MATHEMATISCHE PRINZIPIEN BEI DER HERSTELLUNG VON COMPACT-DISCS

H.K.Kaiser (TU Wien)

1.Einleitung

Unser Leben wird in vielen Bereichen von der Mathematik beeinflußt. Allerdings ist uns dies oft nicht bewußt. Neben den klassischen Anwendungen der Differential-und Integralrechnung, der Differentialgleichungen u.a.m. in Naturwissenschaften und Technik treten in jüngster Zeit verstärkt Methoden der Diskreten Mathematik in den Vordergrund. Ziel der folgenden Ausführung ist es,die vornehmlich algebraischen Methoden der Fehlererkennung und Fehlerkorrektur von digitalisierten Audiosignalen in ihren Grundzügen zu beschreiben. Diese Methoden sind hauptverantwortlich dafür, daß beim Abspielen von Compact-Discs ein authentisches störungsfreies Musikhören möglich ist. Wesentlich bei dieser Anwendung ist, daß die entsprechenden Rechenalgorithmen rasch und einfach durchgeführt werden. Es wird gezeigt , wie dies durch eine Anreicherung der verwendeten algebraischen Struktur immer besser erreicht werden kann.

2.Compact-Discs

Die Geschichte der Schallplatte beginnt 1877 mit Edisons Aufnahme von "Mary had a little lamb" auf einem Wachszylinder. Im Zeitalter der Wachsplatte und der Schellackplatte war das Hauptproblem die kurze Spieldauer. Also wurde das System der Langspielplatte in Hi-fi-Qualität entwickelt, das sich weltweit durchgesetzt hat und in den Jahren nach dem zweiten Weltkrieg bis an die "natürlichen" technischen Grenzen perfektioniert worden ist. Diese natürlichen Grenzen des authentischen Musikhörens im System der konventionellen Schallplatte liegen beispielsweise in Verunreinigungen (Staub, Fingerabdrücke, Kratz-

spuren), geringfügigen Produktionsfehlern, oder in Verschleißerscheinungen durch mechanische Abtastung.
Die Einführung der Digitaltechnik brachte eine "Revolution" auf dem Sektor der Musikwiedergabe mit sich. Diese neue Ära ist jene der Compact-Discs. Die Produktion dieser Compact-Discs wurde möglich, als sich 1979 SONY und PHILIPS auf ein gemeinsames Standardsystem für die Speicherung und Reproduktion von Audiosignalen einigten.
Compact-Discs haben nur einen Durchmesser von 12 cm. Dadurch ist die Möglichkeit von kleinen Abspielgeräten gegeben.
Bei der Reproduktion des Audiosignals kommt es zu keinem mechanischen Kontakt zwischen Disc und Informationsabnehmer. Daher kann es beim Abspielen keine Abnützung geben (ein Laserstrahl liest die Information). Die digitale Information wird auf einer spiralförmigen Spur gespeichert, und zwar als Folge von sogenannten "pits" and "lands" (siehe Abb.1, Teil der Oberfläche einer Compact-Disc, Informationsschicht,Vergrößerung).

Der Abstand der Spuren voneinander beträgt 1,6µm. Der Laserstrahl tastet die Scheibe von unten ab. Der Durchmesser des Lichtflecks beträgt 1µm. An den "lands" wird das Licht fast zur Gänze reflektiert, die "pits" bewirken eine geringe Reflektion, da ihre Tiefe etwa die Hälfte der Wellenlänge des auftreffenden Lichts beträgt (Interferenz). Der Abtaststrahl wird über ein Servosystem gesteuert. Abb.1

