Audiokompression (Datenreduktion)

Alexander Leidinger
Alexander@Leidinger.net

18. Februar 2000

Inhaltsverzeichnis

1 Warum Audiokompression?

Audiodaten haben ein grosses Datenaufkommen. Zur Archivierung von Audiodaten in CD-Qualität (44 kHz, 16 Bit, Stereo) benötigt man ca. 10 Megabyte pro Minute. Will man Audiodaten (CD-Qualität) in Echtzeit übertragen, braucht man ca. 2,7 Megabit pro Sekunde. Wenn man statt eines Stereosignales auch noch die Geräuschkulisse eines Kinofilmes übertragen oder speichern will, verdreifachen sich diese Werte sogar.

Anwendungen wie z.B. Video/Audio on Demand sind aufgrund der benötigten Bandbreite und des benötigten Speicherplatzes deshalb so nicht kostengünstig realisierbar. Es muss also ein Weg gefunden werden das Datenvolumen zu reduzieren, ohne dabei die Qualität der Audiodaten hörbar zu beeinflussen. Die momentane Technik bietet dazu verschiedene Arten der Audiokompression an.

2 Arten der Audiokompression

Man kann die Audiokompression in zwei grosse Teilbereiche untergliedern. Zum einen ist dies die verlustlose Audiokompression (Datenkomprimierung), zum anderen die verlustbehaftete Audiokompression (Datenreduktion).

Bei der Datenkomprimierung benutzt man bekannte Verfahren wie z.B. die Huffman-Kodierung, das LZW-Verfahren oder ähnliche. Diese Algorithmen werden auch von normalen Datenkomprimierungsprogrammen wie z.B. *(Win)zip*, *lha*, oder *gzip* verwendet. Da sie nicht speziell für Audiodaten optimiert sind, wird auch keine optimale Kompressionsrate erziehlt. Der Vorteil dieser Kompressionsverfahren ist jedoch, dass die Audiodaten nach einer Dekompression eine 100%ige Kopie der Originaldaten darstellen. Auf die Datenkomprimierung wird im weiteren nicht mehr weiter eingegangen.

Bei der Datenreduktion hingegen wird ein Datenverlusst in Kauf genommen, nach einer Dekompression entsteht also keine originalgetreue Kopie der Ausgangsdaten. Da jedoch die Premisse besteht, dass die Qualität der Audiodaten nicht hörbar beeinflusst werden darf, darf man nicht einfach beliebig Daten weg lassen (z.B. nur das arithmetische Mittel jeweils zweier aufeinanderfolgender Samples abspeichern), sondern muss das psychoakustische Verhalten des menschlichen Gehöres berücksichtigen.

