Blocklängen: 18 und 6 Samples. Nachfolgende Transformations-Fenster überlappen sich zu 50%, so daß die Fenstergröße 36 und 12 ist. Die lange Blockgröße erlaubt eine bessere Frequenzauflösung für stationäre Audiosignale, während die kurze Blocklänge bessere Zeitauflösung für transiente Signale bietet. Layer 3 schreibt definiert Varianten, in welcher Kombination kurze und lange Blöcke vorkommen können. Der Wechsel zwischen Blocklängen funktioniert nicht unmittelbar, sondern wird durch einen speziellen langen Block eingeleitet.

   Weil die MDCT bessere Frequenzauflösung bietet, ist die Zeitauflösung entsprechend schlechter. Die MDCT operiert auf 12 oder 36 Filterbank-Samples, also ist das effektive Zeitfenster um den Faktor 12 oder 36 Mal größer. Das Quantisierungsrauschen wird also Fehler erzeugen, die über dieses große Zeitfenster verteilt sind, also ist es wahrscheinlicher, daß man sie hört. Bei einem Signal, bei dem sich laute und leise Abschnitte in schneller Folge abwechseln, verteilt sich das Rauschen nicht nur auf die lauten Stellen (wo man es eher weniger stark hört), sondern auch auf die leisen Stellen. Diese Störungen treten gewöhnlich als Pre-Echo auf, weil die temporale Maskierung vor einem Signal schwächer (und kürzer) ist als danach.

   Layer 3 beinhaltet deshalb mehrere Maßnahmen, um Pre-Echo zu reduzieren. Das Psychoakustische Modell versucht, die Voraussetzungen dafür zu erkennen, und Layer 3 kann auch Bits aus dem Bit-Reservoir borgen, um das Quantisierungsrauschen einzuschränken. Schließlich kann der Encoder auch auf eine kleinere MDCT-Blockgröße umschalten, um die effektive Fensterlänge zu kürzen.

   Neben der MDCT hat Layer 3 noch andere Verbesserungen:

   a) Wie erwähnt kann Layer 3 die Aliasing-Effekte der Filterbank wegrechnen.
   b) Der Quantisierer potenziert die Eingabe mit $\frac{3}{4}$ (der Decoder muß das natürlich wieder rückgängig machen), um das Signal-to-Noise Ratio gleichmäßiger auf den Wertebereich der Quantisierungs-Werte zu verteilen.
   c) Die Skalierungsfaktoren werden bei Layer 3 in Bändern zusammengefaßt. Ein Band umfaßt mehrere MDCT-Koeffizienten und hat ungefähr die Breite der Bänder des

Gehörs. So wird das Quantisierungsrauschen ähnlich den Kontouren der Maskierungsschwelle eingefärbt (damit es keine Rausch-Spitzen gibt).

d) Layer 3 benutzt einen statischen Huffman-Code für die quantisierten Samples. Der Encoder sortiert die 576 Koeffizienten (32 Subbänder $\times$ 18 MDCT Koeffizienten) in einer standardisierten, Reihenfolge, die dafür sorgen soll, daß die hohen Werte (von den tiefen Frequenzen) am Anfang stehen und die kleinen am Ende. Dann vergibt er für die hohen Werte am Anfang lange Codewörter und für die kleinen Werte kurze. Der Encoder teilt die Koeffizienten in drei Regionen auf, für die jeweils eine speziell optimierte Huffman-Tabelle benutzt wird. Teilweise werden hier auch mehrere Zahlen auf einen Huffman-Code abgebildet.

e) Das Bit-Reservoir ist auch eine Innovation von Layer 3. Der Encoder darf nur Bits entnehmen, die er vorher reingelegt hat – Kredit wird nicht vergeben.

4. **MPEG 2 AAC**

   AAC ist ein sehr großer Schritt, der den Rahmen dieses Artikels sprengen würde.

   Das Signal wird erst mit einem Polyphase Quadrature Filter (PQF) in vier Subbänder geteilt, für die jeweils ein Verstärkungsfaktor separat übertragen wird. Die vier Subbänder werden jeweils mit einer MDCT der Länge 256 transformiert (oder 32 für sehr dynamische Passagen).

   Die Koeffizienten werden bei AAC aber aus den zwei vorhergehenden Frames vorhergesagt, mit einem speziellen Vorhersage-Mechanismus für jedes Frequenzband. Bei stationären Signalen erhöht das die Effizienz. Die Differenzen werden nicht-uniform quantisiert und mit einem von 12 vordefinierten Huffman-Codes codiert.