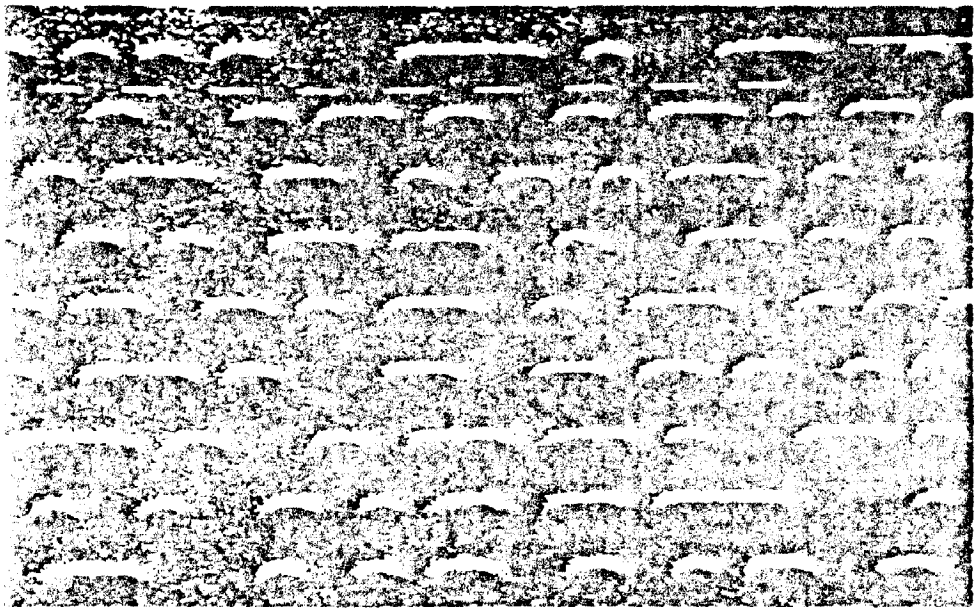


Abb.1

zeigt das System in schematischer Darstellung. Bezüglich der Umwandlung des analogen Signals in digitale Information und die Rückumwandlung in analoge Form beim Abtastvorgang wird auf die angeführte Literatur verwiesen. Wesentlich für uns ist, daß das Audiosignal auf der Disc in digitaler Form gespeichert ist. Dies bietet nämlich die Möglichkeit des Hinzufügens von sogenannten C&D-bits (Kontroll-und Displayinformationsstellen) an die Audiosignalbits. Dadurch wird das System programmierbar, woraus ein erhöhter Bedienungskomfort resultiert. Man kann so etwa die Restspielzeit programmieren oder die Direktwahl von einzelnen, numerierten Musikblöcken ermöglichen.