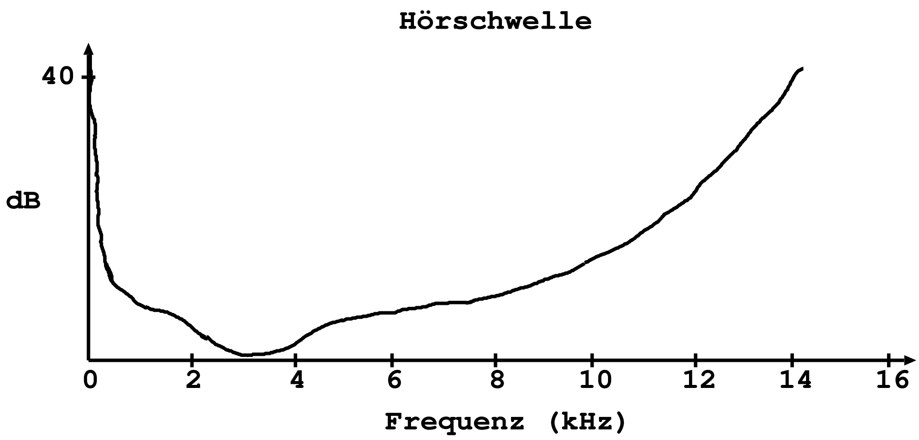


Abbildung 1: Hörschwelle des menschlichen Gehöres

3 Datenreduktionsverfahren

3.1 Vorarbeit

Zunächst wird mittels der Fast Fourier Transformation (FFT) oder der Diskreten Cosinus Transformation (DCT) das Frequenzspektrum der Audiodaten bestimmt. Danach wird das Frequenzspektrum in viele diskrete, nicht notwendigerweise gleich grosse, Intervalle unterteilt. Diese Intervalle sind unter psychoakustischen Gesichtspunkten so ausgewählt, dass für jedes Intervall ein speziell angepasstes psychoakustisches Modell des Ohres verwendet werden kann. Desweiteren wird auch die Zeitachse des Audiosignales in Intervalle unterteilt. Die Intervalle auf der Zeitachse sind jedoch gleich gross und dienen dazu das Audiosignal zu portionieren, also in kleine Pakete zu unterteilen die seperat kodiert und dann abgeschickt oder abgespeichert werden können. Die Grösse der verwendeten Paketlänge hängt vom vorgesehenen Anwendungszweck ab.

3.2 Hörschwellenmaskierung

Zuerst wird in jedem Intervall jedes Paketes eine Hörschwellenmaskierung vorgenommen, d.H., dass die Signale aus dem Frequenzspektrum entfernt werden, die unter der Hörschwelle liegen (Abb. 1) und somit vom Menschen gar nicht wahrgenommen werden können. Wie in Abbildung 1 ersichtlich, würde z.B. ein Ton mit einer Frequenz von 16 kHz und einer Lautstärke von 40 dB entfernt werden, da das menschliche Gehör ihn nicht wahrnimmt.

3.3 Frequenzmaskierung

Nach der Hörschwellenmaskierung folgt eine Frequenzmaskierung. Bei der Frequenzmaskierung werden die Signale entfernt, die von Signalen mit einer anderen Frequenz übertönt werden (Abb. 2). Ein Ton der eine Frequenz von 8 kHz und eine Lautstärke von 60 dB hat, übertönt z.B. einen Ton von 10 kHz/60 dB.

