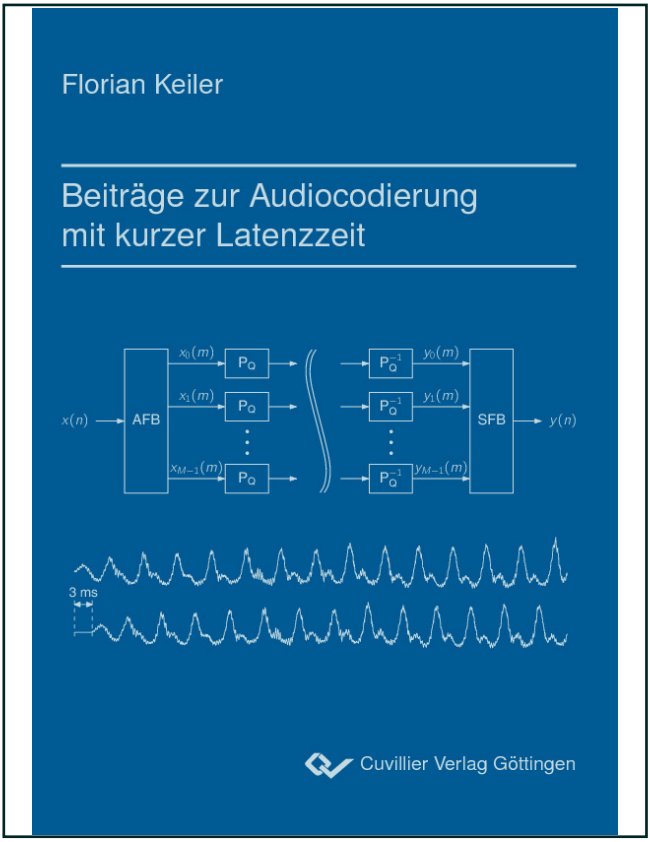
Florian Keiler (Autor)

Beiträge zur Audiocodierung mit kurzer Latenzzeit



https://cuvillier.de/de/shop/publications/2108

1 Einführung

Die verlustbehaftete Audiocodierung hat in den letzten Jahren eine starke Verbreitung sowohl in der Unterhaltungselektronik als auch in der professionellen Studiotechnik gefunden. Die Audiocodierung wird immer von einem *Encoder* und einem *Decoder* durchgeführt. Der Encoder berechnet aus dem Eingangs- bzw. Referenzsignal das in der Datenmenge reduzierte codierte Signal bzw. einen Datenstrom, während der Decoder aus diesem Datenstrom wieder ein hörbares Audiosignal rekonstruiert: das decodierte bzw. rekonstruierte Signal. Die Kombination von Encoder und Decoder wird kurz auch als *Codec* bezeichnet.

Durch die Codierung und Decodierung eines Audiosignals erfährt das betrachtete Signal eine Verzögerung. Diese Verzögerung wird auch als *Latenzzeit* oder *Latenz* bezeichnet und hängt vom benutzten Codierverfahren ab. Vor der Auswahl eines Codierverfahrens muss zunächst eine Obergrenze für die akzeptable Signalverzögerung durch den Codec festgelegt werden. Diese maximal zugelassene Latenzzeit hängt von der betrachteten Anwendung ab.

Jede verlustbehaftete Audiocodierung produziert einen *Codierungsfehler*. Hiermit wird die Differenz zwischen dem decodierten Signal und dem Referenzsignal (dem zu codierenden Signal) bezeichnet. Bei dieser Differenzbildung ist zu beachten, dass ggf. die Verzögerung (Latenz) durch die Encodierung/Decodierung berücksichtigt werden muss, d. h. das Referenzsignal muss bei der Differenzbildung entsprechend verzögert werden. Das Ziel einer Audiocodierung ist immer, dass der durch die Codierung entstandene Codierungsfehler nicht wahrnehmbar ist.

Für die Reduktion der Datenmenge von Audiosignalen gibt es eine Vielzahl von Anwendungen. Zum Einen möchte man Audiodaten mit möglichst wenig Speicheraufwand archivieren und abspielen. Für diesen Anwendungsfall ist keine extrem kurze Latenzzeit nötig. Vielmehr ist hier nur die Signalverzögerung durch den Decoder von Interesse. Für diese Anwendung ist ein psychoakustisches Codierverfahren wie z. B. MP3 (MPEG-1 Layer-3, MPEG = Moving Pictures Expert Group) [BSD+94, BB97, Bra99] gut geeignet. Diese Codecs haben eine Gesamtverzögerung von etwa 100 bis 200 ms, wobei die Verzögerung des Decoders unter 50 ms beträgt.