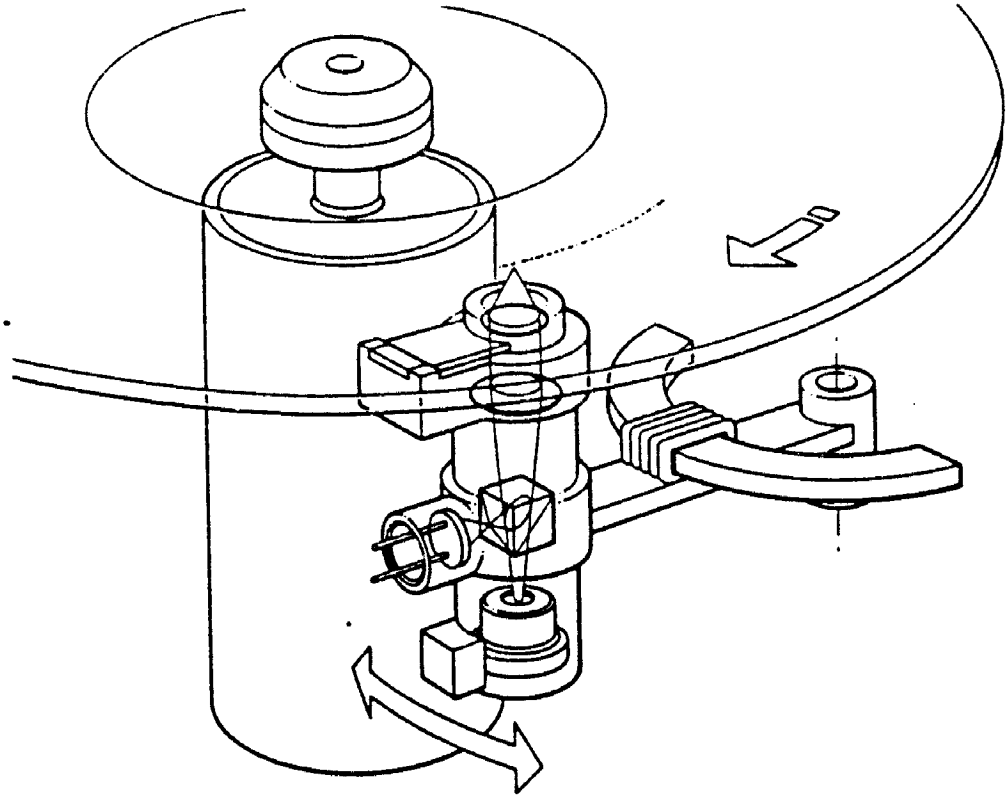
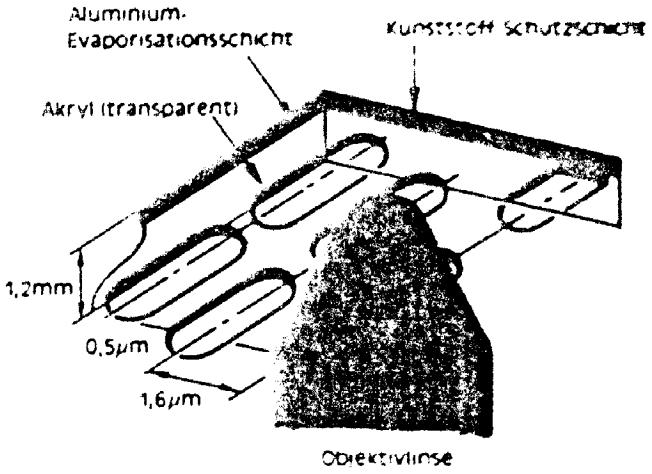


Abb.2

Die Compact-Discs sind weitgehend unempfindlich gegenüber Staub, Kratzern etc. Zum einen wird dies erreicht, daß die Informationsschicht mit einer 1,2 mm dicken Schutzschicht überzogen wird. Die Reflexionsschicht besteht aus Metall und hat eine Dicke von 60 nm. Der Laserstrahl ist stark fokusiert. Er hat beim Auftreffen auf die Schutzschicht noch einen Durchmesser von 0,7 mm. Staub, leichte Kratzer etc. beeinträchtigen also die Reflexion kaum. Siehe schematische Abb.3.

Vergrößerte Darstellung des Plattenaufbaus



Zum andern ist aber für die Unempfindlichkeit gegenüber Verunreinigungen und sogar gegenüber kleinen Produktionsfehlern eine mathematische Methode verantwortlich, die eine Fehlererkennung bzw. eine Fehlerkorrektur der digitalen Signalblöcke (von 16-bit-Länge) ermöglicht. Diese soll nun in ihren Grundprinzipien beschrieben werden. Dazu fassen wir das CD-System als Informationsübertragungssystem auf.

3.Fehlererkennung und Fehlerkorrektur in digitalen Informationsübertragungssystemen.

Ganz allgemein versteht man unter Codierung einer Information deren Verschlüsselung zu einem gewissen Zweck, etwa um sie über einen Kanal übtertragbar zu machen (z.B. Morsezeichen, Rauchsignale etc.), oder sie in einem Medium zu speichern. Meist wird die Information dabei komprimiert (z.B. Inskriptionsnummer, Autokennzeichen etc.). Im Falle der Compact-Disc wird die Audio-Information digitalisiert, speziell in gleichlange 0,1-Folgen zerlegt. Wir nehmen also an, daß unsere Information bereits in digitalisierter Form vorliegt. Die Methoden dieser ersten Codierung sind nicht algebraischer Natur (siehe Literatur). Jedoch kann sie stets so durchgeführt werden, daß jedes Symbol in einem Block und jeder Block mit der gleichen Wahrscheinlichkeit auftritt. Weiters kann man diese erste Codierung, die sogenannte Quellencodierung, stets mit bestmöglicher Informationskomprimierung durchführen.
Die digitalisierte Information nennen wir Nachricht N. Wir wollen sie nun ein weiteres Mal codieren, um Fehlerkorrektur bzw. Fehlererkennung zu erreichen. Als Vorbild dient dabei