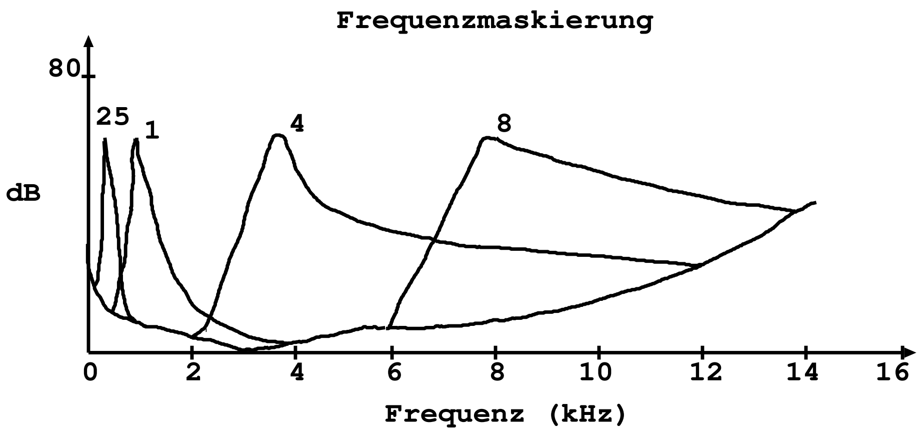


Abbildung 2: Frequenzmaskierung

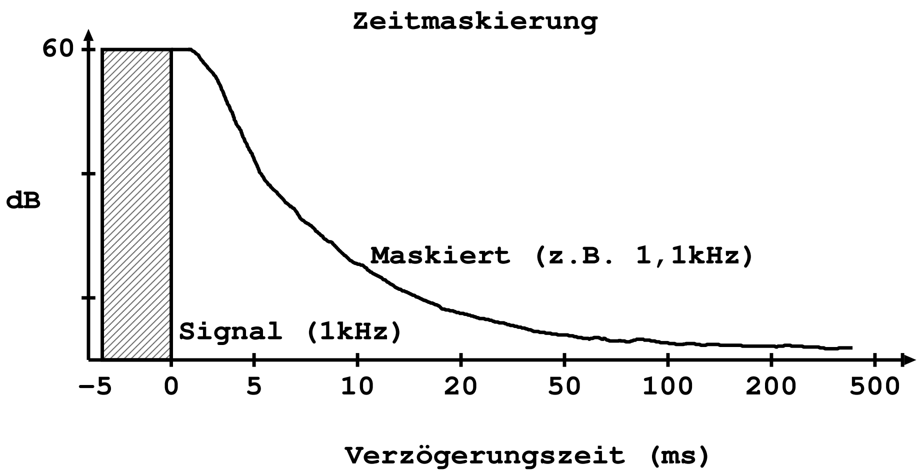


Abbildung 3: Zeitmaskierung

3.4 Zeitmaskierung

Wie bei der Frequenzmaskierung werden bei der Zeitmaskierung ebenfalls Signale entfernt, die von anderen Signalen übertönt werden. Als weitere Komponente wird hierbei jedoch auch die zeitlich Abfolge der Signale beachtet. Wird z.B. ein Ton bei 1 kHz/60 dB abgeschaltet, maskiert dieser noch 5 Millisekunden nach dem Abschalten einen Ton mit 1,1 kHz/40 dB (Abb. 3). Diese Art der Maskierung ist nicht nur nach dem Abschalten eines Signales zu beobachten, sondern auch in einer schwächeren Ausprägung vor dem Anschalten eines Signales.

3.5 Stereoredundanz

Ist in einer Stereoübertragung auf beiden Kanälen das selbe Signal vorhanden, z.B. die Gesangsstimme in einem Muskistück, muss man dieses Signal nicht zweimal übertragen. Da meistens jedoch ausser der Gesangsstimme auch noch Instrumente zu hören sind die nicht auf beiden Kanälen gleich übertragen werden, gibt es verschiedene Verfahren zwischen denen je nach Audiosignal und Verwendungszweck ausgewählt werden kann. Für jedes Intervall in jedem Paket wird deshalb entweder das Intensity-Stereo-, das Mid/Side-Stereo- oder kein Redundanzverfahren benutzt.

Bei Intensity-Stereo (IS-Stereo) wird das Summensignal der Kanäle gebildet und für jeden Kanal ein Skalierungsfaktor bestimmt. Dies führt zwar zu übersprechen zwischen den Kanälen, da man jedoch z.B. bei allen Frequenzen unterhalb von 80 Hz nicht feststellen kann, aus welcher Richtung das Signal kommt, ist dies unter psychoakustischen Gesichtspunkten zu vernachlässigen. Wenn man einen Qualitätsverlusst in Kauf nimmt, kann man auch für alle Intervalle IS-Stereo verwenden um eine grössere Kompressionsrate zu erziehlen. Dies hat jedoch den Nachteil, dass Phaseninformationen verloren gehen. Somit kann keine Dekodierung von Dolby-Surround kodiertem Audiomaterial vorgenommen werden, da zur korrekten Dekodierung der Surroundinformationen diese Phaseninformationen benötigt werden.

Bei Mid/Side-Stereo (MS-Stereo) wird ein Summen- und ein Differenzsignal gebildet, wobei bei der späteren Abspeicherung bzw. Übertragung dem Differenzsignal weniger Speicherplatz eingeräumt wird, als dem Summensignal. Diese Art der Redundanzkodierung benötigt zwar etwas mehr Speicherplatz als die IS-Stereo-Kodierung, hat aber den Vorteil, dass keine Phaseninformationen verloren gehen und dadurch auch die Dekodierung von Dolby-Surround Signalen möglich ist.