Zum Anderen gibt es eine Reihe von Anwendungen, für die eine kürzere Latenzzeit erforderlich ist. Zu diesen Anwendungen zählen die Kommunikation und der Einsatz bei drahtlosen Mikrophonen und Kopfhörern in Echtzeitumgebungen. Hier darf die Verzögerung des Signals durch den Codec nur in einer Größenordnung von etwa 10 ms liegen. In [HL99] wird eine Codec-Verzögerung von 2–5 ms vorgeschlagen. Die vorliegende Arbeit beschäftigt sich mit einem Codierungsansatz für das zuletzt genannte Anwendungsfeld.

1.1 Digitale Übertragung von Audiosignalen

Ein Anwendungsgebiet für die Audiocodierung mit kurzer Latenzzeit ist die Verwendung drahtloser Kopfhörer und Mikrophone im Studio- oder Bühneneinsatz. Insbesondere beim Einsatz drahtloser Mikrophone, z. B. bei Aufführungen oder Konzerten, ist eine kurze Latenzzeit erforderlich, damit das übertragene Signal synchron zum Direktschall und zur Lippenbewegung ist. Ein Großteil der derzeit eingesetzten Systeme für drahtlose Mikrophone arbeitet mit einer analogen Übertragung und nutzt als Übertragungsverfahren Schmalband-FM (FM = Frequenzmodulation). Diese Systeme haben den Nachteil, dass sich das Rauschen der Übertragungsstrecke im demodulierten Signal wiederfindet. Wesentlich robuster gegenüber Kanalstörungen sind digitale Übertragungsverfahren. Bei Nutzung von Kanalcodierung und optimalem Modulationsverfahren sind die Empfangsdaten identisch mit den Sendedaten. Für eine digitale Übertragung bei vorgegebener Kanalbandbreite sollte das zu übertragende Signal eine konstante Bitrate aufweisen, welche zudem die Wahl eines geeigneten Modulationsverfahrens vereinfacht.

Um die Bandbreite der digitalen Übertragung möglichst gering zu halten, sollte die Menge der Audiodaten vorab reduziert werden, ohne den Informationsgehalt des Audiosignals zu verfälschen. Eine verlustlose Codierung [HS01] ist für diesen Zweck ungeeignet, weil die damit erreichbare Datenreduktion signalabhängig ist und typischerweise nur eine Datenreduktion um den Faktor 2–3 erfolgt. Zur Einsparung von Übertragungsbandbreite sollte daher eine verlustbehaftete Audiocodierung zum Einsatz kommen.

1.2 Zielvorgaben für den zu entwerfenden Audiocodec

Für die Übertragung in von der ETSI (European Telecommunications Standards Institute) verwalteten Kanälen beträgt die maximal mögliche Kanalbandbreite 200 kHz [ETSI98]. Als Kanalbandbreite gilt hierbei die Bandbreite des modulierten Signals, dessen Leistung an den Bandgrenzen bezüglich des unmodulierten Trägers um 60 dB abgefallen sein muss. Mit einer Bandbreite-Effizienz von 1 bit/s/Hz folgt eine Bitrate von 200 kbit/s. Ausgehend von der Audiobitrate für ein Monosignal bei CD-Qualität von

$$R_{\mathrm{CD}} = 16\,\frac{\mathrm{bit}}{\mathrm{Abtastwert}} \cdot 44100\,\frac{\mathrm{Abtastwerte}}{\mathrm{s}} = 705{,}6\,\mathrm{kbit/s} \qquad (1.1)$$

ist eine Kompression um den Faktor 4 sinnvoll. Dies entspricht im Mittel einer Darstellung mit 4 bit/Abtastwert.

Wie im vorangegangenen Abschnitt beschrieben, ist für die betrachtete digitale Übertragung von Audiosignalen ein Codierverfahren mit kurzer Verzögerungszeit erforderlich. Der Audiocodec sollte eine Gesamtlatenz von unter 5 ms haben (s. o.) und einen codierten Bitstrom konstanter Bitrate liefern. Außerdem soll der zu realisierende Audiocodec eine transparente Audioqualität liefern, d. h. im decodierten Signal soll gegenüber dem Originalsignal keine Störung wahrnehmbar sein.