in gewissem Sinn die natürliche Sprache. Dort wird die Korrekturmöglichkeit einerseits durch die Gesetzmäßigkeit der Sprache (Grammatik), anderseits durch eine gewisse Redundanz der Information erreicht. Man erweitert also die vorliegende Information durch Hinzufügen von sogenannten Kontrollbits. Eine einfache Möglichkeit wäre die dreimalige Wiederholung jedes Symbols. Man nimmt dann jenes Symbol als Bestandteil der Information, das in den entstehenden Dreiergruppen mindestens zwei Mal auftritt. Liegt etwa der quellencodierte Block 011 vor, so codiert man ihn zu 000 111 111. Nach Übertragung der Information (z.B. Abspielen der Platte) erhält das System beispielsweise 010 111 111. Wenn das System einen Dreifachwiederholungscode benützt, so wird man diese Symbolkette zu 011 (also richtig) decodieren, da dies - sofern das System gewisse Qualitätskriterien erfüllt - die wahrscheinlichste Information ist. Allerdings ist diese Form der "Korrektur von Einfachfehlern" extrem unwirtschaftlich und benötigt entsprechend viel Zeit. Üblicherweise stellt man folgende Forderungen an diese Codierung:

(i) Sie soll möglicht große Sicherheit vor zufälligen Fehlern bieten. Die Wahrscheinlichkeit der "falschen Korrektur" soll möglichst gering sein.

(ii) Die Verfahren zur Codierung und Decodierung sollen möglichst rasch und einfach sein.

(iii) Das Verfahren soll möglichst wirtschaftlich sein. Ein Kriterium dafür ist, daß die Anzahl der benötigten Kontrollsymbole möglichst gering ist.

Um dies zu verwirklichen prägt man dem Code eine Gesetzmäßigkeit auf. Da die Information in einem Computer verarbeitbar sein soll, wird man dabei eine algebraische Struktur wählen. Diese wollen wir schrittweise anreichern. Zunächst diskutieren wir

A. Blockcodes

Die Symbole für die Informationübertragung entnehmen wir

einem festen "Eingabealphabet" A. Im Falle der Compact-Discs ist $A=\{0,1\}$. Im Prinzip muß A nur so beschaffen sein, daß die Information in der quellencodierten Form vom System verarbeitet werden kann. Jede Menge von gleichlangen Blöcken aus Symbolen von A heißt Blockcode über A. Die Elemente der quellencodierten Information N nennen wir Nachrichtenwörter über A. Haben sie alle die Länge m, so ist $N \subseteq A^m$. Die Codierung (d.h. das Anfügen von Kontrollbits) beschreibt man durch eine Funktion $f_C: A^m \to A^n (n>m)$ und $f_C(N) =: C$ heißt Code. Die Elemente von C heißen Codewörter. C ist also ein Blockcode der Länge n über A. Meist führt man die Codierung systematisch durch, d.h. die ersten m Stellen beinhalten die Information, die restlichen n-m Stellen die Kontrollbits. Nach der Übertragung der codierten Information liegen also Elemente von A^n vor. Man nennt sie daher Empfangswörter. Bei der Rückgewinnung der Information bedient man sich bei der Fehlerkorrektur eines sogenannten Korrekturschemas. Darunter versteht man eine Klasseneinteilung von $A^n=\{T_c \mid c \in C\}$ in Klassen T_c. Liegt ein Empfangswort in T_c, so nimmt man c als "korrigiertes" Wort. Nimmt man an , daß die auftretenden Fehler in den einzelnen Symbolen voneinander unabhängig sind und daß im System jeder Übertragungsfehler gleichwahrscheinlich ist, so erhält man ein optimales Korrekturschema, wenn man ein Empfangswort v durch jenes Codewort decodiert, das sich von v an möglichst wenig Stellen unterscheidet. Es kommt also bei der Konstruktion des Codes auf den Abstand zwischen den Codewörtern an. Dies führt auf folgenden Begriff:

Definition: Die Anzahl der Stellen i aus $\{1,\ldots,n\}$, an denen sich zwei Wörter $v,w \in A^n$ unterscheiden, nennt man Hammingdistanz von v und w. In Zeichen: d(v,w).
d ist also ein Funktion von $A^n \times A^n$ in $\mathbb{N}_0$. Sie hat alle Eigenschaften einer Metrik.
Gelingt es nun, C so zu konstruieren, daß verschiedene Codewörter möglichst großen Abstand voneinander haben, so wird nun die Fehlerkorrekturkapazität möglichst groß sein. Kenngröße dafür ist also der Minimalabstand $d_{min}(C)$ des

Codes, d.h. der kleinste Abstand zwischen verschiedenen Codewörtern. Man kann folgenden Satz zeigen: (siehe Abb.4)

Satz: Es gibt genau dann ein Schema K(C), mit dem alle Fehler an höchstens t Stellen korrigiert werden, wenn der Minimalabstand des Codes mindestens 2t+1 ist.

Will man so einen Blockcode maschinell verarbeiten, so bleibt nur die Speicherung des Codes in Listenform, und auch das konstruierte Decodierungsschema muß man als Liste in den Speicher eingeben. Dies ergibt einerseits ein Problem bei der Speicherkapazität, anderseits ein Problem bei der Verarbeitungsgeschwindigkeit. Bevor wir an die Behebung dieser Mängel gehen, sei ein Beispiel angeführt, das das Prinzip des Blockcodes verdeutlicht.

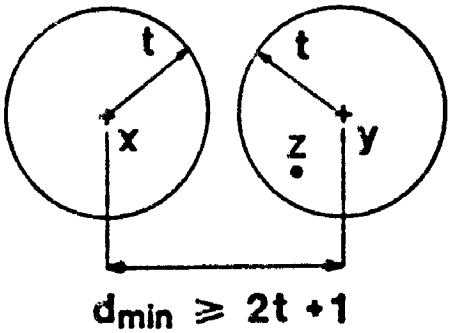


Abb.4

Sei die quellencodierte Information N gegeben durch die Blöcke 001, 110, 100, 101. Wir berechnen die Abstandsverhältnisse.

d	001	110	011	100	101
001	0	3	1	2	1
110		0	2	1	2
011			0	3	2
100				0	1
101					0

Wird N keiner weiteren Codierung unterworfen, so ist keine Möglichkeit einer sinnvollen Fehlerkorrektur gegeben, denn wird etwa 110 verarbeitet und tritt genau ein Fehler an der zweiten Stelle auf, so liegt das Empfangswort 100

vor. Da dieses in N liegt, wird es als richtig angesehen. Daher versucht man die Abstandsverhältnisse zu verbessern, etwa durch folgende Codierung f_C:

$$\begin{aligned} 001 &\rightarrow 001011 \\ 110 &\rightarrow 110001 \\ 011 &\rightarrow 011100 \\ 100 &\rightarrow 100110 \\ 101 &\rightarrow 101101 \end{aligned}$$

Im so erhaltenen Code $C=f_C(N)$ hat man folgende Abstandsverhältnisse:

d	001011	110001	011100	100110	101101
001011	0	4	4	4	3
110001		0	4	4	3
011100			0	4	3
100110				0	3
101101					0

Es ist also $d_{min}(C)= 3$, somit können nach obigem Satz mit C Einfachfehler korrigiert werden.