   Eine sehr innovative Neuerung ist Temporal Noise Shaping, das auf dynamische Signale zugeschnitten ist. Dabei wird die Beobachtung ausgenutzt, daß ein Ton in der Frequenz-Darstellung wie ein Transient aussieht, und umgekehrt sieht ein Transient in der Frequenz-Darstellung wie ein tonales Signal aus. Tonale Signale kann man mit einem LPC-Ansatz gut vorhersagen (siehe Kasten). Es wird dann ein linearer Vorhersager benutzt, um das nächste spektrale Sample vorherzusagen.

5. **TwinVQ**

   TwinVQ wurde von der NTT (jap. Telekom) entwickelt und hat einen anderen Ansatz als MPEG. Beide Verfahren wandeln das Signal in die Kurzzeit-Frequenz-Darstellung, aber TwinVQ codiert die Samples dann nicht, sondern betrachtet jeweils Muster aus Samples (in Form von Vektoren). Der Encoder hat eine Tabelle von Standard-Mustern, mit denen die Eingabevektoren verglichen werden. Der Index des der Eingabe am nächsten kommenden Vektors wird dann übertragen.

   Die Artefakte von TwinVQ sind anders. Es klingt nicht so, als würde Rauschen dazukommen, sondern als würde Signal verlorengehen, das Signal klingt sozusagen ẁeichgespült:

   Bei Sprachübertragung geht dabei gewöhnlich u.a. der Straßenlärm-Teil des Signals verloren, weshalb TwinVQ für Sprachübertragung und für sehr kleine Bitraten prima ist. TwinVQ ist deshalb Teil von MPEG-4.

6. **ATRAC**

   ATRAC steht für *Adaptive Transform Acoustic Coding* und ist der von Sony entwickelte Codec für die MiniDisc.

   Ein ATRAC-Frame hat 512 Samples, die Eingabe wird erstmal mit einem QMF (Quadrature Mirror Filter) in drei Subbänder unterteilt, die jeweils nochmal mit eine MDCT aufteilt werden: 0-5.5 kHz (128 MDCT), 5.5-11 kHz (128 MDCT) und 11-22 kHz (256 MDCT). Es gibt auch einen langen Block-Modus (11.6 ms) und einen Modus mit drei kurzen Blöcken (2.9 ms + 2.9 ms + 1.45 ms). Die Koeffizienten werden nach Wortlänge quantisiert und mit einem Skalierungsfaktor versehen.

   Bei ATRAC schleichen sich bei einer Kaskadierung Artefakte ein. MPEG hingegen hat sich in Hörtests als weitgehend resistent erwiesen.

7. **Dolby AC-3**

   Ebenfalls ein psychoakustischer Transformations-Codec, der eine Filterbank mit Aliasing-Kürzung verwendet. Überlappende Blöcke aus 512 Eingabe-Samples werden mit einer Transformation mit den Eigenschaften der MDCT (die AC-3 Dokumente nennen es nur TDAC, was Time Domain Aliasing Cancellation heißt und die wichtigste Eigenschaft von MDCT beschreibt) zu 256 spektralen Koeffizienten gewandelt.

   Dolby wollte zu Ihren Verfahren keine genaueren Angaben machen, weil es sich um ein Geschäftsgeheimnis handelt, das man unter NDA lizensieren kann.

   Dolby beschäftigt übrigens laut eigenen Angaben doppelt so viele Anwälte wie Programmierer.

8. **QDesign**

   QDesign ist der primäre QuickTime-Audio-Codec. Die Firma QDesign trat erstmal als MPEG2-Codec-Lieferant auf, und sie sagen von sich, ihr psychoakustisches Modell selber entwickelt zu haben. Leider sind über den Algorithmus keine Details bekannt, und auf ihrer Webseite steht dazu nur die Aussage, daß ihr neuer Codec die Audioquelle parametrisch erfaßt, das Signal in „relevante Komponenten“ zerlegt und das Ergebnis als Koeffizienten ablegt. Von der Differenz wird gespeichert, wofür man Platz hat.

   Unsere Anfragen bei QDesign nach einer (auch oberflächlichen) Beschreibung ihres Encoders wurden leider mit dem Hinweis abgewiesen, daß es sich um ein Geschäftsgeheimnis handeln würde.

1 Glossar

Filterbank: Komponente, die das Audio-Signal in der Darstellung als Samples in Frequenzbänder aufteilt

Quantisierung: Das Verringern der Auflösung eines Signals.

Quantisierungsrauschen: Die kantigen Übergänge im Signal durch die Auflösungsverringerung erzeugen ein Rauschen nahe der Zielauflösung.

