Christian Helmrich

Efficient Perceptual Audio Coding Using Cosine and Sine Modulated Lapped Transforms

Effiziente wahrnehmungsorientierte Audiocodierung unter Verwendung kosinus- und sinusmodulierter überlappender Transformationen

---

Der Technischen Fakultät der
Friedrich-Alexander-Universität Erlangen-Nürnberg
zur Erlangung des Doktorgrades

Doktor-Ingenieur

vorgelegt von
Christian R. Helmrich
aus Cuxhaven

Als Dissertation genehmigt
von der Technischen Fakultät der
Friedrich-Alexander-Universität Erlangen-Nürnberg

Tag der mündlichen Prüfung: 18. Mai 2017

Vorsitzender des Promotionsorgans: Prof. Dr.-Ing. Reinhard Lerch

Gutachter: Prof. Dr.-Ing. Bernd Edler
Prof. Dr.-Ing. habil. Rudolf Rabenstein

Printed in Germany

Second edition, May 2017 (first edition: November 2016)

Dedicated to my mother

Angelika

In loving memory of my father

Lutz

Abstract

The increasing number of simultaneous input and output channels utilized in immersive audio configurations primarily in broadcasting applications has renewed industrial requirements for efficient audio coding schemes with low bit-rate and complexity. This thesis presents a comprehensive review and extension of conventional approaches for perceptual coding of arbitrary multichannel audio signals. Particular emphasis is given to use cases ranging from two-channel stereophonic to six-channel 5.1-surround setups with or without the application-specific constraint of low algorithmic coding latency.

Conventional perceptual audio codecs share six common algorithmic components, all of which are examined extensively in this thesis. The first is a signal-adaptive filterbank, constructed using instances of the real-valued modified discrete cosine transform (MDCT), to obtain spectral representations of successive portions of the incoming discrete time signal. Within this MDCT spectral domain, various intra- and inter-channel optimizations, most of which are of linear predictive nature, are employed as a second step to minimize spectral, temporal, and/or spatial redundancy. These processing steps are succeeded by a psychoacoustically motivated and controlled quantization process, with optional simple parametric extensions such as noise substitution or related forms of MDCT coefficient exchange, in order to reach the desired coding bit-rate. The fourth component comprises lossless entropy coding of the quantized spectral coefficients and parameters as well as the compilation of all entropy coded data into a transmittable bitstream. Components five and six, finally, represent low-bit-rate methods for improved high-frequency regeneration for audio bandwidth extension and downmix-based stereo or surround coding, which generally do not operate in the MDCT domain but require an additional pair of complex-valued pseudo-quadrature mirror filter (QMF) banks around the MDCT core infrastructure. The auxiliary filter-banks are shown to notably increase both the algorithmic codec complexity and latency, rendering their usage for low-delay communication applications difficult, especially on battery-powered mobile devices.

The complex-domain coding tools can be regarded as pre- and post-processors to the MDCT core-coder, and it is demonstrated that most algorithmic details of these tools can be integrated directly into the MDCT architecture. Moreover, algorithms for respective encoder-side calculation of the modified spectral coefficients and the associated coding parameters, i. e., analysis, are derived which allow the decoder-side reconstruction, i. e. synthesis, to remain real-valued. More specifically, exclusive utilization of the MDCT can be maintained in the decoder, while the modulated complex lapped transform (MCLT),

whose real part is the MDCT and whose imaginary part is represented by the modified discrete sine transform (MDST), may be employed in the encoder for best audio quality. Phase-related details of the conventional complex-valued coding algorithms, which are difficult to realize using only real-valued transformation, are substituted by an intensity downmix-based but subjectively acceptable encoder-side pre-processing operation.

The characteristics of state-of-the-art MDCT filter-bank designs are the second focus of this thesis. Continuing the above investigation of parametric stereo/surround coding methods, an extension of the MDCT coding paradigm, applying sine modulation by way of the MDST instead of the traditional cosine modulation in some channels, is described. Time domain aliasing cancelation (TDAC) compliant transitions between the MDCT and MDST instances, for perfect reconstruction (PR) in the absence of spectral quantization, are discussed. When used in a signal-adaptive fashion, this so-called "kernel switching" method leads to significant coding quality gains on input material with an inter-channel phase difference (IPD) around ±90°. Thereafter, a so-called "ratio switching" approach is presented. Its purpose is the signal-adaptive variation of the inter-transform overlap ratio based on the input's instantaneous harmonicity and temporal flatness. To this end the definition of the extended lapped transform (ELT), whose overlap ratio exceeds that of the MDCT and MDST, is modified to allow transitions to and from the latter two transforms with PR, i. e., proper TDAC. Using the modified ELT (MELT) with a newly designed window function on tonal quasi-stationary waveform portions, e. g., recordings of single instruments, while resorting to the MDCT or MDST on noise-like and/or non-stationary parts, is shown to yield small but significant improvements in overall coding quality.

For low-delay use cases, where the additional look-ahead due to increased transform overlap ratio is undesirable, long-term predictive (LTP) coding as an alternative to ratio switching is examined as a third and final topic. After reviews of conventional time- and frequency-domain approaches, a new MDCT-domain algorithm with low parameter rate (one periodicity value per time unit) and complexity (a fraction of that of the prior art) is proposed. Supporting intra- and inter-channel prediction, this frequency-domain predictor (FDP) offers coding gains which are close, and orthogonal, to those of the MELT.

The work concludes with comparative objective and subjective evaluation of the presented contributions, when integrated into the MPEG-D USAC based MPEG-H 3D Audio codec. Objective assessment reveals large savings in delay and decoder complexity, and blind subjective testing indicates that, in terms of audio quality, the modified MPEG-H codec matches or outperforms the respective state of the art in both general-purpose and low-delay applications. Most importantly, for both stereo and 5.1-surround channel configurations, more consistent audio quality across the different types of input signals, with fewer observed negative outliers, is achieved in comparison to the state of the art.

Kurzfassung

Die steigende Anzahl gleichzeitig genutzter Eingangs- und Ausgangskanäle in Raumklangkonfigurationen v. a. in Rundfunkanwendungen hat industrielle Forderungen nach effizienten Audiocodiersystemen mit niedriger Bitrate und Komplexität erneuert. Diese Arbeit präsentiert einen umfassenden Überblick über die konventionellen Ansätze zur wahrnehmungsorientierten Codierung beliebiger Multikanal-Audiosignale und stellt im Anschluss Erweiterung bzw. Verbesserungen dieser vor. Besonderes Augenmerk gilt dabei Anwendungsfällen von Zweikanal-Stereo bis Sechskanal-5.1-Surround mit und ohne etwaiger einsatzspezifischer Beschränkung auf niedrige algorithmische Codierlatenz.

Konventionelle wahrnehmungsbezogene Audio-Codecs verwenden sechs vergleichbare algorithmische Komponenten, welche alle in dieser Arbeit untersucht werden. Die erste ist eine signal-adaptive Filterbank aus Realisierungen der reellwertigen modifizierten diskreten Kosinus-Transformation (MDCT), die eine spektrale Darstellung aufeinanderfolgender Abschnitte des eingehenden diskreten Zeitsignals erlaubt. Innerhalb dieses MDCT-Spektralbereichs finden diverse Intra- und Interkanal-Optimierungen, von denen die meisten linearprädiktiver Natur sind, als zweiter Schritt Anwendung mit dem Ziel der Minimierung spektraler, zeitlicher und räumlicher Redundanz. Darauf folgt ein psychoakustisch motivierter und kontrollierter Quantisierungs-Prozess, mit optionalen parametrischen Erweiterungen wie Rausch-Ersatz oder ähnlichen Formen des MDCT-Koeffizientenaustauschs, zur Erzielung der gewünschten Codierbitrate. Die vierte Komponente umfasst die verlustfreie Entropie-Codierung aller quantisierten Spektralwerte und Parameter sowie die Erfassung der codierten Daten im zu übertragenden Bitstrom. Die Komponenten fünf und sechs schließlich repräsentieren Methoden für verbesserte Hochfrequenzrekonstruktion zur Audiobandbreitenerweiterung und downmixbasierte Stereo- oder Surround-Codierung bei niedrigen Bitraten. Diese arbeiten meist nicht in der MDCT-Domäne, sondern benötigen zusätzliche komplexwertige Pseudo-Quadratur-Spiegelfilterbänke (QMF) außerhalb der MDCT-Infrastruktur. Die Zusatz-Filterbänke führen dabei zu deutlich erhöhter algorithmischer Codec-Komplexität und -Latenz, was ihre Verwendung in Kommunikationsanwendungen, v. a. auf Mobilgeräten, erschwert.

Die komplexwertig arbeitenden Codierkomponenten können als Vor- und Nachverarbeitungsschritte um den MDCT-Codierkern angesehen werden, und es wird aufgezeigt, dass die meisten algorithmischen Details dieser Komponenten in die MDCT-Architektur integriert werden können. Außerdem werden Analyse-Algorithmen für entsprechende encoderseitige Berechnungen modifizierter Spektralwerte und zugehöriger Codierpara-

meter entwickelt, welche eine Beibehaltung der Reellwertigkeit der entsprechenden decoderseitigen Rekonstruktion, sprich der Synthese-Algorithmen, ermöglichen. Im Detail bedeutet dies die ausschließliche Nutzung der MDCT im Decoder, während im Encoder eine modulierte komplexe überlappte Transformation (MCLT), deren Realteil die MDCT darstellt und deren Imaginärteil durch die modifizierte diskrete Sinus-Transformation (MDST) gegeben ist, für beste Klangqualität verwendet werden kann. Phasenbezogene Einzelheiten der konventionellen komplexen Algorithmen, welche nur mit reellwertigen Transformationen schwer zu realisieren sind, werden durch Mono-Einkanalmischungsbasierte aber perzeptuell akzeptable Vorverarbeitung auf der Encoderseite ersetzt.

Die Eigenschaften des Stands der Technik bezüglich MDCT-Filterbank-Design bilden den zweiten Schwerpunkt dieser Arbeit. Der vorherigen Untersuchung parametrischer Stereo-/Surround-Codiermethoden folgend wird eine Erweiterung des MDCT-Prinzips beschrieben, in der eine Sinus-Modulation mittels der MDST, statt der üblichen Kosinus-Modulation, in manchen Kanälen verwendet wird. Erhaltung der „time domain aliasing cancelation" (TDAC) bei Übergängen zwischen MDCT- und MDST-Instanzen für perfekte Rekonstruktion (PR) bei fehlender spektraler Quantisierung wird dabei betrachtet. Auf signal-adaptive Weise realisiert führt diese sogenannte „kernel switching"-Methode zu merklicher Verbesserung der Codierqualität bei Eingangsmaterial mit einer Interkanal-Phasendifferenz (IPD) nahe ±90°. Im Anschluss wird ein sogenanntes „ratio switching" präsentiert, dessen Zweck die signal-adaptive Variation des Überlappungsverhältnisses zwischen den Transformationen basierend auf der momentanen Harmonizität und zeitlichen Flachheit des Eingangssignals ist. Hierzu wird die Definition der extended lapped transform (ELT), deren Überlappungsverhältnis das der MDCT und MDST übersteigt, so verändert, dass TDAC-konforme Übergänge von und zu letzteren Transformationen, d. h. mit PR, ermöglicht werden. Bei der Anwendung der modifizierten ELT (MELT), mit einer neuentwickelten Fensterfunktion, auf tonalen quasistationären Wellenformabschnitten z. B. von Einzelinstrument-Aufnahmen, kombiniert mit der üblichen Nutzung der MDCT oder MDST bei rauschartigen und/oder nichtstationären Signalbereichen, lassen sich so geringfügige aber signifikante Verbesserungen der Gesamt-Codierqualität erzielen.

Für Anwendungen mit geringer Latenz, welche zusätzliche zeitliche Vorgriffe bedingt durch verlängerte Transformationen nicht erlauben, wird die langzeit-prädiktive (LTP) Codierung als Alternative zum ratio switching als drittes und letztes Thema untersucht. Nach der Bewertung konventioneller Zeit- und Frequenzbereichsansätze wird ein neuer MDCT-Algorithmus mit niedriger Parameterrate (nur ein Periodizitätswert pro Zeiteinheit) und Komplexität (ein Bruchteil der des Stands der Technik) vorgeschlagen. Dieser sowohl Intra- als auch Interkanalprädiktion unterstützende Spektralbereichs-Prädiktor (FDP) bietet Codiergewinne, die vergleichbar und orthogonal zu denen der MELT sind.

Abschließend werden vergleichende objektive und subjektive Auswertungen der vorgestellten Beiträge, nach Integration dieser in den MPEG-D USAC-basierten MPEG-H 3D Audio-Codec, dokumentiert. Objektive Messungen zeigen deutliche Ersparnisse in der Codierlatenz und Decoderkomplexität, während subjektive Blindtests nahelegen, dass der modifizierte MPEG-H-Codec sowohl bei generischen als auch Low-Delay-Anwendungen mit dem entsprechenden Stand der Technik qualitativ gleichauf liegt bzw. diesen übertrifft. Insbesondere zeigt sich, für sowohl Stereo- als auch 5.1-Surround-Kanalkonfigurationen, eine konsistentere Klangqualität über die unterschiedlichen Arten von Eingangssignalen, die weniger negative Ausreißer aufweist als der Stand der Technik.

x

Contents

1 Introduction

Despite ever-increasing network and storage capacities, “lossy” perceptual coding of digital audio signals, guided by the exploitation of psychoacoustic phenomena, remains ubiquitous. One reason is that the number of end users simultaneously communicating, or otherwise sending data, across current-generation public networks has increased by more than an order of magnitude over the user count in previous-generation networks such as (Enhanced) GPRS [ETSI12]. This implies that, to guarantee some level of quality of service (QoS) even when many users must share one network path, the transmission bandwidth, i. e., speed, must be reduced considerably, rendering “lossless” audio coding impractical. In the case of IP-based music streaming over the Internet, for example, low coding bit-rates are, thus, desirable to minimize the possibility of playback drop-outs.

A comparable situation occurs when network users are relocated to legacy network configurations like the abovementioned EGPRS, either because they move to e. g. a rural area where faster network connection is not available or because the quota for fast data transmission allocated, by contract, to them by their Internet service provider (ISP) has been exceeded. More specifically, it is common practice to “downgrade” users of mobile data contracts to transmission speeds conforming to the Enhanced Data-rates for GSM Evolution (EDGE) standard [ETSI12] once their monthly data quota has been exceeded. This roughly coincides with the infrastructure one can find in some rural areas even of developed countries, where the payload transmission rate (excl. all overhead) is limited to 58.4 kbit/s per timeslot in case of the “best” modulation and coding scheme, MCS-9.

Another reason for the persisting use of perceptual audio codecs (coders/decoders) is the trend toward an increased number of input, i. e. microphone or track, and output, i. e. loudspeaker, signals. In fact, up to 24 channels arranged in a “22.2” surround setup are currently under investigation for introduction into broadcasting markets especially in Asia. Naturally, such multichannel configurations imply stricter requirements on the per-channel bit-rates employed for coding — 22.2-surround material coded, on average, at 48 kbit/s per input waveform already leads to a total bit-rate of more than 1.1 Mbit/s. Paired with full-HD or UHD video, coded at up to 60 Mbit/s [ISO15c], this renders transmission over the Internet difficult even when current-generation network infrastructure such as LTE or the latest digital subscriber line (DSL) is available throughout the path.

The utilization of audio codecs in communication and broadcasting applications also brings about two other practical algorithmic considerations. For bidirectional real-time communication e. g. between two mobile devices or live on-site acquisition and wireless transmission of broadcasting material to remote studio facilities, end-to-end (encoding and decoding) delay, or latency, is a critical aspect. In both cases, the general consensus is that, for minimal perception, such latency must not exceed approximately 33 ms, i. e. two video images when recording at a frame rate of 59.94 or 60 Hz [ETSI16, ISO15c]. Note that algorithmic delay shall be defined as the latency caused by data dependencies of the coding/decoding algorithms, excluding delays due to the particular hardware or software implementation. In other words, infinitely fast signal processing is assumed.

The second issue is algorithmic complexity. Especially when used on mobile battery-powered equipment, a media codec should employ as few computational operations as possible for the encoding and, most importantly, the decoding process. A widely applied rule of thumb is that any new codec should not, at least in terms of decoding complexity for a given input/output signal configuration, substantially exceed the requirements of a comparable codec already established in the respective market. For audio in broadcasting applications, MPEG-4 High-Efficiency Advanced Audio Coding, abbreviated HE-AAC [ISO09], and Dolby Digital Plus, or E-AC-3 in short [ATSC12], arguably represent the two most commonly used coding standards, with the former offering better quality [EBU07]. The same objective can be formulated with regard to low-latency communication, since a low-delay variant of HE-AAC, termed AAC Enhanced Low Delay or AAC-ELD [Schn08], recently gained popularity in IP-based audio and/or videoconferencing applications.

Among these five restrictions resp. requirements — high quality, low complexity and delay, as well as high channel count and low per-channel bit-rate — certain concessions must often be made in order to reach a feasible implementation of a perceptual codec:

- Maximized reconstruction quality must, generally, be abandoned in favor of low algorithmic latency or complexity, especially when, as in mobile communication on resource limited devices, both constraints must be enforced simultaneously.
- An increase in the number of input/output signals usually causes a proportional increase in codec complexity, so tradeoffs between the actual channel count and the total complexity (and, as noted previously, coding quality) are usally made.
- Finally and most evidently, the perceived quality of a codec rises with increasing bit-rate. Determining an optimal average bit-rate for a specific use case, possibly in comparison to legacy codecs, thus represents an inevitable tradeoff in which high subjective quality for at least some rate-demanding, “critical” input material must be sacrificed to some extent. The issues of bit-rate selection, perceptual evaluation, and input signal criticality will be addressed throughout this work.

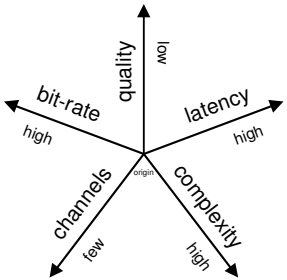


Figure 1.1. Illustration of the tradeoff between the five different requirements for audio coding.

To summarize the above, Figure 1.1 visualizes the tradeoff between quality, latency, complexity, channel count, and bit-rate as a five-dimensional space. Ideally, at least the decoder part of a codec, denoted by a point in that space, should reside near the origin.

1.1 Objective and Outline of this Thesis

The objective of this work is to develop a flexible audio coding framework which can be configured for both regular and low-delay applications as well as virtually arbitrary channel setups, and whose algorithmic decoder complexity, in the regular-latency case, shall not exceed that of HE-AAC. Regarding subjective coding quality, the goal is twofold:

- For regular-latency, i. e., unrestricted, use cases, its overall quality should exceed that of HE-AAC even when the latter uses the best performing encoder available.
- For low-latency, i. e., constrained, communication applications, its overall quality across several items should not be worse than that of HE-AAC and should exceed that of conventional dedicated low-delay codecs like AAC-ELD or Opus [IETF12].

In both cases, "good" perceptual quality after decoding, i. e., a reconstruction fidelity without obvious and possibly annoying coding artifacts, is desirable irrespective of the type of input material or the number of channels. Naturally, this key requirement is not only determined by the utilized coding algorithms and their signal-adaptive activation but also by the coding bit-rate for the specific channel configuration. In past subjective tests the author observed that, given some single-channel bit-rate b_1 providing a certain overall (averaged over many test items) monophonic quality level, a comparable quality level for a target channel configuration *cc* can be achieved using the bit-rate b_{cc} given by

$$b_{\mathrm{cc}} \;=\; b_1 \cdot \{cc \text{ as decimal number}\}^{0.75}, \tag{1.1}$$

where *cc* simply represents the literal expression of the channel count or multichannel speaker configuration as a decimal value, e. g., "5.1" for 6-channel surround including an LFE channel (for low-frequency effects/enhancement) and "2" or "2.0" for two-channel stereo. Table 1.1 enumerates a few monophonic b_1 and their perceptually equivalent b_{cc} counterparts for traditional stereo and 5.1 multichannel as well as 7.1+4 multichannel. The latter, for which playback equipment has been available since 2007 [Yama07] and which recently gained popularity, is a 12-channel surround setup including an LFE and four added height speakers at the front left/right and rear left/right "corners" [Theil11]. It is worth mentioning that 2.0 stereo forms a direct subset of the 7.1+4 configuration.

Several of the bit-rates provided in Tab. 1.1 are widely utilized values at integer multiples of 16 kbit/s (**bold** font), rendering evaluations both realistic and straightforward. Given the practically relevant rate of 58.4 kbit/s noted on page 1, a stereo coding rate of $b_2 = 48$ kbit/s and the qualitatively equivalent 5.1 surround rate will be focused upon.

The remainder of this work is organized as follows. Chapter 2 revisits the state of the art in modern transform-based audio coding by examining the design, implementation, performance as well as advantages and disadvantages of the most commonly employed individual algorithmic tools: overlapped time-frequency mapping by way of filter banks (section 2.1), transform-domain optimization of the spectrotemporal coding resolution as well as joint-stereo or multichannel coding (section 2.2), scalar spectral quantization with coefficient substitution and entropy coding, governed by a rate-distortion loop and psychoacoustic model (section 2.3), as well as parametric extensions for high-frequency regeneration (section 2.4) and downmix-based stereo or surround (section 2.5) at low rates. Whenever possible, a comparison to the new AC-4 codec [Kjör16] will be drawn. Section 2.6 ends the chapter with a brief review of the overall benefits and drawbacks.

Following the abovementioned objective, Chapter 3 then continues with an in-depth presentation and discussion of novel contributions to more flexible and efficient, unified audio transform coding. In doing so, all of the conventional tools examined in Chapter 2 are addressed and, in most cases, improved upon: time-frequency transformation with cosine and sine modulation as well as variable overlap ratio and low-delay block length switching (sections 3.1 and 3.2), frequency-domain prediction with much lower computational complexity than the state of the art (section 3.3), transform-domain intelligent spectral gap filling with complex-valued envelope calculation for semi-parametric high-frequency reconstruction (section 3.4), and semi-parametric enhancements of the joint-stereo and multichannel coding tools for lower bit-rates via Stereo Filling (section 3.5). Section 3.6 completes the chapter with an overview over some relevant entropy coding techniques which can be employed to compress the additional transform-domain side information (i. e., algorithmic parameters) required by the contributed tool proposals.

b_1	$\lfloor b_1 \rceil$	b_2	$\lfloor b_2 \rceil$	$b_{5.1}$	$\lfloor b_{5.1} \rceil$	$b_{7.1}$	$\lfloor b_{7.1} \rceil$	$b_{11.1}$	$\lfloor b_{11.1} \rceil$
18.90	19	31.79	**32**	64.14	**64**	82.21	82	114.94	115
23.64	24	39.76	40	80.23	**80**	102.82	103	143.76	**144**
28.35	28	47.68	**48**	96.21	**96**	123.31	123	172.40	172
37.80	38	63.57	**64**	128.28	**128**	164.41	164	229.87	230
47.28	47	79.52	**80**	160.46	**160**	205.65	206	287.52	**288**
56.70	57	95.36	95	192.42	**192**	246.62	247	344.81	345

Table 1.1. Bit-rates b_{cc} in kbit/s for equivalent channel number (column) and audio quality (row).

The contributions of Chapter 3 are implemented into USAC [ISO12], an extension of HE-AAC, and, subsequently, evaluated to assess their absolute and relative benefits. To this end, Chapter 4 introduces the basic principles behind the objective and subjective evaluation of perceptual audio codecs. Regarding the objective assessment, section 4.1 reports on derivations and measurements of both the algorithmic latency and decoding complexity of the basic and altered USAC system in comparison with the legacy MPEG-2 and MPEG-4 audio specifications. For subjective testing of the proposals, several formal experiments were conducted, whose underlying methodology, preparations, and executions are documented in section 4.2. The chapter concludes, for each experiment, with a respective analysis and discussion of the results in the context of the given use case.

It is worth noting in this context that, although several algorithmic methods for subjective audio quality assessment (i. e., implementations of models of human assessment and judgment of sound quality or degradation) have been developed during the last two decades, such objective approaches will not be utilized herein. The reason for this decision is the prevalent reliance on signal-to-noise ratio (SNR) or comparable measures of the degradation(s) of one (output) waveform relative to a reference (input) waveform, especially in the mono- and stereophonic case. On partially parametric low-rate codecs like those studied here, which do not preserve all aspects of the original waveform, such measures produce results which do not always agree with blind listening impression.

Chapter 5, finally, summarizes the goal of this work as well as the purpose, algorithmic construction, implementation, and performance of each individual tool contributed in pursuance of this goal. It further draws a conclusion on whether, and to what extent, the ultimate twofold objective of page 3 has been achieved. Section 5.1 then concludes the thesis with an outlook on future research and development, considered promising or incomplete by the present author, for the purpose of enabling the developed flexible perceptual transform coding system to perform a bit more consistently in certain cases.

Most of the work towards the contributions of Chapter 3 has been published before. Flexible input-adaptive time-frequency transformation has been addressed in [Helm14, Hel15c, Hel15d, Hel16a, Hel16b] and, to some extent, in [Helm10]. Frequency-domain prediction is discussed in [Hel16c], and major parts of the "semi-parametric" coefficient substitution principles have been reported in [Helm14, Hel15a, Hel15b, Schu16]. The complex-valued stereo prediction approach, from which the joint-channel coding proposals of this work were derived, has been described in [Helm11, Neue13]. Adoptions of the above techniques for efficient low-delay coding in communication scenarios have been presented in [Helm14, Fuch15, Hel15d]. These 13 references are cited repeatedly throughout this thesis, particularly in Chapters 3 and 4. However, an exhaustive citation, being cumbersome, has been avoided whenever deemed possible and acceptable. For all other references, an attempt at thorough and exhaustive citation has been undertaken.

2 Modern Perceptual Audio Transform Coding

The digital storage and transmission of audio signals has been a subject of exhaustive research and development for half a century. Starting with time-domain (TD) broadcast solutions directly operating on the pulse code modulated (PCM) digital audio waveform in the late 1960s [Rout69], using simple uniform or non-uniform (e. g., logarithmic) re-quantization of the PCM samples in order to attain a reduced bit-rate, the focus shifted to frequency-domain (FD) approaches, applying the fast Fourier transform (FFT), in the 1970s. A good overview over this early work on audio transform coding is provided in [JaNo84] for the period until 1983 and in [Gers94] for the following decade until 1993.

Since then, a fundamental set of algorithmic codec components has been established. The resulting architecture, which is illustrated as a block diagram in Figure 2.1, remains in use for high-bit-rate coding with equivalent $b_1 > 60$ kbit/s, i. e., rates higher than the ones listed in Table 1.1. For lower bit-rates such as those of Tab. 1.1, however, the basic architecture was slightly enhanced, as depicted in Figure 2.2 (with TD parameter signals omitted for clarity). In the following, each block of this enhanced design — starting with the outermost ones — will be examined in greater detail. It is worth emphasizing that Figs. 2.1 and 2.2 present only the *essential* components for transform coding of a *single* input channel. Naturally, several extensions for more efficient coding of certain types of audio signals have been proposed over the last years. Moreover, psychoacoustic models have been added to the encoder side in order to enable perceptually motivated coding, i. e., coding for minimally *audible* distortion [John88, Wies90]. Whenever suitable, these extensions will be introduced, classified, and described separately in respective sections along with descriptions of their locations within the signal path of the general scheme.

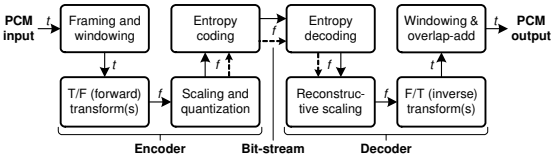


Figure 2.1. Basic architecture of a simple audio transform codec operating on a single channel. (—) audio and (---) parameter signals in (*t*) time and (*f*) frequency domain.

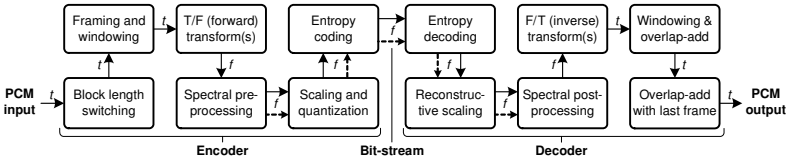


Figure 2.2. Extended single-channel codec framework with block switching and more FD tools.

2.1 Filter Banks for Input-Adaptive Time/Frequency Mapping

Having established the objective of FD coding for best access to, and exploitation of, the spectral properties of the input waveform, the first step in the coding process is the successive conversion of adjacent fixed-length portions of the PCM input signal, called frames, from the time to the frequency domain. Hereafter, let x_i represent the TD audio waveform segment of length N samples for the current frame identified by index i, with the individual samples at indices n running from 0 (inclusively) to N (exclusively), i. e.,

$$x_i(n),\ 0 \le n < N \ \rightarrow\ x_i = [x(iN), x(iN+1), \ldots, x(iN+N-1)]. \tag{2.1}$$

This framing is typically accompanied by a transform length detection algorithm, better known as a block switching detector, which analyzes the incoming instantaneous signal characteristics and, based on present and past frame statistics, selects the number and length of the individual time-frequency (T/F) transforms to be employed for the specific frame and channel [Edler89]. Such input-adaptive processing will be revisited at the end of this section. For now, use of a single fixed-length transform per frame is assumed.

In order to minimize framing artifacts after coding, manifesting themselves as either buzzing or crackling distortion caused by FD quantization induced discontinuities at the frame borders, the transform operation is typically conducted in an overlapped fashion. More specifically, a lapped transform across the current and next frame's audio segment,

$$\bar{x}_i(m),\ 0 \le m < 2N \ \rightarrow\ \bar{x}_i = [x_i, x_{i+1}] = [x(iN), x(iN+1), \ldots, x(iN+2N-1)], \tag{2.2}$$

having a total length of $2N$ samples, is employed in most cases. Naturally, utilizing a FFT for T/F conversion, yielding N unique complex-valued frequency coefficients ($N+1$ real, $N-1$ imaginary) at a transform hop-size of N, causes overcoding by a factor of two since, effectively, $2N$ spectral values are computed for each new length-N x_i. Reducing the FFT to a length between N and $2N$ decreases — but does not eliminate — the overcoding, and a value near N again leads to framing artifacts. This approach is, therefore, undesirable.

Princen and Bradley [Prin86, Prin87] were among the first to address the objective of lapped transform coding with critical sampling, i. e., the coding of N spectral coefficients associated with N time samples. In [Prin86], they presented a solution employing two evenly stacked real-valued transforms, which are applied alternately in adjacent frames. The first of these transforms, named "DCT based" in the cited paper, can be written as

$$X_i(k) = \sum_{m=0}^{2N-1} \hat{x}_i(m) \cos\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \quad 0 \le k \le N, \;\; k_0 = 0, \tag{2.3}$$

for the encoder-side analysis (i. e., forward) case, yielding the length-N spectrum X_i, and

$$\hat{y}_i(m) = \frac{2}{N} \sum_{k=0}^{N} X_i(k) \cos\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \;\; 0 \le m < 2N, \;\; k_0 = 0, \tag{2.4}$$

for the decoder-side synthesis (i. e., inverse) case, transforming X_i back into a TD signal. The second transform, called "DST based", makes use of a sine instead of a cosine kernel:

$$X_i(k) = \sum_{m=0}^{2N-1} \hat{x}_i(m) \sin\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \quad 0 \le k \le N, \;\; k_0 = 0, \tag{2.5}$$

for the analysis (forward) formulation and, respectively for the synthesis (inverse) case,

$$\hat{y}_i(m) = \frac{2}{N} \sum_{k=0}^{N} X_i(k) \sin\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \;\; 0 \le m < 2N, \;\; k_0 = 0. \tag{2.6}$$

In other words, the definitions (2.3)–(2.6) are based on either the discrete cosine transform (DCT) or the discrete sine transform (DST), using either cosine or sine modulation, respectively, in their base functions (hence the above-noted terminology in [Prin86]).

Alternating use of the DCT based transform (2.3), (2.4) and the DST based transform (2.5), (2.6) represents the construction and application of an evenly stacked filter bank, having its base functions located at even integer multiples of the basic angular frequency

$$\omega_\mathrm{b} = \frac{\pi}{2N}. \tag{2.7}$$

Put differently, the frequency offset k_0 in the above definitions takes on an integer value. The coefficients for the sub-band components $X_i(0)$ at direct current (DC, zero Hz) and $X_i(N)$ at the Nyquist frequency, therefore, need to be multiplied by ½ since they exhibit only half the spectral bandwidth of the other coefficients, as illustrated in Figure 2.3(a) (remember these are real-valued transforms similar to the DCT-II/III and DST-II/III).

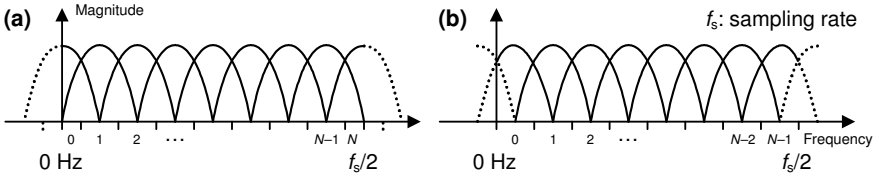


Figure 2.3. Sub-band design of an (a) evenly stacked, (b) oddly stacked real-valued filter bank.

The appropriate scaling of $X_i(0)$ and $X_i(N)$ is trivial: it can be performed either only on the analysis or the synthesis side via a factor of ½, as noted above, or equally on both the analysis and synthesis side using a common factor of $\sqrt{½}$. The evenly stacked filter bank design has been applied in the AC-2 codec developed by Dolby Laboratories as an error-resilient, very-low-complexity, single-channel predecessor (much like the scheme of Fig. 2.1) to the Dolby Digital (AC-3) codec until the early 1990s [Field89, Field96].

Although the alternating usage of the above evenly stacked lapped transforms, with proper scaling of their DC and Nyquist sub-band coefficients, is straightforward, it may be useful to further simplify the filter-bank structure to, e. g., allow cutting and joining of coded bit-streams and avoid having to keep track of each frame's index (odd or even). Moreover, the observant reader will have noticed from (2.3) – (2.6) and Fig. 2.3(a) that the evenly stacked filter bank per se actually comprises N+1 sub-band channels instead of the more intuitive N channels which, in principle, leads to over*sampling* by a factor of $(N+1)/N$ (over*coding*, however, can still be avoided since $X_i(0)$ and $X_i(N)$, being zero in every second frame, can be undersampled by a factor of two [Prin86]). To address these issues, Princen *et al.* [Prin87] modified their prior proposal to develop an oddly stacked N-channel system, in which only a single DCT-IV based transform type is required in all frames. In this design, all sub-bands exhibit the same bandwidth, as depicted in Figure 2.3(b), rendering a dedicated scaling of the DC and Nyquist coefficients unnecessary. Its definition is identical to that of (2.3), (2.4), except that k_0 now introduces an offset of ½:

$$X_i(k) = \sum_{m=0}^{2N-1} \hat{x}_i(m) \cos\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \quad 0 \le k < N, \;\; k_0 = \frac{1}{2}, \tag{2.8}$$

for the analysis (forward) formulation and, respectively for the synthesis (inverse) case,

$$\hat{y}_i(m) = \frac{2}{N} \sum_{k=0}^{N-1} X_i(k) \cos\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \;\; 0 \le m < 2N, \;\; k_0 = \frac{1}{2}. \tag{2.9}$$

As such, the base functions of (2.8), (2.9) are now located at odd integer multiples of the

fundamental frequency given by (2.7), hence the term "oddly stacked". In the following, (2.8) will be referred to as the modified discrete cosine transform (MDCT), and, accordingly, (2.9) will be called the inverse modified discrete cosine transform (IMDCT). The two are widely applied in perceptual audio coding, most prominently in all MPEG audio codecs since MPEG-1 Layer 3 (MP3) [ISO93], (E-)AC-3 [ATSC12], and Vorbis [Xiph15].

Note that there also exists a sine-modulated counterpart to the (I)MDCT, whose specification differs from (2.8), (2.9) only in the choice of the employed trigonometric term:

$$X_i(k) = \sum_{m=0}^{2N-1} \hat{x}_i(m) \sin\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \quad 0 \le k < N, \;\; k_0 = \tfrac{1}{2}, \tag{2.10}$$

for the analysis (forward) definition and, correspondingly for synthesis (inverse) cases,

$$\hat{y}_i(m) = \frac{2}{N} \sum_{k=0}^{N-1} X_i(k) \sin\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \;\; 0 \le m < 2N, \;\; k_0 = \tfrac{1}{2} \tag{2.11}$$

[Mal92b]. However, given that the above (I)MDCT suffices for lapped transform coding, this so-called (inverse) modified discrete sine transform (MDST) is almost never applied (only in E-AC-3, MDSTs are computed for enhanced joint-channel coding [Field04]).