Zur effizienten Realisierung ist eine Anpassung der eingesetzten Signalverarbeitung für eine Echtzeitimplementierung notwendig. Als Zielsystem wird ein digitaler Signalprozessor (DSP) mit Festkomma-Arithmetik benutzt. Ein wichtiger Aspekt ist hierbei, dass der DSP möglichst gleichmäßig ausgelastet wird. Die für die Berechnungen benötigte Anzahl an Instruktionen sollte also wenig über der Zeit schwanken bzw. keine großen Maximalwerte (Peaks) aufweisen. Für die Auswahl eines geeigneten Prozessors ist nur die maximale Anzahl von benötigten Instruktionen pro Abtastwert entscheidend.

1.3 Verwendeter Codierungsansatz

Bild 1.1 zeigt den in dieser Arbeit untersuchten Codierungsansatz der Teilband ADPCM Codierung (ADPCM = adaptive differenzielle Pulse Code Modulation).

Das Eingangssignal $x(n)$ wird mit einer Analyse-Filterbank (AFB) in kritisch abgetastete Teilbandsignale $x_i(m)$ zerlegt. Diese Teilbandsignale werden

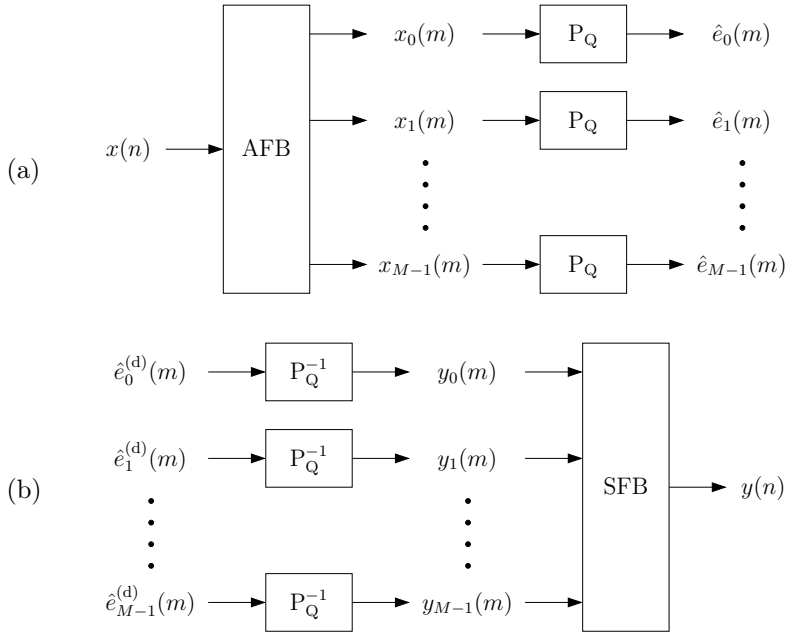


Bild 1.1: *Gesamtstruktur der Teilband ADPCM Audiocodierung: (a) Encoder, (b) Decoder*

– angedeutet durch den Block P_Q – durch eine Prädiktion und Quantisierung codiert. Die Rekonstruktion des Signals $y(n)$ erfolgt durch die Decodierung der Teilbänder (Block P_Q^{-1}) und anschließende Synthesefilterbank (SFB).

Zur Prädiktion wird pro Teilband eine ADPCM-Struktur benutzt und es erfolgt eine adaptive Quantisierung der Prädiktionsfehlersignale. Die ADPCM berechnet aus Vergangenheitswerten des rekonstruierten Teilbandsignals eine Schätzung (Prädiktion) $\hat{x}_i(m)$ des Wertes $x_i(m)$ und es wird nur der Prädiktionsfehler $e_i(m) = x_i(m) - \hat{x}_i(m)$ weiterverarbeitet. Zur Neuberechnung der Prädiktionkoeffizienten wird eine modifizierte Variante des Burg-Algorithmus benutzt, die für eine Echtzeitimplementierung optimiert ist [KAZ00]. Die Filteroperationen werden in einer Lattice-Struktur realisiert, die sehr vorteilhaft für eine Festkomma-Implementierung ist. Hierbei ist beim Ändern der Filterkoeffizienten eine Neuberechnung der Filterzustände nötig, da die Zustände – anders als bei der Direktform – von den Koeffizienten abhängen. Es ergibt

sich damit eine Ähnlichkeit zu den Problemen bei adaptiven rekursiven Filtern [VL98b].