Maskierung: Überdecken eines Signals durch ein anderes, z.B. maskiert ein Preßlufthammer ein Handy-Telefonat. Das alte Signal ist noch vorhanden, es wird aber nicht mehr wahrgenommen.

Frame: Die Fourier-Transformation oder MDCT wird auf mehrere Samples zusammen angewendet. Diese Samples werden Frame genannt.

Fenster: Encoder schauen sich gewöhnlich nicht das ganze Signal an, sondern gehen nur kurz zurück. Dieser betrachtete Bereich heißt Fenster.

Signal-to-Noise Ratio: Das Verhältnis des Signals und des Rauschens. In der analogen Übertragung ist dies ein wichtiges Maß für die Qualität, bei perzeptioneller Kompression nimmt man aber gerade eine Verschlechterung dieses Wertes in Kauf, wenn es an der Stelle nicht wahrgenommen wird.

FFT: Die Fast Fourier Transformation ist die Standard-Methode, um ein Signal in die Frequenz-Darstellung zu transformieren. Sie ist so verbreitet, daß Hersteller von Prozessoren hand-optimierte Implementationen mitliefern (besonders bei DSPs ist das üblich).

MDCT: Die Modifizierte Diskrete Cosinus-Transformation wird gerne benutzt, um ein Signal in die Frequenz-Darstellung zu transformieren. Die MDCT hat 50% Überlappung, so daß ein einzelner hin- und zurücktransformierter Block nicht mit dem Original-Block übereinstimmt. Man muß auch den Block davor und danach partiell dazuaddieren. Dann kürzen sich die Fehler der Transformation weg und damit entfällt auch das sonst hörbare Blockgrenzen-Artefakt zwischen transformierten Blöcken. Die MDCT kann außerdem mit nur einer $n/4$ Punkt FFT und ein paar Pre- und Post-Rotationen der Sample-Punkte berechnet werden.

Artefakt: Ein Fehler, den ein codec nach einem Codieren-Decodieren-Durchlauf hinterläßt. Auch bei Bild- und Videokompression nennt man diese Fehler Artefakte.

Aliasing: Der selbe Wert ist an mehreren Stellen abgelegt oder aus mehreren verschiedenen Datensätzen bestimmbar. Auch das Gegenteil heißt Aliasing, d.h. wenn mehrere Samples durch den gleichen Wert dargestellt werden, wie z.B. wenn man ein Signal mit einer Frequenz größer als der halben Samplingrate abzubilden versucht. Auch die so verursachten Störgeräusche nennt man Aliasing-Effekte.

Sprach-Kompression: LPC (Linear Predictive Coding). LPC geht davon aus, daß Sprache ein Summen am Ende einer Röhre ist. Die Glottis (Platz zwischen den Stimmbändern) produziert das Summen, welches durch Intensität (Lautstärke) und Frequenz (Tonhöhe) charakterisiert wird. Rachen und Mund formen die Röhre, die durch ihre Resonanzen charakterisiert wird, die *Formanten* genannt werden. LPC analysiert das Signal, indem die formants geschätzt werden, deren Effekt auf das Summen rausgerechnet wird, und von dem Rest Frequenz und Intensität bestimmt werden. Das Signal kann dann reproduziert werden, wenn man diese beiden Werte und die Koeffizienten der formants gespeichert werden. Weil sich das Signal wandelt, teilt man es in 30-50 Frames pro Sekunde auf, die man einzeln betrachtet. Das Problem bei LPC ist also,

die formants aus dem Sprachsignal zu bestimmen. Die Lösung ist eine Differentialgleichung, die jedes Sample als Linearkombination voriger Samples darstellt. Diese Differentialgleichung heißt „linear predictor“, die Koeffizienten charakterisieren die formants. LPC wählt diese Koeffizienten so, daß die Abweichung des synthetisierten Formats nach einer einfachen statistischen Methode (mean-square) minimiert wird. Leider lassen sich u.a. nasale Laute nicht so abbilden, so daß ein großer Rest übrig bleibt, der extra encoded werden muß. Der erfolgreichste Lösungsansatz ist, typische Reste in einem vordefinierten Codebuch zu haben. Leider bräuchte man pro Stimmhöhe einen neuen Eintrag im Codebuch, so daß das Codebuch unhandlich würde. Bei einem großen Codebuch muß man lange suchen und der übertragene Code wird auch sehr lang. Man behilft sich damit, daß man zwei Codebücher benutzt, wovon nur das erste fest definiert ist, das zweite ist adaptiv. Diese Methode wird **CELP** (Code Excited Linear Prediction) genannt.