Both the evenly and oddly stacked filter bank designs introduced above avoid overcoding, i. e., achieve critical sampling, by mapping $2N$ samples (in TD) into N coefficients (in FD) on the analysis side, and vice versa on the synthesis side, with a transform hopsize of N. More precisely, each sample $\bar{x}_i(m)$ of the size-$2N$ $\bar{x}_i$ of (2.2) is multiplied by a respective window sample $w_\mathrm{a}(m)$, obtained via a weighting function, and the resulting windowed sequence $\hat{x}_i$ is subjected to an analysis transform as specified in (2.3), (2.5), (2.8), or (2.10). After application of the corresponding inverse transform according to (2.4), (2.6), (2.9), or (2.11), the output values $\hat{y}_i(m)$, again with $0 \le m < 2N$, are scaled by synthesis window samples $w_\mathrm{s}(m)$, yielding an intermediate vector $\bar{y}_i$ to be combined with the previous frame's vector $\bar{y}_{i-1}$ by means of an overlap-and-add (OLA) procedure:

$$y_i(n) = \bar{y}_{i-1}(N + n) + \bar{y}_i(n) = \hat{y}_{i-1}(N + n) \cdot w_\mathrm{s}(N + n) + \hat{y}_i(n) \cdot w_\mathrm{s}(n). \tag{2.12}$$

Thereby, the initial length-N waveform portion associated with frame i can be obtained, as will be clarified on the next pages. Regarding the choice of w_a and w_s, it can be shown [Prin86, Prin87] that, irrespective of which of the above lapped transform types is used, perfect reconstruction (PR) of $x_i(n)$ in (2.1) is possible in the absence of quantization if

$$w_\mathrm{a}(2N - 1 - n) \cdot w_\mathrm{s}(n) + w_\mathrm{a}(N - 1 - n) \cdot w_\mathrm{s}(N + n) = \text{constant}. \tag{2.13}$$

Assuming identical and symmetrical analysis and synthesis windows, i. e., $w_a = w_s = w$ with $w(2N-1-n) = w(n)$, constraint (2.13) reduces to the well-known requirement

$$w^2(n) + w^2(N+n) = \text{constant, usually } w^2(n) + w^2(N+n) = 1, \tag{2.14}$$

also referred to as "Princen-Bradley" or power complementarity (PC) condition. Several such lapped-transform compliant windows have been published. The most common are

- the sine window originally presented by Edler [Edler89], defined by the function

$$w_{\text{sine}}(m) = \sin\left(\frac{\pi}{2N}\left(m+\frac{1}{2}\right)\right),\ 0 \le m < 2N. \tag{2.15}$$

  Its usage with (2.8) and/or (2.9) has been adopted in the MP3 standard [ISO93]. This combination is also called the modulated lapped transform (MLT) [Mal90b].

- Fielder's Kaiser-Bessel derived (KBD) window [Field89, Field96], specified as

$$w_{\text{kbd}}(n) = \sqrt{\frac{\sum_{j=0}^{n} v(j,\alpha)}{\sum_{j=0}^{N} v(j,\alpha)}},\ w_{\text{kbd}}(2N-1-n) = w_{\text{kbd}}(n),\ 0 \le n < N, \tag{2.16}$$

  in its simplest form [Span07], where the sequence $v(j,\alpha)$ denotes the symmetric Kaiser-Bessel kernel, configurable via a shape parameter α (typically, $4 \le \alpha \le 6$).

- the Vorbis window function used in the open-source Ogg-Vorbis codec, given by

$$w_{\text{vorbis}}(m) = \sin\left(\frac{\pi}{2}\cdot\sin^2\left(\frac{\pi}{2N}\left(m+\frac{1}{2}\right)\right)\right),\ 0 \le m < 2N, \tag{2.17}$$

  as in [Xiph15]. This window function is also utilized in the Opus codec [Valin13], where it determines the slopes of the symmetric low-overlap "flat-top" windows.

Figure 2.4(a) illustrates the temporal shapes of the three windows along with a fourth, sum-of-sines derived (SoSD) function developed by the present author [Helm10] based on work by Prabhu [Prab85]. It can be observed that w_{kbd} with $\alpha = 4$, as used in the AAC family [Bosi97, ISO97, ISO09, ISO12], tapers more quickly to zero at its boundaries than the other windows (i. e., it exhibits the most compact TD support of the four), whereas w_{sine} decays faster from unity gain at its center than the other three but almost linearly approaches a level of zero at its borders (i. e., offers the least compact TD support).

The temporal boundary properties of a window function influence the attenuation of the high-frequency side lobes — also known as far-field stop-band rejection — observed in that window's Fourier transform [Nutt81, Smith11] (assuming window values of zero

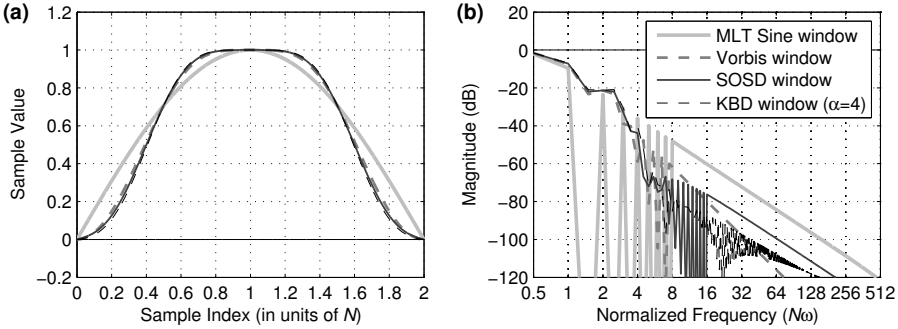


Figure 2.4. Properties of PC window functions: (a) temporal shapes, (b) Fourier power spectra.

outside the specified range, i. e., for $m < 0, m \geq 2N$). This is visualized in Figure 2.4(b) in a similar manner as in [Helm10, Fig. 5], where the near-field stop-band attenuation (frequency response of the first few side lobes) of w_{sine}, w_{vorbis}, and w_{sosd} is depicted.

It is evident from Fig. 2.4(b) that the aforementioned characteristics of the windows' shapes are also reflected in the windows' transfer functions. Assuming zero-padding, as noted earlier, the functions for w_{sine} and w_{sosd} are continuous at their borders, leading to a side-lobe decay rate of 12 dB per octave. The Vorbis window is continuous as well, not only in (2.17) but also in the derivative of this function. Hence, its side lobes fall off at 18 dB per octave. However, with its only moderately compact TD support, it neither reaches the level of near-field stop-band rejection (at $4 \lesssim N\omega \lesssim 11$) attained by w_{kbd} and w_{sosd}, nor is its main lobe width (i. e., pass-band selectivity) as narrow as that of w_{sine}. It is also worth noting that w_{kbd} of (2.16), although being similar in shape to w_{sosd}, is discontinuous at its borders and, thus, its far-end side lobes decay at only 6 dB per octave.

Returning to the transform definitions (2.3)–(2.11), two more aspects shall be noted:

- The common normalization factor of $\sqrt{2/N}$, which is needed to reach a constant gain of 1 in (2.13) and (2.14), is independent of frequency and can, therefore, be integrated into w to save a sample-wise multiplication for each transform. Moreover, if orthonormality of the forward and inverse transforms is not necessary (which is usually the case), a factor of $2/N$ may be applied instead on only either the analysis or synthesis side. The latter approach is, e. g., used in MPEG codecs.
- The mapping (undersampling) of $2N$ time to N frequency values introduces time domain aliasing (TDA) which, inevitably, remains in $\hat{y}_i$ and $\bar{y}_i$ after the synthesis transform. Assuming that (2.13) is enforced, this aliasing is canceled via (2.12) if

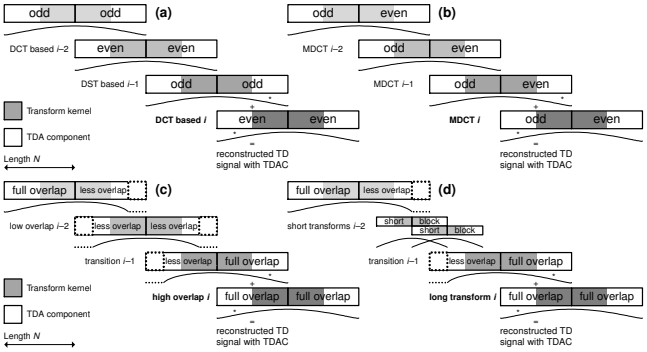


Figure 2.5. Illustration of inverse transforms, OLA, and window sequence. (a) invariant evenly stacked and (b) oddly stacked filter bank, (c) overlap switching, (d) block switching.

$$w_{\mathrm{a}}(N+n)\cdot w_{\mathrm{s}}(n) - w_{\mathrm{a}}(n)\cdot w_{\mathrm{s}}(N+n) = 0,\ 0 \le n < N, \tag{2.18}$$

[Smar94] or, for identical, symmetrical analysis/synthesis windows w, as above,

$$w(n)\cdot w(N-1-n) - w(N+n)\cdot w(2N-1-n) = 0,\ 0 \le n < N, \tag{2.19}$$

[Edler89], which is guaranteed with all four window designs introduced earlier and which, further, gives the TDA cancelation (TDAC) filter banks their name.

Figure 2.5 clarifies the operation of the TDAC process by way of a schematic illustration and, at the same time, summarizes the different aspects of the lapped T/F mapping in the construction of the discussed filter banks. The basic TDAC principle is identically applied in the evenly and oddly stacked filter bank realizations; only the symmetries of the specific TDA components — even or odd — vary, as shown in Figs. 2.5(a) and (b).

To complete this subject, Figures 2.5(c) and (d) visualize the modifications required for realization of the block switching design introduced at the beginning of this section. Assuming identical analysis and synthesis windows w in the OLA region of two adjacent transforms, Edler [Edler89] demonstrated, utilizing (2.14) and (2.19), that both window overlap adaptation (Fig. 2.5(c), also abbreviated overlap switching) as well as transform length adaptation (Fig. 2.5(d), also called block switching, requiring overlap switching), can be achieved, with PR, on a per-frame basis. Note that these two input-adaptive filter bank extensions can also be implemented using non-identical analysis/synthesis windows, or even asymmetric functions, by way of respective generalizations of (2.13) and (2.18) instead of (2.14) and (2.19) [Phili08, Viret08]. For the sake of brevity, however, such cases, which are intended for low-delay applications, will not be examined here.

2.2 Reduction of Spectrotemporal Redundancy and Irrelevance

The previous section introduced the filter bank components — framing, windowing, and T/F mapping using lapped transforms, with optional block switching — required to convert the TD input samples into FD coefficients (which is desirable since coding gain, i. e., energy compaction, can be achieved thereby). The following process, as depicted in Fig. 2.2, comprises several pre-processing steps, consecutively applied on the transform coefficients prior to their perceptually and rate-distortion (RD) motivated quantization in the encoder, with corresponding reconstructive post-processing procedures (carried out in reverse order) before the inverse transform(s) at the decoder side. The common objective of these pre-processing algorithms is the minimization of residual correlation between the coefficients of the transform instances, and/or psychoacoustic irrelevance contained within these coefficients, as a means to further increase the transform coding performance with regard to audio quality. The numerous pre/post-processing solutions developed in the course of codec standardizations can, generally, be classified into four categories based on the type of inter-coefficient dependency which they address:

- Correlation across frequency, i. e., autocorrelation within the same transform X_i, which indicates significant non-stationary *temporal* structure in the associated $\hat{x}_i$ (analogously to the fact that a highly autocorrelated TD signal exhibits a non-flat *spectral* structure) [Herr96]. This aspect is discussed in subsection 2.2.1.
- Correlation across time, i. e., multiple transforms or frames X_i, X_{i-1}, etc., which is a "long-term" TD counterpart of the preceding "short-term" FD correlation issue that is observed on strongly stationary input waveforms. Approaches reducing this type of inter-transform dependency are investigated in subsection 2.2.2.
- Correlation across space, i. e., two or more channels in a stereo or multichannel application. Recent research and development towards the minimization of this second type of inter-transform correlation, including "joint-channel" work which the present author contributed to [Helm11], is summarized in subsection 2.2.3.
- Combinations of the above aspects, including the previously unmentioned intra-transform "short-term" correlation across time, are outlined in subsection 2.2.4.

2.2.1 FD Redundancy/Irrelevance: Temporal Noise Shaping, Sub-Band Merging

A straightforward and well-known method for the decorrelation of digital waveform signals is the application of linear predictive coding (LPC) principles to the signal. More specifically, a linear filter, typically with finite impulse response (FIR), is run across the waveform samples prior to their (re)quantization, and corresponding inverse filtering, usually with infinite impulse response (IIR), is applied to the quantized samples during

decoding, i. e., signal reconstruction [JaNo84]. The real-valued coefficients of the LPC filter are derived from an estimate of the input's instantaneous autocorrelation function

$$\mathrm{ACF}(x) = \int x(\tau) \cdot x^*(\tau - t)\ d\tau \tag{2.20}$$

via Levinson-Durbin or a related recursion [Hayk14]. This is possible since there exists, by way of Fourier transformation (denoted by $\mathcal{F}$ below), a direct relationship between the TD autocorrelation function (ACF) and the signal's FD power spectral density (PSD):

$$\mathrm{PSD}(X) = \mathcal{F}\{\mathrm{ACF}(x)\} \ \leftrightarrow\ \mathrm{ACF}(x) = \mathcal{F}^{-1}\{\mathrm{PSD}(X)\}, \tag{2.21}$$

with the real-valued time signal *x*, as earlier, *X* specifying the associated complex-valued spectrum, and stationarity assumed [Herr96]. An analogous relationship can be formulated between the FD ACF for all positive frequencies of the single-sided spectrum $\tilde{X}$ of *x*,

$$\mathrm{ACF}\big(\tilde{X}\big) = \int \tilde{X}(\vartheta) \cdot \tilde{X}^*(\vartheta - f)\ d\vartheta, \tag{2.22}$$

and the TD equivalent of the FD PSD, namely, the square of the Hilbert envelope *H* of *x*:

$$\mathrm{ACF}\big(\tilde{X}\big) = \mathcal{F}\{H^2(x)\} \ \leftrightarrow\ H^2(x) = \mathcal{F}^{-1}\big\{\mathrm{ACF}\big(\tilde{X}\big)\big\}, \tag{2.23}$$

In other words, the signal's time envelope is directly connected to its autocorrelation in the spectral domain—its square H^2 is the inverse Fourier transform of its (single-sided) spectral ACF. From this observation it follows that power spectral density and squared Hilbert envelope are dual concepts and that predictive coding techniques similar to the ones utilized in static or adaptive differential PCM (DPCM) systems [JaNo84] may also be employed on the coefficients of a T/F transform like the MDCT [Her97a, Her97b].

Traditional linear predictive waveform coding allows for three approaches [JaNo84]:

- Closed-loop LPC, as used in the original DPCM approach shown in Figure 2.6(a), performs the linear prediction using the (re)quantized sample values in both the encoder and decoder. Thus, the prediction can be fully inverted in the absence of transmission errors, and the error due to the quantizer (block Q) is not shaped but — sufficiently fine quantization assumed — white or flat. In spectral-domain LPC, where the FD quantization can be modeled as an error spectrum *E* added to the LPC residual or difference spectrum *D*, this implies that the quantization distortion *E* is temporally white or flat, just as in cases without any prediction, and that the prediction gain directly translates into a proportional SNR increase.

- Open-loop DPCM, also abbreviated D*PCM, employs the same decoder design as its closed-loop variant but uses unquantized instead of quantized input samples in the encoder-side prediction. As such, the prediction values in the encoder and decoder differ even in the absence of transmission errors (the data loop through the quantizer is not closed), and E is subjected to potentially non-white/non-flat predictive synthesis filtering, as depicted in Figure 2.6(b). In FD LPC, this means that E represents an additive noise-like spectrum which is temporally shaped by the decoder-side filter according to the instantaneous short-time envelope H of the input waveform, as discussed previously. Such a realization of a spectral predictor is known as Temporal Noise Shaping or TNS [Herr96]. Its advantage over closed-loop FD LPC, beside a notably simpler encoder design (no error feedback around the quantizer), is that E, distributed across the transform window range, can be temporally shaped such that psychoacoustic pre- and post-masking can be exploited, i. e., that it follows the temporal masking threshold specified by the actual input signal if the coding bit-rate is sufficiently high. Note, however, that as in D*PCM, the shaping is attained at the cost of no effect of the prediction gain on the SNR, meaning no SNR advantage over predictionless coding [JaNo84].
- As indicated above and in Fig. 2.6(b), open-loop predictive coding subjects D and E to the same filter, which is suboptimal for a minimum mean squared error criterion [JaNo84]. By modifying the DPCM system such that X and E are processed by separate filters, as in Figure 2.6(c), an intermediate behavior between that of the closed-loop scheme (no shaping of E) and open-loop method (no SNR gain), exhibiting both noise shaping capability and prediction gain to some extent, can be achieved. This generalized noise feedback coding, having the former designs as special "boundary" cases, is similar to open-loop prediction with a bandwidth expanded LPC filter (having its poles/zeros moved towards the center of the z-plane unit circle) in terms of noise shaping but provides higher prediction gains.

Hence, the choice of LPC approach for the given application depends on the intended system behavior with regard to redundancy and irrelevance reduction: the closed-loop method enables the minimization of intra-transform redundancy, the open-loop design, here in the form of TNS, allows for the reduction of short-term perceptual irrelevance, and the generalized third scheme yields a tradeoff between the former two objectives.

Forward-adaptive TNS is used in all MPEG audio codecs since MPEG-2 AAC [ISO97]. Up to three non-overlapping TNS filters, whose up to 7–20 (depending on the standard) Parcor/lattice coefficients are conveyed as side-information, are allowed per transform [Her97a]. It is worth noting that, due to identical synthesis filter realization, the MPEG decoder specifications also offer the means for generalized FD noise feedback coding.

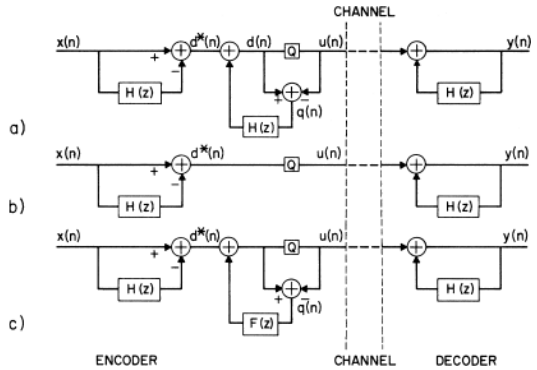


Figure 2.6. Different types of linear predictive coding: (a) closed-loop, no quant. error shaping, (b) open-loop, error shaping by prediction filter H(z), (c) generalization [JaNo84].

The combination of lapped transform coding, using the MDCT and/or MDST, and LPC in the transform domain, by way of input-adaptive TNS via convolution of the transform samples, can be interpreted as a continuously adaptive filter bank in terms of temporal and spectral sub-band resolution [Her97b]. In other words, a TNS enhanced filter bank can produce for a given time instant, via said FD convolution, more strongly overlapping (widened) sub-band samples with spectrotemporal resolutions between those of fixed long and short transforms according to the previous section. A comparable effect can be achieved by combining (merging) adjacent sub-band samples prior to quantization and by splitting them again (undoing the merge) before the synthesis filter bank [Mau95].

Sub-band merging, initially devised by Mau *et al.* for use in image coding applications [Mau95], denotes the procedure of replacing a specified tuple of spectrally neighboring transform coefficients by the same number of different weighted combinations of these coefficients. In [Mau95] and the Constrained Energy Lapped Transform (CELT) core of the Opus codec [IETF12, Valin13], Hadamard matrices, which are recursively defined as

$$\boldsymbol{H}_0 = 1,\ \ \boldsymbol{H}_m = \frac{1}{\sqrt{2}}\begin{bmatrix} \boldsymbol{H}_{m-1} & \boldsymbol{H}_{m-1} \\ \boldsymbol{H}_{m-1} & -\boldsymbol{H}_{m-1} \end{bmatrix},\ \ m > 0\ \ \rightarrow\ \ \boldsymbol{H}_m = \boldsymbol{H}_1 \otimes \boldsymbol{H}_{m-1},\ \ m > 1, \quad (2.24)$$

with merge level m, determine the weights. For a level-one sub-band combination, the spectral value pair $X_i(k)$, $X_i(k+1)$ of transform (or frame) i is mapped onto a new pair

$$\begin{aligned} \dot{X}_i(k) &= \tfrac{1}{\sqrt{2}}[X_i(k) + X_i(k+1)], \\ \dot{X}_i(k+1) &= \tfrac{1}{\sqrt{2}}[X_i(k) - X_i(k+1)], \end{aligned} \quad (2.25)$$

which, alongside the adjacent pairs $\dot{X}_i(k-2), \dot{X}_i(k-1), \dot{X}_i(k+2), \dot{X}_i(k+3)$, and so on,

is subjected to possible further spectral pre-processing and quantization in place of the associated values of X_i. The respective splitting operation inside the decoder is given by

$$\begin{aligned} X_i^q(k) &= \tfrac{1}{\sqrt{2}}[\dot{X}_i^q(k) + \dot{X}_i^q(k+1)], \\ X_i^q(k+1) &= \tfrac{1}{\sqrt{2}}[\dot{X}_i^q(k) - \dot{X}_i^q(k+1)], \end{aligned} \tag{2.26}$$

where superscript q indicates the spectral quantization, as in case of TNS. Merge/split matrices other than such constructed via (2.24) — including DCT- or DST-like (with $\boldsymbol{H}_1$ equaling a DCT-II matrix of the same size), MDCT- or MDST-like (with overlap between neighboring merge tuples), or even biorthogonal ones (where the encoder/analysis and decoder/synthesis matrices differ) — may also be used [Mau95, Niam03, Yoon06]. In addition, it is possible to employ matrix sizes other than 2^m as in (2.24). This property makes it clear that the process of sub-band combination actually represents the application of an intermediate transform on the coefficients of the filter-bank transform, i. e., a partial inverse transform synthesizing an intermediate spectrotemporal resolution.

The matrix size defines the increase in temporal resolution — and, thus, the decrease in spectral resolution and TD support — of the underlying filter bank sub-bands and can be selected on a per-frame or -transform basis depending on the instantaneous signal characteristics. Thereby, a nearly continuously input-adaptive filter bank design much like that incorporating TNS, with transmission of the matrix parameters to the decoder (and, possibly, separate configurations for different frequency regions), can be realized. The combination of such a design and the opposite approach, sub-band merging in time instead of frequency direction for adaptively increased short-transform spectral resolution, is available in the CELT codec under the name *T/F adjustment* [IETF12, Valin13].

Figure 2.7 illustrates the effect of level-two (4-tuple) sub-band merging applied to a 64-channel MLT. The increased and unequal temporal localization of the merged IMDCT outputs in Figs. 2.7(b) and 2.7(c) in comparison with the unprocessed IMDCT results of Fig. 2.7(a) for the same spectral range is evident. The benefit of sub-band merging in an audio codec, like that of FD LPC, can be exploited in two ways, or a combination thereof:

- Redundancy reduction can be achieved for relatively long transforms applied to non-stationary "transient" input. The spectral coefficients are densely populated and of similar magnitude in such cases [Her97a], and thanks to a partial inverse transform such as (2.25), a sparser FD representation (i. e., better energy compaction into a few coefficients) of the given frequency region can be reached.
- Irrelevance reduction can be obtained by quantizing each sample set of tuple $\dot{X}_i$ associated with a specific time location, as in Fig. 2.7, using a separate strategy guided by a psychoacoustic temporal masking model [Fastl07]. See also sec. 2.3.

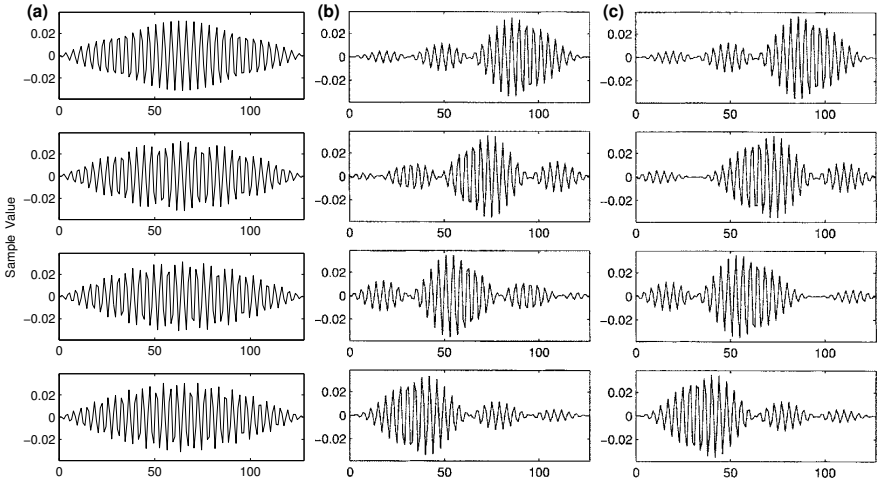


Figure 2.7. Effect of sub-band merging on the quantization error after inverse MLT for (top to bottom) $X_i(k)-X_i(k+3)$: TD output (a) without, (b, c) with different merging [Niam03].

2.2.2 TD Redundancy/Irrelevance: Temporal Prediction, Sub-Band Splitting

The objective of the previously described FD coding tools is the exploitation of inter-coefficient dependencies (across frequency) within the same transform X_i as a means to improve the audio quality for non-stationary "transient" input. The opposite effect can be achieved by addressing inter-transform dependencies (across time) between same-frequency (i. e., same-k) samples of X_i and its predecessors X_{i-1}, X_{i-2}, etc. It was already noted that CELT's T/F adjustment comprises both the sub-band merging scheme and its opposite, namely, sub-band *splitting* for increased frequency resolution [Field04]. For a level-one split of $X(k)$ across the current and last transform (or frame), this results in

$$\begin{aligned} \ddot{X}_{i-1}(k) &= \tfrac{1}{\sqrt{2}}[X_{i-1}(k) + X_i(k)], \\ \ddot{X}_i(k) &= \tfrac{1}{\sqrt{2}}[X_{i-1}(k) - X_i(k)], \end{aligned} \tag{2.27}$$

when applying a Hadamard or DCT-II matrix. In other words, for each k, the temporally separate X_{i-1} and X_i are mapped to the spectrally (almost) separated $\ddot{X}_{i-1}$ and $\ddot{X}_i$, overlapping substantially in time, for further (often joint) pre-processing and quantization. Accordingly, the corresponding inverse (recombinatory) operation at the decoder side is

$$\begin{aligned} X^q_{i-1}(k) &= \tfrac{1}{\sqrt{2}}[\ddot{X}^q_{i-1}(k) + \ddot{X}^q_i(k)], \\ X^q_i(k) &= \tfrac{1}{\sqrt{2}}[\ddot{X}^q_{i-1}(k) - \ddot{X}^q_i(k)], \end{aligned} \tag{2.28}$$

similarly to (2.26). Analogously to the dual concept of sub-band merging and splitting, it is possible to also apply the FD LPC principle in time instead of frequency direction.

TNS, as described above, represents an open-loop realization of a FD linear sub-band predictor along a region of X_i. Specifically, upon decoding/synthesis, a prediction value

$$X_i'(k) = \sum_{f=1}^{F} tns_i^q(f) \cdot X_i^q(k-f), \;\; F > 0, \;\; X_i^q(k-f<0) = 0, \tag{2.29}$$

summed up according to the TNS prediction order F using the filter coefficients tns_i^q, is added to each quantized residual $D_i^q(k)$ of said region to reconstruct $X_i^q(k)$. "Turning" the same "short-term" predictor such that time instead of frequency neighbors are used,

$$X_i^-(k) = \sum_{t=1}^{T} fdp_i^q(t) \cdot X_{i-t}^q(k), \;\; T > 0, \;\; X_{i-t<0}^q(k) = 0, \tag{2.30}$$

with respective filter weights fdp_i^q, yields an order-T temporal predictor which, like FD LPC, can reduce both the redundancy and irrelevance, depending on the encoder design:

- Redundancy reduction, i. e., coding gain, can be reached via a closed-loop design in which the same reconstructed MDCT values X_{i-1}^q, X_{i-2}^q, etc., as in the decoder are utilized in the subtraction yielding D_i. The TD equivalent of this scheme is a long-term predictor (LTP) applied directly to the input waveform x_i (or $\bar{x}_i$). LTPs for transform coding have been developed by Nokia [YinS97, Ojan99] for MPEG-4 AAC [ISO09], Ramprashad [Ramp03] at Bell Laboratories, Valin *et al.* [Valin10] for an early variant of CELT, and Song *et al.* [Song10, Song11] during MPEG work.
- Irrelevance reduction by way of fine spectral noise shaping can be achieved via open-loop prediction, which adopts unquantized past input values X_{i-1}, X_{i-2}, … instead of quantized ones to obtain the prediction values X_i^- of (2.30). A similar comb-filter-like effect can be attained using TD pre-/post-filtering, an invertible extension of Chen's pitch post-filter [Chen95] used e.g. in Opus [IETF12, Valin13].

Closed-loop FD prediction was first introduced by Mahieux *et al.* in the context of an oversampling FFT-based audio codec (employing complex spectral coefficients) in the late 1980s [Mahi89] and later adapted to real-valued MDCT-based coding by Fuchs as a contribution to the standardization work of the MPEG audio subgroup [Fuch93, Fuch95, Bosi97]. A backward-adaptive intra-channel lattice implementation of Fuchs' predictor, where the fdp_i^q are not transmitted in the bit-stream but are synchronously derived in the encoder and decoder from past decoded spectral data (only a band-wise activation indicator is added as side-information), has been integrated into MPEG-2 AAC [ISO97].

Concluding this subsection, it is noted that the differential codec by Paraskevas and Mourjopoulos [Paras95] can be regarded as a simplified LTP-like adaptive FD predictor with $D_i \stackrel{\text{def}}{=} 0$, which can be controlled on a per-frame and per-sub-band basis. Moreover, the author of this thesis is not aware of any realization of an open-loop FD predictor.

2.2.3 Spatial Correlation: Mid/Side, Intensity/Error, Predictive Joint-Stereo

When coding recordings of isolated direct sound sources (as opposed to immersive diffuse sources) with multiple channels, a considerable amount of correlation typically remains between the individual transforms of a given frame, even when some reverb or ambience can be heard in the recordings. Johnston [John89] was among the first to discover that cross-channel linear transformation similar to that for sub-band merging and splitting can yield quality gains in two-channel perceptual coding of near-monophonic signals with a small inter-channel level difference (ILD). In the following, it is assumed that two transform spectral vectors L_i and R_i of equal length N, possibly pre-processed by one or more of the tools described in the previous two subsections, are available.

On near-monophonic (i. e., correlated) center-panned (i. e., low-ILD) stereo signals, a transform coder may quantize a sum (mid) and a difference (side) value defined, e. g., as

$$\begin{aligned} M_i(k) &= \tfrac{1}{\sqrt{2}}[L_i(k) + R_i(k)], \\ S_i(k) &= \tfrac{1}{\sqrt{2}}[L_i(k) - R_i(k)], \end{aligned} \tag{2.31}$$

instead of $L_i(k)$ and $R_i(k)$ whenever, for the given frame at index i, the psychoacoustic model indicates a perceptual benefit. Obviously, in the decoder, this process is inverted,

$$\begin{aligned} L_i^q(k) &= \tfrac{1}{\sqrt{2}}[M_i^q(k) + S_i^q(k)], \\ R_i^q(k) &= \tfrac{1}{\sqrt{2}}[M_i^q(k) - S_i^q(k)], \end{aligned} \tag{2.32}$$

to obtain the initial channel spectra [ISO93]. The choice between $L_i(k)$, $R_i(k)$ or $M_i(k)$, $S_i(k)$ can be made globally for all coded k of the channel pair or separately for each contiguous subset of k constituting a parameter band. Combinations other than the above Hadamard/DCT-II based ones are also feasible. An asymmetric variant of (2.31), (2.32), where, for easy implementation, the common scalar $\sqrt{½}$ is replaced by ½ in the encoder and by 1 in the decoder [John92], has been adopted in Vorbis [Xiph15], Opus [IETF12], (E-)AC-3 [ATSC12], AC-4 [ETSI14, Kjör16], as well as all MPEG codecs since AAC [ISO97, ISO09, ISO12]. A rotation-based transform, yielding so-called *intensity* and *error* values

$$\begin{aligned} I_i(k) &= \quad \cos\alpha_i^q \cdot L_i(k) + \sin\alpha_i^q \cdot R_i(k), \\ E_i(k) &= -\sin\alpha_i^q \cdot L_i(k) + \cos\alpha_i^q \cdot R_i(k), \end{aligned} \tag{2.33}$$

with per-frame or -band angle $\alpha_i \in [-\frac{\pi}{2}, \frac{\pi}{2})$, has been proposed in [Vand91] as a general form of the mid/side (M/S) paradigm of (2.31) which also works well on out-of-phase (negatively correlated) or panned (non-zero ILD) channel pairs. Notice that (2.33), with

$$\begin{aligned} L_i^q(k) &= \cos\alpha_i^q \cdot I_i^q(k) - \sin\alpha_i^q \cdot E_i^q(k), \\ R_i^q(k) &= \sin\alpha_i^q \cdot I_i^q(k) + \cos\alpha_i^q \cdot E_i^q(k) \end{aligned} \tag{2.34}$$

as corresponding inverse rotation in the decoder, is a Karhunen-Loève transform (KLT) of length two that, via $\alpha_i = \frac{\pi}{4}$ or $\alpha_i = 0$, becomes equivalent to the M/S matrix of (2.31) or a left/right (L/R) "bypassing" identity matrix, respectively. Furthermore, due to good signal power concentration into I_i (i. e., optimal decorrelation in a mean-squares sense, regardless of the instantaneous ILD and, thus, the value of α_i), "intensity stereo" coding with $E_i(k) \stackrel{\text{def}}{=} 0$ at high frequencies, to save coding bit-rate, can be realized [Vand91].

Naturally, increasing the KLT length allows for combined "joint" coding of more than two channel spectra, which is especially useful in low-rate surround sound applications [Yang00, Yang06]. However, as the Hadamard, or any trigonometric, transform can also be increased in size, the multichannel pre-/post-processors do not need to be limited to KLT-based designs. In fact, a three-channel M/S approach has recently been presented [ShiR14], and AC-4's *Stereo Advanced* (or *Audio*) *Processing* tool for MDCT-domain joint coding of up to five channels [ETSI14, Kjör16] builds upon the M/S principle as well.

The M/S and KLT rotary transformations, like those for sub-band merges and splits, are advantageous in both an objective and a subjective way when applied appropriately:

- Objectively, as indicated above, energy compaction into fewer output than input channels, i. e., reduced redundancy leading to coding gain, can be achieved. This is particularly true for the strongly correlated channel transforms representing a two- or three-dimensionally vector panned signal portion [ShiR14]. It must be emphasized, though, that for M/S-based coding, i. e., a rotation by $\alpha_i = \frac{\pi}{4}$ (or $-\frac{\pi}{4}$), maximum compaction into M_i (or S_i) can only be attained in case of zero ILD.
- Subjectively, the joint-channel matrix operations around the spectral quantizers affect the statistical properties of the quantization error, modeled as an additive transform-wise noise spectrum, in a perceptually beneficial manner. Specifically, proper adaptive selection of the matrix renders the spatial direction (angle) and width (correlation) of the combined quantization noise in the decoded channel spectra identical to those of the input signal itself. Hence, spatial noise shaping toward the dominant sound source in the stereophonic image can be performed [Kjör16], which minimizes binaural unmasking of the coding distortion [John92] due to, e. g., binaural masking level difference effects [Blau96, Moor12, Bran13].

- Irrelevance reduction can additionally be obtained via intensity stereo coding at frequencies above approximately 5 kHz where, due to the missing phase-locking capabilities of the human auditory system [Moor12], only the spectrotemporal magnitude (but not phase) envelope at each ear is psychoacoustically relevant.

In Opus/CELT and [VanS08], M/S stereo matrixing is conducted after spectral energy normalization, i. e., after each input spectrum has been divided by its parameter-band-wise L_2 norm given by the square root of the band energy [IETF12, Valin13]. In doing so, any ILD between the spectral pair is compensated for prior to the stereo pre-processor, and spatial decorrelation approaching that provided by the KLT of (2.33) and (2.34) can be realized for any arbitrarily panned signal (with the ILD being a substitute for α_i).

A comparable approach is the prediction of the smaller of the two M/S outputs, computed from the non-normalized input transforms, using the larger of the two vectors:

$$\begin{aligned} \acute{S}_i(k) &= S_i(k) - \rho_i^q \cdot M_i(k), \quad \|S_i\| \le \|M_i\|, \\ \acute{M}_i(k) &= M_i(k) - \rho_i^q \cdot S_i(k), \quad \|S_i\| > \|M_i\|, \end{aligned} \tag{2.35}$$

with $\|\cdot\|$ denoting the L_2 norm and $\rho_i \in [-1, 1]$ being the prediction coefficient, in order to further decorrelate the mid and side spectra. This technique is employed in the AC-4 codec [ETSI14] under the name *Enhanced M/S coding* [Kjör16] and the MPEG-D Unified Speech and Audio Coding (USAC) standard [ISO12] as a "real prediction" subset of the complex-valued stereo prediction tool [Helm11, Neue13], illustrated in Figure 2.8.

Complex-prediction stereo, in short, is an extension of (2.31) with (2.35) allowing to compensate for inter-channel phase difference (IPD), in addition to the previously noted ILD, to maximize the joint-channel coding efficiency [Helm11]. The necessary complex representation of the MDCT downmix (see also section 2.7) is directly derived from past and current values of M_i (or S_i, if that has more energy) via Cheng's method [Chen04].