3.6 Koeffizientenquantisierung

Nach der Reduzierung der Audiodaten werden die Koeffizienten des Frequenzspektrummes noch quantisiert, wobei sich ein sogenannter Quantisierungsfehler bemerkbar macht (6 dB Rauschen pro weggelassenem Bit). Hierbei wird jedoch darauf geachtet, dass der Quantisierungsfehler unter den oben beschriebenen Maskierungsschwellen bleibt, damit das Gehör diese Fehler nicht bemerkt.

4 Gängige Übertragungsformate und ihre Einsatzorte

Alle folgenden Formate verwenden die oben beschriebenen Reduktionsverfahren, wobei nicht alle Formate alle Reduktionstechniken verwenden. Die Zeitmaskierung wird z.B. nur bei MPEG 2 AAC bzw. bei Sound VQ verwendet. Zu den Formaten Real Audio, MS-Audio (nicht zu verwechseln mit der MS-Stereo-Kodierung), Sound VQ und ATRAC können in dieser Hinsicht leider keine Angaben gemacht werden, da dies proprietäre Formate sind und Informationen darüber deshalb zahlenden Firmen vorbehalten sind. Die Formate MPEG 2 AAC und Dolby AC-3 sehen desweiteren auch mehr als nur zwei Audiokanäle vor.

- Radio-TV / ADR / DAB:
  - MPEG 2 Layer 2
- Film / DVD / DF1:
  - MPEG 2 Layer 2
  - MPEG 2 AAC, bis zu 48 Audiokanäle
  - Dolby AC-3 (Dolby Digital), die Bedeutung der Kanäle findet sich in Abbildung 4
- Internet:
  - MPEG 1 (Layer 1-3)
  - Real Audio
  - MS-Audio (WMA)
  - Sound VQ (VQF)
- Stereoanlage
  - Sony Minidisc (ATRAC)
  - Digital Compact Casette (DCC): MPEG 1 Layer 2 (192kbits/s)

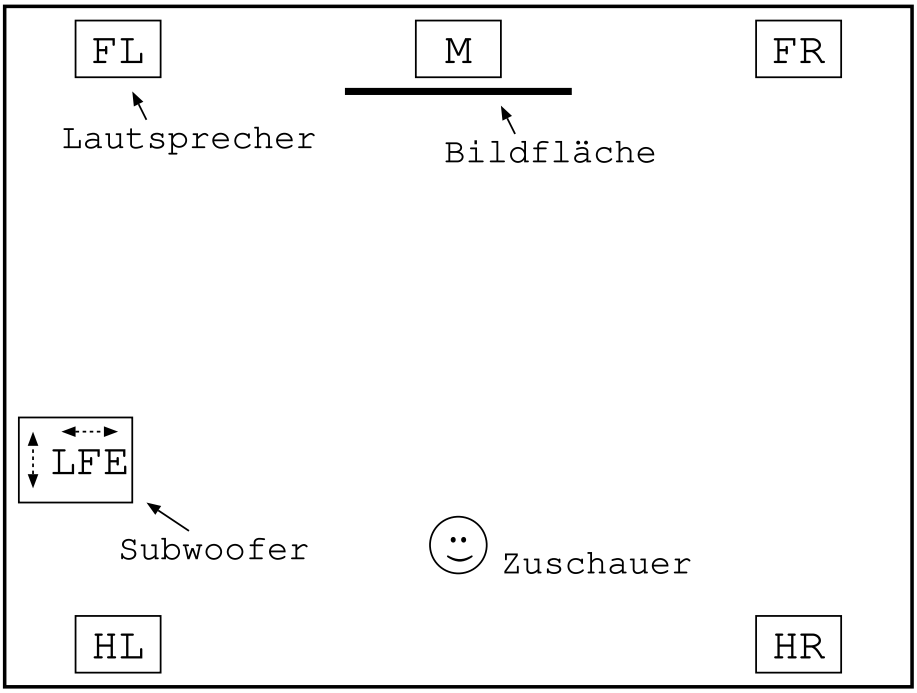


Abbildung 4: Aufteilung der Kanäle bei Dolby Digital (Heimkino)