Zur Behebung der oben angesprochenen Mängel verbessern wir die Struktur unseres Codes:

B.Gruppencodes

Wir wählen als Alphabet eine abelsche Gruppe $\langle A,+\rangle$. Im Fall von {0,1} nimmt man die Addition modulo 2. Nun wählen wir den Code C der Blocklänge n so, daß $\langle C,+\rangle$ eine Untergruppe des n-fachen Produkts $\langle A^n,+\rangle$ von $\langle A,+\rangle$ ist (die Operation + in $\langle A^n,+\rangle$ wird bekanntlich komponentenweise definiert). Wir nennen C dann einen Gruppencode. Wir nützen die algebraische Struktur, indem wir die Codierung so vornehmen, daß f_C ein Gruppenhomomorphismus ist. Dadurch kennt man f_C, wenn die Werte auf einem Erzeugendensystem bekannt sind (bringt geringeren Speicheraufwand). Ist $x \in A^n$ ein Codewort, das nach Durchgang durch das

System zum Empfangswort $y \in A^n$ wird, so gilt x+e=y. Man nennt e Fehlerwort (auch Fehlermuster). Die Möglichkeit des Rechnens mit den Code- und den Empfangswörtern wird vielfach ausgenützt. Definiert man beispielsweise das Hamming-Gewicht w(x) von $x \in A^n$ als die Anzahl der in c auftretenden Symbole ungleich 0 (0 bezeichne das neutrale Element der Gruppe $<A,+>$), so sieht man: $d(x,y) = w(x-y)$. Daraus folgt:

Satz: Ist C ein Gruppencode, so ist der minimale Abstand zwischen verschiedenen Codewörtern gleich dem minimalen Gewicht der vom Nullwort verschiedenen Codewörter.

Also wird der Aufwand zur Bestimmung von $d_{min}(C)$ und damit die Bestimmung der Korrekturkapazität wesentlich vereinfacht, und damit in kürzerer Zeit möglich.

Weiters kann man zeigen, daß die Konstruktion des optimalen Korrekturschemas nichts anderes ist als die Zerlegung der Gruppe $<A^n,+>$ nach der Untergruppe $<C,+>$ in Nebenklassen. Damit kann die Konstruktion algorithmisch durchgeführt werden.

Dazu ein Beispiel: Sie C ein Gruppencode der Blocklänge 6 über {0,1}. Wir schreiben alle Wörter von C in eine Reihe, beginnend mit dem Wort Null:

000000 100110 010101 001011 110011 101101 011110 111000

Nun bestimmen wir ein Wort x von minimalem Gewicht unter den Elementen von $\{0,1\}^6 \setminus C$, bilden die Nebenklasse x+C und schreiben x+c unter das Wort c. So führen wir die gesamte Nebenklassenzerlegung von $<\{0,1\}^6,+>$ nach $<C,+>$ durch. Das erhaltene Korrekturschema nennt man Standardschema:

000000	100110	010101	001011	110011	101101	011110	111000
100000	000110	110101	101011	010011	001101	111110	011000
010000	110110	000101	011011	100011	111101	001110	101000
001000	101110	011101	000011	111011	100101	010110	110000
000100	100010	010001	001111	110111	101001	011010	111100
000010	100100	010111	001001	110001	101111	011100	111010
000001	100111	010100	001010	110010	101100	011111	111001
001100	101010	011001	000111	111111	100001	010010	110100

Man korrigiert jedes empfangene Wort zu dem Wort an der Spitze jener Spalte, in der es im Standardschema steht. Wie man aus der ersten Zeile des Standardschemas sieht, ist $d_{min}(C) = 3$. C kann also Einfachfehler korrigieren.
Um eine Verbesserung der Codierungs-und Decodierungsverfahren zu erreichen, prägt man C eine reichhaltigere Struktur auf:

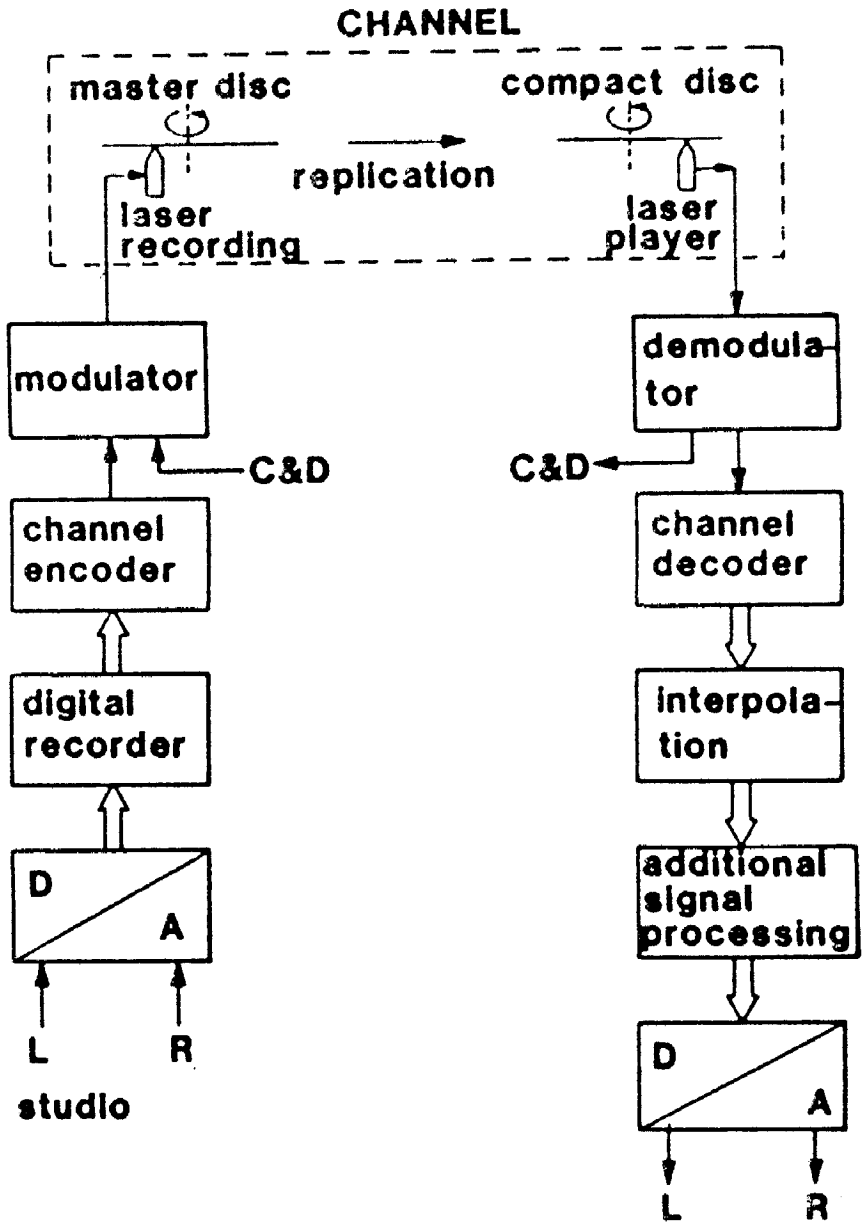
C.Lineare Codes

Wir wählen für das Alphabet einen endlichen Körper $<A,+,\cdot>$. Dann kann man $<A^n,+,A>$ als Vektorraum über A auffassen. Ein Blockcode C der Länge n über A heißt nun linearer Code über A, wenn C Unterraum des Vektorraumes $<A^n,+,A>$ ist. Ist die Dimension von C gleich k, so bezeichnet man C als (n,k)-Linearcode über A.
Wählt man die Codierung f_C als lineare Abbildung, so kann man f_C über eine Matrix (die sogenannte Basismatrix) beschreiben. Somit hat man die Algorithmen der Matrizenrechnung für die Durchführung der Codierung zur Verfügung. Weiters ergeben sich Vorteile für die Decodierung und für die Konstruktion des Korrekturschemas. Für Details siehe [3],[6].
Eine weitere Verbesserung der Codierungs-und Decodierungsalgorithmen erreicht man in sogenannten zyklischen Linearcodes. Ein (n,k)-Linearcode heißt zyklisch, wenn mit $c=c_1 \ldots c_n$ auch alle jene Wörter in C liegen, die durch zyklische Vertauschung der Symbole von c entstehen. Bezeichnen wir das von (x^n-1) erzeugte Ideal im Polynomring A[x] (A endlicher Körper) mit S. Man zeigt leicht, daß $<A^n,+>$ isomorph zu $<A[x]/S,+>$ ist. Man kann nun zeigen, daß C genau dann zyklisch ist, wenn C ein Ideal des Ringes A[x]/S ist. Nun weiß man aus der Algebra, daß A[x]/S ein Hauptidealring ist (A ist ja endlicher Körper), also wird C von einem Poynom, dem sogenannten erzeugenden Poynom, erzeugt. Somit kann man jedes Codewort des zyklischen Linearcodes als Polynom deuten und diese Darstellung zur Verbesserung