The observant reader might be tempted to conclude that, as in subsections 2.2.1 and 2.2.2, it is also feasible to apply forms of "direct" prediction similarly to (2.29) or (2.30) instead of coefficient remapping analogously to (2.25) or (2.27). However, such designs, in which one channel's value $B_i(k)$ is converted into a prediction residual by subtracting

$$B_i'(k) = \beta_i^q \cdot A_i(k) \quad \text{or} \quad B_i'(k) = \beta_i^q \cdot A_i^q(k), \tag{2.36}$$

computed from another channel's value $A_i(k)$ or $A_i^q(k)$ and a channel prediction weight β_i, are rarely used in lossy audio coding. In fact, only implementations based on Fuchs' work [Fuch93, Fuch95] and some lossless coders [Lieb02] seem to incorporate this type of explicit inter-channel predictor as a special (trivial) case. Moreover, notable evidence against cross-channel prediction at high frequencies has been presented [KuoJ01].

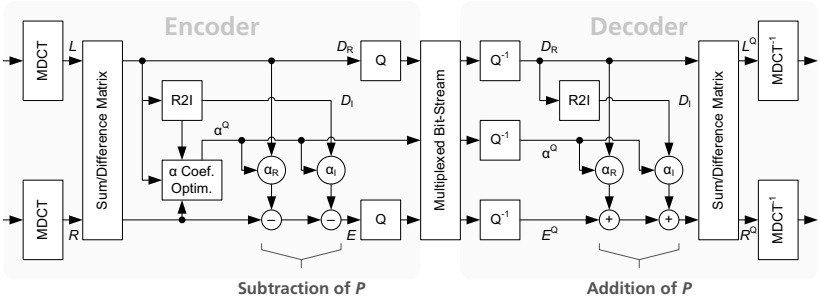


Figure 2.8. Complex predictive stereo coding and decoding in USAC. α: complex predictor (*P*) coefficient, R2I: real-to-imaginary, *D*: downmix (*M*), *E*: residual (*Ś*), Q: quantization.

2.2.4 Combinations of the Above: Intra-/Inter-channel, T/F Prediction, FDNS

To complete this section, some further approaches, which combine some of the pre-/post-processing methods described in the preceding three subsections, are introduced.

- Joint intra-/inter-channel prediction, proposed by Fuchs [Fuch93, Fuch95] in the early 1990s, combines the FD temporal and cross-channel predictors presented individually in subsections 2.2.2 and 2.2.3, respectively, into a single algorithm.
- Opus, as noted, permits simultaneous (but non-overlapped) sub-band merging and splitting in short-transform frames using its *T/F adjustment* tool [Valin13].
- In AC-4 [ETSI14], "efficiently tabulated periodic signal model based" prediction [Kjör16] is used in the *Speech Spectral Front-end* tool. Closer inspection reveals that this *T/F predictor*, as it shall be called herein, applies a two-dimensional FD filter in the coding loop (around the quantizer) that may also be regarded as the unification of a TNS-like frequency-direction and LTP-like time-direction filter.
- Frequency-domain noise shaping (FDNS) is supported in the MDCT-based transform coded excitation (TCX) path of MPEG-D USAC [ISO12]. This type of spectral processing addresses the objective of short-term irrelevance reduction by multiplicative application of an LPC filter envelope [Mori96], computed on the frame's waveform input (thereby avoiding TD filtering of the signal), in conjunction with first-order TNS [Neue13]. Short-term irrelevance will be revisited in section 2.3.

It is noteworthy that an alternative intra-/inter-channel predictor can be constructed by extending the TNS filter to a cross-channel design. However, no explicit realization of such two-dimensional filtering (along frequency and space) is known to the author; the closest approximation, or simplification, is the application of the TNS tool on the M/S (downmix/residual) instead of the L_i and R_i spectra, which is allowed in USAC [ISO12].

2.3 Scaling, Quantization, Substitution, and Entropy Coding

The pre-processing steps described in the last section prepare the individual channel spectra associated with the given frame i for quantization. To reach the desired overall and/or instantaneous bit-rate for the signal configuration (number of channels, sample rate, and audio bandwidth) at hand, the quantization process applied on each transform vector $\bar{X}_i$ (with the bar denoting pre-processing) is generally divided into the following steps, which are often carried out iteratively in a rate-distortion (RD) loop [Bosi97]:

- grouping, then scaling by means of one or multiple gains for coding SNR control
- the actual quantization process in a uniform, i.e., "linear", or non-uniform fashion
- parametric substitution of spectral coefficients (or regions) quantized to zero.

For the sake of brevity, only recently standardized codecs utilizing forward-adaptive scalar quantization (SQ), where the scaling parameters are sent to the decoder, shall be investigated. Readers interested in older codecs, or such implementing other methods like block floating point and vector quantization (VQ), are referred to [Bran97, Span07] or [Field04, ATSC12] (E-AC-3), [Mäki05, Sala06] (AMR-WB+), [Vaill08, Jelín09] (G.718), [IETF12, Valin10, Valin13, Xiph15] (Opus, Vorbis), and [Mori96] (TwinVQ), respectively.

2.3.1 Grouping, Scaling Using Global or Local Gain Factors

The main objective of spectral quantization in audio transform coding is a significant reduction of the bit-rate needed to represent the waveform input. By multiplying every $\bar{X}_i$ with the inverse of a specific global gain factor g_i, or a quantized version g_i^q thereof, the spectral entropy after quantization — and, thus, the instantaneous bit consumption required to convey the encoded version of $\bar{X}_i$ — can be precisely controlled. In order to reconstruct the initial amplitudes before post-processing in the decoder, each transmitted quantized $\bar{X}_i^q$ is multiplied with its associated g_i^q, also included in the bit-stream.

Given that the quantization introduces distortion whose audibility should be limited to a minimum, it is desirable to spectrally shape this distortion on a frame-by-frame or even transform-by-transform basis according to the instantaneous frequency envelope of the signal and the corresponding simultaneous masking characteristics of the human auditory system [John88, Wies90]. When employing LPC-based multiplicative FDNS (or the equivalent TD short-term predictive filtering, as noted in the previous section), the spectral envelope of the transform input is accounted for, and no extra weighting other than by a global gain and, optionally, some low-frequency (de-)emphasis is necessary to accomplish subjectively adequate quantization noise shaping [Fuch15, ETSI16, ISO12].

When FDNS or TD LPC filtering is not utilized, the global gain design must be extended to a frequency-dependent band-wise scheme in order to enable spectral noise shaping. To this end, all $\bar{X}_i$ are partitioned into B disjoint scaling bands indexed by b, each comprising $c_i(b)$ consecutive transform samples, and a separate *local* gain $l_i(b)$ is applied to each band. The width sequence c_i is typically isotone, increasing with center frequency,

$$0 < c_i(0) \leq \cdots \leq c_i(b-1) \leq c_i(b) \leq c_i(b+1) \leq \cdots \leq c_i(B-1) \leq c_{\max}, \qquad (2.37)$$

to reflect the "critical" auditory filter bandwidths of the human ear [Fastl07, Moor12].

In E-AC-3 [ATSC12], AC-4 [ETSI14], and all MPEG audio coders since MP3 [ISO93], the bandwidths $c_i(b)$ are inspired by the equivalent rectangular bandwidth (ERB) model of human hearing [Moor12], and each $l_i(b)$ represents the quantization scale factor for b, which is quantized logarithmically for transmission (see also the next subsection). The chosen values of $l_i(b)$ determine the SNR in each band after quantization and are, thus, governed signal-adaptively by the psychoacoustic model and the (mean) target bit-rate. In the MPEG AAC family [ISO97, ISO09, ISO12], each gain band, called *scale factor band* (SFB), is additionally limited in width (its $c_i(b)$ does not exceed a certain sampling rate dependent maximum $c_{\max}$) due to several reasons [Bosi97], as exemplified in figure 2.9. Most notably, b, B, and c_i are shared with the joint-stereo tools, that benefit from a uniform narrow bandwidth sequence on signals with, e. g., inter-channel time delay. Opus also makes use of ERB-like logarithmically quantized gains in its CELT codec, but these gains convey the band variances rather than the step sizes for quantization. Using said variances, band-wise power normalization is performed at an early stage in the encoder (see also page 24), and fine perceptual noise shaping is obtained via more or less static but frequency dependent weighting of each band prior to VQ [IETF12, Valin10, Valin13]. CELT's energy bands, unlike the SFBs in AAC, are not width-limited at high frequencies.

In frames and channels with activated block switching (section 2.1) and/or sub-band merging (section 2.2), for which multiple short-size transforms are to be quantized, the scaling can be applied in three different ways. First, all transform vectors could be multiplied (or divided) independently using individual global gains g_i or, if applicable, scale factors l_i. This scheme is easy to implement and allows for fine temporal distribution of the quantization error according to the psychoacoustic masking model (i. e., irrelevance reduction, see also page 19). Second, and oppositely, all vectors could share a common g_i or set of l_i. This approach is useful at very low coding rates because it reduces the bit consumption of the quantizer parameters in the bit-stream, but it renders the temporal distribution of the quantization distortion more difficult or even impossible. Therefore, a third method, known as block grouping [Bosi97] and also depicted in Fig. 2.9, is used in AAC to compromise between temporal resolution and bit-rate of the scaling data.

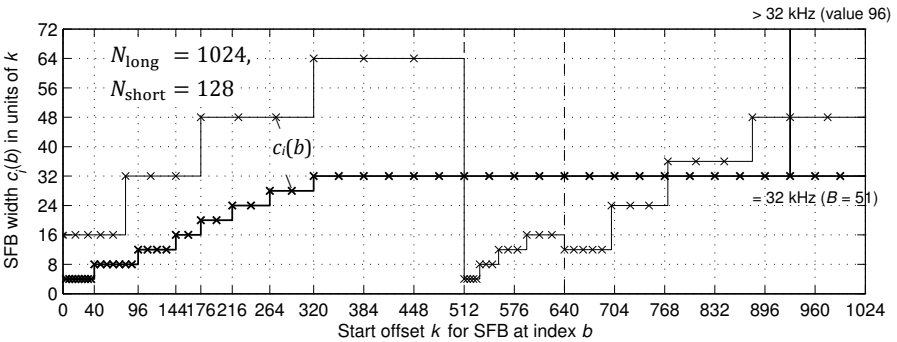


Figure 2.9. SFB configuration in AAC for sample rates between 32 and 48 kHz (inclusive) and (–) 1 long transform, (–) 8 short transforms with exemplary 4-1-3 grouping [ISO97].

2.3.2 Uniform or Non-Uniform Scalar Quantization

Scalar quantization (SQ) rounds the "continuous" amplitude of each spectral sample after scaling (or power normalization) to one of a limited set of discrete values [JaNo84, Span07], thereby introducing the abovementioned quantization distortion manifesting itself as an additive noise-like error signal E. The process and effect of SQ is well studied and documented, so only aspects relevant to recent transform coders are discussed.

Historically, both mid-rise and mid-tread quantizer designs have been used [JaNo84]. The former, however, do not provide a reconstruction value, or *level*, of $\bar{X}_i^q = 0$, which is considerably suboptimal in transform coding, where spectral coefficient values around zero are much more likely to occur after scaling than other coefficient magnitudes. The scaled frequency data can, in fact, be modeled as a random variable with a probability density function (PDF) that is symmetric around zero and smoothly decaying to its outskirts. For such input, a mid-rise quantizer always produces a mean output entropy of at least one bit per sample, which is too large for low-bit-rate coding applications. Only mid-tread quantization can, therefore, be found in modern audio codecs. Aside from the above binary classification, the SQ process can be performed in three different ways:

- "Linear", uniform quantization, the most common type of input-output mapping for a reduced-entropy representation, assigns each input value $\hat{X}_i(k) = \bar{X}_i(k)/g_i$ (or g_i^q, $l_i(b)$, $l_i^q(b)$ as applicable) to one of the equidistant output indices $q_i(k)$:

$$q_i(k) = Q\left(|\bar{X}_i(k)|/g_i\right) \cdot \text{sgn}\left(\bar{X}_i(k)\right),\ X_i^q(k) = q_i(k) \cdot g_i. \tag{2.38}$$

  In other words, the step size between adjacent q_i is constant, but the SNR is not.

- Logarithmic quantization, a form of "non-linear" non-uniform mapping utilized, e. g., in a-Law and μ-Law compression [JaNo84], is achieved by subjecting $\bar{X}_i(k)$ to a logarithmic function and by quantizing the result uniformly as in (2.38):

$$q_i(k) = Q\big(\log(|\bar{X}_i(k)|)/\log(g_i)\big) \cdot \mathrm{sgn}\big(\bar{X}_i(k)\big),\ X_i^q(k) = g_i^{|q_i(k)|} \cdot \mathrm{sgn}(q_i(k)),\quad (2.39)$$

  assuming that the output of $Q(\cdot)$ is shifted, or linearized, to be positive [JaNo84]. Hence, the SNR is constant (i. e., independent of the input variance), but the step sizes between adjacent q_i are not (i.e., they grow exponentially with magnitude). Note that g_i is applied in the logarithmic domain; it defines the logarithmic base.

- Power-law quantization can be employed to achieve intermediate non-uniform behavior between that of a linear and a logarithmic quantizer. In this approach, introduced in MP3 [ISO93], $\bar{X}_i(k)$ is exponentiated before and after the mapping:

$$q_i(k) = Q\big((|\bar{X}_i(k)|/g_i)^{1/\gamma}\big) \cdot \mathrm{sgn}\big(\bar{X}_i(k)\big),\ X_i^q(k) = |q_i(k)|^{\gamma} \cdot \mathrm{sgn}(q_i(k)) \cdot g_i,\quad (2.40)$$

  with $\gamma \geq 1$. A value of $\gamma = \frac{4}{3}$ has been adopted in MPEG audio coding, and $\gamma = 1$ reduces (2.40) to the uniform quantizer of (2.38). Using $\gamma > 1$, both the SNR and the step sizes vary; they rise with input variance and magnitude, respectively.

In all 3 cases, $Q(\cdot) = \lfloor \cdot + \delta \rfloor$, where $\delta \geq 0$ defines the deadzone width [Bosi97, Fuch15]. Due to the input compression before and output expansion after the execution of $Q(\cdot)$, the logarithmic and power-law quantizers are also known as *companding* quantizers. In addition, the term *re*quantizer is often used to stress that the PCM input to the codec is a discrete signal whose samples have already been quantized during A/D conversion.

Figure 2.10 compares the effect of the above methods on the maximum magnitude of the distortion E for an input signal s decaying exponentially in magnitude. The SNRs of the different quantizers are chosen arbitrarily for this example. Considering s a transform spectrum or a portion thereof (i. e., SFB), which rolls off towards high frequencies, it can be observed that, as indicated earlier, uniform quantization leads to spectrally flat "white" E, whereas logarithmic quantization causes E to adopt the shape of s itself. The power-law schemes provide a tradeoff between the former two, with the special case of square-root quantization producing exactly a half-way "half-shaped" error spectrum.

The motivation behind the use of a $\gamma = \frac{4}{3}$ non-uniform quantizer in MP3 and all AAC variants is subtle *intra-band* spectral noise shaping. The example of Fig. 2.10 is chosen deliberately since a magnitude response tapering off at higher frequencies represents a typical spectral shape in audio coding. The coarser *inter-band* noise shaping is attained by the partitioning of $\bar{X}_i$ into SFBs and the use of SFB-wise scale factors for SNR control. Strong shaping via square-root or logarithmic companding is, therefore, not necessary.

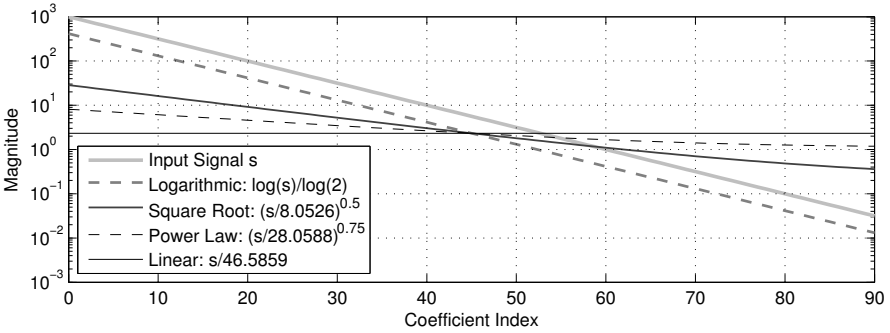


Figure 2.10. Quantization noise shaping: signal and maximum error magnitude for $Q(\cdot)=[\cdot+½]$.

2.3.3 Substitution of Zero-Quantized Spectral Coefficients

At very low coding rates, SQ according to the previous subsection generally leads to most $q_i(k)$ being zero, which implies that large parts of $\bar{X}_i^q$ and, thus, X_i^q are also zeroed out. This property leads to energy loss especially in high-frequency (HF) regions of the reconstructed channel spectra — known as *spectral holes* or *gaps* — and/or only isolated non-zero coefficients standing out as short tonal bursts — or *birdies* — in these regions at varying frequencies over the course of a few frames, even if the quantizer input has a spectrotemporally flat and noise-like structure. The consequence is an often unpleasant dullness or excessive tonality that can be heard in the decoded waveforms [Valin10].

To minimize the appearance of such artifacts in low-rate coding, a number of related solutions have been proposed. The most fundamental techniques, which exploit the fact that the issue at hand arises primarily on noise-like signals, are examined hereafter.

- Perceptual Noise Substitution (PNS), introduced as an additional coding tool for the MPEG-4 General Audio specification [Herr98, ISO09], is based on the observation that "one noise source sounds like the other": the exact spectrotemporal structure of a noise signal is perceptually irrelevant, and only the parameters of the noise, i. e., the coarse temporal and spectral envelopes, are needed to reconstruct the respective signal [Schul96]. In the context of SFB-based audio coding, it, therefore, suffices to communicate only said temporal structure and the band energy if a SFB is detected to be noise-like, and the decoder can recreate the SFB spectrum with a (pseudo-)random number generator. All transform coefficients of the noisy SFBs can be exempt from "expensive" coding and transmission (i. e., set to zero), leaving more bits for the coding of the demanding, e. g., tonal, bands.

- Noise filling (NF) is a simplification of the PNS paradigm which considers every spectral hole after quantization to be noise-like in nature. This assumption can be regarded as realistic since tonal frequency regions are not flat, and their prominent spectral peaks, representing the individual harmonics of the input, often "survive" the quantization even at low bit-rates (thereby leaving non-zero $q_i(k)$ in the corresponding parameter band). Hence, the conclusion is that, given tonal $q_i(k) \neq 0$, the $q_i(k) = 0$ are noisy with sufficient probability. This alleviates the need for difficult and computationally intensive band-wise prediction and noise detection [Schul96] as it is required for pre-quantizer SFB classification (tonal/noisy) in PNS. NF, like PNS, can be applied on a band-wise basis by determining whether all $q_i(k)$ of a specific parameter band are zero after quantization and, if affirmative, by transmitting a respective NF energy or root-mean-square (RMS) value. In this approach, which is implemented in MPEG USAC [ISO12] and AC-4 [ETSI14, Kjör16], the band-wise RMS values for NF are conveyed in substitution for the "empty" bands' scale factors, which are not needed because all associated $q_i(k) = 0$. USAC additionally supports NF of individual coefficients in non-zero SFBs having at least one $q_i(k) \neq 0$, as long as k is located at or above a specified transform length dependent *noise filling start offset*. For such "non-empty" SFBs, a scale factor is required for decoder-side reconstructive scaling of the non-zero $q_i(k)$, so to limit the NF parameter rate, only a frame-wise global noise level nl_i, applied relatively (multiplicatively) to the scale factor gain, is added per channel.

To summarize, both PNS and NF exploit temporal and spectral *de*correlation of specific frequency regions in order to attain efficient *parametric* coding (instead of discrete waveform preserving coding) of the corresponding transform coefficients. A converse apparatus, with comparable effect, can be constructed by considering inter-coefficient correlation, either in time direction using differential perceptual coding [Paras95] or in frequency direction via spectral translation, as in E-AC-3 [Field04] and CELT [Valin10]. In the former, zero-quantized transform coefficients are substituted with past decoded non-zero values at the same frequency index k using a predictor-like algorithm (see also subsection 2.2.2). The latter prevents spectral sparseness by way of direct or reversed (folded) copy-up of decoded low-frequency (LF) sub-band vectors to empty HF bands, a scheme which will be revisited in section 2.6 in the context of bandwidth extension.

The joint execution of scaling, quantization, and substitution completes the process of irrelevance reduction — via "noise" insertion and shaping — on the coded transform coefficients which, in turn, have been subjected to redundancy reduction — via filtering and/or transformation — by the closed-loop pre-processing and the filter bank itself.

The next subsection addresses the lossless coding of the q_i and all codec parameters.

2.3.4 Entropy Coding of the Spectral Coefficients and Parameters

The quantized spectra X_i^q are typically very sparse in comparison with the TD input waveforms, especially at low coding rates (where a low output entropy is targeted) and on spectrally strongly shaped signals (where large regions of X can be quantized to zero due to simultaneous masking). Moreover, the side information which must be conveyed to the decoder in order to invert the pre-processing and scaling consumes a significant percentage of the total bit-rate when stored in plain PCM form. To this end, individually customized entropy coding schemes are applied to the quantization indices q_i and gains g_i or l_i as well as each set of pre-processing parameters, as described in the following.

In MP3 [ISO93] and AAC [ISO97, ISO09], the subset vector of q_i covering the desired audio bandwidth, with SFB granularity, is compressed using multi-coefficient Huffman coding [Huff52], trained on value pairs or quadruples of neighboring coefficients. Thus, 2- or 4-tuples of successive $q_i(k)$ are represented by a single Huffman code word. MP3 allows said vector to be divided into five partitions, with variable Huffman code books, tabulated in both the encoder and decoder, assigned to three of the partitions [Bran97]. AAC improves the efficiency of this method by providing more code books and flexible dynamic partitioning of q_i into variable-length (in SFB units) Huffman sections. The size of each section and the selected Huffman table index are sent to the decoder [Bosi97].

Most recent audio transform coders abandoned the low-complexity Huffman coding of the q_i in favor of slightly more efficient (but also more resource intensive) arithmetic coding techniques. Opus [IETF12] uses an early variant termed range coding [Mart79], with tabulated individually trained symbol probabilities, on most of its bit-stream components, including the VQ coded spectral values. MPEG USAC [ISO12] and AC-4 [ETSI14, ETSI15] employ more advanced multi-tuple arithmetic coders, with spectrotemporally dynamic probability contexts in case of the former codec [How94, Mein05, Fuch11]. The basic operation of such *context-adaptive* arithmetic coding, exploiting the higher-order conditional spectral entropy by determining the symbol context for each subset of q_i (a 2-tuple in USAC) from past and/or LF neighbors, is visualized in Figure 2.11. From the context, a state is derived and associated with a tabulated cumulative frequency model, which, in turn, is used to generate or decode the variable-length code for said subset.

The real-valued $l_i(b)$, quantized logarithmically in steps of 1.5 or 3 dB via (2.39), i. e.,

$$r_i(b) = Q\big(\log(l_i(b)\,)/\log(\varepsilon)\big),\ l_i^q(b) = \varepsilon^{r_i(b)},\ \ \varepsilon = \sqrt[4]{2}\ \text{for 1.5 dB},\ \sqrt{2}\ \text{for 3 dB}, \quad (2.41)$$

are coded differentially using a dedicated Huffman table. More precisely, the differences between adjacent $r_i(b)$ are computed (representing first-order predictive filtering), and the resulting integers (prediction residuals) are mapped to variable-length code words.

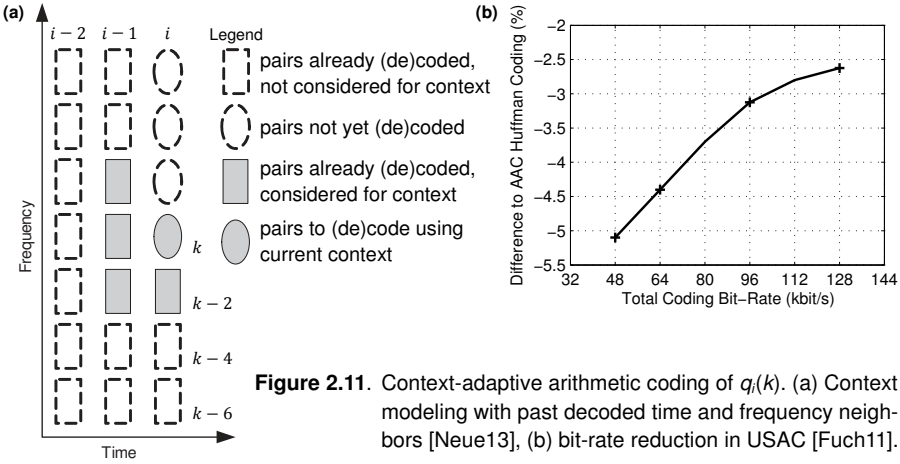


Figure 2.11. Context-adaptive arithmetic coding of $q_i(k)$. (a) Context modeling with past decoded time and frequency neighbors [Neue13], (b) bit-rate reduction in USAC [Fuch11].

The band-wise RMS or energy levels for NF/PNS are also quantized according to (2.41), with a resolution of 1.5 dB in AAC and USAC and 3 dB in AC-4. They replace the corresponding $l_i(b)$, as noted above, and are incorporated into the delta-frequency coding. The first scale factor index $r_i(0)$ (or noise fill level, if the first SFB is fully quantized to zero), representing g_i in the differential scheme, is, along with the index for level nl_i, stored as an absolute PCM value. The possibly band-wise joint-stereo parameters α_i^q, β_i^q, and ρ_i^q, quantized uniformly in their respective coordinate domains, are entropy coded like the scaling and NF/PNS data, optionally with delta-time coding [Helm11, ISO12, ETSI14].

The code vectors returned by the separate entropy coding operations are combined, or *multiplexed*, into the bit-stream representation to be stored or transmitted, including

- the sectioned-Huffman, range, or arithmetically coded quantization indices $q_i(k)$
- the global gains g_i and/or predictive-Huffman coded SFB-wise scale factors $l_i(b)$
- the NF or PNS indicators as well as the global nl_i and/or the SFB-wise RMS levels
- the SQ or VQ coded predictor coefficients and/or matrix indicators (usually with Huffman coding, except for TNS) for single and joint-channel FD post-processing
- the filter bank parameters such as the number of transforms and groups (i.e., the block switching and grouping data) and the window shape types for each frame
- the long-term prediction and/or gain coefficients when using TD post-processing
- global frame and/or channel data, e. g., element identifiers and coded bandwidth.

In addition, the bit-stream header contains the global audio information like bit-stream identification, output sampling rate, channel count, location, and grouping, frame length N, and so on. Payload data requiring only a few bits are usually included in PCM form.

2.4 Extensions for Parametric High-Frequency Regeneration

Subsection 2.3.3 described how coefficient substitution algorithms such as PNS can be applied to fill spectral gaps and to reduce birdies in low-rate perceptual coding. Not all input signals, however, are noise-like, and inserting (pseudo-)random values in place of tonal frequency components — like those belonging to harmonics in an instrumental recording — again leads to annoying artifacts. As with previous-generation codecs like MP3 and AC-3, this drawback still makes it frequently necessary to introduce low-pass filtering of the audio input in the encoder [Bran13], which minimizes the occurrence of birdies but, naturally, maximizes the perception of dullness in the decoded waveform.

To compensate for a low-pass induced HF loss in audio coding, bandwidth extension (BWE) methods have been developed [Makh79]. These parametric techniques achieve high-frequency regeneration (HFR) in the decoded output up to the desired — possibly complete — signal bandwidth via weighted spectral folding or translation from lower-frequency regions, often in combination with NF. The necessary information about the spectrotemporal structure (energy and tonality) of the original HF content is compactly coded and transmitted to the receiver as an auxiliary data-stream. Two low-complexity transform-domain implementations have already been mentioned in subsection 2.3.3:

- E-AC-3 employs adaptively weighted mixtures of pseudo-random and translated MDCT values, controlled via banding and noise blending data [ATSC12, Field04].
- CELT uses simple spectral folding into energy bands for which no bits have been allocated for MDCT coefficient coding, without extra NF in such bands [IETF12].

In both cases, band-wise energies are additionally coded for HFR level control, as in NF and PNS. Such basic approaches, although leading to notably increased reconstruction quality, tend to cause artifacts such as harshness in the BWE region [Valin10] especially in very-low-rate cases, where only narrowband (NB) waveform coding yields acceptable audio quality. To solve this issue, three advanced BWE algorithms have been developed:

- Spectral Band Replication (SBR), an amendment to MPEG-4 AAC [Wolt03, ISO09]
- Enhanced SBR, an extensive toolset designed for MPEG-D USAC [ISO12, Neue13]
- Advanced Spectral Extension (A-SPX) coding, included in AC-4 [ETSI14, Kjör16].

Similar schemes have also been introduced for the 3GPP Enhanced Voice Services (EVS) [ETSI16], G.718 [Vaill08, Tam09, Jelín09], and AMR-WB+ standards [Mäki05, Sala06] as well as for further work [Anna06, Ferre05, Laak05, LeeC13, Neuk13, Sinha06, Tsuji09]. In the following, the common building blocks of these HFR designs — analysis/synthesis filter banks, control parameter extraction, HF generation, envelope estimation/adjustment, and additional HF post-processing, as shown in Figure 2.12 — will be investigated.

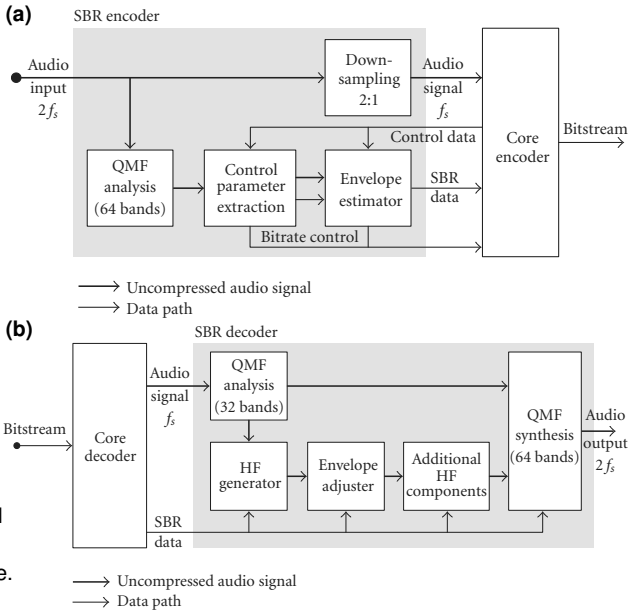


Figure 2.12. General block diagram for high-frequency regeneration in perceptual audio coding. Shown is the dual-rate SBR scheme, but USAC's Enhanced SBR and AC-4's A-SPX operate very similarly. f_s: core sample rate. Illustrations from [DenB09].

2.4.1 Analysis and Synthesis Filter Banks, Resampling

A dedicated pair of complex-valued pseudo-quadrature mirror filter (QMF) banks for T/F mapping, realized using 128 polyphase quadrature filter (PQF) instances [Roth83] to construct 64 complex-valued sub-bands, is utilized in the decoder in order to obtain a quasi-spectral representation with high temporal resolution for HFR. A corresponding equally designed analysis QMF bank is employed in the BWE encoder for the parameter acquisition. With these filter banks, each input channel is processed separately, and the resulting domain is oversampled by two (a real and an imaginary sub-band coefficient is computed per TD input sample). Alternatively, only the 64 cosine-modulated PQF instances are used to avoid the added complexity due to oversampling [DenB09, ISO09].

In the *dual-rate* HE-AAC design of Fig. 2.12, the pseudo-QMF banks are also used for downsampling by two — and, concurrently, the desired low-pass filtering — of the input in the HFR encoder and upsampling by two for reproduction in the decoder. This allows the MDCT core codec (AAC, in this case) to run at half the input sampling rate, which is beneficial at very low rates [Wolt03]. Note that, for this reason, a 32-sub-band analysis QMF bank suffices in the HFR decoder, since the upper 32 sub-bands of a 64-band filter bank are all zero. The resampling is omitted in *downsampled* HE-AAC and AC-4 [Kjör16].

2.4.2 Extraction and Coding of HFR Control Parameters

The HFR parameter acquisition in the encoder, as noted, comprises the measurement of energy and tonality data attributed to the original HF spectral content for the current frame. More precisely, temporal and spectral envelope and flatness information for the decoder-side pseudo-QMF-domain regeneration is collected. The temporal envelope is quantified through a combination of transient detection (to distinguish between quasi-stationary and nonstationary audio segments, much like the core-coder block switching detection introduced in section 2.1) and — dependently thereon — time/frequency grid selection (to decide upon the number, temporal support, and spectral width of the HFR parameter bands in which the decoder-side processing will be performed). The spectral envelope is then determined by energy or RMS calculation in each of said HFR bands as in "empty" SFBs for PNS or NF, but in a potentially complex-valued domain [Wolt03].

The assessment and parametric reconstruction of the frame-wise spectrotemporal flatness constitutes the primary focus of scientific research in BWE related work during the last decade [ISO12, Neue13, ETSI14]. For HFR, the spectral flatness information can be regarded as the fine details, or "peakiness", in each parameter band which is missing in the relatively coarse envelope parameterization. This aspect can be governed by the computation of a spectral flatness measure (SFM) — or "noisiness" value — for each HFR band. The band-wise SFM values can then be used to, e. g., control the level of a pseudo-random noise signal mixed into the HF bands during the decoder-side BWE process, as in E-AC-3's Spectral Extension [Field04, ATSC12] and SBR [DenB09, ISO09]. This type of additional HF component (see also Fig. 2.12(b) for the location within the SBR decoder) is, along with some recently introduced alternatives or extensions, examined in greater detail in subsection 2.4.5. A temporal flatness measure (TFM), on the contrary, indicates characteristics like the fine temporal "buzziness" of the HF input which, usually, are not modeled by the coarse time/frequency grids, especially during quasi-stationary signal passages. Introducing a temporal "smoothness" parameter in the BWE side-information allows to manipulate the fine temporal structure of the HRF output to better match the original waveform. Again, a detailed description will be provided in subsection 2.4.5.

The collected HFR control data for each frame and channel — typically comprising a transient indicator, the T/F grid layout, the logarithmically quantized band energy/RMS values associated with that grid, noise level information, and optional spectrotemporal flatness parameters like quantized normalized SFM and TFM values — are differentially coded in either time or frequency direction, possibly with M/S-like treatment for stereo signals. The "expensive" parameters are entropy coded using dedicated Huffman tables, as introduced in subsection 2.3.4 [DenB09, ISO09, ETSI14]. Along with the delta coding, this minimizes the per-channel BWE data rate to about 1–3 kbit/s on average [Wolt03].

2.4.3 Generation and Flattening of High-Frequency Content

After reception of the BWE-extended bit-stream and waveform decoding of the contained TD core signal, the HFR control information is entropy decoded, and all potential delta coding is undone. The HFR then commences with the pseudo-QMF transformation of the core waveform and a "generator" process, wherein low-band PQF coefficients are selected based on predefined rules controlled by the transmitted HFR parameters and are copied or mirrored up to the (still empty) high-band PQF coefficients [DenB09]. The copy and mirror operations are also known as transposition and folding, respectively.

At low bit-rates and/or crossover frequencies between the core and BWE range, very tonal audio signals, such as single-instrument recordings, often benefit from harmonic transposition avoiding dissonance due to the copy-up. In other words, it is subjectively advantageous to preserve the harmonic structure of the input as accurately as possible in such cases, even at and above said crossover frequency. In order to enable harmonic continuation after HFR, the USAC specification [ISO12] allows harmonic transposition, by means of QMF-based spectral stretching, in addition to traditional copy-up [Neue13]. The stretching, which can be applied alternatively to legacy linear transposition via the transmission of an appropriate SBR bit-stream payload header, is achieved using phase vocoder techniques. More precisely, time stretching and pitch shifting are performed at certain ratios on the LF QMF coefficients to obtain the desired HF coefficients, and cross product terms are optionally added to generate missing harmonic partials [Zhon11].

In (Enhanced) SBR and A-SPX, precise reconstruction of the original spectrotemporal flatness in the BWE region is carried out via inverse filtering and sub-band smoothing, both of which are applied in temporal direction and separately for each T/F grid interval [ISO09, ETSI14]. The former represents second-order linear predictive analysis filtering (i. e., with FIR) in each of the HFR QMF sub-bands, thereby acting as an in-band spectral whitening pre-processor prior to the envelope adjustment step. The filter strength is a function of a chirp factor between 0 and 1, which is controlled by the transmitted SFM parameter [DenB09]. The latter is a temporal whitening procedure intended to remove (or at least soften) subtle peaks in the time structure of the generated HF waveform. As such, it is the TD equivalent of the LPC-based whitening filter. Because of the high time resolution of the QMF bank, temporal smoothing can be applied multiplicatively by normalizing the HFR time-slot energies (the individual energies across all QMF coefficients belonging to the same time instance). The normalization factor, ranging from 0 to 1 like the chirp factor, can be determined from the TFM data in the bit-stream. Naturally, such an algorithm is only useful in quasi-stationary frames, as indicated by the transient flag. For unknown reasons, temporal smoothing can only be activated globally, via the BWE payload header, in (Enhanced) SBR and A-SPX. A frame-wise off/on flag is not provided.