Bei der adaptiven Quantisierung wird der von der ADPCM gelieferte Prädiktionsfehler $e_i(m)$ basierend auf quantisierten Vergangenheitswerten an den Aussteuerungsbereich des Quantisierers angepasst. Außerdem wird eine optimierte Quantisierungskennlinie benutzt, die auf die Amplitudenverteilung des zu quantisierenden normierten Prädiktionsfehlers angepasst ist; hierzu werden die Ansätze aus [Max60, Llo82] benutzt.

Die Teilbandverarbeitung ist latenzfrei, da sowohl die Prädiktionskoeffizienten als auch der Skalierungsfaktor im adaptiven Quantisierer aus vergangenen, rekonstruierten Abtastwerten berechnet werden. Diese Coderstruktur hat den weiteren Vorteil, dass außer den quantisierten Teilband-Prädiktionsfehlern keine Nebeninformation zu übertragen ist. Im Decoder werden die benötigten Werte in gleicher Weise wie im Encoder aus den quantisierten Teilband-Prädiktionsfehlersignalen zurückgewonnen. Die Gesamtlatenz des präsentierten Coders resultiert bei diesem Ansatz also nur aus der Signalverzögerung durch die Analyse-/Synthese-Filterbank.

Die Verarbeitung von Teilbandsignalen hat mehrere Vorteile. Zum Einen ist die Bandbreite der Teilbandsignale gegenüber der Gesamtbandbreite reduziert. Die Filterordnung der Prädiktion kann deutlich kleiner sein als bei breitbandiger Prädiktion. Da die Teilbandsignale in der reduzierten Abtastrate verarbeitet werden, ist der Aufwand vergleichbar zu breitbandiger Prädiktion der gleichen Ordnung. Zum Anderen kann in jedem Teilband für die Quantisierung eine unterschiedliche Wortbreite – angepasst an die Eigenschaften des menschlichen Gehörs – gewählt werden. Die durchschnittliche Wortbreite aller Teilbandsignale muss dabei die angestrebten 4 bit erreichen. Bei der Codierung mit ADPCM entsteht durch die Quantisierung weißes Rauschen pro Teilband. Der Pegel dieses Codierungsfehlers hängt von der Qualität der Prädiktion sowie von der benutzten Bitzahl ab. Durch den beschriebenen Ansatz der Teilbandverarbeitung wird also eine grobe spektrale Formung des Codierungsfehlers erreicht.

Das untere Teilband besitzt bei 8 Teilbändern und der Abtastrate $f_\mathrm{A} = 44{,}1\,\mathrm{kHz}$ eine Bandbreite von etwa 2,7 kHz und es ist wünschenswert, den entstehenden Codierungsfehler in diesem Frequenzbereich zusätzlich spektral zu formen. Als Verbesserung ist somit vorgesehen, im unteren Teilband eine adaptive Vor- und Nachfilterung durchzuführen.

1.4 Überblick über folgende Kapitel

In diesem Abschnitt wird ein kurzer Überblick über die folgenden Kapitel gegeben.

In Kapitel 2 wird zunächst die psychoakustische Audiocodierung nach dem MPEG-Standard kurz beschrieben. Anschließend werden verschiedene Ansätze und Aspekte der Audiocodierung mit kurzer Latenzzeit aus der aktuellen Literatur näher erläutert.

Kapitel 3 beschreibt die Grundlagen der Psychoakustik sowie die Eigenschaften des menschlichen Gehörs. Es wird ein später eingesetztes Verfahren zur objektiven Qualitätsbeurteilung von Audiocodecs vorgestellt.

Die eingesetzte Filterbank wird in Kapitel 4 behandelt. Hier werden zunächst die benötigten Grundlagen vorgestellt, bevor eine effiziente Realisierung einer Polyphasenfilterbank beschrieben wird. Eine Modifikation eines Entwurfsverfahrens für linearphasige Tiefpass-Prototypen wird erläutert, und eine Erweiterung der Polyphasenfilterbank für die Nutzung nicht-linearphasiger Prototypen wird dargestellt.

Die in den Teilbändern eingesetzte Prädiktion wird in Kapitel 5 erläutert. Verschiedene Coderstrukturen sowie eine effiziente adaptive Neuberechnung der benötigten Filterkoeffizienten werden dargestellt.

Die eingesetzte adaptive Quantisierung der Teilband-Prädiktionsfehler beschreibt Kapitel 6. Hier werden die erforderliche Anpassung des Signalpegels und eine Optimierung der Quantisierungskennlinie erläutert.

In Kapitel 7 wird die spektrale Rauschformung mit Benutzung von adaptiven Vor-/Nachfiltern beschrieben. Es wird insbesondere eine Adaption von parametrischen Bewertungsfiltern (Shelving-Filtern) in einer Reihenschaltung betrachtet.