der benötigten Algorithmen zur Fehlerkorrektur einsetzen.
Für spezielle Aufgaben - wie etwa die Korrektur von Fehlerbündeln - wurden spezielle zyklische Linearcodes entwickelt. Für Details sei auf die Literatur verwiesen.
Die entwickelten Verfahren kann man auch gleichzeitig zur Fehlererkennung einsetzen, allerdings wird dabei die Fehlerkorrekturkapazität verringert.
Treten in den Wörtern bei der Übertragung bzw. beim Systemdurchgang blanke Stellen (sogenannte erasures) auf, so kann man versuchen, wenn genügend "benachbarte" Information vorhanden ist, durch Interpolation das Signal zu rekonstruieren. Man hat bereits mit dem Einsatz der linearen Interpolation gute Resultate erzielt. Erasures entstehen, wenn einige bits ein Empfangswort als unzulässig erkannt werden. Der Decoder betrachtet sie als ausgelöscht.

4.Compact Discs als digitales Audio-System

Nachdem wir nun die verschiedenen Prinzipien der Fehlerkorrektur vorgestellt haben, soll noch der Systemablauf in Compact-Discs kurz skizziert werden. Auf der linken Seite von Abb.5 ist jener Teil des Produktionsvorganges von Compact-Discs, der im Studio abläuft,schematisch dargestellt, auf der rechten Seite der Vorgang im Abspielgerät. Das Compact-Disc-System verwendet zur Fehlerkorrektur zwei sogenannte Reed-Solomon-Codes, das sind spezielle zyklische Linearcodes, die sich besonders gut zur Korrektur von Fehlerbündeln eignen.
Man kann zeigen, daß t Fehler und e erasures zugleich (pro Code) korrigiert werden können, wenn $2t+e \leq 4$ gilt.
Das ist der Hauptgrund für das authentische Hörerlebnis, das uns das Abspielen von Compact-Discs vermittelt.



The compact disc digital audio system, considered as a transmission system

Abb.5

Literatur

[1] B.A.Blesser: Digitization of Audio.
J.Audio Eng.Soc. 26, 739-771 (1978).

[2] M.G.Carasso - J.B.Peek - J.P.Sinjou: The Compact Disc Digital Audio System.
Philips tech.Rev. 40, 151-155 (1982).

[3] G.C.Clark - J.B.Cain: Error-Correcting Coding for Digital Communications.
Plenum Press, 1981.

[4] D.Goedhart - R.J.Van de Plassche - E.F. Stikvoort: Digital-to-analog Conversion in Playing a Compact Disc.
Philips tech.Rev. 40, 174-179 (1982).

[5] H.Hoeve - J.Timmermans - L.B.Vries: Error Correction and Concealment in the Compact Disc System.
Philips tech.Rev. 40, 166-172 (1982).

[6] H.Kaiser - R.Mlitz - G.Zeilinger: Algebra für Informatiker.
Springer-Verlag, Wien-New York, 2.Auflage (1985).

[7] J.B.H.Peek: Some Features of the Compact Disc Digital Audio System.
Proc. ICIAM 87, 215-234 (1987).

Institut für Algebra und Diskrete Mathematik
TU Wien
Wiedner Hauptstr. 8-10
1040 WIEN