2.4.4 Estimation and Adjustment of High-Frequency Envelope

The generated and, possibly, spectrotemporally pre-flattened QMF signal in the BWE range now needs to be scaled to match the parameter-band-wise envelope of the input signal at the encoder. To this end, the band energy or RMS values are transmitted to the decoder, as already noted in subsection 2.4.2. In the default case of complex-exponential modulated pseudo-QMF banks, the sub-band samples can be interpreted as the analytic versions of the samples obtained from the real (i. e., cosine-modulated) part of the filter bank. This feature leads to a sub-band representation which is suitable for aliasing-free modifications such as the envelope scaling, and also inherently allows for measurement of the instantaneous energy for the sub-band signals [DenB09]. Hence, for each T/F grid interval and parameter band, the HF input energy can simply be acquired by averaging the per-time-slot sums of the squared real and imaginary PQF coefficients. In case of a real-valued filter bank, the imaginary values are not available, so the "true" energies are typically approximated by doubling the average squared real PQF samples [Field04].

The HFR decoder applies the envelope data, consisting of the quantized energy/RMS averages, by first computing the same values on the respectively regenerated sub-band coefficients, i. e., using the same algorithm. Again, an accurate estimate is possible when a complex-valued filter bank implementation is available, otherwise an approximation using only the real-valued information, as above, can be performed. The resulting mean *source* estimate $S_h(v)$ for each time grid interval h and HF band index v is reciprocalized and multiplied by the transmitted respective mean *target* value $T_h^q(v)$ to obtain a scalar:

$$ts_h(v) = \sqrt{\frac{T_h^q(v)}{S_h(v)}} \text{ for energies, } \quad ts_h(v) = \frac{T_h^q(v)}{S_h(v)} \text{ for RMS values.} \tag{2.42}$$

This scalar, finally, is multiplied onto all pseudo-QMF coefficients associated with band v and interval h in order to impose the desired energy onto the regenerated HF content.

2.4.5 Additional Post-Processing of Adjusted HF Content

The observant reader might have noticed that the temporal whitening, or smoothing, optionally applied during the HF generation process only allows to flatten — but not to sharpen — the time envelope in the BWE region. To address this shortcoming, temporal envelope shaping (TES) functionality was added to the Enhanced SBR toolset during the standardization of MPEG-D USAC [ISO12, Neue13]. The inter-sub-band-sample TES, or "Inter-TES", applied after the HFR envelope adjustment step, alleviates the need for fine T/F grids, with inevitably expensive transmission of a large number of $T_h^q(v)$, on highly transient frames. By temporally sharpening the HF signal, with sub-interval resolution,

based on the LF core content (thereby exploiting correlation between their envelopes) and a few bits of side information (an activation flag and the TFM parameter), Inter-TES can sufficiently minimize pre-/post-echo distortion in the HFR signal upon decoding.

Speaking of correlation, it is worth mention that the spectrotemporal fine structure of the generated HF signal, regardless of whether copy-up or harmonic transposition is employed, always remains somewhat correlated with the LF waveform. Several natural audio stimuli, however, are noisier at high than at low frequencies since their contained harmonic components decay faster with increasing frequency than the noise-like "background" components [Field04, DenB09]. Failure to decorrelate the LF and HF samples may, therefore, lead to artifacts such as harshness on some material after BWE. A trivial synthetic example is given in Figure 2.13(a). To properly reconstruct the ratio between the dominant "foreground" tones and the residual "background" noise, also often called tonal-to-noise ratio, pseudo-randomly generated white noise is blended into the transposed HFR signal. The encoder determines the component weights in the decoder-side tone-noise mixture by way of an SFM-like noise floor parameter, which is quantized and coded (per frame i or band v) as part of the BWE side-information [Wolt03, Field04].

Analogously to the temporal sharpening approach, spectral sharpening for increased HF tonality, in comparison with the LF source signal for HFR, can be performed. Figure 2.13(b) presents a use case for this process. Here, the spectral sharpening is realized by inserting "missing harmonics" not present in the copy-up content, implemented using a sinusoidal oscillator at the center of every affected pseudo-QMF band [ISO09, ETSI14]. The necessity for the insertion of such a tonal component is assessed, and transmitted, by the encoder, having access to the original HF spectral representation and the source range for the copy-up, in which the desired tone is missing. The control data needed for the decoder-side synthesis comprises the frequency (i. e., the band index v covering two or more sub-band indices), start offset (i. e., index h if the frame is divided into multiple intervals), and relative level (e. g., a quantized SFM value) of the sinusoid to be added.

Figure 2.14 summarizes the different algorithmic parts of HFR post-processing, after core decoding, as a block diagram for the case of Enhanced SBR. Note that, alternatively to SQ and differential Huffman coding, USAC also supports predictive VQ and coding of T_h. Moreover, in BWE bands in which transposed content, missing harmonics, and noise are combined, a (not depicted) re-normalization to the respective $T_h^q(v)$ is carried out. The regenerated high-band signals and the delay-compensated (resulting from the HFR process) low-band signals are finally supplied to the 64-channel synthesis pseudo-QMF bank, which usually operates at the sampling frequency of the original PCM signal. The synthesis filter bank is, just like the analysis bank, generally complex-valued, however, the imaginary part of its TD output is discarded to obtain a real-valued signal [DenB09].

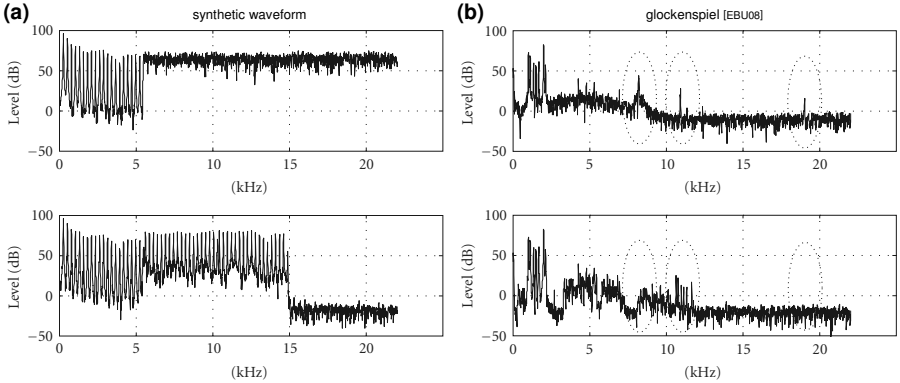


Figure 2.13. Illustrative examples for the necessity of (a) noise addition and/or inverse filtering, (b) insertion of missing harmonics. Top: encoder input, bottom: decoder output in case of HFR, ranging from 5.5 to 14.8 kHz, but without respective tools [DenB09].

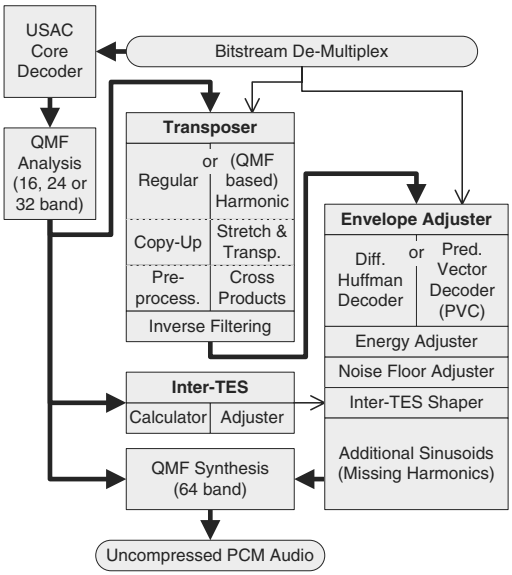


Figure 2.14. Complete overview of the individual components of the Enhanced SBR decoder in USAC [ISO12]. (—) audio signal, (—) side information or control data [Neue13].

2.5 Extensions for Parametric Stereo or Multichannel Coding

The previous section described how parametric high-frequency reconstruction may be utilized in order to lower the audio bandwidth to be waveform coded — by means of transform-domain quantization and entropy coding — at low target bit-rates. However, for two-channel stereophonic or even multichannel immersive input, HFR methods for BWE alone do not sufficiently reduce the burden on the perceptual core coder, causing a rapid loss in subjective audio quality toward very low average per-channel bit-rates.

An early proposal to ameliorate this issue, termed "intensity stereo coding", has been introduced already in subsection 2.2.3 [Vand91]. The basic paradigm of this parametric technique is the exploitation of reduced sensitivity of human hearing to IPDs at higher frequencies (where only the spectrotemporal envelopes, or ILDs, are most relevant) by calculating a single downmix I_i of a channel pair at these frequencies and an associated, usually band-wise panning angle $\alpha_i \in [-\frac{\pi}{2}, \frac{\pi}{2})$. The original channel spectra are then modeled perceptually from the quantized and entropy coded I_i^q, α_i^q via (2.34) with $E_i \stackrel{\text{def}}{=} 0$.

Given that the quantized angles α_i^q for each frame i can be coded with a much lower bit-rate than that required for waveform coding of the second channel, i. e., the residual HF spectral content, more data rate is available for transform coding of the monophonic downmix and the LF two-channel signal. As a result, the overall (mean) coding quality notably improves at very low bit-rates, which is why this technique is used extensively in, e. g., (E-)AC-3 [ATSC12, Field04] and Opus [IETF12, Valin13]. However, as mentioned by v. d. Waal and Veldhuis [Vand91], a robust psychoacoustic model is required for safe artifact-free use of intensity stereo coding at frequencies below 10 kHz. The increase in auditory phase sensitivity toward LFs does not fully explain this observation, as studies indicate that its perceptual relevance only turns significant below 4–5 kHz [Moor12].

Baumgarte and Faller [Falle03] and Breebaart *et al.* [Bree05] investigated the issue of low-rate stereo coding further and noticed that proper reconstruction of inter-channel cross-correlation (ICC), closely related to inter-aural correlation [Blau96] and virtually identical to the latter in case of playback via headphones [Baum03], is just as important as accurate ILD (and, at LFs, IPD or inter-channel time difference) reproduction. An ICC parameter serves to control the apparent width and, for loudspeaker playback, distance of the downmixed source when upmixed to multiple channels in the decoder, as shown in Figure 2.15. A decrease in ICC magnitude is perceived as an increase in spatial width or diffuseness until the downmix splits into two signals, one at each output channel.

Based on the abovementioned work by Baumgarte, Faller, and Breebaart *et al.*, three highly efficient parametric stereo and multichannel coding approaches were developed:

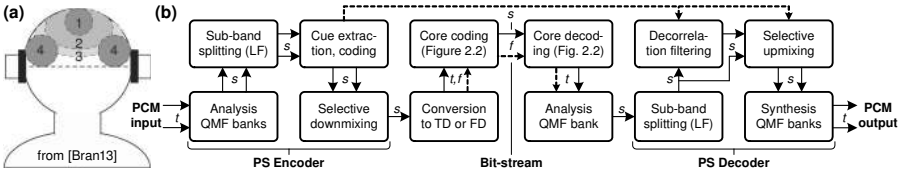


Figure 2.15. Reconstruction of ICC in Parametric Stereo (PS) coding. (a) The perceived width of an auditory event increases with decreasing ICC magnitude (1–3), until distinct events appear at each ear (4). (b) ICC control by mixing in a decorrelation result.

- MPEG-4 Parametric Stereo (PS), an amendment to the MP4 HE-AAC specification [ISO09] whose standardization was finished in 2003, shortly after SBR [Schui04]
- MPEG Surround (MPS), an enhanced PS variant standardized as an independent generic coding tool [ISO07] as well as a two-channel component of USAC [ISO12]
- Advanced Coupling [ETSI14] and Advanced Joint Channel Coding [ETSI15], representing evolutions of E-AC-3's channel coupling [Field04] for the AC-4 system.

Like SBR and A-SPX, all of these implementations operate in a complex-valued pseudo-QMF domain, which can be configured identically (including similar ERB-like parameter band partitioning) so that, as shown hereafter, the necessary extra filter bank instances can largely be shared among the HFR/BWE and the PS based tools [Schui04, Bree05]. It is also fair to note that alternative but simpler (and, therefore, less efficient) parametric joint-channel schemes are in use in other low-rate codecs as well, e. g. [Lind05, Mäki05]. For the sake of brevity, however, these approaches will not be examined in this thesis.

The remainder of this section introduces those common algorithmic components of the above three state-of-the-art solutions which have not already been presented in the course of the HFR discussion. Following the outline of the last section, these comprise, in order of execution, the spatial/binaural cue extraction and coding, the downmixing, the derivation and spectrotemporal shaping of the decorrelation signal(s), and the upmixing process resynthesizing the initial stereo or multichannel configuration. Each of these four steps, illustrated in Fig. 2.15(b), is described in a separate subsection below.

2.5.1 Sub-Band Splitting, Extraction and Coding of Spatial Parameters

The first operation performed in the sub-band domain (*s* in Fig. 2.15) of a parametric multichannel pre-/post-processor after the channel-wise analysis pseudo-QMF banks is an increase of the spectral coefficient resolution at low frequencies to better match the auditory resolution [Schui04, Bree05]. This is done by cascading the complex QMF out-

puts with a second, oddly modulated "subfilter" bank which, in fact, represents the separate application of a real-valued sub-band splitting according to subsection 2.2.2 on the real and imaginary PQF coefficients. [Schui04] explains this process in greater detail.

The resolution-enhanced "hybrid" pseudo-QMF samples for each input channel now enter the parameter acquisition, quantization, and entropy coding algorithm, where the spatial information between the channel pairs (or, in general, tuples) to be restored at the receiver is extracted. Since the decoded output is often played over headphones, the term "binaural cues" instead of "spatial parameters" is employed as well here [Baum03, Falle03], but for reasons of generality (i. e., an applicability also in case of two or more loudspeakers), the word "spatial" will be preferred hereafter. As in SBR or A-SPX, these parameters are determined in ERB-like bands consisting of multiple sub-bands — now covering the low and mid frequencies — with similar input adaptation — a combination of transient detection and dependent T/F grid selection — described in subsection 2.4.2.

- The band-wise ILD values are obtained by computing the average or total power (i. e., energy, intensity) for each channel in said band v and by nonlinearly quantizing the ratio between pairs of these power values on a logarithmic scale. Thus,

$$i_{\mathrm{ild},h}(v) = Q_{\mathrm{n}}\left(10 \cdot \log_{10}\left(\frac{\sum_{s \in h,v} A_i(s)A_i^*(s)}{\sum_{s \in h,v} B_i(s)B_i^*(s)}\right)\right), \tag{2.43}$$

  where h is the T/F grid interval, A_i and B_i are the complex QMF channel signals, superscript $*$ denotes complex conjugation, and $Q_{\mathrm{n}}(\cdot)$ maps to a value with non-uniform step size, e. g., a signed integer with magnitude $|Q_{\mathrm{n}}^{-1}(i_{\mathrm{ild},h}(v))| \in [0, 2, 4, 6, 8, 10, 13, 16, 19, 22, 25, 30, 35, 40, 45, 50]$ as in MPEG-4 PS or MPS [Bree05]. The individual energy sums are computed identically to the $S_h(v), T_h(v)$ in HFR. MPS codes two ILDs in its *three-to-two* boxes used on channel triples [Bree07].

- The IPD for band v represents the relative phase rotation angle used for optimal phase alignment, producing maximum ICC, between the channels' sub-band signals in v before downmixing, as described in the next subsection. It is defined as

$$i_{\mathrm{ipd},h}(v) = Q\left(\tfrac{P}{\pi} \cdot \measuredangle\left(\textstyle\sum_{s \in h,v} A_i(s)B_i^*(s)\right)\right), \tag{2.44}$$

  i. e., yields a constant phase *difference* (instead of a constant phase *slope*, as it is caused by an inter-channel time delay) within each v. As indicated, uniform SQ with $P = 4$ is typically used; see also page 29. PS additionally codes an identically quantized overall phase difference (OPD) per band v [Schui04, Bree05], which is avoided in MPS 2-1-2 to lower the side information rate [Hyun12, Neue13]. The standalone MPS specification [ISO07] omits phase coding to simplify the system.

- The ICC parameter, acquired either globally per frame or locally per v, specifies an inter-channel coherence which, here, equals the normalized cross-correlation coefficient after phase alignment of the complex A_i and B_i according to the IPDs:

$$i_{\mathrm{icc},h}(v) = Q_{\mathrm{n}}\left(\frac{\left|\sum_{s\,\in\,h,v} A_i(s)B_i^*(s)\right|}{\sqrt{\sum_{s\,\in\,h,v} A_i(s)A_i^*(s)\cdot\sum_{s\,\in\,h,v} B_i(s)B_i^*(s)}}\right) \tag{2.45}$$

  with A_i and B_i here being the phase rotated pseudo-QMF inputs for best possible channel alignment. When omitting the transmission of IPDs, $i_{\mathrm{icc},h}(v)$ is given by

$$i_{\mathrm{icc},h}(v) = Q_{\mathrm{n}}\left(\frac{\mathrm{Re}\left\{\sum_{s\,\in\,h,v} A_i(s)B_i^*(s)\right\}}{\sqrt{\sum_{s\,\in\,h,v} A_i(s)A_i^*(s)\cdot\sum_{s\,\in\,h,v} B_i(s)B_i^*(s)}}\right) \tag{2.46}$$

  [Bree05, Bree07]. This implies that, in case of perfectly IPD-compensated inter-channel time difference, the unquantized ICC data is equivalent to the maximum value of the normalized cross-correlation as a function of the relative time delay between A_i and B_i [Bree05], again determined either frame- or band-wise. In PS or MPS, the ICC is mapped non-uniformly to an integer with $Q_{\mathrm{n}}^{-1}(i_{\mathrm{icc},h}(v)) \in$ [–1, –0.589, 0, 0.368, 0.601, 0.841, 0.937, 1], offering higher resolution near the peak at 1 than at 0 or the minimum at –1. In *three-to-two* MPS boxes, (2.46) is adjusted appropriately to account for 3 instead of 2 input channels [Bree07, Hoth08].

- Moreover, residual coding in all sub-bands below an encoder-defined frequency (specified as a parameter band index v_{rc}) is supported in MPS [Herr08, Neue13] and AC-4 [ETSI14]. For the affected LF pseudo-QMF samples in the band range $0 \le v < v_{\mathrm{rc}}$, values parameterizing the error between the channel downmix and the initial channel coefficients are computed in place of $i_{\mathrm{icc},h}(v)$. More precisely, complex pseudo-QMF signals $E_i(s)$ with $s \in h, v$, similar to the real-valued $E_i(k)$ obtained in the KLT rotation of (2.33), are determined, as will be clarified later.

  In USAC, the combination of MPS 2-1-2 with IPD coding and a core-coded band-limited residual is called Unified Stereo (UniSte) [Neue13]. This scheme allows for the downmix and residual to be jointly transform coded according to sections 2.1 to 2.3, potentially with further redundancy or irrelevance removal by way of the FD stereo coding techniques of subsection 2.2.3. Similar methods are applied in the Advanced Coupling, Joint Object, and Joint Channel tools in AC-4 [ETSI14]. Due to its standalone design, MPS codes $E_i(s)$ independently of the downmix.

When not employing residual coding, the abovementioned parametric spatial audio encoders, in summary, collect vectors of band-wise ILD indices i_{ild} and, optionally, IPD indices i_{ipd} as well as band- or just frame-wise ICC indices i_{icc}, attained via perceptually

motivated, mostly non-uniform quantization. Only a coarse parameter band resolution is required, both spectrally (modeling the ERB-like auditory selectivity) and temporally (reflecting a binaural "sluggishness" to spatial changes of at least 30 ms, except on transient events) [Baum03, Bree05]. As such, the combination of these spatial cues carries roughly two orders of magnitude less information than what would be contained in the fullband residual needed for waveform preserving coding [Bran13]. Complemented by a selective time- or frequency-differential scheme (whichever returns a lower entropy) and dedicated Huffman coding, as known from the predictive joint-stereo coding tools and SBR or A-SPX (subsections 2.3.4 and 2.4.2, respectively), it is, therefore, possible to achieve a spatial side information rate of only 1.5–8 kbit/s (stereo) for PS or MPS 2-1-2 [Bree05, Neue13] and 3–32 kbit/s (5.1, etc.) for MPS or the AC-4 tools [Herr08, Kjör16].

2.5.2 Calculation of Spatial Downmix and Residual Signals

The substantial irrelevance reduction reached in parametric spatial coding, as noted, is realized by downmixing the pseudo-QMF-domain input channel signals to a reduced number of output channel signals after the ILD, IPD, and ICC parameter extraction. The downmix process, a band-wise linear combination with weighting coefficients collected in a downmix matrix $\boldsymbol{H}_{\mathrm{d}}$, can be carried out in various ways. The most common two are outlined hereafter, with a focus given to stereo downmixing from two channels to one, as is common in MP4 PS, USAC MPS 2-1-2, and *two-to-one* boxes in standalone MPS. The necessary adaptations for the *three-to-two* MPS box are described in [Bree07, Hoth08].

In principle, the downmix operation may be a simple summation of the two channel sub-band signals $A_i(s)$ and $B_i(s)$, again for each $s \in h, v$, with some scaling of the result to enforce a specific power criterion on the combined downmix signal, similarly to the (classic or predictive) M/S stereo formulation of subsection 2.2.3. MPS in its standalone or USAC 2-1-2 flavor, for example, requires the band-wise downmix energy to equal the sum of the respective band energies of A_i and B_i [Bree07]. This leads to the definition

$$M_i(s) = \frac{A_i(s) + B_i(s)}{c_h(v)} = \frac{\big(A_i(s) + B_i(s)\big) \cdot \sqrt{\sigma_{A,h}^2(v) + \sigma_{B,h}^2(v)}}{\sqrt{\sigma_{A,h}^2(v) + \sigma_{B,h}^2(v) + 2\,\sigma_{A,h}(v)\sigma_{B,h}(v)\; Q_{\mathrm{n}}^{-1}\big(i_{\mathrm{icc},h}(v)\big)}} \tag{2.47}$$

for the sub-band downmix signals $M_i(s)$, where $\sigma_{C,h}^2(v) = \sum_{s \in h,v} C_i(s) C_i^*(s)$ equals the band/grid energy for the given channel C_i, as in (2.43)–(2.46), and $Q_{\mathrm{n}}^{-1}(\cdot)$ indicates non-uniform inverse mapping of the ICC cue index $i_{\mathrm{icc},h}(v)$ of (2.45) or (2.46) back to a real reconstruction value. Note that, in case of IPD coding, $A_i(s)$ and $B_i(s)$ denote the phase aligned sub-band channel signals as in (2.45), by way of which $i_{\mathrm{icc},h}(v)$ is then acquired.

In (2.47), the same weight $c_h(v) = a_h(v) + b_h(v)$ is applied to both $A_i(s)$ and $B_i(s)$ in the downmix process. Clearly, more control over the individual contributions of the two audio signals in M_i can be attained by employing a separate scalar for each channel:

$$M_i(s) = c_{A,h}(v) \cdot A_i(s) + c_{B,h}(v) \cdot B_i(s), \quad s \in h, v, \tag{2.48}$$

which is comparable to the rotation-based transform of (2.33) producing a downmix I_i. It is worth repeating, however, that all values in (2.48), including $c_{A,h}(v)$ and $c_{B,h}(v)$, are complex-valued, while the spectral vectors and scalars in (2.33) are purely real-valued. (2.48) allows to avoid cancelation (i. e., destructive interference) between the downmix sources in M_i which may occur especially on out-of-phase (IPD = 180°) signals, thereby improving the spatiotemporal stability of the regenerated multichannel waveform after upmixing [Neue13]. A fixed $c_h(v)$ as in (2.47), on the contrary, bears the risk that the v-wise energy of M_i strongly depends on the v-wise ICC between A_i and B_i [Bree05]. Deriving the $c_{A,h}(v)$ and $c_{B,h}(v)$, analogously to the definition of $c_h(v)$, lies beyond the scope of this work; it shall only be noted that both are defined by the ILD, IPD, and ICC data.

The KLT-like rotation of (2.33) also yields an error spectrum E_i alongside the downmix I_i by means of a summation of the two input signals using different weights. Similar error signals — better known as *residuals* — can be obtained in case of the pseudo-QMF representations of (2.47) and (2.48). *Two-to-one* MPS models the two input channels by

$$\begin{aligned} A_i(s) &= [a_h(v) \cdot M_i(s) + E_i(s)], \\ B_i(s) &= [b_h(v) \cdot M_i(s) - E_i(s)], \end{aligned} \tag{2.49}$$

i. e., a binary decomposition into a common *in-phase* component (the downmix M_i) and an *out-of-phase* component E_i (exhibiting equal magnitude but opposite signs in the two channels). The latter represents the difference between M_i and the input channel signal at hand or, in other words, the desired residual signal which can be utilized along with M_i to perfectly reconstruct said input channel waveform in the absence of quantization. USAC's UniSte module, as a special case, computes E_i in the following M/S-like fashion:

$$E_i(s) = \frac{A_i(s) - B_i(s)}{c_h(v)} - \rho_h(v) \cdot M_i(s), \quad s \in h, v, \tag{2.50}$$

where $\rho_h(v)$ is a complex prediction coefficient similar to the predictive M/S parameter of subsection 2.2.3. The motivation behind this approach is to minimize the power of E_i for maximum input signal compaction into (and, hence, efficient joint core coding with) M_i [Neue13]. As noted earlier, E_i is usually limited to only a few LF bands, and this band count v_{rc} is conveyed to the decoder. Moreover, an $i_{\mathrm{icc},h}(v)$ is only transmitted for those bands $v \geq v_{\mathrm{rc}}$ and grid intervals h for which the associated $E_i(s), s \in h, v$, are not coded.

2.5.3 Coding of M_i and E_i, Generation of Decorrelation Signals

The downmix M_i and, if employed, the (full- or low-band) residual E_i are subjected to critically sampled perceptual transform coding to achieve a low overall bit-rate. To this end, the pseudo-QMF signals after downmixing via $\boldsymbol{H}_{\mathrm{d}}$ as a function of $a_h(v)$, $b_h(v)$, and $\rho_h(v)$, are converted from the complex-valued sub-band domain into a real-valued PCM TD waveform representation for subsequent MDCT processing according to section 2.1. Alternatively, a direct pseudo-QMF-to-MDCT transform, as indicated by the "Conversion to TD or FD" block in Fig. 2.15(b), can be used [Bree07], which avoids the intermediate TD representation and, as such, the added algorithmic complexity associated therewith. Irrespective of the exact realization of this complex-to-real conversion, a "hybrid" QMF synthesis process, involving an inversion of the LF sub-band splitting of subsection 2.5.1 as the first step, must be carried out. In UniSte MPS 2-1-2, both M_i and, if applied, E_i are passed to the USAC core coder [ISO12], while in the basically core coder agnostic stand-alone MPS, the intended perceptual coding scheme is only specified for E_i (conformance with the MPEG-2 AAC Low Complexity (LC) profile [ISO97] is dictated here) [Herr08].

Having forwarded the quantized and entropy coded M_i along with the quantized and entropy coded spatial parameter set (comprising the code vectors for i_{ild}, i_{icc}, and, when applicable, i_{ipd} and E_i) to the receiver, the multichannel signal configuration can be re-synthesized. To prepare the inputs for this procedure in the parametric decoder, which is performed by way of a cross-channel matrix operation with upmix matrix $\boldsymbol{H}_{\mathrm{u}}$ (similar to $\boldsymbol{H}_{\mathrm{d}}$, see subsection 2.5.4), the following initial algorithmic steps must be carried out:

- M_i^q and E_i^q must be reconstructed by the appropriate core decoder(s) and transformed to the complex sub-band domain using the combination of pseudo-QMF banks and LF sub-band splitting known from the encoder (see subsection 2.5.1).
- Reconstructive scaling, also often called "dequantization", must be applied to the ILD, ICC, and IPD index vectors, yielding the band/interval-wise quantized values

$$ild_h^q(v) = 10^{Q_{\mathrm{n}}^{-1}\left(i_{\mathrm{ild},h}(v)\right)/10}, icc_h^q(v) = Q_{\mathrm{n}}^{-1}\left(i_{\mathrm{icc},h}(v)\right), ipd_h^q(v) = \frac{\pi}{P}\cdot Q^{-1}\left(i_{\mathrm{ipd},h}(v)\right), \quad (2.51)$$

  respectively, where Q_{n}^{-1}, Q^{-1}, and P are defined as in the encoder (subsec. 2.5.1).
- For the high-frequency bands at $v \geq v_{\mathrm{rc}}$, a substitute for $E_i^q(s), s \in h, v$, must be derived from the parameters available at the decoder, namely, M_i^q and $icc_h^q(v)$. Note that proper selection of $\boldsymbol{H}_{\mathrm{d}}$ allows to minimize the coherence (maximum of the normalized cross-correlation) between signals $M_i(s)$ and $E_i(s)$. Thus, $E_i^q(s)$ can, perceptually, be approximated relatively closely by convolving $M_i^q(s)$ with a decorrelation filter and by scaling the result depending on $icc_h^q(v)$. A cheap filter design can be realized using frequency dependent delay lines [Schui04, Bree05].

The derivation and application of the decorrelated signals $E_i'(s)$, discussed in further detail in [Engd04, Bree07, Herr08], constitutes the major advantage of the parametric spatial coding schemes of this section over the intensity stereo coding technique of subsection 2.2.3. For utilization in, e.g., MPS, the following requirements can be established:

- The filtered output signal, E_i', shall be incoherent with the input signal, M_i^q. This can already be attained with a simple delay, as proposed by Lauridsen [Laur54].
- In case of multiple simultaneous decorrelated signals E_i', these shall be independent of each other, i.e., mutually incoherent. MPS ensures this by combining the parameter-band dependent sub-band sample delays with decorrelator-instance dependent lattice all-pass filtering, also in the QMF domain [Bree07, Herr08].
- The spectrotemporal envelopes of each E_i' shall closely follow those of M_i^q. This is, on average, true even for simple Lauridsen-type decorrelators but may not be the case on transient signal parts, where temporal smearing of the signal attacks may occur due to the delay or filter operation. To address this issue, standalone MPS allows to adjust E_i' using one of two methods, "sub-band domain temporal processing" and "guided envelope shaping". The activation of these tools, which, in short, adjust E_i' by way of scaling (with different amounts of side information) so that its band/interval-wise energy levels match those of M_i^q, is controlled by the encoder [ISO07, Bree07, Herr08]. USAC's MPS 2-1-2 pursues an alternative approach by extending the decorrelation filtering process itself [ISO12, Neue13].

2.5.4 Dynamic Upmixing to Original Channel Configuration

After the preparatory steps of the last subsection, the downmix M_i^q and disjoint sets of residuals E_i^q and decorrelator signals E_i', all of which are mutually orthogonal (exhibit maximum cross-correlations around zero), are available in the pseudo-QMF domain. In conjunction with the reconstructed ILD, ICC, and (optionally) IPD data, the initial spatial waveform configuration can be restored exactly (with E_i^q) or approximately (with E_i').

The first algorithmic process in the upmix yielding, e.g., A_i^q and B_i^q, which is common to all abovementioned parametric codecs, is the derivation of the individual coefficients of the upmix matrix $\boldsymbol{H}_{\mathrm{u}}$ from the decoded ild_h^q, icc_h^q, and (if available) ipd_h^q values. Thus, the rows and columns of $\boldsymbol{H}_{\mathrm{u}}$ are obtained separately for each parameter band v and T/F grid interval h. The general formulation for each $s \in h, v$ can be summarized as follows:

$$\begin{bmatrix} A_i^q(s) \\ B_i^q(s) \end{bmatrix} = \boldsymbol{H}_{\mathrm{u}} \cdot \begin{bmatrix} M_i^q(s) \\ E_i^{q,\prime}(s) \end{bmatrix} = \sqrt{2}\boldsymbol{L}_h(v)\boldsymbol{P}_h(v)\boldsymbol{R}_h(v) \cdot \begin{bmatrix} M_i^q(s) \\ E_i^{q,\prime}(s) \end{bmatrix}, \tag{2.52}$$

where the real-valued diagonal level-scaling matrix $\boldsymbol{L}_h(v)$ enables relative weighting in the upmix process and the also real-valued but full-rank matrix $\boldsymbol{R}_h(v)$ provides rotation in the two-dimensional signal space constructed by the (roughly orthogonal) $M_i^q(s)$ and $E_i^{q,\prime}(s)$. Both $\boldsymbol{L}_h(v)$ and $\boldsymbol{R}_h(v)$ depend on $ild_h^q(v)$ as well as $icc_h^q(v)$ in a manner which is examined more thoroughly in [Bree05]. The remaining array $\boldsymbol{P}_h(v)$ in (2.52) denotes a complex-valued matrix allowing modifications of the phase relationships between the output signals and, as such, additionally depends on $ipd_h^q(v)$ [Bree05, Neue13]. In bands where IPD/OPD coding is not employed, a simple alternative solution can be applied:

$$\begin{bmatrix} A_i^q(s) \\ B_i^q(s) \end{bmatrix} = \boldsymbol{H}'_{\mathrm{u}} \cdot \begin{bmatrix} M_i^q(s) \\ E_i'(s) \end{bmatrix} = \begin{bmatrix} l_1\cos\big(\iota_h(v)+\kappa_h(v)\big) & l_1\sin\big(\iota_h(v)+\kappa_h(v)\big) \\ l_2\cos\big(\iota_h(v)-\kappa_h(v)\big) & l_2\sin\big(\iota_h(v)-\kappa_h(v)\big) \end{bmatrix} \cdot \begin{bmatrix} M_i^q(s) \\ E_i'(s) \end{bmatrix} \quad (2.53)$$

with all elements of $\boldsymbol{H}'_{\mathrm{u}}$ being real-valued, and where l, κ, ι in $\boldsymbol{H}'_{\mathrm{u}}$ are specified as follows:

$$l_1 = \sqrt{\frac{ild_h^q(v)}{1+ild_h^q(v)}}, \quad l_2 = \sqrt{\frac{1}{1+ild_h^q(v)}}, \quad \kappa_h(v) = \frac{1}{2}\arccos\big(icc_h^q(v)\big), \quad (2.54)$$

and, dependently thereon,

$$\iota_h(v) = \tan\left(\frac{l_2-l_1}{l_2+l_1}\arctan\big(\kappa_h(v)\big)\right) \approx \kappa_h(v)\cdot\frac{l_2-l_1}{\sqrt{2}} \quad \text{[Bree05, Bree07]}. \quad (2.55)$$

Note that (2.53) is primarily used in higher-frequency bands, where $E_i^q(s)$ is not coded. UniSte MPS 2-1-2 with residual coding, on the contrary, applies the linear combinations

$$\begin{bmatrix} A_i^q(s) \\ B_i^q(s) \end{bmatrix} = \tfrac{c_h(v)}{2} \cdot \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} M_i^q(s) \\ S_i^q(s) \end{bmatrix} = \tfrac{c_h(v)}{2} \cdot \begin{bmatrix} 1+\rho_h(v) & 1 \\ 1-\rho_h(v) & -1 \end{bmatrix} \cdot \begin{bmatrix} M_i^q(s) \\ E_i^q(s) \end{bmatrix}, \quad (2.56)$$

representing a complex-valued M/S matrix where, as in the encoder-side calculation of $E_i(s)$ in (2.50), $c_h(v)$ and $\rho_h(v)$ are functions of the ILD, IPD, and ICC values [Neue13].

After upmixing, the output signals A_i^q and B_i^q, in summary, exhibit a cross-correlation obeying icc_h^q, a power ratio obeying ild_h^q, and a cross-channel power sum which, in each band and grid interval, equals the power of M_i^q [Bree07]. At low frequencies, the phase relationship between the original input signals can additionally be recovered with ipd_h^q and/or residual coding. To reconstruct the final PCM channel waveforms, the complex sub-band-domain A_i^q and B_i^q are fed separately through "hybrid" pseudo-QMF synthesis banks. In HE-AAC PS, the basic complex-modulated filter banks are shared with SBR, as shown in Figure 2.16, so only the LF sub-band splitting and its inversion must be added.

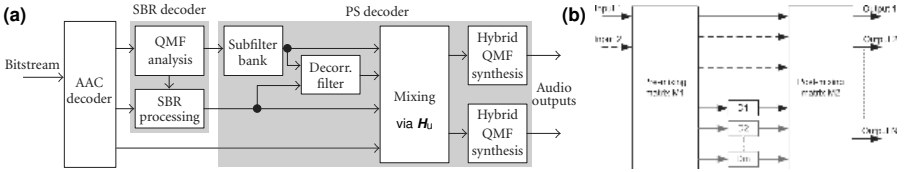


Figure 2.16. Structural overview of (a) combined SBR and PS decoding in HE-AAC [DenB09], (b) parallelized single-stage decoding in MPS with (D)ecorrelation filters [Herr08].

In conclusion of this section, it is worth noting that, for surround configurations with more than two or three channels, a tree-like cascade comprising multiple channel-pair (two-to-one, one-to-two) or channel-triple (three-to-two, two-to-three) modules can be utilized. These, however, typically introduce additional delay and audible reverberation due to the sequential decorrelation and matrix processing, which can be circumvented by means of parallelization [Bree07, Herr08]. More specifically, the decorrelation filters can be combined into a single multi-dimensional algorithmic operation and enclosed by pre- and post-mix matrices, as illustrated in Figure 2.16(b). This solution, in which the matrix elements are derived from the transmitted spatial information given the specific tree structure used for the parameterization, "flattens" the decorrelation and upmixing into separate, efficiently implementable stages. Furthermore, low-power decoding with reduced computational complexity, as in SBR [DenB09], can be realized by employing real instead of complex-valued hybrid QMF banks and decorrelation filters [Herr08].