Das Gesamtsystem aus Filterbank, Teilbandprädiktion und adaptiver Quantisierung der Teilband-Prädiktionsfehler wird in Kapitel 8 behandelt. Auf die Wahl der Anzahl der Teilbänder und der Wortbreiten zur Quantisierung der Teilbandsignale wird eingegangen. Mit der objektiven psychoakustischen Bewertung gemäß Kapitel 3 erfolgt ein Vergleich verschiedener Codierungsstrukturen. Anschließend wird das eingesetzte Echtzeitsystem mit einer Anbindung an eine Funkstrecke beschrieben.

Die Erkenntnisse dieser Arbeit werden zusammenfassend in Kapitel 9 präsentiert. Dort werden auch Hinweise auf mögliche Verbesserungen und zukünftige Forschungsbereiche gegeben.

2 Stand der Forschung

In diesem Kapitel wird zunächst die psychoakustische Audiocodierung nach dem MPEG-Standard kurz beschrieben. Obwohl diese Verfahren für den betrachteten Ansatz mit kurzer Latenzzeit nicht in Frage kommen, dienen die Verfahren als Referenz bezüglich der erreichten Audioqualität.

Anschließend wird ein Überblick über bereits veröffentlichte Ansätze zur Audiocodierung mit kurzer Latenzzeit gegeben. Zunächst werden Ansätze basierend auf linearer Prädiktion dargestellt, dann Ansätze mit einer Teilbandzerlegung beschrieben. Abschließend werden Verfahren zur spektralen Formung des Codierungsfehlers mit Nutzung einer Vor-/Nachfilterung analysiert.

Es wird sich zeigen, dass die Aspekte der Rauschformung (spektrale Formung des Codierungsfehlers) bei der Audiocodierung viele Gemeinsamkeiten mit den Ansätzen der Kompandertechnik [Sch89] besitzen. Diese Verfahren der Kompandertechnik werden zur Rauschunterdrückung analoger Speichermedien (z. B. Compactcassette) bzw. zur analogen Übertragung benutzt.

2.1 Psychoakustische Audiocodierverfahren

Mit Kenntnis der psychoakustischen Eigenschaften des menschlichen Gehörs und Nutzung dieser Kenntnisse für die Audiocodierung kann der Speicherbedarf von Audiodaten drastisch reduziert werden, so dass entsprechende Verfahren sehr gut zum Speichern bzw. Archivieren von Audiodaten geeignet sind. Das bekannteste Kompressionsformat, das diesen Ansatz verfolgt, ist MP3 [BSD$^+$94, Bra99].

Psychoakustische Audiocodierverfahren wie z. B. MP3 produzieren einen Codierungsfehler, der spektrale Eigenschaften hat, so dass dieser Fehler bei entsprechender Qualitätseinstellung des Codecs nicht wahrnehmbar ist. Im Idealfall ist der Pegel des Codierungsfehlers kleiner als die Maskierungsschwelle. Die Maskierungsschwelle ist signalabhängig und gibt für jede Frequenz den minimalen Pegel eines Tons an, damit dieser Ton zusätzlich zum betrachteten Signal wahrnehmbar ist, s. a. Kapitel 3. In der Regel kann bei psychoakustischen Audio-Codierverfahren durch Vorgabe der gewünschten (mittleren) Bit-

rate die zu erreichende Qualität gewählt werden.

Es ist bei einigen Algorithmen auch möglich, die gewünschte Qualität in mehreren Stufen zu wählen, so dass dann ein codierter Bitstrom mit variabler Bitrate erzeugt wird. Der Einsatz einer variablen Bitrate ist nicht für eine Übertragung bei konstanter Sendebandbreite geeignet, da in diesem Fall mit Hilfe eines Zwischenspeichers eine Anpassung an das benutzte Modulationsverfahren erfolgen müsste. Dieser Zwischenspeicher würde die Gesamtlatenz des Audiocodecs erhöhen.