5 Kompressionsraten

5.1 Ereichbare Werte

Im folgenden sind die ereichbaren Werte für verschiedene MPEG-Formate aufgelistet (MPEG 2 erlaubt Layer 1-3 und AAC als Kompressionsverfahren, MPEG 1 lediglich Layer 1-3). Die Tabellen sind folgendermassen zu lesen:

Bei MPEG 1 sind 32, 44.1 und 48 kHz als Samplingfrequenz erlaubt. Die höchste Kompressionsrate ist bei 32 kbits pro Sekunde erzielbar (niedrigste Audioqualität). Wird Layer 3-Kodierung verwendet, wird die niedrigste Kompressionsrate bei 320 kbits pro Sekunde erreicht (höchste Audioqualität). Das Delay bei der Dekodierung ist hierbei mindestens 59 ms, typisch jedoch ca. 150 ms. Um bei einer Übertragung CD-Qualität mit Layer 3 zu erreichen, muss das Audiosignal mit 128 kbits pro Sekunde enkodet sein. Man erziehlt hierbei keine Kompressionsrate von 1:12.

Die Angaben bezüglich der CD-Qualität sind lediglich Richtwerte, da dies sehr stark vom verwendeten Audiosignal und der Sensitivität des Zuhörers abhängt (näheres dazu in Kapitel 6).

Wird anstatt einer konstanten Bitrate (bestmögliche Qualität bei gegebener Kompressionsrate, zur Übertragung der Audiodaten in Echtzeit gedacht) variable Bitrate (abgekürzt VBR, bestmögliche Kompression bei gegebener Qualität, zur reinen Archivierung von Audiodaten gedacht) verwendet, kann natürlich keine Aussage über die resultierende Dateigrösse gemacht werden, da immer jeweils pro Paket (siehe Kapitel 3.1) versucht wird eine optimale Bitrate zu bestimmen.

MPEG	Frequenz (kHz)	min. kbits/sec
1	32, 44.1, 48	32
2	16, 22.05, 24	8
2 AAC	8-96	

Layer	kbits/sec	delay (ms)	CD-Qualität (kbits/sec)
1	-448	19 (<50)	384 (1:4)
2	-384	35 (100)	192 (1:8)
3	-320	59 (150)	128 (1:12)

AAC:	˜1,4 * Kompressionsrate Layer 3
AC-3:	32-640 kbit/sec

5.2 Konkrete Zahlen bei MP3

In folgenden Tabellen werden anhand von Musikstücken aus verschiedenen Stilrichtungen konkrete Zahlen für Enkodezeit und resultierende Dateigrösse angegeben. Als Format wurde das verbreitete MPEG 1 Layer 3 (MP3) Format gewählt. Ermittelt wurden die Zahlen auf einem Intel Celeron Prozessor mit 400 MHz unter dem Betriebssystem FreeBSD 4.0-current[1]. Als Enkoder kam hierbei der freie LAME[2] in der Version 3.55beta zum Einsatz, da dieser qualitativ sehr nahe an den kommerziellen Enkoder des Fraunhofer Instituts für integrierte Schaltungen[3] heranreicht. Kompiliert wurde er mit dem EGCS in der Version 2.91.66 und den Kompilierparametern "-Os -march=pentiumpro -ffast-math -funroll-loops -fschedule-insns2".

[1] FreeBSD Webseite `http://www.freebsd.org`, ein freies UNIX-artiges Betriebssystem. Wie Linux kann man es kostenlos aus dem Internet beziehen.

[2] LAME Webseite `http://www.sulaco.org/mp3/`

[3] Das Fraunhofer Institut für integrierte Schaltungen war massgeblich an der Entwicklung des MPEG 1 Audioformates beteiligt. Der Enkoder des Fraunhofer Instituts gilt als Qualitätsreferenz.