2.6 Discussion of Quality, Delay, Advantages, Disadvantages

This section completes Chapter 2 with a brief discussion of the advantages as well as disadvantages which are encountered when using the coding schemes of the preceding sections in regular or low-latency applications. Particular emphasis is given to quality and delay considerations — and their interdependency — in the respective use cases.

2.6.1 Advantage of High Subjective Reconstruction Quality

The overall benefit of using the core coding tools of sections 2.1–2.3 in combination with the parametric HFR/BWE and stereo/multichannel extensions of sections 2.4 and 2.5, respectively, is a comparatively high subjective audio quality of the complete system especially at relatively low bit-rates. The most recently standardized ISO/MPEG-D USAC [ISO12] and ETSI/EBU AC-4 [ETSI14, ETSI15] codec frameworks deliver *excellent* signal reconstruction quality on every input item of a diverse test set at the following bit-rates:

- 96 kbit/s stereo (signal configuration 2.0) without parametric MPS 2-1-2 coding,
- 160 kbit/s surround (configuration 5.1) in case of HE-AAC as the predecessor of USAC, except on applause input, as evaluated by the EBU in a large test [EBU07],
- 256 kbit/s for immersive content (configuration 7.1+4) in case of AC-4 [Kjör16].

Moreover, *good* overall audio quality after (partially parametric) decoding is achieved at

- 32–48 kbit/s stereo with USAC and MPS 2-1-2, though not on all items [Neue13],
- 96 kbit/s surround 5.1 via HE-AAC with MPS, again except on applause [EBU07],
- 144–192 kbit/s (channels–objects) immersive 7.1+4 via AC-4 [Kjör16, Purn16].

Results for AC-4 in the range 32–48 kbit/s stereo, or for any ISO/MPEG codec through a dedicated 7.1 or 7.1+4 evaluation have, to the author's knowledge, not been published.

The subjective performance of USAC was additionally compared with that of HE-AAC [ISO09], AMR-WB+ [Mäki05, Sala06], and a virtual codec (VC) constructed from the per-item maximum mean listener scores attributed to the former two codecs [Neue13]. The results, depicted in Figure 2.17, demonstrate that the tested USAC implementation from 2011 outperforms even the highly challenging (theoretical) VC reference, and it does so at every tested operating point (i. e., bit-rate and channel configuration, where the latter is limited to 2.0 stereo in this case). The methodology behind this set of formal listening tests will be described in more detail in Chapter 4. Unfortunately, no comparative inter-codec studies of AC-4's performance were publicly available at the time of writing. It is, however, already safe to conclude that the abovementioned specifications represent the state of the art in perceptual coding of stereophonic and multichannel audio content.

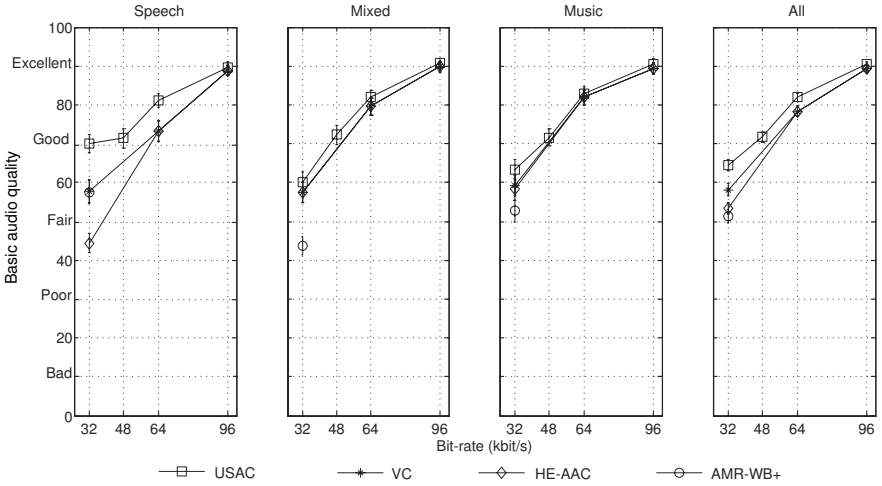


Figure 2.17. Mean absolute scores across all 8 items (per category, 24 total) and 25 listeners for USAC verification test 3 [Neue13]. Vertical bars are 95% confidence intervals.

2.6.2 Advantage of Extensibility to Low-Latency Operation

Both the transform-domain core coding of sections 2.1–2.3 as well as the parametric pre-/post-processing of sections 2.4 and 2.5 can be modified to reduce the algorithmic latency, or delay, exhibited by a corresponding combined implementation. In particular,

- the core frame length N can be reduced, as done in Low Delay (LD) AAC [Alla99],
- the sample rate f_s can, e. g., be doubled, as recently applied to AAC-ELD [Schn16],
- the maximum transform window length can be reduced by decreasing the inter-frame overlap, as realized in the CELT module of Opus [IETF12, Valin10, Valin13],
- the transforms and/or filter banks themselves can be designed differently, e. g., with shorter or asymmetric dedicated base functions [Schul00, Schn08, Vaill08],
- the delay-introducing FD algorithmic parts can be omitted, or replaced by causal (minimum-phase), open-loop, and/or TD alternatives [Schn08, LuVa10, Valin13].

The last two bullet points are especially relevant to the pseudo-QMF-domain parametric HFR and spatial coding tools. In the original AAC-ELD specification, the flexibility of the LD SBR module is restricted over that of HE-AAC (a frame-locked "fixed" time grid h is mandated), and the complex-QMF banks with symmetric prototype filters are replaced by special LD filter banks with asymmetric prototypes [Schn08]. The LD MPS extension for AAC-ELD employs a similar LD filter bank and additionally minimizes the latency via

- TD downmixing of the input channels, allowing parallel core and LD MPS coding,
- omission of the LF "hybrid" sub-band decomposition and the residual E_i coding,
- halving of the encoder-side lookahead window used in the parameter extraction,
- adaptation of the all-pass decorrelation filters to the LD operating environment,
- prohibition of the low-power decoding noted on page 50 (requires extra delay),
- FD interfacing between LD MPS and the AAC-ELD LD SBR tool [Schn08, LuVa10].

By obeying these restrictions, a total encoder-decoder delay of 37.7 ms at $f_s = 48$ kHz is achievable, which is close to the benchmark value of 33 ms introduced in Chapter 1.

2.6.3 Disadvantage of Signal-Dependent Quality vs. High Complexity

Subsection 2.6.1 already indicated that, although modern perceptual codecs deliver surprisingly high *overall* reconstruction quality even upon low-rate coding, the output sounds considerably worse for *specific* input stimuli than for other types of signals. One recording that was found to be challenging in the multichannel evaluations of [EBU07], besides the "Applause" and "R_Plant_Rock" items (where the latter also includes concert applause), was the single-instrument and mostly single-tone "Harpsichord". In general, two classes of difficult input waveforms for perceptual audio codecs can be identified:

- highly stationary recordings of sustained notes with fixed or slowly varying pitch played by brass, wind, or electronic instruments, e.g., horn, harpsichord, trumpet
- highly non-stationary signals with distinct isolated or densely spaced transients especially in the HF region, e. g., applause, claps, rain, castanets, drums, cymbals.

Among these, the stationary signals are usually characterized by a non-flat fine *spectral* envelope, containing the (often closely spaced) individual harmonics of the tonal waveform, whereas the non-stationary ones exhibit a non-flat fine *temporal* envelope, with a (very low or high) number of individual transient peaks standing out in front of a noisy background waveform. At higher bit-rates, the stationary signal class tends to be more problematic due to relatively short frames (e. g., $N = 1024$ in MPEG audio codecs since MPEG-2 AAC [ISO97], leading to a frame length of 21.3 ms at $f_s = 48$ kHz). At low rates and dual-rate SBR (and, possibly, PS or MPS) coding, however, the core frame length is, effectively, increased. This renders the non-stationary inputs more challenging than the tonal ones, as the short transforms for block switching grow proportionally in length.

Naturally, a trivial solution to the issue of frame length increase with a dual-rate SBR system is to switch to a *single-rate* scheme as in downsampled SBR or A-SPX, potentially in combination with a slight reduction of f_s to, say, 32 kHz. While this method raises the temporal coding resolution to a satisfactory level, it brings up another issue: complexity.

To understand why the transition from dual-rate to single-rate coding increases the total computational complexity of the system, it shall be mentioned that such a change, given the higher f_s, leads to more samples being passed to the core coder per second. As the core codec's workload is, typically, approximately proportional to the sampling rate, it, therefore, grows accordingly. The same is true for the auxiliary filter banks (at least some instances thereof) — regardless of whether they are complex-modulated or real-valued — and most of the algorithmic operations of the parametric coding extensions.

It is worth noting, in this context, that in the most commonly applied "level 2" of the HE-AAC v2 profile [ISO09], only a "baseline" version of the PS tool is implemented in an attempt to limit the computational complexity. This baseline version includes a simpler variant of the hybrid filter bank and does not implement IPD/OPD synthesis [DenB09], which has two implications in the present discussion. First, this design decision implies that both the unrestricted hybrid filter bank as well as the phase difference coding are, algorithmically, quite expensive. Given that MPS 2-1-2 supports IPD coding again, it can be concluded that USAC's parametric channel coding is more expensive than the PS tool in HE-AAC v2 "level 2". The preparation and outcome of actual comparative complexity evaluations will be examined later in Chapter 4. Second, the omission of phase coding in MPEG-4 PS implies a performance compromise: for some "phasy" input signals, with a phase difference around, say, 90 – 270° ($\frac{\pi}{2} - \frac{3\pi}{2}$) in certain parameter bands across several frames, the reconstruction quality will be lower with PS than, e.g., IPD-aware MPS 2-1-2 coding. Put differently, the subjective performance is more item-dependent for a codec with "level 2"-like PS coding than with a codec whose parametric spatial coding scheme is phase-aware, i. e., more difficult-to-code inputs are likely to exist in case of the former. In fact, the USAC verification test results depicted in Fig. 2.17 indicate that this is true.

2.6.4 Disadvantage of Lower Limit on Delay vs. Reduced Quality

Subsection 2.6.2 noted that LD SBR [Schn08] and MPS [LuVa10] employ customized complex-valued low-latency filter banks in order to minimize the overall codec delay. The basic shapes and transfer functions of the prototype weighting utilized in such LD filter bank designs are visualized in Figure 2.18. By using these asymmetric prototypes, the total encoding-decoding latency added to a codec by the parametric coding modules is reduced to 256 TD samples, or 5.3 ms at $f_s = 48$ kHz. However, complete *elimination* of this additional source of delay — which is highly desirable in LD applications — is not feasible since this would necessitate the omission of the pseudo-QMF *banks* themselves. This aspect can be regarded as one of the reasons why QMF-domain tools similar to SBR or A-SPX and PS, MPS, or Advanced Coupling cannot be found in very-low-delay codecs.

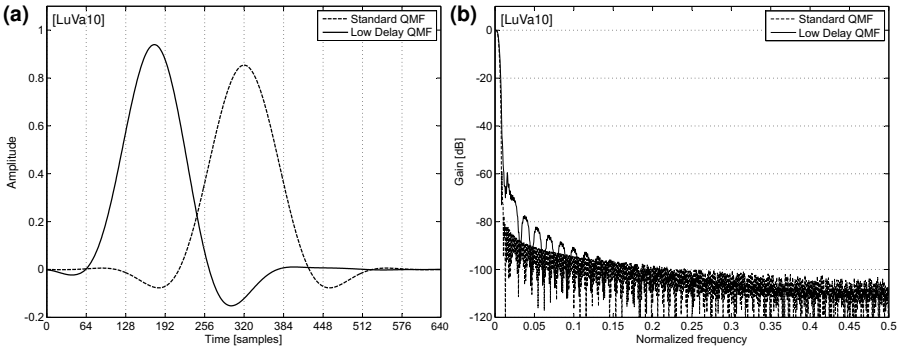


Figure 2.18. Comparison of regular (symmetric) and LD (asymmetric) pseudo-QMF prototypes used in MPS. (a) impulse response, (b) frequency response (Fourier magnitude).

Having established that parametric HFR or spatial coding using auxiliary filter banks imposes a lower limit on the algorithmic delay of a codec in practical implementations, the chapter will conclude with a comparison to one of said codecs which do not employ any QMF-domain techniques: the Opus codec standardized by the Internet Engineering Task Force (IETF) for file-based storage and IP-based streaming or real-time communication [IETF12, Valin13]. This excursion, which only addresses CELT, i. e., the transform coding part of Opus based on [Valin10], has been published by the author in [Helm14].

CELT makes use of the MDCT like HE-AAC, but differs from the latter in some details:

- inter-transform overlap: HE-AAC applies a maximum overlap of 43 ms at $f_s = 48$ kHz (100% of the dual-rate frame length, see also subsection 2.6.3), while CELT utilizes a fixed 2.5 ms overlap regardless of the frame or transform length used.
- block switching for transform length adaptation: both codecs support switching between frames of either one “long” or eight “short” MDCTs, but HE-AAC needs to apply transitory frames, as already noted in section 2.1, while CELT does not.
- algorithmic delay: unlike in CELT, which only needs 2.5 ms of lookahead for the transform overlap, further delay sources exist in HE-AAC, as already described.

In principle, CELT may also be used in offline, non-realtime situations. However, as will be described hereafter, it has a certain drawback which limits the achievable subjective quality on specific audio material to a level below that of the general-purpose HE-AAC and USAC systems, particularly at relatively low bit-rates. Aside from a lack of efficient parametric stereo coding and bandwidth extension, CELT’s performance falls short on very tonal stationary signals due to a low-overlap, near-rectangular transform window.

Beside the bitrate used for coding, the "outer" blocks depicted in Fig. 2.2 exhibit the largest influence on the quality of the decoded audio signals. A graphical comparison of the framing, MDCT, and time-resolution optimization modules of a CELT and AAC-(E)LD encoder is shown in Figure 2.19. It is evident that the shapes of the windows processed by the MDCT differ considerably: a CELT encoder forms nearly rectangular Tukey-like windows, whereas AAC-(E)LD uses much smoother bell-shape windows. Consequently, CELT's window exhibits more spectral leakage than the (E)LD windows, which explains (at least partially) the fidelity and efficiency problems CELT shows on very tonal signals such as recordings of trumpets. However, it has a significant advantage over the (E)LD windows: it enables switching between a *long* and eight *short* transforms like HE-AAC, but without intermediate *start* or *stop* transition windows, thus avoiding an additional block-switch lookahead. In (E)LD, block switching is very difficult to integrate without violating the Princen-Bradley condition [Prin86] for PR in the absence of quantization, which is why the latter codecs only offer one *long* transform length. Although AAC-LD allows to reduce the transform overlap upon detecting a transient [Alla99], the shortest possible time span of its windows still is about 2.5 times longer (e. g., 13.3 ms) than that of CELT's *short* windows (e. g., 5.3 ms), assuming both codecs operate at the same f_s (48 kHz in this case). Since coding errors due to spectral line quantization extend over the complete duration of a reconstructed window [Prin87], pre-echo artifacts caused by a temporal unmasking of coding noise are more likely to arise in AAC-(E)LD than CELT.

It must be restated in this regard that, as illustrated in Fig. 2.19, both coding systems provide means for improving the time resolution of a frame *after* the MDCT stage. In an AAC-(E)LD encoder, FD linear predictive filtering via TNS (section 2.2) can be applied to each spectrum, while in CELT, adjacent lines are optionally subjected to T/F adjustment by means of Hadamard transformation (also section 2.2). As both methods should yield similar levels of pre-echo reduction, CELT's advantage on transient frames due to block switching support remains since T/F adjustment can also be used on *short* windows.

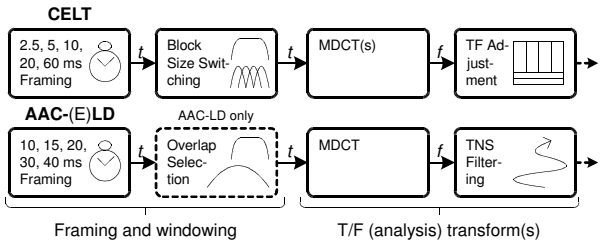


Figure 2.19. Illustration of the differences between CELT and AAC-LD or AAC-ELD in terms of framing, windowing, filter bank design, and T/F resolution optimization [Helm14].

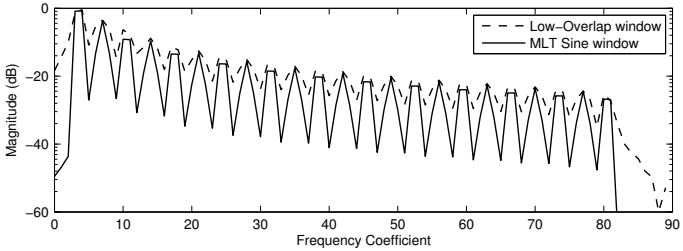


Figure 2.20. MDCT magnitude spectrum of low-pitch harmonic signal using different windows.

In *long* windows, CELT aims at preventing musical noise by means of "spreading", a line-wise Givens rotation as a pre- and post-processor around the spectral VQ [IETF12, Valin13]. The activation of this tool is based on a FD measure of frame tonality and a bit sensitive to misdetection, potentially leading to audible noise in the decoded signal. As the frequency selectivity (i. e., stopband rejection) of the near-rectangular *long* windows is rather low, this is not surprising: as depicted in Figure 2.20, a MDCT of closely spaced harmonics resembles one of noise, implying that the transform coding gain (i. e., amount of decorrelation) is lower than with a full-overlap window. Even a subtle application of spreading to tonal inputs — which, due to the lower transform selectivity, appear somewhat noisy in the spectral domain — will, thus, most likely lead to audible quality loss.

It is worth emphasizing that LD coding with long inter-transform overlap and good overall quality is possible, as demonstrated by AAC-LD [Alla99] and AAC-ELD [Schn08]. These two transform coding schemes, however, do not offer block switching capability, and TNS alone can, in the experience of the present author, not fully compensate for the lack of *short* transforms on some strongly transient signal passages. To summarize, the following two statements can be made regarding the previously described observations:

- Using QMF-domain parametric coding tools in LD applications ensures, overall, *good* audio quality even at low rates but imposes additional delay on the codec. Omission of said pseudo-QMF tools avoids the extra delay but implies a quality compromise typically leading to reduced subjective performance at low bit-rates.
- Low-overlap transform windows simplify block length switching and render an additional lookahead unnecessary, but cause reduced coding gain on stationary waveforms, especially tonal ones. High-overlap and potentially asymmetric long transform windows, in turn, are more efficient on stationary signal portions but complicate TDAC-compliant LD block switching without extra input lookahead. The aspect of block detection lookahead will be revisited in the next chapter.

3 Contributions for Flexible Transform Coding

The last chapter introduced the fundamentals and motivations behind the individual components of modern perceptual transform codecs and identified four shortcomings:

- a difficulty of coding both highly tonal and highly transient material equally well,
- a difficulty of utilizing block switching techniques without additional lookahead,
- a difficulty of applying QMF-domain parametric coding tools in LD applications,
- a general tendency of the parametric tools towards high algorithmic complexity.

These issues can be observed, to varying extents, on state-of-the-art transform codecs, namely, MPEG-4 HE-AAC with PS [ISO09, DenB09] and AAC-ELD with LD MPS [Schn08, LuVa10], MPEG-D USAC with MPS 2-1-2 [ISO12, Neue13], and CELT as part of the IETF Opus codec [IETF12, Valin13]. It is likely that some of the drawbacks also apply to the recently standardized AC-4 core codec, but due to the present lack of publicly available encoder/decoder implementations, this cannot be verified by the author of this work.

To address the four abovementioned disadvantages of recent codec specifications, a number of scientific and engineering contributions were developed or co-developed by the author, which are presented, in respective sections, in this chapter. These include

- an alternative block switching design requiring almost zero lookahead (sec. 3.1),
- a fully flexible filter bank supporting kernel and overlap ratio switches (sec. 3.2),
- an intra- and inter-channel FD prediction method with low complexity (sec. 3.3),
- transform-domain *semi-parametric* SBR-like (sec. 3.4) and PS-like (sec. 3.5) tools.

The basic objective of these contributions, when used in combination, is the realization of a fully flexible perceptual transform coding architecture providing both conventional and LD block switching capability as well as fundamental parametric HFR and MPS-like spatial coding techniques directly within the transform domain. Thereby, any additional algorithmic delay — and, possibly, some of the computational complexity — due to said codec components can be avoided while, hopefully, the corresponding perceptual benefits can be largely maintained. An appropriate evaluation will be discussed in Chapter 4.

3.1 Low-Latency Block Switching with Minimum Lookahead

Since the development of MP3 [ISO93], block switching capability has found its way into virtually every general-purpose perceptual transform codec. In CELT and all MPEG audio codecs since MPEG-2 AAC [ISO97], the encoder can choose between single-MDCT frames, comprising the coefficients of one *long* transform spanning across the entire TD support of the frame, and multiple-MDCT frames containing the interleaved coefficients of eight temporally successive *short* transforms [Bosi97]. The window shape definitions for, and temporal locations of, these *long* and *short* MDCTs, providing $N_{\text{long}} = 1024$ and $N_{\text{short}} = 128$ spectral coefficients, respectively, in AAC and later codecs, are documented in [Edler89] and illustrated in Fig. 2.5. The interleaving, in contrast, is described in the context of grouping in [Bosi97] and exemplified in Fig. 2.9. To switch from single-MDCT *long* frames to an eight-MDCT *short* frame in AAC requires an intermediate single-MDCT *start* transition frame utilizing an asymmetric window to prepare for the short overlap (of length N_{short}) of the first *short* MDCT. Hence, to allow the insertion of said transitory frame in a timely manner, i. e., before the arrival of the non-stationary signal portion to be coded using *short* transforms, an encoder-side "block detection" lookahead of length

$$N_{\text{lookahead}} = \frac{N_{\text{long}} + N_{\text{short}}}{2} \tag{3.1}$$

becomes necessary [Alla99, Lutz04]. To revert to symmetric *long* transforms after eight *short* ones, a transitory *stop* frame applying a single MDCT with the temporal reverse of the asymmetric *start* window is inserted. MPEG-D USAC [ISO12] additionally supports a *stop-start* single-MDCT frame in between two eight-*short* frames, making use of a symmetric low-overlap window whose ends are PR-compatible to length-N_{short} transforms. Specifically, this low-overlap shape represents the "outer" envelope of eight consecutive overlapping *short* windows of size $2N_{\text{short}}$ each, with a non-zero center of size $9N_{\text{short}}$.

It is worth noting that the exclusive usage of *stop-start* and eight-*short* frames avoids the necessity of transitory frames (since direct TDAC-compliant switching between the two frame types is possible) and, thus, eliminates the need for block-switch lookahead (reducing $N_{\text{lookahead}}$ to zero). Such a LD configuration is utilized in CELT, with the *short* window slopes specified by the Vorbis function w_{vorbis} of (2.17) in Chapter 2. However, the application of a low-overlap window shape on quasi-stationary signals — especially tonal waveform portions — is, as explained in subsection 2.6.4, inefficient and should be avoided whenever possible. In other words, high-overlap windows such as AAC's *long* w_{sine} of (2.15) or w_{kbd} of (2.16) are preferable in this case. In LD scenarios, a minimum block switching lookahead around $N_{\text{lookahead}} = 0$ is still desirable, so "fast" transitions from high-overlap *long*-transform to *short*-transform frames are worth an investigation.

In the following, an alternative to AAC's block transition technique, applying the same basic frame, transform, and window types, is devised. This LD switching method requires

$$N_{\text{lookahead}} = \frac{N_{\text{short}}}{2} \tag{3.2}$$

TD samples of additional encoder-side lookahead, which is much less than the value of (3.1) exhibited by the conventional scheme and which amounts to only 1.33 ms for the AAC transform sizes and $f_s = 48$ kHz. At the same time, the proposal is TDAC compliant like the conventional transitions, thus allowing PR in the absence of quantization. Parts of the following discussion have been presented by the author in [Helm14, Hel15d].

To develop the LD block switching proposal, consider the temporal shape of the *start* window noted on the previous page and illustrated in Figure 3.1(a) in frame $i-1$. With this asymmetric transitory shape, the overlap range between frames $i-1$ and i can be reduced to the length necessary for TDAC compatibility with the first of the eight *short* transforms of frame i. For the exemplary location of a transient (i. e., non-stationarity), depicted in Fig. 3.1 by a dashed vertical line, the maximum pre-echo duration (gray left-pointing arrow) is restricted to the TD support of the first *short* window containing the transient, i. e., the sixth of the sequence in Fig. 3.1(a). Setting $N_{\text{lookahead}}$ (black bar) to a value near zero with this AAC-type scheme, the transient can only be detected in frame i, which is too late for switching to eight *short* transforms. Hence, to maintain full TDAC, only the *long* and *start* frame types can be utilized in i in this case, as shown in Figures 3.1(b) and (c), respectively, so the maximum pre-echo duration grows considerably.

Closer inspection of the *start* window in Fig. 3.1(c) reveals a flat unity-gain portion of size $(N_{\text{long}} - N_{\text{short}})/2$ within which the transient onset is located and which, by design, does not overlap with the upcoming window in $i+1$. Note that three overlapping *short* transforms can be placed at the location of the flat portion such that the right-half slope of the last of the three windows is at the exact same position as the falling slope of the start transform. In other words, a TDAC-compliant separation of the *start* frame into a leading *medium*-sized transform of length $5N_{\text{short}}$ and three trailing *short* transforms of length N_{short} each can be accomplished. In this way, $N_{\text{lookahead}}$ can safely be reduced to a value of $3N_{\text{short}}/2$, and the pre-echo is restored to the desirable range of Fig. 3.1(a).

The proposal now is to add a fourth *short* transform to the left of the previous three, as shown in Fig. 3.1(d), thereby reducing the *medium* transform to a convenient power-of-two size of $N_{\text{long}}/2$ and $N_{\text{lookahead}}$ to the target of (3.2). Obviously, the left half of this additional *short* transform overlaps with not only the *medium* window but also the *long* window of frame $i-1$. However, full TDAC can still be guaranteed by executing the filter bank operations (windowing, TDA, transform, OLA) in a specific order, as shown below.

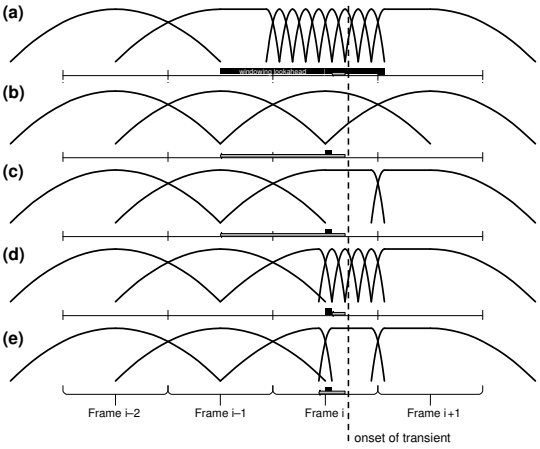


Figure 3.1. Block switching designs for AAC and USAC using different durations of (■) the required $N_{\text{lookahead}}$ and (■) the resulting maximum pre-echo for the exemplary transient location. (a) default block switching, (b) no block switching, (c) window shape switching, (d) and (e) LD proposals.

The crucial aspect to ensure TDAC (and, thereby, the possibility of PR) in the "double overlap" region at the center of the proposed LD *start* window is a separation into *outer* and *inner* filter bank processing. The former provides TDAC on an *inter*-frame level, i. e., between the overall *start*-shape windowed PCM waveform of the current frame at i and the *long*-shape weighted waveform at $i - 1$. The latter part, in turn, deals with TDAC on an *intra*-frame level, i. e., between all individual transforms of i which, in the present LD investigation, comprise five instances of *medium* or *short* length. Another characteristic which is essential in this context is the separability of both the direct and inverse transform operations into temporal mirroring, or "folding", processes for the purpose of TDA handling (with proper symmetry, more on this in the next section) and non-overlapped DCT- or DST-type *core* transforms of the aliased "folded" signals. The necessary order of the filter bank operations for the analysis and synthesis case can be specified as follows:

- In the encoder-side analysis process shown in Figure 3.2, asymmetric windowing as for a typical *start* frame is performed first, followed by the introduction of the *outer* TDA by means of "folding-in" of the outer waveform parts (dashed). In a regular frame, the procedure would now complete with the actual length-N_{long} (i. e., N) DCT or DST of the resulting TDA signal to acquire the FD coefficients, as illustrated for the *long* frame at $i - 1$. In the LD frame proposal at i, however, the *start*-windowed "fold-in" result is further divided into the desired five segments by applying the same operations — windowing and TDA generation — again, but on the smaller intra-frame scale. This is the step labeled "*inner* fold-in aliases" in Fig. 3.2, which is finally succeeded by the individual non-overlapped transforms.

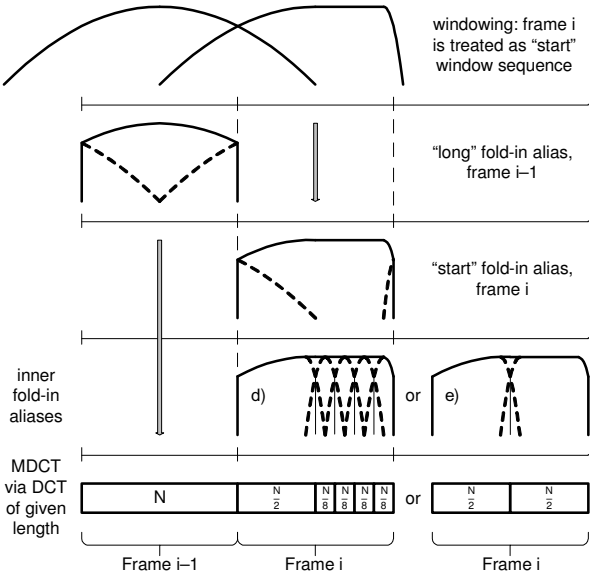


Figure 3.2. TDAC-preserving order of operations for the LD *start* frame proposals when employed in frame *i* on the encoder (analysis) side. *N* indicates the frame length.

- In the decoder-side synthesis algorithm depicted in Figure 3.3, all encoder steps are inverted (or undone, depending on the process) in reverse order. This means that the non-overlapping inverse core transforms are performed first, with one *medium* and four *short* synthesis transforms being applied in case of the LD *start* frame. Thereafter, intra-frame TDAC is carried out using "folding-out", synthesis windowing, and OLA between the four *short* and the *medium* transform outputs, as summarized by the "*inner* fold-out aliases" step in Fig. 3.3. The result in i is a preliminary signal which is free of *inner* TDA, i. e., that represents a conventional *start*-frame output of a synthesizing length-*N* core transform step. To cancel the remaining *outer* TDA, legacy *long* "fold-out" aliasing, followed by *start* synthesis windowing and, finally, OLA with the past frame at $i - 1$ can now be employed.

From this description it should be evident that the separate filter bank steps remain exactly the same as in state-of-the-art AAC or USAC processing; they are merely carried out in a nested fashion and a specific order. Furthermore, the optimized *long* and *short* window slopes — whose coefficients do not exceed unity — can be reused, with the *start* (and *stop*) shape specified as a combination of the former, as known from the literature [Edler89, Bosi97]. These two properties represent a clear advantage over an alternative proposal [Phili08, Viret08], revealing two drawbacks which the present design does not:

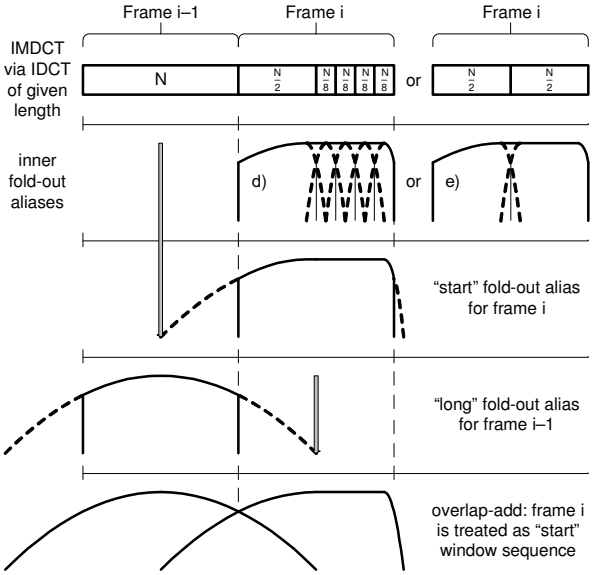


Figure 3.3. TDAC-preserving order of operations for the LD block switching proposals on the decoder (synthesis) side.

- two extra window shapes, whose coefficients require static memory, are utilized,
- one of the added window shapes (w_1 in [Phili08]) exceeds a value of one, leading to unwanted inefficiency due to amplification of the coding error in the decoder.

Concluding this section, it is noted that the *medium* and *short* transform coefficients of the presented LD *start* frame (which, again, appears as a regular *start* transition from the perspective of adjacent frames) can be processed jointly like the spectral samples of a *short* block. More specifically, the window grouping approach of subsection 2.3.1, also described in [Bosi97], can be applied, with the $4N_{\mathrm{short}}$ samples of the *medium* transform treated as a fixed "short" group of length 4. The remaining four transforms can be combined arbitrarily into either one (of length 4), two (1-3, 2-2, or 3-1), three (1-1-2, 1-2-1, or 2-1-1), or four groups (1-1-1-1), depending on the input waveform and/or bit-rate.

In some cases, it may be desirable to simplify the LD block switching design so that, in the LD *start* frame, only transforms of equal length are employed. This can be accomplished by replacing the four *short* instances with a single low-overlap variant of length $4N_{\mathrm{short}} = N_{\mathrm{long}}/2$, similar to USAC's *stop-start* window, as illustrated in plots (e) of Figs. 3.1–3.3. By way of this substitution, the left- and right-half transform lengths of the LD *start* frame are synchronized at the cost of a slight increase in the worst-case pre-echo duration (Fig. 3.1). This type is integrated into the MPEG-H 3D Audio standard [ISO15a].

3.2 A Flexible Cosine- and Sine-Modulated TDAC Filter Bank

The previous section demonstrated that, by separating and recursively applying the windowing, TDA, core transform, and OLA steps of the filter bank operation, an efficient LD-optimized alternative to the conventional block switching approach can be realized. This section provides a detailed examination of the TDA and core transform algorithms themselves and develops therefrom a more flexible, generalized filter bank design. The motivation behind this modified design is improved coding gain on some "phasy" stereo signals and highly harmonic input, as described by the author in [Hel15c] and [Hel16a], respectively. A consolidation of this work has been published by the author in [Hel16b].

3.2.1 Signal-Adaptive Transform Kernel Switching for Stereo Audio Coding

As noted earlier, all contemporary perceptual audio codecs, including Opus [IETF12], the (Extended) HE-AAC family [ISO09, ISO12], and the new MPEG-H 3D Audio [ISO15a] and 3GPP EVS [ETSI16] codecs, apply the MDCT for FD quantization and coding of one or more channel waveforms. Utilizing the MPEG notation, the synthesis version of this lapped transform, given a length-N decoded spectrum $X_i(k)$, can be specified as in (2.9) of section 2.1, where $M = 2N$ shall indicate the time-window length. After the synthesis windowing process, the first half of the TD result $\bar{y}_i$ is combined with the second half of the last frame's result $\bar{y}_{i-1}$ using OLA, as in (2.12), yielding a TDA-free waveform for i.

Input signals comprising more than two channels are often separated into individual single-channel elements (SCEs) or channel-pair elements (CPEs), which are processed independently. The CPEs, containing, e. g., the left and right channels of a stereophonic signal, support joint-channel coding of the two (possibly grouped) MDCT frame spectra for greater efficiency, based on the intensity and M/S stereo coding paradigms [Vand91, John92] introduced in subsection 2.2.3. The 3D Audio codec, in particular, provides the already described complex-prediction stereo coding tool [Helm11] which unifies — and extends — the former two methods for high-quality transform coding even at low rates.

At 96 kbit/s stereo, the combination of M/S and complex-prediction stereo coding in USAC, also known as Extended HE-AAC [Neue13] or just xHE-AAC, was shown to enable *excellent* audio quality on every signal tested [Helm11, Neue13]. Lowering the bit-rate to about 48 kbit/s, it is desirable to maintain at least *good* audio quality on the same set of signals, but it was found that, for some material, the quality dropped below the *good* range, i. e., to *fair*. Further investigation identified the predictive stereo tool as the likely origin of this issue; within its algorithm, the MDST estimates computed for the complex-valued predictor only *approximate* the actual MDST downmix, as depicted in Figure 3.4.

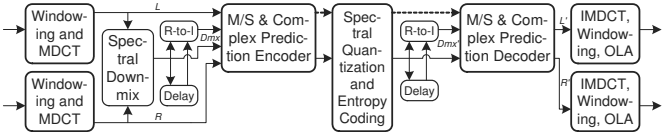


Figure 3.4. Signal flow in USAC with modules, inputs, and outputs of predictive M/S stereo encoder and decoder.