Bild 2.1 zeigt den Aufbau des MP3 Audiocoders. Das Eingangssignal wird zur Ermittlung der Maskierungsschwelle (psychoakustisches Modell) einer Frequenzanalyse unterzogen, und parallel dazu wird das Signal mit einer fein auflösenden Filterbank in seine Frequenzanteile zerlegt. Das Eingangssignal wird zunächst mit einer Polyphasenfilterbank in 32 Teilbänder zerlegt. Jedes dieser Teilbänder wird dann mit einer MDCT (modifizierte diskrete Cosinus Transformation) in 18 noch kleinere Teilbänder zerlegt. Jedes dieser fein aufgelösten Teilbänder wird dann mit einer optimalen Bitanzahl quantisiert, die Anzahl der Bits hängt für jedes Band von dem Abstand des Signalpegels von der Maskierungsschwelle ab (Signal-Maskierungs-Abstand SMR). Im Decoder erfolgt die Rekonstruktion des breitbandigen Signals aus den quantisierten Teilbandsignalen mit der inversen MDCT (IMDCT) und der Synthesefilterbank. Aufgrund der durchgeführten Frequenzanalyse und der feinen Teilbandzerlegung wird eine sehr genaue spektrale Formung des Codierungsfehlers erreicht.

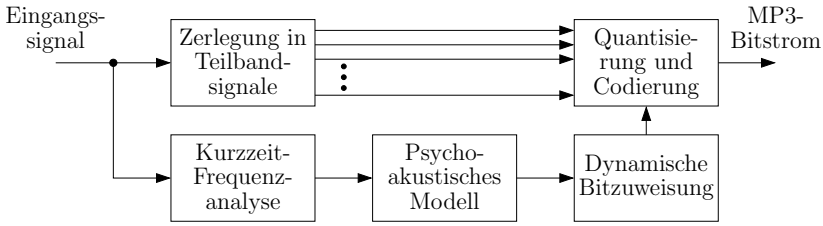


Bild 2.1: *Aufbau eines MPEG-1 Layer-3 (MP3) Audio-Encoders*

2.1.1 Latenzzeit von psychoakustischen Audiocodecs

Bei psychoakustischer Audiocodierung nach dem MPEG-Standard hängt die entstehende Signalverzögerung von verschiedenen Faktoren ab. Die folgende Betrachtung basiert auf [AGHS99, Pur99], die jedoch MPEG-2 AAC (Advan-

ced Audio Coding) und MPEG-4 behandeln. Übertragen auf MPEG-1 Layer-3 [BSD+94] ergeben sich folgende Faktoren:

- Blocklänge der Blockverarbeitung: Es muss zunächst eine bestimmte Anzahl von Abtastwerten bekannt sein, bevor die Transformation durchgeführt werden kann.
- Laufzeit durch die Analyse- und Synthese-Filterbank sowie Transformation per MDCT
- Analyse zukünftiger Blöcke (engl. look-ahead, block switching) für die Umschaltung des Analysefensters zur Verbesserung der Codierung transienter Anteile (Vermeidung von Pre-Echos)
- Bit Reservoir für eine variable Bitrate, da das Vorgehen zur Vermeidung von Pre-Echos temporär eine höhere Bitrate benötigt

Die Blocklänge beträgt dabei $36 \cdot 32 = 1152$ Abtastwerte, und die Länge des Filterbank-Prototyps beträgt 512 Abtastwerte [BSD+94]. Damit ergibt sich für die Analyse- und Synthese-Filterbank inklusive MDCT eine Verzögerung von 2816 Abtastwerten entsprechend etwa 64 ms bei einer Abtastrate von $f_A = 44{,}1\,\text{kHz}$. Die Gesamtlatenz vergrößert sich weiter durch das Block Switching sowie das Bit Reservoir. Insbesondere bei kleinen Bitraten steigt die Latenzzeit durch das Bit Reservoir beträchtlich. Insgesamt ergibt sich eine Latenzzeit von mindestens rund 100 ms.

Beim MPEG-4 Low Delay Audio Coder (AAC-LD) [AGHS99, Pur99, Fra05a] wird die Latenzzeit im Vergleich zu MPEG-1 Layer-3 durch folgende Maßnahmen reduziert:

- Statt der hybriden Filterbank (Polyphasen-Filterbank kombiniert mit einer MDCT) wird nur eine MDCT eingesetzt
- Reduktion der Block- und Transformationslänge auf 512 oder 480 Abtastwerte
- Umschaltung des Analysefensters wird durch Temporal Noise Shaping (TNS) [HJ97, Her99] ersetzt
- Verkleinerung oder vollständiges Weglassen des Bit Reservoirs

Somit ergibt sich eine Minimalverzögerung von 960 Abtastwerten, was bei $f_A = 44{,}1\,\text{kHz}$ etwa 21,7 ms entspricht. MPEG-4 AAC-LD ist damit für eine Zweiwegekommunikation zwar bedingt geeignet, für eine Anwendung im