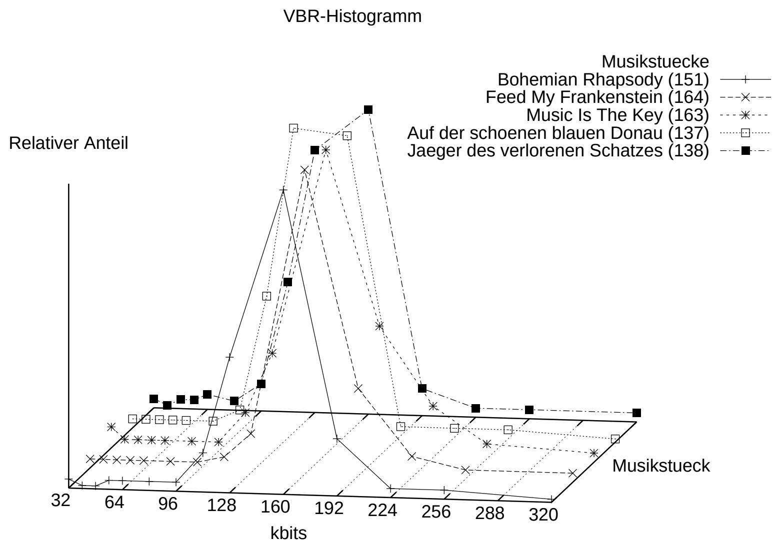
Lied / Artist	bekannt aus/von	Spielzeit	Originalgrösse
Bohemian Rhapsody Queen	Wayne's World	5:48	60,3 MB
Feed My Frankenstein Alice Cooper	Wayne's World	4:46	48,2 MB
Musik Is The Key Dr. Motte and Westbam	Loveparade '99 (Berlin)	3:35	36,3 MB
Auf der schönen blauen Donau Strauss Jr.	2001, Elite	10:06	102,1 MB
Jäger des verlorenen Schatzes	Indiana Jones	5:14	53,0 MB

Die Prozentzahlen in den Klammern der folgenden Tabelle geben an, auf wieviel Prozent der Originalgrösse herunterkomprimiert wurde, die Zeitangaben geben an, wieviel Rechenzeit der Enkoder gebraucht hat. Die Angaben mit dem Prefix "VBR-" sind Angaben bezüglich der variablen Bitraten Kodierung, die mit dem Prefix "128k-" bezüglich der Kodierung mit der festen Bitrate von 128 kbits pro Sekunde.

VBR-Zeit	VBR-Grösse	128k-Zeit	128k-Grösse
6:52	6,5 MB (10,7%)	2:58	5,5 MB (9%)
5:40	5,6 MB (11,6%)	2:39	4,4 MB (")
4:12	4,2 MB (11,5%)	1:57	3,3 MB (")
10:10	9,9 MB (9,6%)	4:50	9,3 MB (")
5:27	5,2 MB (10%)	2:42	4,8 MB (")

Aus der Tabelle ist ersichtlich, dass bereits mit für heutige Verhältnisse relativ moderater Hardwareausstattung die Erzeugung von MP3-Streams mit fester Bitrate in Echtzeit möglich ist.

Das folgende Histogramm zeigt die Verteilung der verwendeten Bitraten innerhalb eines Musikstückes im VBR-Fall. Da die Kurven zur übersichtlicheren Präsentation einzeln skaliert sind, ist ein vergleich der einzelnen Kurven untereinander nicht möglich.



Wie man sieht kommen bei der VBR-Kodierung überwiegend Bitraten über 128 kbit pro Sekunde zum Einsatz. Das bedeutet, dass bei einer festen Bitrate von 128 kbit pro Sekunde ein Qualitätsverlusst in kauf genommen werden muss (siehe dazu auch das folgende Kapitel). Desweiteren hat die VBR den Vorteil, dass bei gegebener Qualität nicht unnötig Speicherplatz verschwendet wird (im Gegensatz z.B. zur Kodierung mit mit einer festen Bitrate von 160 oder 192 kbit pro Sekunde).