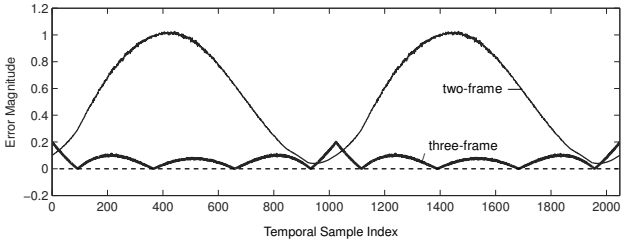


Figure 3.5. Error in MDST approximation via (—) two-frame and (—) three-frame R-to-I method in complex-valued stereo prediction on white Gaussian noise input as a function of in-frame location, averaged across 20 frames. MLT with N=1024.

More specifically, a real-to-imaginary (R-to-I) procedure using the current and previous frame's downmix of the left (L) and right (R) MDCT spectra is carried out to obtain the corresponding current MDST downmix as the imaginary part of the predictor; see also [Helm11] and subsection 2.2.3. The slight inefficiency of this approximation, illustrated in Figure 3.5 via an assessment of the average estimation error across a frame, becomes most noticeable when said imaginary part strongly contributes to the coding, i. e., when its associated predictor coefficient (or gain) ρ_i^{im} lies near ± 1. In fact, simulations show that the estimation — and, thereby, the coding quality — improves considerably in such cases if the *next* frame's MDCT downmix is included in the R-to-I process. This solution, however, increases both the algorithmic delay (by one frame due to the necessary extra lookahead) and computational complexity (by a factor of 1.6 in the complex prediction module) of the codec [Helm11]. A more practical alternative is, therefore, desirable.

In the following, a low-complexity (LC) amendment to the complex-valued predictive joint-stereo coding technique is proposed which, as a welcome side effect, can be implemented without increasing the codec delay. To this end, the foundation of modern audio coding — the lapped transform coding approaches according to Princen *et al.* — as well as the abovenoted stereo coding issue are revisited. Thereafter, a generalization of said lapped transform paradigm is formulated and discussed, and, lastly, it is demonstrated how the proposed generalized scheme can be employed in a signal-adaptive fashion to improve the joint-stereo coding of critical two-channel input. The designs and results of blind listening tests conducted in the course of this study will be examined in Chapter 4.

The MDCT, specified in (2.9) for the inverse and (2.8) for the forward (analysis) case, respectively, is based on a 1987 paper by Princen *et al.* [Prin87]. Therein, an overlapped usage of the transform in successive signal frames was shown to yield an *oddly* stacked filter bank design, named such due to the offset $k_0 = ½$ applied in its base functions; see section 2.1. A similar transform is realized by the MDST of (2.10) and (2.11), mentioned previously in connection with complex stereo prediction, which differs from the MDCT of (2.8) and (2.9) only in that it makes use of the $\sin(\cdot)$ instead of the $\cos(\cdot)$ function.

There also exist two further transforms which can be used in an alternate fashion to construct an *evenly* stacked filter bank system [Field96]. As described in section 2.1, the inverse (synthesis) variants of these transforms, referred to as *DCT based* and *DST based* in [Prin86], can be defined such that they differ from the MDCT and MDST, respectively, only in the choice of k_0 which, as will be examined in greater detail below, now takes an integer value. For better distinction and due to their correspondence to the type-II DCT and DST utilized in image and video coding, the evenly stacked filter bank transforms of [Prin86] will hereafter be referred to as MDCT-II and MDST-II, respectively. Accordingly, the oddly stacked transforms of [Prin87, Mal90b] will be called MDCT-IV and MDST-IV, due to their relation — and reducibility (by separating the TDA "folding" and core transform operations, as in the last section) — to the type-IV DCT and DST, respectively.

In CPEs instead of SCEs, windowing and MDCT-IV processing is typically performed separately for each channel in the encoder before a M/S stereo operation is carried out to attempt channel compaction, i. e. the derivation of a downmix spectrum *Dmx* having maximized energy and a corresponding residual spectrum *Res* with minimized energy. This approach, which in case of real-valued stereo prediction (no imaginary component, i. e., $\rho_i^{\mathrm{im}} \stackrel{\mathrm{def}}{=} 0$) is defined by (2.35) or, maintaining the notation of the present discussion,

$$Dmx_i = \frac{L_i \pm R_i}{2}, \quad Res_i = \frac{L_i \mp R_i}{2} - \rho_i^{\mathrm{re}} \cdot Dmx_i, \tag{3.3}$$

with *L* and *R* representing the left and right channel's MDCT-IV coefficients [Helm11], is indicated in Figure 3.4. It works well for stereo input signals whose channel spectra are roughly in-phase (0 degrees of IPD) or out-of-phase (180 degrees of IPD). However, for frames with approximately 90 or 270 degrees of IPD in dominating LF spectral regions, the joint-stereo coding approach of (3.3) fails to provide any useful channel compaction into *Dmx* (and, thus, does not offer any quality advantage over separate coding of *L* and *R* in said frequency regions). Moreover, the two-frame complex predictor, being a compromise between complexity, delay, and efficiency as illuminated earlier, does not yield much improvement either in this situation. In other words, the energy of *Res* will not be much lower than that of *Dmx* because the correlation between *L* and *R* approaches zero.

To ameliorate the previously described issue without having to resort to three-frame prediction, it is noteworthy that the four lapped transform types — MDCT-IV, MDST-IV, MDCT-II, and MDST-II — are merely realizations of a unified, general formulation where

$$X_i(k) = \frac{1}{s(k)} \sum_{m=0}^{M-1} \hat{x}_i(m) \, \mathrm{cs}\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \quad 0 \le k < N, \tag{3.4}$$

specifies the forward (analysis) case and where the inverse (synthesis) case is defined as

$$\hat{y}_i(m) = \frac{2}{N} \sum_{k=0}^{N-1} X_i(k) \, \mathrm{cs}\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \quad 0 \le m < M. \tag{3.5}$$

In other words, all four transforms can be derived from (3.4), (3.5) by proper selection of $\mathrm{cs}(\cdot)$ — representing the $\cos(\cdot)$ or $\sin(\cdot)$ function — and $k_0, s(k)$. Specifically, by using

- $\mathrm{cs}(\cdot) = \cos(\cdot)$, $k_0 = ½$, and $s(k) = 1$ to obtain an MDCT-IV,
- $\mathrm{cs}(\cdot) = \sin(\cdot)$, $k_0 = ½$, and $s(k) = 1$ to obtain the MDST-IV,
- $\mathrm{cs}(\cdot) = \cos(\cdot)$, $k_0 = 0$, and $s(k) = 1 + \delta(k)$ for an MDCT-II,
- $\mathrm{cs}(\cdot) = \sin(\cdot)$, $k_0 = 1$, and $s(k) = 1 + \delta(k')$ for an MDST-II,

with $\delta(\cdot)$ being the Kronecker delta and $k' = N - 1 - k$ (for scaling of the Nyquist bin), the four transforms can be made TDA compatible. This means that, with proper choice of the kernel parameters between successive coding frames, the TDA can be canceled in the OLA step, thereby allowing PR of the PCM waveform input in the absence of spectral quantization. For completeness it is noted that this generalization is independent of the applied window shapes and lengths according to previous discussions herein, and that fast implementations of each transform type are possible [Mal90b, Field96, Brita03].

Figure 3.6 illustrates the TDAC property of the generalized transform scheme by way of the depicted TDA symmetries for each transform's left and right half. An upward TDA slope indicates even symmetry (bump, no sign change), while a downward slope stands for odd symmetry (valley, sign change). It can be stated that, generally, TDAC is attained between two consecutive lapped transform instances when, within their overlap region, one of them is evenly and the other is oddly symmetric [Prin86, Prin87, Mal90b]. It is easy to notice that this characteristic is the case between all frames shown in Fig. 3.6.

Going back to the IPD issue examined on the previous page, it is worth noting that a modulated complex lapped transform (MCLT), having similar properties as the discrete Fourier transform (DFT), can be constructed by employing the MDCT-IV as its real and the MDST-IV as its imaginary component [Malv99]. Utilizing this MCLT in each channel for frame-wise spectral analysis of the input waveform, the following can be observed:

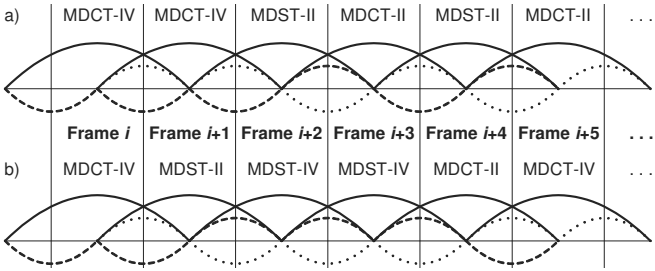


Figure 3.6. Illustration of two lapped transform sequences with TDAC. (—) window shape and TD support, associated (- -) left- and (···) right-half TDA symmetries.

- For IPDs 0 or 180 degrees, the FD coefficient correlations between the real parts and between the imaginary parts of the two MCLTs exceed the cross-correlations between one channel's real MCLT and the other channel's imaginary MCLT part.
- For IPDs near 90 or 270 degrees, the reverse is true, i. e., said cross-correlation is larger in magnitude. This is expectable, as a phase shift of $\frac{\pi}{2}$ also exists between the cosine-modulated MDCT-IV and the sine-modulated MDST-IV base functions.

Although sporadically used for lapped transform coding [Yoon08], the MCLT is not very attractive for this purpose due to the inherent oversampling by two. However, the above findings indicate how the properties of the MCLT can be exploited in the present study. To be specific, if the real-part/imaginary-part cross-correlations exhibit greater magnitude than the real-only or imaginary-only correlations (i. e., the frame's overall IPD lies closer to 90 or 270 than to 0 or 180 degrees), it seems appropriate to try to "adjust" one channel's transform so that its real and imaginary MCLT parts are switched. Given that the conventional MCLT has the MDCT-IV as its real part, this implies that the "adjusted" MCLT takes the MDST-IV as the real part and the MDCT-IV (possibly with FD sign flips) as the imaginary part. Hence, in the context of transform coding, an overall frame IPD of $\pm\frac{\pi}{2}$ can be converted into an IPD of $\pm\pi$ by simply employing the MDCT-IV in one channel and the MDST-IV in the other channel. A respective modification of the block diagram of Fig. 3.4, integrating this "transform kernel switching" — or, as it shall be abbreviated in the remainder of this work, kernel switching (KS) — concept, is depicted in Figure 3.7.

Note that, as described on the previous page and by way of Fig. 3.6 above, transitions from MDCT-IV to MDST-IV coding in one channel, triggered by the KS detector shown in Fig. 3.7, can be chosen TDAC compatibly so as to maintain the possibility of PR (proper windowing according to section 2.1 assumed). Moreover, conversion from a 90- or 270-degree "phasy" channel pair into a mostly 180-degree (i. e., out-of-phase) transform pair can be prevented through an appropriate selection of the channel in which the switch to MDST-IV coding is to be carried out. Both of these specifics will be clarified hereafter.

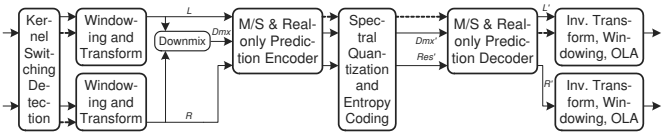


Figure 3.7. Joint-stereo transform coding using real-only prediction and kernel switching proposal in filter bank module.

It must also be pointed out that the complex-valued stereo prediction of Fig. 3.4 has been substituted by the simpler real-valued prediction variant in Fig. 3.7 to reduce the overall codec complexity. This change could, in principle, compromise the performance of the codec in terms of subjective coding quality, especially on "phasy" input. However, the KS design is expected to (at least partially) compensate for the reduced flexibility of the simplified predictive M/S coding, and it provides the option of joint operation with full complex predictive stereo coding without adding much decoding complexity. Actual perceptual evaluation of both designs has been conducted, as reported in Chapter 4.

Implementations of the KS concept should be signal-adaptive with per-frame or even per-transform resolution so that sudden changes in the signal's IPD characteristics can be followed. Furthermore, they should be compatible with other coding tools like block switching (transform length selection) and window switching (overlap range selection) without requiring modifications to these codec tools. For the sake of clarity and brevity, the latter aspect is omitted in the following discussion, and focus is laid on the situation of "one long transform per frame", i. e., only *long* frames. Block switching compatibility was, however, implemented during this work, as will be described in subsection 3.2.3.

- Decoder specification and processing. Since the KS proposal requires the transmission of the kernel types in each frame at index i in order to communicate the appropriate inverse transforms, a generic decoder architecture for codecs such as USAC [ISO12] may be constructed as follows. To indicate switching between the four kernel modes, it suffices to transmit one additional bit per channel and i that signals whether the right-side TDA symmetry $symm_i$ is even (value 0) or odd (value 1). The left-side symmetry does not need to be conveyed explicitly as it depends on the right-side symmetry $symm_{i-1}$ of the already transmitted and decoded last frame at $i-1$. The decoder reads the extra bit of each channel and, utilizing a mapping function implementing Table 3.1, derives the required mode parameters $\mathrm{cs}(\cdot)$ and k_0. Once these mode values have been obtained (and $s(k)$ has been determined therefrom), spectral decoding — including any noise filling or stereo processing — is applied as usual. Then, generalized inverse transforms according to (3.5) are being applied for each channel, followed by the traditional final steps of TD synthesis windowing and OLA processing, as seen in Fig. 3.7.

current frame → **last frame ↓**	right-side symmetry even ($symm_i = 0$)	right-side symmetry odd ($symm_i = 1$)
right-side symmetry even ($symm_{i-1} = 0$)	MDCT-IV: $\mathrm{cs}(\cdot) = \cos(\cdot)$ $k_0 = 0.5$	MDST-II: $\mathrm{cs}(\cdot) = \sin(\cdot)$ $k_0 = 1.0$
right-side symmetry odd ($symm_{i-1} = 1$)	MDCT-II: $\mathrm{cs}(\cdot) = \cos(\cdot)$ $k_0 = 0.0$	MDST-IV: $\mathrm{cs}(\cdot) = \sin(\cdot)$ $k_0 = 0.5$

Table 3.1. Mapping signaled symmetry data of current (i) and last ($i-1$) frame to kernel modes.

- Encoder design and implementation. Figure 3.6 indicated how the lapped transform kernels must be selected in successive frames so that all TDA components can be canceled during the OLA step after inverse transformation in the decoder. More specifically, an MDCT-IV in i must be followed by an MDCT-IV or MDST-II in $i + 1$, whereas an MDST-IV in i must be succeeded by an MDST-IV or MDCT-II in $i + 1$. Moreover, the MDCT-II and MDST-II are allowed to alternate between consecutive frames, as in the original evenly stacked filter bank design [Prin86], but they may also be followed by a type-IV transform, as is evident from Fig. 3.6. The latter property represents the key functionality enabling KS from MDCT-IV to MDST-IV coding, and vice versa. Clearly, the sequences applied in the decoder (i. e., synthesis filter bank) must also be used in the encoder (i. e., analysis filter bank). In other words, the transform kernel sequencing of Fig. 3.7, signaled via the *symm* vector, needs to be utilized identically in the forward transforms. The remaining issue to address is robust, rarely "toggling" kernel mode detection, as KS between MDCT-IV and MDST-IV necessitates a type-II transitory transform. Based on this detection and a look-up via Tab. 3.1, said *symm* vector can then be obtained and coded into the bit-stream using 2 bit per frame (one per channel).

A simple transform kernel detector can be constructed in the encoder by applying a DFT (with a rectangular window) or, alternatively, MCLT (with a sine or a KBD window, see also section 2.1) on the input signal of each channel and frame i and by determining

- a sample correlation r_i^0 between the real parts of the DFTs/MCLTs as a measure of in- and out-of-phase strength for traditional MDCT coding (0° and 180° shifts),
- a sample correlation r_i^{90} between the real part of one DFT/MCLT and the imaginary part of the other DFT/MCLT as a measure of "KS phase" (90° and 270° IPD).

However, more temporally stable results, with less value fluctuation on quasi-stationary input, can be achieved by utilizing both the real and imaginary parts in r_i^0 as well as r_i^{90}:

$$r_i^0 = \frac{\sum_{k=4}^{K}(\mathrm{Re}\{L_i(k)\}\cdot\mathrm{Re}\{R_i(k)\}) + \sum_{k=4}^{K}(\mathrm{Im}\{L_i(k)\}\cdot\mathrm{Im}\{R_i(k)\})}{\sqrt{\sum_{k=4}^{K}(\mathrm{Re}\{L_i(k)\})^2\cdot\sum_{k=4}^{K}(\mathrm{Re}\{R_i(k)\})^2} + \sqrt{\sum_{k=4}^{K}(\mathrm{Im}\{L_i(k)\})^2\cdot\sum_{k=4}^{K}(\mathrm{Im}\{R_i(k)\})^2}}$$

$$r_i^{90} = \frac{\sum_{k=4}^{K}(\mathrm{Re}\{L_i(k)\}\cdot\mathrm{Im}\{R_i(k)\}) - \sum_{k=4}^{K}(\mathrm{Im}\{L_i(k)\}\cdot\mathrm{Re}\{R_i(k)\})}{\sqrt{\sum_{k=4}^{K}(\mathrm{Re}\{L_i(k)\})^2\cdot\sum_{k=4}^{K}(\mathrm{Im}\{R_i(k)\})^2} + \sqrt{\sum_{k=4}^{K}(\mathrm{Im}\{L_i(k)\})^2\cdot\sum_{k=4}^{K}(\mathrm{Re}\{R_i(k)\})^2}}$$

where L_i and R_i denote the DFT/MCLT spectra of the left and right channel, respectively. Note that the summations start at $k = 4$ to suppress undesired effects due to DC offsets or LF hum often present in natural or musical recordings. Furthermore, it is beneficial to apply a limit $K \ll N$ to focus the detection onto spectral regions in which the human auditory system is most sensitive to phase. In the present study with $N = 1024$ [ISO12], a value of K representing a bandwidth between 2.25 and 3 kHz was found to work well.

Having acquired r_i^0 and r_i^{90}, the necessity of switching to MDST-IV processing in one channel can be determined from a conditional difference between their absolute values:

$$d_i = \left|r_i^{90}\right| - \left|r_i^0\right| \quad \text{if} \quad \left|r_i^{90}\right| > \beta, \;\; d_i = 0 \quad \text{otherwise,} \tag{3.6}$$

with $\beta = \frac{9}{16}$ chosen empirically for the present investigation. In frames for which $d_i > 0$, MDCT-MDST coding, i. e., MDCT in one, MDST in the other channel, is applied as follows:

- If $r_i^{90} > 0$, MDST coding is selected in the left channel: the MDST-IV kernel mode is utilized upon odd left-side symmetry, otherwise the MDST-II mode is chosen. Hence, if an MDCT-IV was used in frame $i-1$, an MDST-II is enforced in frame i.
- If $r_i^{90} < 0$, MDST coding is selected in the right channel: again the kernel mode is dependent on the left-side window symmetry (type IV if odd, type II otherwise).

Accordingly, for the channel in which MDST coding is not applied, an MDCT-II kernel is used in case of odd left-side symmetry, otherwise the (default) MDCT-IV configuration is employed. Hence, if an MDST-IV was utilized in frame $i-1$, an MDCT-II is now taken in frame i. This algorithm ensures maximum signal compaction into *Dmx* (i. e., in-phase operation) and minimum reversion of the stereo prediction direction [Helm11, ISO12] (i. e., out-of-phase occurrences). Figure 3.8 illustrates the behavior of the KS detection algorithm on a concatenated set of PCM test material sampled at 48 kHz. The temporal characteristics of d_i lead to the conclusion that the detector succeeds in identifying the phase critical sections of the depicted input signals. The same set of signals is also used for subjective evaluation, which, as noted earlier, will be described in Chapter 4. The KS approach is integrated into the "Phase 2" amendment of the 3D Audio standard [ISO16].

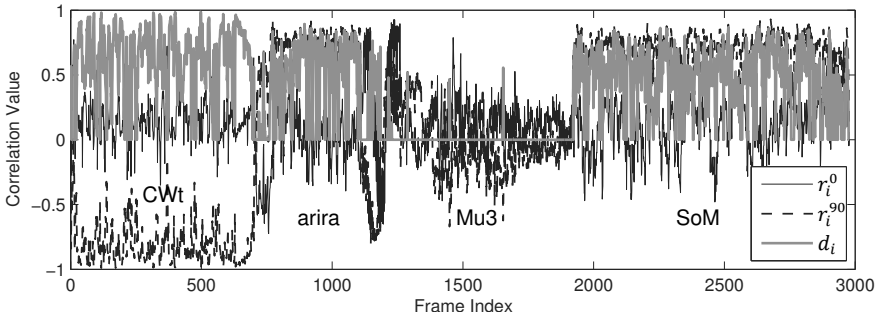


Figure 3.8. Exemplary operation of correlation based kernel switching detector on four signals.

3.2.2 Signal-Adaptive Switching of Overlap Ratio in Audio Transform Coding

The previous subsection proposed an extension of the TDA filter bank scheme which enables improved coding performance in case of "phasy" two-channel frame input with an overall IPD near $\pm\frac{\pi}{2}$ (90 or 270 degrees). This subsection develops another enhancement to the filter bank design addressing the efficient coding of single- or multichannel stationary tonal signals with relatively short frames, as discussed in subsection 2.6.3.

During the last two decades, especially since the development of the MPEG-1 Layer 3 (MP3) and AC-3 (Dolby Digital) systems, perceptual audio coding has relied exclusively on the MDCT, developed by Princen *et al.* [Prin86, Prin87] as a "TDAC filter bank" design and further investigated, under the acronym MLT, by Malvar [Mal90b]. Using the MDCT or MLT, as specified by (2.8) for the analysis and (2.9) for the synthesis case, waveform preserving quantization can be achieved in a spectral domain. Applying a maximum TD window length M being twice that of the transform length N (the number of FD coefficients), i. e., $M = 2N$, the inter-transform overlap ratio is 50%. In recent standards based on MPEG-2 AAC [ISO97, ISO09, ISO12], the MDCT coding principle has been extended to allow parametric noise filling in the transform domain, examined in subsection 2.3.3.

Subsection 2.6.3 examined observations that *dual-rate* SBR/MPS configurations are able to code quasi-stationary harmonic signals with higher perceptual quality than the same codec operating in a *downsampled* mode or without the QMF-domain parametric tools. The effective doubling of the core frame length — and, thus, of the number N and spectral resolution of the transform coefficients — in the *dual-rate* case was identified as a plausible cause. The latter setting, however, is impractical at higher bit-rates, so an alternative solution for improved spectral resolution of the transform coefficients which, ideally, maintains the relatively short frame length of the higher-rate setup, is desirable.

A viable measure for increased spectral efficiency on quasi-stationary audio parts is the extended lapped transform (ELT) of Malvar [Mal90a, Mal92a] and Vaupel [Vaup90], whose inverse (synthesis) formulation is identical to (2.9), except that $0 \le m < L$ with $L \ge 4N$ instead of M. Unfortunately, as will be shown below, its inter-transform overlap ratio is fixed to at least 75% instead of the MDCT's 50%, which tends to produce audible pre-echo artifacts for transient waveform parts like drum hits or tone onsets. Moreover, practical solutions for block switching between ELTs of different lengths — or between an ELT and MDCT/MLT — similarly to the technique applied in conventional transform codecs for precisely such non-stationary frames have, apparently, not been presented in the literature (only theoretical work has been published [Teme93, Teme95, Schul00]).

To address this shortcoming, a simple modification of the ELT definition of (2.9) with L, allowing PR transitions (i. e., with complete TDAC) between transforms with 50% and with 75% overlap ratio, are proposed in the following, along with a newly designed ELT window. Using this modified ELT (MELT), a signal-adaptive coding scheme applying the switched-ratio principle in the context of MPEG-style audio coding is then introduced.

The ELT, MLT, or MDCT, as indicated above, can be considered specific realizations of a general lapped transform definition, with (2.9) for the inverse and with $0 \le k < N$ and

$$X_i(k) = \sum_{m=0}^{L-1} \hat{x}_i(m) \cos\left(\frac{\pi}{N}\left(m + \frac{N+1}{2}\right)(k + k_0)\right), \quad k_0 = \frac{1}{2}, \tag{3.7}$$

for the forward (analysis) case, where, as previously, $\hat{x}_i(m)$ denotes the windowed PCM input sample of the given waveform section of length L. For TDAC and PR, at least in the absence of modifications to the spectrum X_i (e. g., its quantization), all analysis and synthesis windows w must fulfill particular design constraints. For the MDCT and MLT, the constraints are defined by the Princen-Bradley condition of (2.14), derived from (2.13), and (2.19), obtained via (2.18), as introduced in section 2.1. A generalized formulation, which is also applicable to the ELT, is the following one proposed by Malvar in [Mal92a]:

$$\sum_{j=0}^{\frac{L}{N}-2l-1} w(n + jN) \cdot w(n + jN + 2lN) \;=\; \delta(l), \quad 0 \le l < \frac{L}{M}, \quad 0 \le n < N, \tag{3.8}$$

assuming an even ratio L/N and identical, symmetric analysis and synthesis windows w. For the MLT, MDCT, or MDST ($L = M = 2N$), the TDA is canceled by combining the first temporal half of the windowed output signal $\bar{y}_i$ with the second half of the last frame's windowed result $\bar{y}_{i-1}$ via OLA, as in (2.12). The corresponding inter-transform overlap ratio is, thus, $(2 - 1)/2 = 50\%$, and (3.8) reduces to the equation pair (2.14) and (2.19).

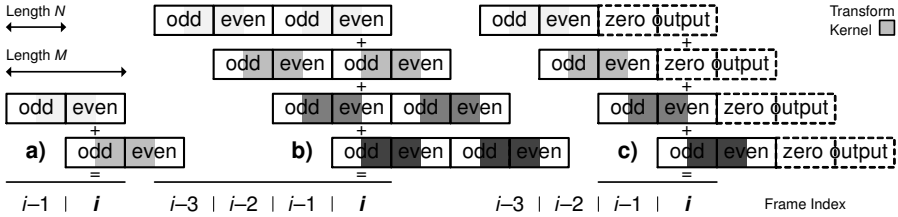


Figure 3.9. Cancelation of evenly and oddly symmetric TDA upon OLA of overlapped transform outputs for (a) MDCT, (b) ELT, (c) MDCT via ELT. (—) Maximum pre-echo duration.

In case of ELT coding using $L = 2M = 4N$, the OLA step must combine the first quarter of $\bar{y}_i$ with the second quarter of $\bar{y}_{i-1}$, the third quarter of $\bar{y}_{i-2}$, and the fourth quarter of $\bar{y}_{i-3}$ to attain the final TDA-free output waveform for frame i, so the overlap ratio grows to $(4-1)/4 = 75\%$. Figure 3.9 illustrates this difference and the associated worst-case temporal spread of FD induced coding errors. Compared to the MDCT or MDST, the ELT clearly leads to stronger pre-echos on transients. More detailed discussions of TDA and PR in transform coding are provided in [Teme93, Teme95, Shlie97, Schul00]. Note, also, that evenly stacked linear-phase ELTs based on the DCT-II, or odd-length ELTs with, e. g., $L = 3N$ are also feasible [Padm92, Hame05], but such designs will not be studied here.

Focusing on the length-$4N$ ELT in the remainder of this thesis, one can observe that, as shown in Figure 3.10(a), TDAC and PR cannot be achieved during switchovers to and from MDCT/MLT coding because the TDA symmetries are incompatible. Simply spoken, the necessity of adjacent even-odd aliasing combinations [Prin87, Hel15c] — e. g., even TDA symmetry in i overlapping with odd symmetry in $i-1$, or vice versa — is violated between frames $i-4$ and $i-3$. To correct this issue and achieve complete TDAC in all frames, including those with a three-part OLA, one transform type needs to be redefined such that its TDA symmetries complement those of the other, e. g., as in Figures 3.10(b) and (c). Since it is preferable to avoid modifications to existing MDCT and MDST implementations, the ELT shall be addressed. Furthermore, to easily acquire PR steady-state and transitory windows for all transforms, respective analytic expressions are desirable.

- Modifications for adaptation of overlap ratio. In order to equip the ELT with the depicted TDA compatibility for PR transitions to and from the 50% overlapping transforms, it suffices to adjust the temporal phase offset in its base functions:

$$X_i(k) = \sum_{m=0}^{4N-1} \hat{x}_i(m) \cos\left(\frac{\pi}{N}\left(m + \frac{3N+1}{2}\right)(k + k_0)\right), \quad k_0 = \frac{1}{2}, \tag{3.9}$$

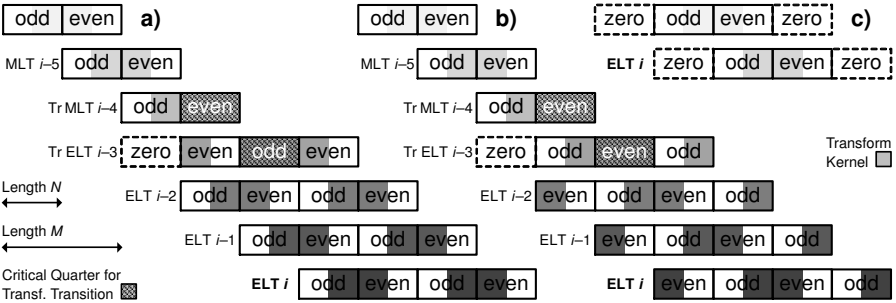


Figure 3.10. Switches from MLT/MDCT to ELT filter bank by way of two (Tr)ansition transforms: (a) incorrect, without PR, (b) desired, with PR. (c) MLT/MDCT using modified ELT.

with k as in (3.7) and the inverse ELT of (2.9) with $m < 4N$ changed accordingly:

$$\hat{y}_i(m) = \frac{2}{N}\sum_{k=0}^{N-1} X_i(k)\cos\left(\frac{\pi}{N}\left(m+\frac{3N+1}{2}\right)(k+k_0)\right), \quad k_0 = \frac{1}{2}. \tag{3.10}$$

These modifications of the traditional ELT formulation will be referred to as the *modified ELT* (MELT) definitions. It can further be shown [Schul00] that, as Figs. 3.10(a) and (b) indicate, the four temporal quarters of the transitory MLT/MDCT and (M)ELT windows are based on the associated steady-state windows w, with the first and/or fourth quarter (depending on the transform and transition type) set to zero to switch the overlap ratio and with the critical quarters described by

$$w_{\mathrm{tr}}(n) = \sqrt{1 - w_{\mathrm{elt}}(k)^2 - w_{\mathrm{elt}}(N+k)^2}, \quad 0 \le k < N, \tag{3.11}$$

with $n = \frac{L}{2} + k$ for ratio-increasing switches as in Fig. 3.10, or $n = \frac{L}{2} - 1 - k$ for the reverse, i. e., ratio-decreasing ELT-to-MDCT transitions, shown in Figure 3.11 on the next page. Utilizing (3.11) to acquire the TDAC-critical quarters for both the MLT/MDCT and MELT transition weightings completes the definition of the transitory windows, leaving only the selection of steady-state window functions.

- Steady-state PR lapped-transform windows. Several PC windows enforcing the Princen-Bradley condition for TDAC have been proposed. Figure 3.12(a) depicts the shapes and corresponding oversampled transfer functions, obtained by way of Fourier transformation of the TD weighting as in Fig. 2.4, of the windows used in MPEG audio codecs, i. e., the sine and KBD functions introduced in section 2.1.

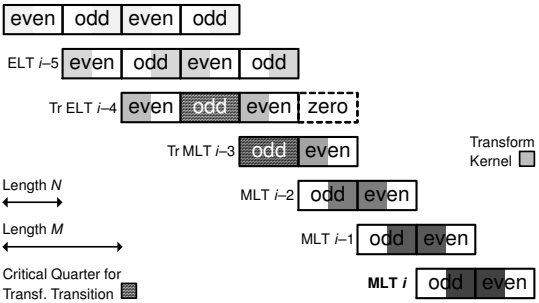


Figure 3.11. TDAC preserving switch-back from modified ELT to MDCT filter bank by utilizing transforms applying (Tr)ansitory windows in frames *i*–4 and *i*–3. As with the last two figures, the darker the kernel shade, the more recently applied the transform, i. e., the higher the frame index.

Also shown is the sum-of-sines derived window constructed by the author in [Helm10], whose shape is very similar to that of the KBD weighting but which, as can be observed, exhibits lower first (near-field) side lobes. Finally, a sine window for the doubled frame length, as it is effectively available in *dual-rate* SBR, serves as a reference and illustrates that longer windows can notably reduce both pass-band width and stop-band level.

Ideally, a MELT or ELT window, subject to the PR constraints of (3.8), which reduce to

$$\begin{aligned} w_{\mathrm{elt}}(n)^2 + w_{\mathrm{elt}}(N+n)^2 + w_{\mathrm{elt}}(2N+n)^2 + w_{\mathrm{elt}}(3N+n)^2 &= 1, \\ w_{\mathrm{elt}}(n) \cdot w_{\mathrm{elt}}(2N+n) + w_{\mathrm{elt}}(N+n) \cdot w_{\mathrm{elt}}(3N+n) &= 0, \end{aligned} \tag{3.12}$$

for $L = 4N$, should exhibit a frequency response comparable to that of the *dual-rate* sine window so as to provide similarly high levels of frequency selectivity and, thus, spectral compaction. However, it can be observed that, due to the orthogonality restrictions for TDAC and PR, main-lobe width can only be minimized by allowing less side-lobe attenuation. Malvar's window [Mal90a] with $p = 1$, for instance, was found to offer the lowest possible main-lobe width of all ELT designs but also undesirably high stop-band levels, as shown in Figure 3.12(b). Its temporal boundaries are strongly discontinuous (since all samples outside the window's TD support are considered zero-valued), resulting in a side-lobe decay of only –6 dB/octave [Nutt81, Helm10] and audible framing artifacts in preliminary experiments toward this work. Temerinac and Edler [Teme95] presented a recursive design technique, which they utilized to obtain the ELT window also shown in Fig. 3.12 (note that the value –0.038411 is missing in column "L = 4N" of their table 1). This window, which can be approximated relatively closely using Malvar's formulations with $p = 0.14$, provides somewhat better but still quite weak stop-band attenuation.

It is worth mentioning that, for $p = 1$, Malvar's window function in [Mal90a] can be simplified to a notation which is remarkably similar to that for a Hann(ing) window:

$$w_{p=1}(l) = a_0 - 0.5\cos\left(2\pi \cdot \frac{l+0.5}{L}\right), \quad L = 4N, \tag{3.13}$$

where $0 \leq l < L$ denotes the time samples of the window and $a_0 = 2^{-3/2}$ is selected to enforce constraints (3.12). Intuitively, a function with more stop-band rejection, such as

$$w_{3\text{term}}(l) = \sum_{k=0}^{2} b_k \cos\left(2k\pi \cdot \frac{l+0.5}{L}\right), \quad b_1 = -\frac{1}{2}, \tag{3.14}$$

with $b_2 > 0$, which can be utilized to derive Blackman's window [Helm10], would seem applicable as well. Unfortunately, it can be shown that PR cannot be achieved with such a window class regardless of the value of b_0. However, accumulating more cosine terms,

$$w_{\text{elt}}(l) = w_{3\text{term}}(l) - \sum_{k=1}^{K} c_k \cos\left(8k\pi \cdot \frac{l+0.5}{L}\right), \tag{3.15}$$

with b_k as above, the resulting shape for any choice of $b_2 \lesssim \frac{3}{8}$ can be corrected such that PR is approached arbitrarily closely. In the present study, let the design target be a low side-lobe level so as to avoid the abovementioned framing artifacts, combined with the additionally imposed restriction of an isotone left-half and, therefore, antitone right-half window slope in order to obtain a smooth shape of the weighting. Then, it is possible to approximate PR with an error below $4 \cdot 10^{-6}$ (–108 dB) using a relatively low $K = 3$ and

$$b_2 = 0.176759 \rightarrow b_0 = 0.3303, c_1 = 0.02366318, c_2 = 0.00042436, c_3 = 0.00001521. \tag{3.16}$$

This simple ELT window function, depicted in Fig. 3.12(b), is notably less discontinuous at its borders than the proposals of [Mal90a, Teme95] and, as a result, achieves roughly the same level of side-lobe attenuation as the double-length sine window of Fig. 3.12(a). Concurrently, its main lobe remains narrower than that of the sine function for equal N (MLT Sine 1 in the figure). Interestingly, it also resembles the latter window in shape.

To complete this discourse on practical window functions for switched MDCT-MELT coding, Figure 3.12(c) illustrates the temporal and spectral responses of the asymmetric MDCT/MDST and (M)ELT transition windows needed for overlap ratio adaptation. In this case, they are based on the PC sum-of-sines design of [Helm10] for the 50% overlap and on w_{elt} of (3.15) with (3.16) for the 75% overlap case. For comparison, the double-length *start* window used in HE-AAC, HE-AAC v2, and HE-AAC/USAC with MPS is shown. Due to the asymmetry and shortened overlap on one side, all transitory windows reach only moderate stop-band rejection, with the side-lobe attenuation of the new MDCT and MELT transition windows being comparable to those of the traditional sine weightings.

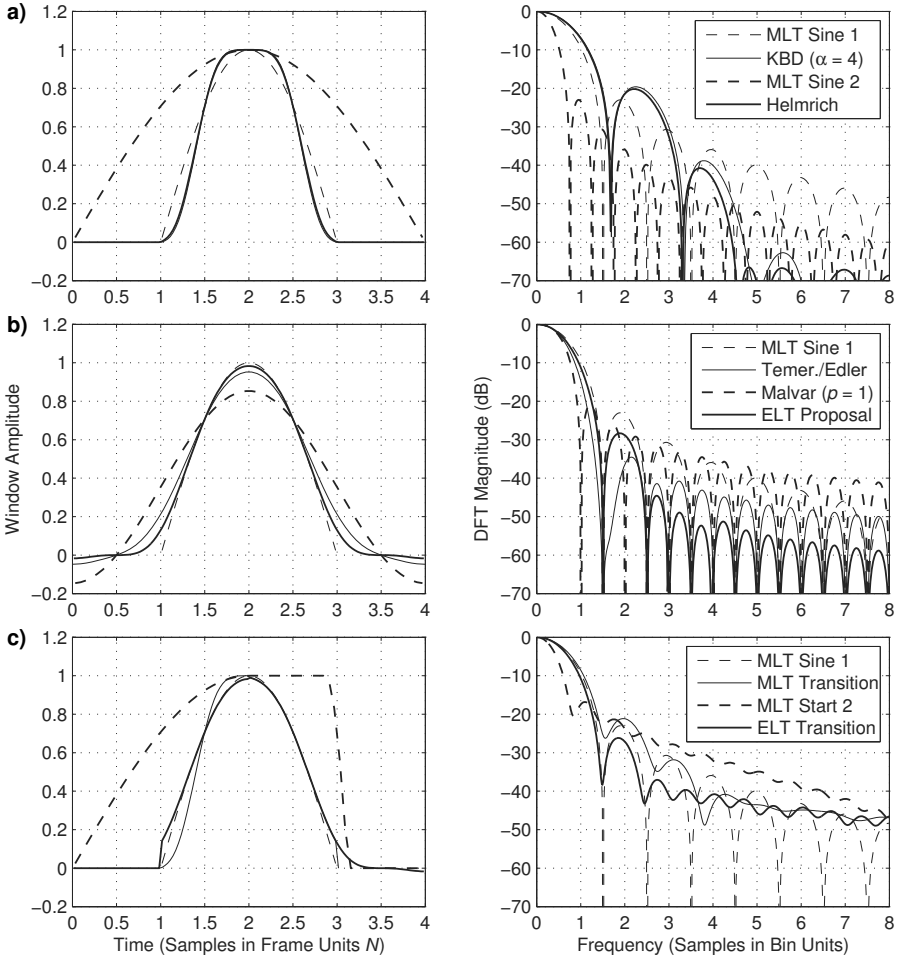


Figure 3.12. PR window designs for (a) MLT or MDCT, (b) ELT or MELT, (c) transitions. See text.

Now that the MDCT and ELT kernels as well as all required windows have been prepared, an input-adaptive ratio switching (RS) architecture can be constructed. In order to verify its expected subjective benefit on tonal input and as a proof of concept, this RS design, utilizing the novel MELT kernel and steady-state or transitory windows, may be integrated into an MPEG-style perceptual transform codec as follows. For brevity, only high-level aspects shall be addressed. More details are discussed in the next subsection.

- Decoder specification and processing. An additional bit, signaling application of the MELT, is received per channel and/or frame i in which a *long* transform (i. e., no block switching) has been utilized by the encoder. In case of MPEG coding the *window_shape* bit may be reused for this purpose (1: MDCT using KBD window of [Field96] or PC window of [Helm10], 0: MELT with novel window). Based on this bit and the *window_sequence* flag (transform mode or frame type), both for the current and last frame, the decoder can then deduce and apply the appropriate inverse transform with the correct overlap ratio and weighting, as described.
- Encoder design and implementation. The transmitter, as in case of the KS design according to the last subsection, must apply and convey the per-channel/frame MDCT-MELT selection so that the encoder and decoder side are synchronized. It is, furthermore, the encoder's task to identify quasi-stationary harmonic frames, for which the MELT shall be used, and to detect non-stationary transient events early enough to revert to 50% overlap ratio (or less, in case of block switching). For instance, by obtaining a 16th-order linear prediction residual of the half-rate downsampled input, as done in speech coders [Neue13], and deriving therefrom
  - a temporal flatness f_t as the ratio between the next and current frame's residual energy, measured non-overlapped, with stationarity specified as $f_\mathrm{t} < \frac{55}{8}$,
  - a spectral flatness f_s, also known as Wiener entropy, obtained from the DFT energy (i. e., power or squared magnitude) spectrum of the current and next frame's concatenated residual signal, with strong tonality indicated by $f_\mathrm{s} < \frac{3}{8}$,

  the encoder can distinguish, for each i, between the utility of MELT coding or the necessity for MDCT coding. Figure 3.13 depicts the resulting per-i MELT (0) and MDCT (–1) selection for five concatenated input items. Except for the sixth tone onset in the harpsichord arpeggio, the RS decision appears surprisingly reliable.

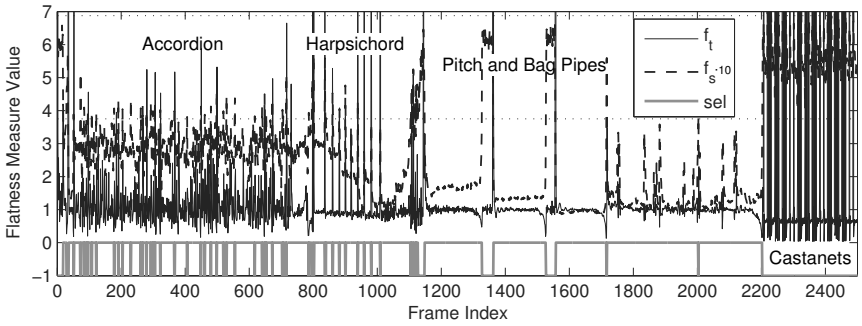


Figure 3.13. Temporal and spectral flatness based RS (sel)ection of MELT or MDCT coding.

3.2.3 Combined Block, Kernel, and Ratio Switching for Fully Flexible Coding

A logical final undertaking in the present investigation is to merge the traditional or low-delay block switching of section 2.1 respectively 3.1, the KS of subsection 3.2.1, and the RS proposal of subsection 3.2.2 into a single flexible filter bank design to allow joint usage of the techniques within the same, or at least adjacent, frames. Some algorithmic details and codec modifications needed to achieve this objective are discussed below.

In [Hel15c], the authors noted that the presented KS scheme merely alters the transform definition without affecting the windowing and OLA steps, hence maintaining full support for the window shape adaptation of [Alla99, Edler89]. Compatibility with the block length switching of [Edler89, Bosi97] was mentioned as well but not described in detail. More recently, the usefulness of a modified ELT, or MELT, was reported [Hel16a]. Based on Malvar's ELT formulation [Mal92a], the MELT constructs an oddly stacked PR filter bank with 75% inter-transform overlap, depicted in Figure 3.14(b), yielding better frequency selectivity than a MDCT or MDST filter bank with 50% overlap, illustrated in Figure 3.14(a), at the same frame length N. Unlike the ELT, said MELT enables straightforward transitions — requiring only special transitory windows — to and from MDCTs. In [Hel16a], a corresponding frame-wise signal-adaptive RS method was developed, but as in [Hel15c], operation within a block switching system was not discussed in detail.

To realize a fully flexible coding framework, the following three additional structural capabilities are desirable and will be investigated individually on the next five pages:

- KS also in case of MELT coding for harmonic input with frame IPDs around $\pm 90°$,
- for LD cases, greater flexibility in the transitions from and to MELT coded frames,
- transform sequencing and windows for combined block switching and KS or RS.

For consolidated KS and MELT processing, it is assumed that the activation of the latter transform type is synchronized between the two channels of the pair. Moreover, a sine-modulated counterpart to the cosine-based MELT, delineating a 75%-overlap equivalent of the 50% overlapping type-IV MDST, can be trivially derived from (3.9) and (3.10):

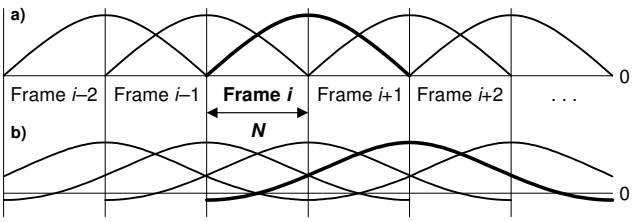


Figure 3.14. Basic TDAC filter banks using lapped transforms: (a) MDCT or MDST, (b) ELT or MELT. The descriptors and window shapes for frame index i are emphasized.

$$X_i(k) = \sum_{m=0}^{4N-1} \hat{x}_i(m) \sin\left(\frac{\pi}{N}\left(m + \frac{3N+1}{2}\right)(k + k_0)\right), \quad k_0 = \frac{1}{2}, \tag{3.17}$$

with $0 \leq k < N$ for the analysis definition and, for the synthesis case with $0 \leq m < 4N$,

$$\hat{y}_i(m) = \frac{2}{N} \sum_{k=0}^{N-1} X_i(k) \sin\left(\frac{\pi}{N}\left(m + \frac{3N+1}{2}\right)(k + k_0)\right), \quad k_0 = \frac{1}{2}. \tag{3.18}$$

It is worth repeating that, even though the applied *window* length ($L = 4N$ or $M = 2N$) varies between the cosine-/sine-based MELT and the MDCT/MDST, the *transform* length N and, thereby, the number of per-frame spectral samples and the inter-transform step size shown in Fig. 3.14 remains identical. This explains the difference in overlap ratio.

The cosine- and sine-modulated MELT definitions of (3.9), (3.10) and (3.17), (3.18), respectively, are, as demonstrated in [Hel15c] and the previous subsection, insufficient for realizing KS — and, thus, efficient coding of stereo signals with $\pm 90°$ of IPD — even in case of 75% inter-transform overlap. Specifically, type-II transition transforms adopted from the initial Princen-Bradley design [Prin86] are required for TDAC when switching between type-IV MDCTs and MDSTs: a MDST-II is needed when changing from MDCT-IV to MDST-IV coding, while a MDCT-II is necessary when reverting to MDCT-IV coding.

Hameed and Elias [Hame05] described that, beside the oddly stacked type-IV (M)ELT instances introduced in [Mal90a] and herein, an ELT-based filter bank allowing for fast implementations using the DCT-II can also be constructed, thereby proving that type-II filter banks with more than 50% inter-transform overlap are feasible. An alternative but equivalent approach following the TDAC filter bank design [Prin86, Hel15c] is to devise an evenly stacked architecture via alternating usage of a type-II cosine modulated MELT

$$X_i(k) = \frac{1}{1+\delta(k)} \sum_{m=0}^{4N-1} \hat{x}_i(m) \cos\left(\frac{\pi}{N}\left(m + \frac{3N+1}{2}\right)(k + k_0)\right), \quad k_0 = 0, \tag{3.19}$$

with Kronecker delta $\delta(0) = 1$ to scale the DC coefficient, and a type-II sine based MELT

$$X_i(k) = \frac{1}{1+\delta(k')} \sum_{m=0}^{4N-1} \hat{x}_i(m) \sin\left(\frac{\pi}{N}\left(m + \frac{3N+1}{2}\right)(k + k_0)\right), \quad k_0 = 1, \tag{3.20}$$

where $k' = N - 1 - k$ is employed to scale the Nyquist coefficient as in subsection 3.2.1. Proving that the application of (3.19) and (3.20) on the analysis side and, respectively,

$$\hat{y}_i(m) = \frac{2}{N} \sum_{k=0}^{N-1} X_i(k) \cos\left(\frac{\pi}{N}\left(m + \frac{3N+1}{2}\right)(k + k_0)\right), \quad k_0 = 0 \tag{3.21}$$

and

$$\hat{y}_i(m) = \frac{2}{N}\sum_{k=0}^{N-1} X_i(k)\,\sin\left(\frac{\pi}{N}\left(m+\frac{3N+1}{2}\right)(k+k_0)\right),\quad k_0 = 1 \tag{3.22}$$

on the synthesis side leads to TDAC after OLA, as indicated in Figure 3.15 (where i is the frame index), is relatively straightforward and will be omitted for the sake of brevity.

Unfortunately, regarding the combination of RS and KS, it can be shown that TDAC is impossible when, analogously to the process for 50% inter-transform overlap, a transitory type-II instance of (3.19, 3.21) or (3.20, 3.22) is employed when switching between type-IV cosine and sine-modulated MELTs. Since it is desirable to keep the architectural codec complexity (number of signal paths, tables in memory, etc.) low when allowing KS regardless of the instantaneous overlap ratio, the following workaround is proposed.

To switch from cosine-modulated MELT-IV coding according to (3.9, 3.10) to the sine-modulated MELT-IV of (3.17, 3.18), a transitory MDST-II frame as in [Hel15c], combined with a necessary temporary reduction of the overlap ratio to 50% on both the analysis and synthesis side, can be utilized. Likewise, an intermediate MDCT-II can be employed when reverting back from sine to cosine-based type-IV MELT coding. Complete TDAC is guaranteed in both cases since, as visualized in Figure 3.16, the overlap length between each type-II transition and its type-IV MELT neighbors is restricted to $N = \frac{M}{2}$ (therefore, there is no TDA-bound overlap between a cosine- a sine-modulated MELT-IV requiring TDAC). Proper windowing, however, remains crucial: a special *stop-start* window must be applied to said type-II transforms, as depicted in Figure 3.17(a). This symmetric window is based on the asymmetric transitory weightings of [Hel16a] and examined below.

To maintain TDAC during RS, dedicated MDCT/MDST and MELT transition windows derived from the steady-state windows were devised in [Hel16a]. These are defined as

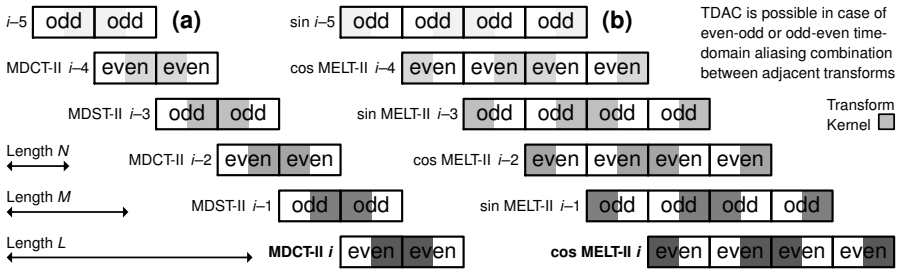


Figure 3.15. TDA in evenly stacked filter banks: (a) Princen-Bradley [Prin86], (b) type-II MELT.

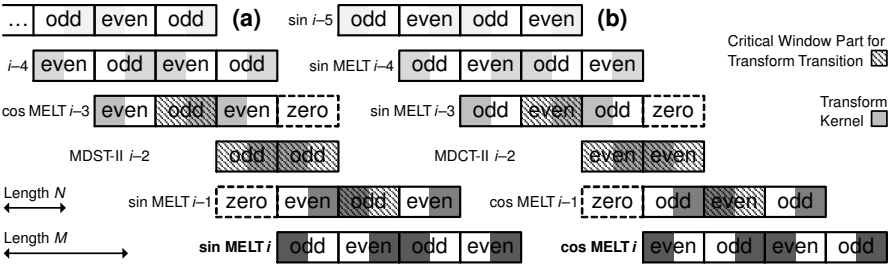


Figure 3.16. Proposed TDAC-compliant kernel switching for type-IV (oddly stacked) MELT filter banks: transition from (a) cosine to sine modulation, (b) sine to cosine modulation.

$$w'_{\mathrm{elt}}(l) = \begin{cases} 0\,, & 0 \le l < N, \\ w_{\mathrm{elt}}(l), & N \le l < M, \\ d\sqrt{1 - w_{\mathrm{elt}}(k)^2 - w_{\mathrm{elt}}(N+k)^2}, & M \le l < 3N, \\ w_{\mathrm{elt}}(l), & 3N \le l < L, \end{cases} \tag{3.23}$$

for the first MELT window upon an overlap ratio increase from 50% to 75% (bold-lined shape depicted in Fig. 3.17(a) for frame i) and

$$w'_{\mathrm{mlt}}(m) = \begin{cases} d\sqrt{1 - w_{\mathrm{elt}}(M+k)^2 - w_{\mathrm{elt}}(3N+k)^2}, & 0 \le m < N, \\ w_{\mathrm{mlt}}(m), & N \le m < M, \end{cases} \tag{3.24}$$

for the first MDCT or MDST window when reducing the overlap ratio to 50% (bold-lined shape in Fig. 3.17(b) for the same frame). Parameter d, introduced for greater flexibility, is studied and optimized in [Hel16b]. Here and in [Hel16a], it is assumed to equal 1.

The complements for w'_{elt} and w'_{mlt} — the last MELT window when switching to 50% overlap, and the last MDCT/MDST window during switchbacks to 75% overlap ($i - 2$ in Fig. 3.17) — are simply the temporal reversals of w'_{elt} and w'_{mlt}, respectively. Value k in the critical window parts (see also Fig. 3.16) is specified as earlier, while w_{elt} resp. w_{mlt} indicate the underlying window functions for a steady-state MELT and MDCT/MDST. For the former, which is also applicable to the ELT, a novel design was devised in [Hel16a]. It is worth mentioning that, when using this novel design, a noticeable jump remains at the *center* of the transitory window shapes, i. e., the first differential of the weightings is discontinuous at $l = M$ (see Fig. 3.17). The discontinuity, however, is comparatively minor and more pronounced when using, e. g., Malvar's window function [Mal90a]. In fact, the jumps to zero at the transitory window *borders*, i. e., towards the zero-valued quarters, are much more likely to cause audible distortion such as clicks in low-rate coding. For this reason, a minimization of the occurrences of w'_{elt} and w'_{mlt} is recommended.

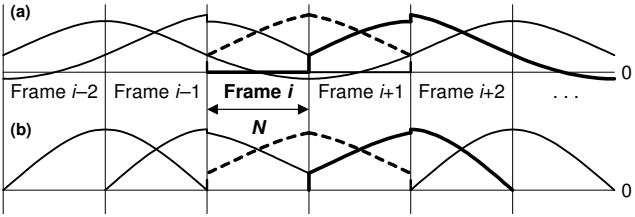


Figure 3.17. Correct windowing with (---) custom *stop-start* window shape during temporary transitions from (a) 75% to 50% overlap ratio for KS, (b) 50% to 75% overlap ratio.

On page 83, the usage of a dedicated transitory *stop-start* window in MELT-based KS was introduced. This window, depicted by a dashed line in Fig. 3.17(a) and denoted by w_{ss} hereafter, can be easily derived from the critical window quarters of w'_{elt} and w'_{mlt}:

$$w_{\mathrm{ss}}(m) = \begin{cases} d\sqrt{1 - w_{\mathrm{elt}}(M+k)^2 - w_{\mathrm{elt}}(3N+k)^2}, & 0 \le m < N, \\ d\sqrt{1 - w_{\mathrm{elt}}(k)^2 - w_{\mathrm{elt}}(N+k)^2}, & N \le m < M. \end{cases} \tag{3.25}$$

More specifically, w_{ss} is a symmetric window with critical parts in both halves, therefore allowing overlap ratio transitions on both sides (i. e., toward the past and future frame). Note, also, that w_{ss} can be applied to the MDCT and MDST as well as the different MELT variants (assuming the outer quarters of the length-L weighting are set to zero). In fact, its usage for analysis-side windowing renders the MDCT-IV and the cosine-modulated MELT-IV coefficients identical apart from sign differences, as indicated by Fig. 3.10(c).

Besides facilitating KS, w_{ss} may also be utilized to make the overlap ratio adaptation scheme more flexible. For instance, a configuration with temporary switches from 50% to 75% overlap, as shown in Fig. 3.17(b), can be achieved therewith. Such a short-term overlap increase is useful when an objective is to minimize the encoder-side lookahead used for ratio detection (after all, the MELT itself already increases the codec delay due to the increase in windowing lookahead; this issue will be revisited in the next section). To provide an example, let us assume that a total encoder lookahead of $L - N = 3N$ may not be exceeded. As can be observed in Fig. 3.17(b), if the RS detector chooses a switch back to MELT coding in frame $i - 2$, e. g., after a transient (thin-lined asymmetric transitory MDCT/MDST window), but in the next frame at $i - 1$ detects a new transient at the end of the lookahead region, i. e., the length-N segment labeled "Frame $i + 2$," it needs to instantly revert to 50%-ratio coding in order to minimize the inevitable pre-echo which will occur [Hel16a]. Since the first half of the MELT window for $i - 1$ (dashed line) has already been defined by the complementary shape for $i - 2$ and, to preserve TDAC, may not be changed anymore, only the second window half can be freely selected. Mirroring the first to the second half produces the TDAC compliant w_{ss} illustrated in Fig. 3.17(b), which can be followed by the w'_{mlt} shape (bold line) in frame i. Note, however, that this

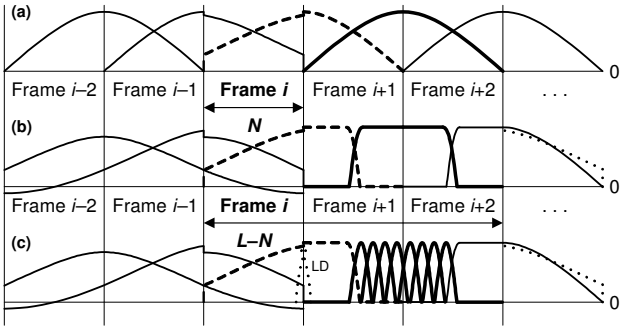


Figure 3.18. Proposed combinations of MELT coding and window or block length switching. See the text for details.

solution exhibits obvious window border discontinuities at three places — frames $i-1$, i, and $i+1$ — causing reduced spectral compaction and, potentially, suboptimal coding quality over this period due to audible distortion, as mentioned previously. It therefore seems preferable to, as an alternative, avoid the symmetric w_{ss} in this case and apply the asymmetric w'_{mlt} already in $i-1$. In doing so, the steady-state w_{mlt} can be employed in i, as shown in Figure 3.18(a), and the boundary discontinuity at $i+1$ can be avoided.

To complete this section, the usage of the block switching technique in combination with RS is addressed. It is well established that the analysis and synthesis windows of a MDCT-based filter bank can be adapted to the instantaneous signal characteristics on a per-transform basis without violating the TDAC principle [Edler89, Alla99]. The same, naturally, holds for the KS approach since, as noted earlier in this subsection, transitions between a MDCT and a MDST do not affect the windowing and OLA processes [Hel15c]. Consequently, switching to a low-overlap window shape during transient input portions is also possible while KS is being utilized (assuming that valid transform sequences are still employed) and/or with MELT coding in case of RS (assuming the overlap ratio has been reduced to 50% in the preceding frame, as proposed earlier for a transform kernel change). Figure 3.18(b) shows a complete sequence for such a window shape switch.

The length-M dashed window shape for frame $i-1$ in Fig. 3.18(b) is a concatenation of the first half of w'_{mlt} or w_{ss} and the second half of the long *start* window known from MPEG-2 AAC [Bosi97, ISO97]. Accordingly, the first half of the MDCT/MDST window for i must equal the first N samples of AAC's length-M *stop* shape. The latter is depicted for frame $i+1$ to denote a switch back to the full 50% overlap ratio (and possibly 75%, as indicated by the dotted line). Notice that block length switching can now be realized by simply replacing the low-overlap, *stop-start* windowed long transform at i by eight successive 50% overlapped short transforms, as in AAC/USAC. The resulting configuration is visualized in Fig. 3.18(c), including an "early" LD block switch according to section 3.1.

3.3 Frequency-Domain Prediction with Very Low Complexity

The flexible filter bank design proposed in section 3.2, allowing various adaptations (regular and LD block and window length switching, KS via MDSTs, and RS via MELTs), lays the foundation for improved coding performance on certain input such as strongly tonal and quasi-stationary recordings. The support for RS, however, comes at the cost of increased algorithmic codec delay since the TD support of the MELT windows is greater that that of the MDCT/MDST windows. More precisely, the windowing lookahead must be extended from $M - N = N$ to $L - N = 3N$ time samples in case of MELT coding, thus doubling the minimum codec latency from M to L samples, i. e., from 42.7 to 85.3 ms for a sample rate of 48 kHz. Since this increase is typically unacceptable in LD applications, it is worth investigating alternatives to the RS principle for improved coding of the tonal quasi-stationary waveform parts, preferably operating directly on an MDCT/MDST-only filter bank. The logical alternative is long-term prediction (LTP), as discussed hereafter. The entirety of this study has been submitted for publication by the author [Hel16c].

Conventional perceptual and lossless audio codecs divide the incoming TD waveform into successive frames which are transformed (using a filter bank or a linear predictive filter), quantized, and coded separately and largely independently. For quasi-stationary input signals, however, there, naturally, remains some residual temporal redundancy in the transformed samples between adjacent frames for a given channel or even between channels. This is especially true for recordings of sustained isolated instrumental notes.

To minimize such intertransform redundancy, two approaches have been pursued in the past. The first, proposed by Mahieux *et al.* for a DFT based coding system [Mahi89], is to apply LPC techniques across time to individual transform coefficients. This method was later adapted for real-valued MDCT coding and extended to include joint-stereo (JS) cross-channel functionality [Fuch93, Fuch95, Lieb02, Krüg08, Helm11]. The second approach is to account for temporal redundancy during the entropy coding stage, i. e., after TD or FD quantization. Most recently, this was addressed in the MPEG-D USAC standard [ISO12] by way of arithmetic coding with intra-/inter-frame signal-adaptive probability contexts for each quantized MDCT value [Fuch11, Neue13]; see also subsection 2.3.4.

The latter technique can be implemented with relatively low algorithmic complexity since the quantized MDCT values are readily available at both the encoder (transmitter) and decoder (receiver) side. The LPC based methods, in turn, often yield higher coding (and quality) gains but require much greater encoder-side complexity: given that they operate on the initial *uncoded* MDCT coefficients using previously *decoded* values, an entire FD decoding path inside the encoder becomes necessary. This is especially the case with TD LTP [Ojan99, Song10], which may even require additional MDCT processing to

prepare the prediction signal. For this predictor type, a complexity of $22 \cdot 672 = 14784$ algorithmic operations per frame and channel was reported [Ojan99], which represents a significant improvement over prior work [Fuch95, YinS97]. With typical stereo input sampled at 48 kHz and coded at a frame length of $N = 1024$ samples, however, this still causes an unwanted 1.4 million operations per second (MOPS) of added complexity.

Contribution and organization of this section. In order to realize very-low-complexity prediction for transform coding, the following structural constraints are to be enforced:

- The predictor's computation and application should reside as closely around the codec's spectral quantizer as possible. This reduces the amount of FD coefficient decoding infrastructure which needs to be implemented at the encoder side.
- The prediction should be bandlimited to further minimize the workload at both the encoder and decoder. Thus, a good tradeoff between coding gain due to, and maximum frequency range of, the predictor must be determined empirically.
- Only those signal components exhibiting (potentially) temporal correlation, e. g., the individual harmonics of a tonal waveform, should be subjected to prediction. This not only minimizes the complexity but also increases the prediction gain.

The remainder of this section devises a novel frequency-domain predictor (FDP) which supports both perceptual and lossless coding (although the former aspect shall be emphasized here) and which, as will be described in subsection 3.3.1, adheres to all of the above three constraints. Extensions of the FDP design for low-bitrate perceptual coding of monophonic and stereophonic content are investigated in subsections 3.3.2 and 3.3.3, respectively. Reports on the preparation and results of objective and subjective tests for the basic FDP scheme will be provided along with those for the other tools in Chapter 4.

3.3.1 Low-Complexity Frequency-Domain Prediction on a Harmonic Grid

The fundamental architecture of a two-channel codec employing temporal prediction as a pre- and post-processor is illustrated in Figure 3.19. The depicted block diagrams, which revisit the subject overview of section 2.2 and in which the lower-case and upper-case letters indicate TD and transformed signals, respectively, apply to both perceptual (e. g., MDCT for transformation, requantization) and lossless coding (e. g., LPC or integer MDCT [Geig02, Yoko06] for transformation, no requantization). The left or right source signal subtracted, after scaling, from the respective target signal during LTP analysis and added again, after equivalent scaling, upon LTP synthesis, is the predictor memory z_p. It equals a delayed portion of the previously *decoded* l' or r' (index i is omitted for clarity):

$$z_p(m) = \begin{cases} l'(m - p_l), & p_l \geq N, \quad \text{for the left channel,} \\ r'(m - p_r), & p_r \geq N, \quad \text{for the right channel,} \end{cases} \tag{3.26}$$

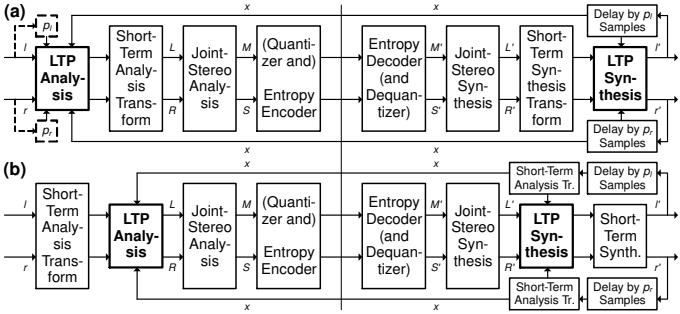


Figure 3.19. Long-term predictive audio coding in the (a) time domain or (b) transform domain.

where $p_{\{l,r\}}$ denotes the per-frame/channel pitch or periodicity lag and $0 \leq m < M$ with $M = 2N$ indicating twice the frame length, as earlier. Note that, in the encoder (left side of thin vertical line) of a "lossy" perceptual codec, the *uncoded* inputs l and r can also be used (dashed paths) instead of l' and r', yielding an open-loop pitch pre-/post-filter for quantization noise shaping [Valin13] instead of a conventional closed-loop LTP. For the sake of brevity, though, this solution will not be examined in the present publication.

Two issues can be observed here. First, as seen in Fig. 3.19(b) and indicated previously, transform-domain LTP application for best achievable selectivity [Ojan99, Kjör16] necessitates an extra analysis transform in each channel since the (re)construction of l' and r' involves an inevitable OLA process between temporally adjacent synthesis transform results. This is particularly true for MDCT based coding, where up to N samples of inter-frame overlap are typically utilized, and is also the reason why the minimum pitch lag in (3.26) must generally be larger for MDCT than for LPC based coding [Song10].

The expensive OLA related requirement for additional transforms can be circumvented by moving the calculation and application of z_p into the MDCT domain [YinS97],

$$Z_P(k) = \begin{cases} L'_{i-P_L}(k), & P_L \geq 1, \quad \text{for the left channel,} \\ R'_{i-P_R}(k), & P_R \geq 1, \quad \text{for the right channel,} \end{cases} \tag{3.27}$$

as in Figure 3.20(a), with $P_{\{L,R\}}$ representing the channel-wise transform-domain (subband) delay for the given frame at index i, restricted to an integer value, and $0 \leq k < N$. Assuming that identical symmetric TDAC-compliant analysis and synthesis windows w according to the relevant previous sections are applied in said frame at i, it follows that

$$L' = \mathrm{MDCT}[l' \cdot w], \; R' = \mathrm{MDCT}[r' \cdot w] \; \rightarrow \; Z \equiv \mathrm{MDCT}[z \cdot w], \tag{3.28}$$

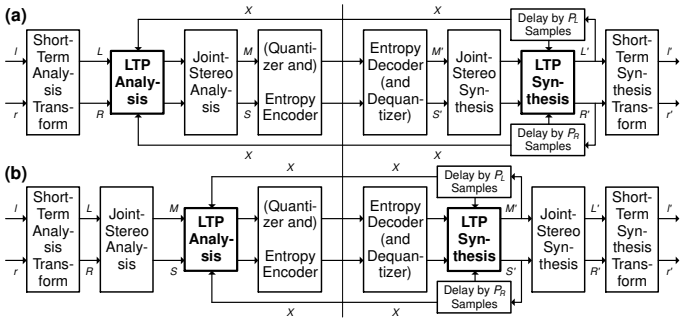


Figure 3.20. FD long-term prediction without TD components: (a) conventional, (b) proposed.

which is easily adaptable, accordingly, in case of MDST coding. This approach, however, still exhibits a second issue: for every algorithmic tool, a corresponding decoder is also required in the encoder, thereby increasing the computational complexity of the latter.

In order to minimize this additional encoder-side workload (i. e., the amount of analysis by synthesis), the LTP may be implemented as closely around the FD entropy coder (and quantizer, in case of perceptual codecs) as possible. Fig. 3.20(b) illustrates that, at most, only reconstructive scaling, via so-called "dequantization", and the obligatory LTP decoder must then be added to the encoder. In the examples of Figs. 3.19 and 3.20, this means that the two predictors now operate in the JS (e. g., downmix/residual) domain.

The complexity can be further reduced by restricting the LTP to a bandwidth lower than the 15 kHz utilized in [Fuch95, Ojan99, YinS97]. The tonality and/or harmonicity of most natural sound sources as well as the ability for phase locking in human hearing [Moor12] diminish above approximately 4–5 kHz, which coincides with the frequency of the highest musical note, C8 $\approx$ 4.19 kHz playable on a piano or piccolo flute. A predictor range from 70 Hz (to avoid DC offset or LF hum) to 4.2 kHz should, therefore, suffice. In fact, informal empirical investigations in the preparation of this study indicate that even a bandwidth of 3 kHz does not restrict the FDP performance in a significant manner.

Frequency-domain prediction on a harmonic grid. In order to address the remaining third bullet point on page 88, the forward-adaptive design of [Ojan99] is pursued, where the LTP parameters and activation flags are transmitted for every frame and channel as part of the coded bit-stream (instead of deriving them at both the encoder and decoder, as in backward-adaptive schemes), and a line-wise spectral second-order predictor

$$\bar{Y}_i(k) = Z_1(k) \cdot t_1 + Z_2(k) \cdot t_2, \;\; 0 \le k < K, \tag{3.29}$$

as in [YinS97], with k as an indicator of the MDCT line index and the reconstructed $Z_{\{1,2\}}$ given by (3.27) via $P = 1$ and 2, respectively. This linear combination of Z_P with weights t_1 and t_2 allows for high temporal resolution in the prediction since pitch lags including fractions of the frame length N can be used. However, the costly spectral band-wise signaling/activation and line-wise prediction for *any* k below limit $K < N$ is undesirable.

To determine t_1 and t_2 in (3.29), consider a conventional MDCT spectrum of length N,

$$Y_i(k) = \sum_{m=0}^{M-1} y_i(m)\, w(m) \cos\left(\omega_k \left(m + \tfrac{N+1}{2}\right)\right), \quad 0 \le k < N, \tag{3.30}$$

where the orthonormality factor $\sqrt{2/N}$ shall be included in window w, and the modulation frequency is $\omega_k = \frac{\pi}{N} \cdot \left(k + \frac{1}{2}\right)$; see also pages 11 and 13. Assume, further, that the input waveform y is a harmonic signal composed of several sinusoids at frequencies ω_s,

$$y_i(m) = \sum_{s=1}^{\lfloor N/s_0 \rfloor} y_s(m) = \sum_{s=1}^{\lfloor N/s_0 \rfloor} \cos(\omega_s(m + iN) + \varphi), \quad 0 \le m < M, \omega_s = \tfrac{s\pi}{N} s_0, \tag{3.31}$$

where φ denotes an arbitrary phase offset and where each harmonic at index $s > 1$ lies at an integer multiple of the fundamental frequency s_0, for which $s = 1$. This s_0 is a FD equivalent of the LTP periodicity lag $p_{\{l,r\}}$ in (3.26) and, conveniently, indicates the spectral spacing between the individual harmonics in units of the line index k. As long as the minimum value for s_0 exceeds the main lobe width exhibited by w for sufficient FD harmonic separation, which is the case for $s_0 \ge 3$ (70.3 Hz for $N = 1024$ and 48 kHz sample rate), an efficient line-selective FDP can be realized. Specifically, for the above example, only the spectral coefficients at $k < K$ whose ω_k are close enough to the nearest ω_s, e. g.,

$$|\omega_k - \omega_s| < \tfrac{3\pi}{M}, \tag{3.32}$$

are to be subjected to the FDP of (3.29), with the appropriate s of (3.31). Note that the predictor coefficients $t_{\{1,2\}}$ only need to be computed once for each ω_s (not for each ω_k), as demonstrated hereafter. For all k satisfying (3.32) and assuming adequate side lobe attenuation due to window w at $|\omega_k - \omega_s| \ge \frac{3\pi}{M}$ as well as the ideal, distortion-free case $Y_i(k) = Z_0(k)$, the predictor weights for each ω_s can be obtained from a system of equations. In particular, the following dependencies, which take into account the hop size N,

$$Z_0(k) = \mathrm{MDCT}[y_s \cdot w](k) = A_k \cos(iN\omega_s + \varphi_k) \tag{3.33}$$

with $\varphi_k = \varphi_s - \omega_s\left(N - \frac{1}{2}\right) + \omega_k\left(\frac{sN}{2}\right)$ and A_k depending on w and the difference $\omega_s - \omega_k$, can be made use of. From these dependencies, first $Z_1(k)$ and $Z_2(k)$ can be determined:

$$Z_1(k) = A_k \cos((i-1)N\omega_s + \varphi_k) = Z_0(k)\cos(N\omega_s) + A_k \sin(iN\omega_s + \varphi_k)\sin(N\omega_s),$$
$$Z_2(k) = A_k \cos((i-2)N\omega_s + \varphi_k) = Z_0(k)\cos(2N\omega_s) + A_k \sin(iN\omega_s + \varphi_k)\sin(2N\omega_s). \tag{3.34}$$

Solving the FDP condition $\bar{Y}_i(k) \stackrel{\text{def}}{=} Y_i(k)$, i. e., $\bar{Y}_i(k) \stackrel{\text{def}}{=} Z_0(k)$, then leads to the equations

$$t_1 \cos(N\omega_s) + t_2 \cos(2N\omega_s) = 1, \;\; t_1 \sin(N\omega_s) + t_2 \sin(2N\omega_s) = 0, \tag{3.35}$$

and, thereby, to the final solution $t_1 = 2\cos(N\omega_s)$, $t_2 = -1$ or, with an extra scalar g_{opt},

$$t_1 = 2g_{\text{opt}} \cos(N\omega_s), \;\; t_2 = -(g_{\text{opt}})^2. \tag{3.36}$$

Conveniently, both t_1 and t_2 are independent of A_k and φ_k, i.e., the waveform's amplitude and phase offset, respectively. Hence, (3.36) is also applicable to MELT based coding.