6 Qualitätsvergleich

Bei einem Qualitätsvergleich zwischen originalem und datenreduziertem Signal ist darauf zu achten, dass die Qualität vom verwendeten Enkoder, dem Musikstück, der Bitrate und der Testperson die den Qualitätsvergleich macht (jede Person hört anders) abhängig ist. Da z.B. der MPEG Standard nicht vorschreibt wie die Daten enkodet werden sollen, sondern nur wie die Daten nach dem enkoden auszusehen haben damit sie dekoded werden können, kann jeder Hersteller eines Enkoders eigene Verfahren (eigene psychoakustische Modelle) benutzen. Dadurch ist jeder Enkoder einzigartig und produziert eine andere MPEG-Datei, die dadurch auch anders klingen kann (es gibt bis jetzt kein allgemein anerkanntes psychoakustisches Modell). Dies bedeutet auch, dass ein Enkoder z.B. besser für klassische Musikstücke geeignet sein kann wärend ein anderer besser für Popmusik geeignet ist. Je nach Know-How eines Herstellers kann es auch sein, dass ein Enkoder zwar bei hohen Bitraten sehr gut ist, bei niedrigen Bitraten jedoch schlecht abschneidet (und umgekehrt). Dies bedeutet auch, dass der Enkoder eines Herstellers bei 128 kbits pro Sekunde CD-Qualität liefern kann, wärend der Enkoder eines Konkurenten dazu nicht in der Lage ist.

Da ein Computer voller hochfrequenter "Störstrahlung" ist, sollte man auch darauf achten, dass Hörtests nicht über den normalen Audioausgang einer Soundkarte gemacht werden. Ein besserer Ansatz ist die Verwendung eines digitalen Ausgangs und eines externen D/A-Wandlers an der Stereoanlage (welche auch keine Mini/Micro-Kompaktanlage sein sollte, da diese nicht das volle Klangpotential ausschöpfen kann). Falls diese digitale Möglichkeit nicht gegeben ist, kann man auch das enkodete Musikstück wieder dekodieren, es auf eine CD brennen und diese dann für die Hörtests benutzen.

Um zu objektiven Ergebnissen bei einem Qualitätsvergleich zu gelangen, wurde der "triple stimulus, hidden reference", oder auch ABC-Blindtest entwickelt. Hierbei hört die Testperson als Signal A immer das Original und als Signal B und C im zufälligen Wechsel das originale und das datenreduzierte Signal. Die Signale B und C sind von der Testperson dann auf einer Skala von 1 bis 5 zu bewerten, die 5 ist dabei zu verwenden wenn kein Unterschied zum Original festzustellen ist, die 1 wenn der Unterschied besonders deutlich ist.

7 Fazit

Der momentane Stand der Technik erlaubt schon eine ausgiebige Datenreduzierung, ist jedoch noch nicht soweit, dass z.B. Audio on Demand flächendeckend in CD-Qualität angeboten werden kann. Desweiteren sind auch noch weitere Qualitätsverbesserungen an den Enkodern möglich. Da es jedoch kein Gerät gibt welches zur Qualitätsmessung benutzt werden kann und deshalb ausgiebige Hörtest mit Testpersonen gemacht werden müssen, sind diese Qualitätsverbesserungen jedoch sehr zeitaufwendig und kostspielig.

Literatur

[1] Matthias Carstens; Report Audiokompression; c't 1998, Heft 21, Seite 242ff

[2] Webseite des LAME Enkoders; `http://www.sulaco.org/mp3/`

[3] Davis Pan; A Tutorial on MPEG/Audio Compression; IEEE Multimedia Journal, Summer 1995 issue

[4] Webseite des Fraunhofer Instituts für integrierte Schaltungen; `http://www.iis.fhg.de/`

[5] MPEG Audio Webseite; `http://www.tnt.uni-hannover.de/project/mpeg/audio/`