The factor $0 \le g_{\text{opt}} \le 1$ in (3.36) is the optimal FDP gain, which can be used to maximize the prediction gain G, i.e., the ratio of the MDCT variances before and after the FDP,

$$G(s_0) = 10 \log_{10}\left(\sum_{k=3}^{K-1} \big(Y_i(k)\big)^2 \Big/ \sum_{k=3}^{K-1} \big(Y_i(k) - \bar{Y}_i(k)\big)^2 \right) \tag{3.37}$$

specified in units of dB. Here, k starts at 3 to account for the lower FDP limit of 70 Hz at 48 kHz sample rate. s_0 and $g = g_{\text{opt}}$ can be chosen such that $G(s_0)$ is maximized, which coincides with a minimum prediction error (denominator in G) in a least-squares sense.

It is worth noting that a fixed gain of $g \approx 0.9$ reduces the FDP analysis workload (i. e., parameter search) and side-information (i. e., parameter rate) since g_{opt} does not need to be calculated or transmitted. Nevertheless, for natural tonal signals, it works almost as well as g_{opt}, especially when g depends on ω_s. For instance, solving a constraint like

$$(1 - t_1 \cos(N\omega_s) - t_2 \cos(2N\omega_s))^2 + (t_1 \sin(N\omega_s) + t_2 \sin(2N\omega_s))^2 \stackrel{\text{def}}{=} \frac{1}{256} \tag{3.38}$$

for $g(\omega_s)$ instead of g_{opt} in t_1, t_2 of (3.36) to enforce a constant attenuation (prediction effect) at the prediction filter's notch frequency for each ω_s, irrespective of ω_s, yields an expression which can be approximated quite closely by the even-exponent polynomial

$$g(\omega_s) \approx \frac{1983}{2048} - \frac{87}{2048} \cdot \cos(N\omega_s)^2 - \frac{360}{2048} \cdot \cos(N\omega_s)^6. \tag{3.39}$$

Note that (3.39) can be computed quite efficiently given that the term $\cos(N\omega_s)$ already occurs in t_1 of (3.36) and given that the divisions by 2048 can be realized using shifts by 11 in integer implementations. With regard to $\bar{Y}_i$ of (3.29), the usage of $g(\omega_s)$ results in a constant *application* of g, with a strength near 24.1 dB for each instance of ω_s, instead of a constant value of g but a varying — and often quite low — strength at said instances.

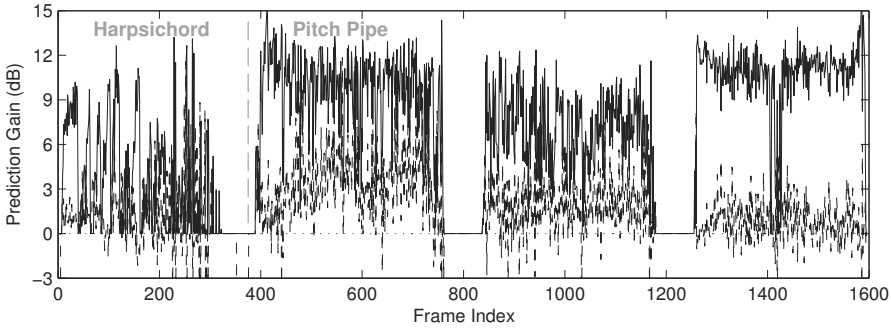


Figure 3.21. Relative gain of (—) ω_s based FDP of subsection 3.3.1 vs. coding without any LTP, (- -) enhanced FDP of subsec. 3.3.2 vs. basic FDP of subsec. 3.3.1, on tonal input.

Figure 3.21 shows the frame-wise value of $G(s_0)$ (solid line) in a single-channel AAC based codec operating at 0.5–1 bit per TD sample (constrained variable bit-rate, CVBR), 48 kHz sampling rate, $N = 1024$ samples, and $K = 132$ MDCT spectral coefficients (for a prediction bandwidth of 3.1 kHz), with $g(\omega_s)$ according to (3.39). For this evaluation, s_0 was quantized to an 8-bit index on the following ERB-like [Moor12] nonlinear scale:

$$s_0^q = \frac{298 \cdot 3}{298 - pi_{\text{opt}}}, \quad 0 \le pi_{\text{opt}} < 2^8, \tag{3.40}$$

with pi_{opt} representing the pitch/periodicity index which is to be transmitted to the FDP decoder. Despite such (moderately coarse) quantization, it is evident from Fig. 3.21 that prediction gains of 7 dB or more are achieved on the depicted exemplary input, namely, throughout the three-tone pitch pipe signal and the first seven tones of the harpsichord arpeggio. Details on these single-instrument recordings will be provided in Chapter 4.

3.3.2 Enhanced Frequency-Domain Prediction for Low-Rate "Lossy" Coding

The spectral quantization in a perceptual transform coder leads to an inevitable loss of information which, as noted in Chapter 2, is often modeled as an added noise signal propagating into (3.26) and (3.27). For $|\cos(N\omega_s)|$ in (3.36) approaching 1 and a typical gain of $g \approx g_{\text{opt}} \approx 0.9$, the FDP design of subsection 3.3.1 amplifies the quantizer noise variance in the predictor memory by a factor of 4, which reduces the achievable $G(s_0)$.

For the two "center" coefficients $Y_i(k)$ near each harmonic at ω_s, the predicted MDCT (or MELT) values $\bar{Y}_i(k)$ can, alternatively to the exclusively *temporal* approach of (3.29), be obtained from the last frame's coefficient $Z_1(k)$ and its *spectral* neighbors, $Z_1(k \pm 1)$:

$$\bar{Y}_i(k) = \begin{cases} Z_1(k) \cdot f_1^+ + Z_1(k+1) \cdot f_2^+, & \omega_k < \omega_s, \\ Z_1(k) \cdot f_1^- + Z_1(k-1) \cdot f_2^-, & \omega_k \geq \omega_s. \end{cases} \tag{3.41}$$

Given the window's real-valued oddly-stacked DFT (ODFT) derived frequency response,

$$W_0(\omega) = \mathcal{F}\{w(m)\} \cdot \mathrm{e}^{j\omega(N-½)}, \quad 0 \leq m < M, \tag{3.42}$$

for the current frame and channel and again solving the abovementioned FDP condition $\bar{Y}_i(k) \stackrel{\text{def}}{=} Y_i(k)$, i. e., $\bar{Y}_i(k) \stackrel{\text{def}}{=} Z_0(k)$, but this time for f_1^+, f_2^+ and f_1^-, f_2^-, respectively, yields

$$f_1^+ = g\cos(N\omega_s), \quad f_2^+ = g\sin(N\omega_s) \cdot \frac{W_0(\omega_s - \omega_k)}{W_0(\omega_s - \omega_{k+1})} \tag{3.43}$$

for $\omega_k < \omega_s$ (taking the upper neighbor) and, accordingly, for $\omega_k \geq \omega_s$ (lower neighbor),

$$f_1^- = g\cos(N\omega_s), \quad f_2^- = -g\sin(N\omega_s) \cdot \frac{W_0(\omega_s - \omega_k)}{W_0(\omega_s - \omega_{k-1})}. \tag{3.44}$$

Employing FDP (3.43) or (3.44), ω_s-selectively, instead of (3.29) whenever the condition

$$\left(f_1^{\pm}\right)^2 + \left(f_2^{\pm}\right)^2 < \left(t_1\right)^2 + \left(t_2\right)^2 \tag{3.45}$$

is true reduces the maximum noise variance amplification to a factor of 2 for $g \approx 0.9$.

In practice, however, this algorithmic extension only results in a *best-case* — but still barely audible — prediction gain increase of about 4 dB (to 15 dB on the first pitch pipe tone, dashed line in Fig. 3.20) even for bit-rates as low as 0.5 bit per sample and channel. It also comes at the cost of additional computational and implementational complexity (more comparison operators with (3.41) and (3.45), more static memory usage because of more look-up tables, and frame-wise dependency of (3.41) on W_0). Due to this lack of practical benefit given the increased "effort", this method was not evaluated in detail.

3.3.3 Extended Frequency-Domain Prediction for Joint-Channel Coding

Analogously to the joint-channel extension of traditional intra-channel FD predictors [Fuch93, Fuch95], the periodicity-based line-selective FDP principle can also be applied to improve state-of-the-art JS coding approaches such as the complex stereo prediction tool described in Chapter 2. More precisely, the approximation of the MDST of downmix D — the imaginary part of the JS predictor, whose limited accuracy was revealed in subsection 3.2.1 — from the current and previous frame's MDCT downmixes as in [Helm11, ISO12] can be replaced with a variant of the low-rate enhanced FDP discussed above for the FD coefficients residing on the determined (joint-channel) harmonic grid [Hel16c]. However, the practical objective or subjective advantage over legacy complex-predictive JS coding is too small to justify the inevitable increase in algorithmic complexity. Hence, like the low-rate enhanced FDP, the stereo extended FDP was not investigated further.

3.4 Transform-Domain High-Frequency Gap Filling

Moving on to parametric coding tools, the efficient reconstruction of HF components directly in the transform domain — thereby rendering separate pseudo-QMF-like filter banks around the core-codec unnecessary — is addressed on the following pages. More precisely, an HFR coding extension for MPEG-style perceptual audio coders is developed which, in terms of algorithmic complexity, is only slightly more expensive than the SPX scheme of E-AC-3 [Field04] or the spectral folding in CELT [Valin13]. At the same time, all Enhanced SBR [Neue13] or A-SPX [Kjör16] functionality described in section 2.4, plus

- spectrotemporal envelope shaping as well as flattening on a signal-adaptive grid,
- NF, frequency-selective control of the tonality/noisiness in each parameter band,
- partial waveform preservation in the HF range to enable *semi-parametric* coding,

is supported by the proposed method, hence setting it apart from the former designs as well as the alternative approaches mentioned in section 2.4 [Ferre05, Laak05, Anna06, Sinha06, Tsuji09, LeeC13, Neuk13]. This is attained by way of a tight integration of the technique into the spectral pre-/post-processing module of the coder (see also Fig. 2.2), as explained hereafter, which allows the full exploitation of the existing FD coding tools for the desired purpose and, hopefully, the prevention of audio quality shortcomings.

For consistency, the remainder of this section is organized analogously to section 2.4. Subsection 3.4.1 revisits the codec design and the differences and similarities between the pseudo-QMF and TDAC core-coder filter bank instances, and subsection 3.4.2 shows how appropriate T/F grid selection and a complex domain for the parameter extraction can be achieved in a trivial way. Subsection 3.4.3 then examines the transform-domain generation and spectrotemporal flattening of the fundamental HF content, followed by brief descriptions of the proper HFR envelope extraction and adjustment procedures in subsection 3.4.4 given the specific context of a TDA-prone filter bank. Subsection 3.4.5, finally, presents HFR post-processing algorithms which can be applied as equivalents to E-AC-3's noise blending, SBR's missing harmonics, and USAC's Inter-TES techniques.

Most of the following discussions have been published in a distributed fashion. Some early studies toward this work were documented in [VanO10], the HF spectral envelope estimation and reconstruction aspect was addressed in [Hel15a], and the overall coding architecture was presented in [Hel15b] and, in the context of a LD scenario, [Hel15d].

Note, also, that an earlier, slightly less tightly integrated variant of the MDCT-domain HFR proposal is supported by the "enhanced noise filling" tool of the MPEG-H 3D Audio coding specification [ISO15a]. Whenever necessary, the difference between the MPEG-H version, also called Intelligent Gap Filling (IGF), and the proposal herein will be noted.

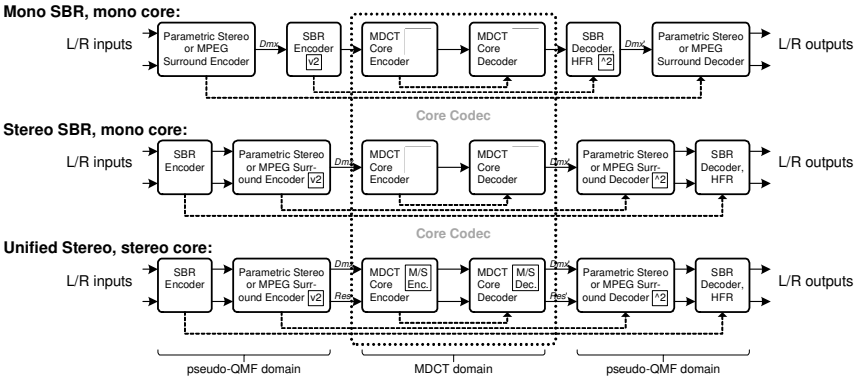


Figure 3.22. Simplified block diagrams of the supported coding-decoding chains in (Extended) HE-AAC with their QMF- and MDCT-domain tools. (—) Signals, (- -) parameters.

3.4.1 Parametric HFR Using TDAC Analysis and Synthesis Filter Banks

Subsection 2.4.1 explained that the utilization of auxiliary pseudo-QMF banks in HFR schemes like (Enhanced) SBR and A-SPX is motivated by two essential properties which these additional filter bank instances provide: a complex-valued T/F representation as well as high temporal resolution for analysis (i. e., parameter acquisition) and synthesis (i. e., parameter application for BWE). Figure 3.22 revisits the overall codec architecture of HE-AAC or USAC in the presence of the pseudo-QMF-based pre-/post-processors. The illustrated signal paths indicate that in the encoder, the MDCT core-coder needs to wait for the result of the QMF-domain pre-processing, whereas in the decoder, the QMF tools require the output of the MDCT core-decoder. Moreover, as mentioned in Chapter 2, the core signal is generally downsampled. Since the "inner" MDCT and "outer" SBR and PS or MPS codecs operate in separate filter bank domains, such sequential operation leads to an accumulation of the individual domains' algorithmic delays. In fact, the sum of all delays equals more than 200 ms at 44.1 or 48 kHz input sampling rate in both USAC and HE-AAC, with the core-codec alone exhibiting a latency of up to 120 ms [Lutz04] due to the downsampling. Semi-parametric coding is difficult as well, as demonstrated later.

Relocating the HFR tool into the MDCT domain — or, in the present work, a switched real-valued TDAC filter bank according to sections 3.1 and 3.2—as shown in Figure 3.23 implies the loss of the complex-valued signal representation with high time resolution at first glance. However, the following subsection will illuminate how the existing block size and window shape switching functionality as well as the possibility of an MCLT-like encoder-side T/F mapping can compensate for this apparent drawback to a large extent.

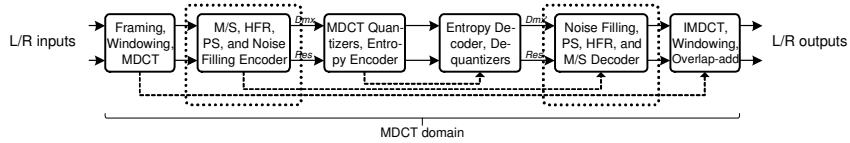


Figure 3.23. Block diagram of the proposed modification to USAC's Unified Stereo architecture depicted in Fig. 3.22. The dotted boxes mark the location of the HF gap filling tool.

3.4.2 Modified FD Extraction of the HFR Control Parameters

Subsection 2.4.2 noted that the HFR parameter acquisition, comprising the measurement of the spectrotemporal envelope (i. e., energy) and flatness information associated with the input signal(s), must be performed with high time resolution since it involves

- a transient detection algorithm based on which the T/F grid for BWE is selected,
- the calculation of a TFM value to determine a temporal "smoothness" parameter.

Conveniently, sufficiently accurate versions of both the T/F grid and the TFM value are readily available from the existing core coder tools preceding, in the encoder, the set of FD pre-processing tools to which the proposed "gap filling" BWE module is intended to be added. Specifically, these "outer" core coder tools around the newly proposed "inner" HFR module (see Fig. 3.23) include the adaptive TDAC filter bank itself, with its window shape, block length, transform kernel (KS), and overlap ratio (RS) selection examined in sections 2.1, 3.1, and 3.2, as well as the TNS filtering or sub-band merging functionality described in subsection 2.2.1. The use of these tools for HFR is summarized hereafter.

- T/F grid selection. A transient detector, typically operating directly on the PCM input samples for each channel, is utilized to select the inter-transform overlap width (by way of window shape and, possibly, the proposed overlap ratio adaptation) and, in highly non-stationary transient signal parts, the transform length (by means of *long*-vs.-*short* block length switching). In a well-designed encoder, the final selection of the window and transform sequences is governed not only by channel-wise *statistical* data like temporal sample variance progressions but also takes relevant *psychoacoustic* information, namely, spectrotemporal masking patterns associated with the instantaneous input signal(s), into account. The set of window shape, overlap ratio, transform size, and kernel data representing the transmittable T/F grid parameters for every frame can, thus, be regarded as the objectively (in terms of redundancy reduction) and subjectively (in terms of irrelevance reduction) optimal choice. In fact, it can be argued that this applies not only to waveform-preserving but also to the desired parametric FD coding.

- TFM computation. A temporal flatness indicator, as outlined in subsection 2.4.2, describes the *fine* TD envelope or "buzziness" of the analyzed waveform which is typically not modeled by the relatively *coarse* T/F grid data examined previously. Assuming, again, a well-designed encoder, this information is readily accessible, in varying representations, in three ways. First, the TD transient detector makes use of a high-pass filtered version of each channel's input waveform to obtain a reliable quantifier of the instantaneous (non-)stationarity. The PCM domain, by definition, exhibits the highest possible time resolution for analysis, so it serves well to derive the desired TFM data, e. g., analogously to Wiener's entropy (SFM, see also page 80), particularly when the high-pass filter isolates the HFR region well. Second, a TNS filter, computed on the FD coefficients representing the HFR range prior to quantization, contains information on the fine temporal envelope of said frequency range with sub-transform resolution. Hence, a "non-flat" filter for which a high prediction gain is obtained (see subsection 2.2.1) indicates high non-stationarity, within the transform's TDA-domain support, for the examined spectral region. Third, the same applies when analyzing and parameterizing the temporal power/variance/L_2 norm evolution for the same frequency range after sub-band merging, albeit in a less continuous "step-function-wise" manner: if it varies notably across time, the HF signal can be considered transient or "buzzy".

The spectrotemporal envelope, being the most expensive and perceptually important HFR parameter, is preferably computed from a complex-valued FD representation since only the latter, but not a real-valued one, allows correct input energy measurements. It is worth repeating, in this context, that for each cosine-modulated filter bank transform of sections 2.1 or 3.2, a sine-modulated counterpart can be constructed, and vice versa. This can be achieved by simply exchanging the trigonometric term cs(·) and, if needed, the spectral offset k_0 and scalar $s(\cdot)$ in the respective analysis and synthesis definitions, and by identically applying the resulting "imaginary" filter bank part on the input signal. Using these modulation pairs, the desired complex-valued variants of the MDCT, MDST, or MELT, exhibiting the form MDCT + jMDST (known as MCLT [Malv99] for the type-IV case, see subsection 3.2.1), MDST − jMDCT (analytical reversal of the MCLT), cos-MELT + $j$sin-MELT, or sin-MELT − $j$cos-MELT, respectively, can be assembled. The parameter-band-wise HF envelope information is then obtainable as described in subsection 3.4.4.

The TNS filter coding scheme, which has remained mostly unchanged since [ISO97], supports up to three filters per *long* transform. Hence, by enabling dedicated TNS filter data for the HFR range, it readily allows to convey a TFM parameter for the gap filling.

A transient flag, the T/F grid, and SFB data are already coded as part of the per-frame filter bank configuration. Entropy coding of the HFR envelope is outlined in section 3.6.

3.4.3 Transform-Domain Generation and Flattening of HF Content

In principle, the proposed transform-domain gap filling algorithm may generate the basic HFR signal components identically to the QMF-based BWE schemes, i. e., via transposition (copy-up) or folding (mirror-up). However, at the decoder side, where the HFR is to be applied, the core-decoded coefficients only explicitly exist in a real-valued form. A potential conversion to a complex-modulated MCLT-like representation as described on the last page — either directly via additional analysis transforms or indirectly by way of Cheng's R-to-I method (see also [Chen04, Helm11] and page 66) — is undesirable due to the resulting increase in algorithmic complexity and, possibly, latency (by one frame). Thus, it was decided to, in this study, perform the gap filling directly in the TDA-afflicted core-transform domain, which, fortunately, only poses a minor and unproblematic extra constraint on the HF generator: the transposer distance d must be even, and the mirror axis must lie between two FD indices, i. e., the folding distance $2(k-d)+1$ must be odd,

$$\text{copy-up: } X_i^q(k) = X_i^q(k-d), \text{ mirror-up: } X_i^q(k) = X_i^q(2d-1-k),\ d \le k < 2d, \quad (3.46)$$

with X_i^q being the reconstructed FD values for frame i after quantization, as previously.

Having adopted the core coder's T/F grid also for parametric gap filling, it is assumed that the SFB partitioning for the spectral quantization and JS pre-/post-processing can be reused as well. Informal evaluation indicates that this assumption is reasonable, and since the SFB offsets and widths, in units of the spectral index k, are integer multiples of 4 (see subsection 2.3.1 and Fig. 2.9), they guarantee that (3.46) is satisfied completely.

It is worth noting, in this regard, that, when loosening the requirement of a low HFR decoding complexity, harmonic TDA-domain transposition analogously to the pseudo-QMF-domain approach in USAC [Zhon11, ISO12, Neue13] is possible as well [Neuk13]. The perceptual benefit of this technique — which, as a side effect, circumvents the above restrictions of (3.46) — over the previously discussed simple transform-domain copy-up was, however, found to be much too small in typical use cases (IGF start frequencies of more than 4–5 kHz) to justify the required additional decoding complexity and delay.

As in Enhanced SBR and, to some extent, A-SPX, the generated HF content, inheriting its fine spectrotemporal structure from the LF source region, can be subjected to band-wise flattening in frequency and time direction. The former, realized by inverse filtering or pre-flattening in the QMF tools, can be imitated via multiplicative *whitening* methods operating on the transposed/folded transform coefficients X_i^q of (3.46) as these exhibit a considerably higher spectral resolution than the pseudo-QMF samples [Schm16]. The latter, applied multiplicatively by way of temporal sub-band smoothing in QMF-domain parametric processing, requires TNS-like analysis filtering at the decoder side [ISO15a].

3.4.4 Estimation and TDA-Domain Adjustment of HF Envelope

IGF is employed for parametric signal reconstruction of a spectral band b comprising multiple transform coefficients set to zero by the encoder, either deliberately a-priori or by hitting the dead-zone of the FD quantizer. For each IGF band — which, in the present study, is equivalent to a core-coder SFB in the HF range — an envelope value in the form of an energy scale-factor, similar to those used in HE-AAC's PNS [ISO09] and USAC's or AC-4's NF [ISO12, ETSI14], is coded and transmitted. As noted earlier, complex energy calculations are preferable, but for completeness, the simpler real-valued computation, which can be useful in case of activated TNS [Hel15a], shall also be documented herein.

Let $X_i \in \mathbb{R}^N$ denote the real-valued transform-domain spectral representation of the windowed audio signal of window length M or L for frame index i, as previously. Given the IGF start frequency as a coefficient index k_s, with the SFB partitioning into intervals

$$p_i(b) = \left[\, k_s + \sum_{l=1}^{b-1} c_i(l),\ k_s + \sum_{l=1}^{b} c_i(l) \,\right[,\quad b_s \le b < B,\quad 1 \le b_s < B, \tag{3.47}$$

based on the widths c_i of the HF SFBs (see also page 27), the envelope per b is defined as

$$E_i(b) = \sqrt{\frac{1}{c_i(b)} \cdot \sum_{l \in p_i(b)} X_i(l)^2},\quad b_s \le b < B,\quad 1 \le b_s < B. \tag{3.48}$$

To prepare $E_i(b)$, parametrizing the RMS of b, for transmission, quantization is applied,

$$e_i(b) = \left\lfloor 4 \cdot \log_2\left(s \cdot E_i(b)\right) \right\rceil,\quad b_s \le b < B,\quad 1 \le b_s < B, \tag{3.49}$$

similarly to the scale factor related formulation of (2.41). The scalar s serves to shift the $E_i(b)$ to a convenient range for the logarithmic quantization (a comparable approach is pursued in legacy MPEG audio coding [ISO97–ISO12]; there, s is applied additively as an offset in the logarithmic domain). The envelope levels e_i, indicating the SFB-wise "real-only" HFR target energies, can now be subjected to lossless entropy coding (details will follow in section 3.6) and, afterwards, multiplexed into the IGF-related bit-stream part.

The receiver, after demultiplexing and entropy decoding all bit-stream components, applies reconstructive scaling to obtain the quantized spectral coefficients X_i^q according to subsection 2.3.2, both in the LF core and HF gap filling region. Thereafter, traditional FD core NF is carried out up to k_s exclusively. Due to the psychoacoustically motivated coarse FD quantization of X_i in the encoder, X_i^q largely or completely consists of zeroed transform coefficients within the HFR bands and, because NF is not used in the range at and above k_s, exhibits large spectral gaps. The IGF envelope E_i^q is now reconstructed via

$$E_i^q(b) = \frac{1}{s} \cdot 2^{e_i(b)/4},\quad b_s \le b < B,\quad 1 \le b_s < B, \tag{3.50}$$

i. e., by inverting (3.49) analogously to (2.41) in subsection 2.3.4, with e_i being the transmitted entropy decoded IGF energies. With X_i^q, the *survived* waveform coded energy S_i, accounting for the power loss due to the zeroing of transform coefficients, is computed:

$$S_i(b) = \sum_{l \in p_i(b)} \left(X_i^q(l)\right)^2, \quad b_s \le b < B, \quad 1 \le b_s < B. \tag{3.51}$$

Again based on X_i^q, the associated respective *tile* (or source) energies T_i are determined:

$$\text{transposer: } T_i(b) = \sum_{l \in z_i(b)} \left(\bar{X}_i^q(l-d)\right)^2, \quad \text{folder: } T_i(b) = \sum_{l \in z_i(b)} \left(\bar{X}_i^q(D-l)\right)^2, \tag{3.52}$$

where $D = 2d - 1$, $z_i(b)$ is the subset of all indices l in $p_i(b)$ at which the corresponding quantizer level $q_i(l) = 0$ (see section 2.3), and transposition distance d is predefined as in (3.46). The bar accent on the X_i^q vector indicates the possible (but optional) temporal and/or spectral pre-flattening of the IGF source coefficients, as described in the last subsection. Since $E_i^q(b)$ reflects the *target* (i. e., input) RMS, the *missing* energy is given by

$$M_i(b) = c_i(b) \cdot \left(E_i^q(b)\right)^2 - S_i(b), \quad b_s \le b < B, \quad 1 \le b_s < B, \tag{3.53}$$

from which, for every b, a tile gain $t_i(b) = \sqrt{\max(M_i(b), 0)/(T_i(b) + \varepsilon)}$ can be obtained. Note the addition of ε, a tiny value, in the denominator to avoid divisions by zero. Then, the content in the HF spectral gaps can, finally, be reconstructed by way of substitution:

$$\text{transposer: } \tilde{X}_i^q(l) = \begin{cases} \bar{X}_i^q(l-d) \cdot \min(t_i(b), 10), & q_i(l) = 0, \\ X_i^q(l), & \text{otherwise}, \end{cases} \tag{3.54}$$

$$\text{folder: } \tilde{X}_i^q(l) = \begin{cases} \bar{X}_i^q(D-l) \cdot \min(t_i(b), 10), & q_i(l) = 0, \\ X_i^q(l), & \text{otherwise}, \end{cases} \tag{3.55}$$

for all $l \in p_i(b)$ and all $b_s \le b < B, 1 \le b_s < B$, as above. This procedure, which is to be executed individually for each transform group (i. e., subscript i represents both a frame index and, for *short*-block frames, a group index within the frame), effectively replaces the zero-quantized, not noise filled HF coefficients in each SFB subset $z_i(b)$ via tile copy or mirror-up, leaving the non-zero quantized spectral values unaffected. Hence, precise coefficient-wise selection of the coding paradigm (waveform-preserving or parametric) is achieved without excessive increase in the side information rate — only the envelope information is slightly redundant because, for each b, separate scale factors are needed for the surviving transform coefficients (as a band-wise "local" quantization scalar) and the IGF RMS values (to convey the band-wise target energies used to compute vector t_i). Moreover, the HFR via (3.51)–(3.55) remains in the real-valued domain of X_i, as desired.

Complex-valued envelope calculation. The previous section demonstrated that IGF can preserve the fine-structure of a HF signal via LF tile copying/mirroring, or via waveform coding by allowing to encode non-zero HF transform coefficient levels. Moreover, it was shown that the IGF envelope information is computable from only the real-valued transform samples. However, these real-valued samples represent only part of the analytic signal obtained using a complex transform [Malv99] and, thus, do not give access to the actual, or "true", input energy in a given spectral region [Zhan13]. In case of a spectrally flat, relatively coarse band partition $p_i(b)$ comprising many bins, this disadvantage has only little impact, and the implicit averaging in the measurement of $e_i(b)$ still leads to a good estimate of the "true" target energy for said band b. For a narrow $p_i(b)$ with small $c_i(b)$ and a tonal signal as input, especially a high-pitched one where only one harmonic falls into each band, the effect is more obvious and, as demonstrated in [Hel15a], causes temporal amplitude modulation in the HFR decoded output for some configurations.

Two changes to the real-valued HFR envelope calculation of (3.48) and (3.49) serve to avoid the described modulation in an IGF processed decoding (reaching a strength of a few dB in some cases) when the HF target signal is quasi-stationary. Firstly, a complex MCLT-based RMS value $E_i'(b)$ can be acquired in place of the real-valued $E_i(b)$ of (3.48),

$$E_i'(b) = \sqrt{\frac{1}{c_i(b)} \cdot \sum_{l \in p_i(b)} \left| X_i(l) + jX_i'(l) \right|^2}, \quad b_s \le b < B, \quad 1 \le b_s < B. \tag{3.56}$$

with X_i' being the imaginary counterpart of X_i, e. g., the MDST in case of MDCT coding or vice versa. This approach delivers stable measurements and has been used in [Field04] in the envelope parameter derivation for the SPX tool of Dolby Digital Plus [ETSI12]. In fact, when substituting $E_i'(b)/\sqrt{2}$ for $E_i(b)$ in (3.49) — which, ignoring any quantization differences and assuming $S_i(b) = 0$, is equivalent to the method in [Field04] — the HFR decoder-side modulation is reduced in strength, albeit only very moderately [Hel15a].

Secondly, a complex compensation for usage with (3.56) is, therefore, proposed that accounts for the TDA inherent in the real-only transform-domain HFR processing in the decoder: given the complex-domain *target* powers E_i', the real-to-complex *source* ratios

$$\text{transposer: } r_i(b) = \sqrt{\sum_{l \in z_i(b)} \left(X_i(l-d) \right)^2 \Big/ \sum_{l \in z_i(b)} \left| X_i(l-d) + jX_i'(l-d) \right|^2}, \quad \text{folder: resp.} \tag{3.57}$$

are determined from the LF input X_i (or, if available, X_i^q or $\bar{X}_i^q$), and $e_i(b)$ is replaced by

$$e_i'(b) = \left\lfloor 4 \cdot \log_2\left(s \cdot r_i(b) \cdot E_i'(b) \right) \right\rfloor, \quad b_s \le b < B. \tag{3.58}$$

Transmitting e_i' as IGF envelope notably reduces the modulation on some tonal material.

3.4.5 Transform-Domain Post-Processing of Adjusted HF Content

A TNS filter computed for some frequency region represents, as noted in subsection 3.4.2, a specific variant of a TFM parameter for that spectral region, which can be coded and transmitted to the decoder. Since, when TNS filtering is activated, the proposed gap filling is preferably applied in the TNS residual (i. e., prediction error) domain, temporal sharpening with sub-group resolution, similar to USAC's Inter-TES [ISO12, Neue13], can simply be realized using conventional TNS synthesis filtering. Specifically, TNS analysis filters are calculated and applied as usual on the encoder-side input spectra, where, for best performance, at least one filter is employed for the core transform range below the HFR start frequency, and at least one filter is used at and above this start frequency. The spectrotemporal envelope data and other gap filling parameters are then derived from the TNS residual spectra and, in the decoder, applied to the (re)quantized and inversely scaled instances of these residuals. The resulting "gap filled" reconstructed spectra are converted into the final output spectra by way of TNS predictive synthesis filtering, i. e., by adding a prediction value to each transform coefficient according to the transmitted TNS filter data. The output spectra are then processed by the synthesis filter bank(s).

Figure 3.24 illustrates the order of the codec operations in the presence of the envisioned gap filling. Notice that, in the HFR spectral region, the TNS module is renamed to Temporal Tile Shaping (TTS) to emphasize that the synthesis filtering is performed on a tile-based semi-parametrically reconstructed HF signal [Disch15]. There is, however, no algorithmic difference between TNS and TTS, especially considering the fact that TTS is also applied on the non-parametric waveform coded HF spectral values. Put differently, the "artificially" regenerated gap filling content may be regarded as a special form of FD quantization error, i. e., spectral noise, which does not affect or influence the TNS tool.

In SBR and A-SPX, spectral sharpening, as a counterpart of the previously discussed temporal sharpening, is realized by inserting "missing harmonics" in the pseudo-QMF domain. In transform-domain gap filling, sinusoidal HF components — or, in general, FD coefficients constituting distinct spectral peaks which cannot be reconstructed by copy-up or mirror-up of LF content — can be waveform coded directly as "survived lines", as noted in the last subsection. In AC-4, where the core transform coder can operate at the full input/output sample rate like in downsampled HE-AAC, similar interleaving of non-parametrically and parametrically coded elements can be achieved [Kjör16]. However, since A-SPX is QMF-based, the spectral selectivity for the waveform/parametric coding decision is very low and, hence, only allows for relatively coarse *mixing* of the two paradigms within a pseudo-QMF sub-band. The proposed core-codec gap filling, in contrast, supports disjoint *non-mixing* operation, i. e., either waveform or parametric coding, with both high spectral resolution (in *long* blocks) and high time resolution (in *short* blocks).

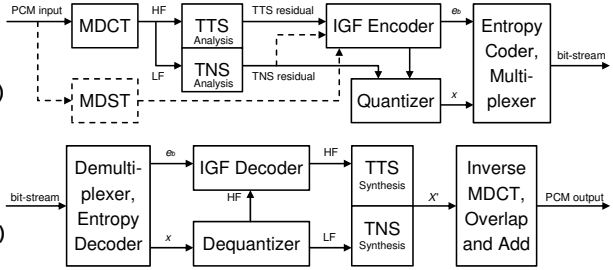


Figure 3.24. Transform-domain semi-parametric gap filling in combination with TNS. (a) Encoder with scaling quantizer, (b) decoder with inversely scaling dequantizer, (- -) optional paths [Hel15a]. e_b, x abbreviate $e_i(b)$, $q_i(l)$.

In the IGF implementations of EVS [ETSI16] and 3D Audio [ISO15a], a tonal-to-noise ratio control by means of pseudo-random noise blending, as present in (Enhanced) SBR and (A-)SPX, only exists in a binary form: in case of disabled or "medium" whitening, no noise is mixed in at all, and in the remaining case of "strong" whitening, the associated HFR content (except for the survived transform coefficients) is synthesized exclusively from noise, i. e., no copy- or mirror-up is performed. If the quantized LF source content does not contain any noise (e. g., because the FD core noise filling is deactivated) but the original HF spectrum is partially noisy, this may lead to frequent and potentially audible *on-off* toggling of the "strong" whitening mode or to incorrectly regenerated HF tonality. To complete this section, a straightforward combination — and, thereby, unification — of the proposed gap filling and the existing FD noise filling is, hence, described hereafter.

Maintaining USAC's NF principle of subsection 2.3.3 with its low parameter rate, the noise level $nl_i \geq 0$ for each channel and frame i shall be divided into two instances: one for the LF spectrum below the HFR start frequency, and one for the HF gap filling range. In the encoder, the former value nl_i^{LF} is determined as usual, while the latter value nl_i^{HF} is derived from an SFM-based noise floor parameter, just like the noise blending data in SBR, E-AC-3, and AC-4. After quantization and transmission of the two levels, and prior to the gap filling process, core NF is performed in the decoder, but now also in the HFR region and with potentially different noise magnitudes in the LF and HF portions. Then,

$$\text{transposer: } \tilde{X}_i^q(l) = \begin{cases} X_i^q(l) + \bar{X}_i^q(l-d) \cdot \min(t_i(b), 10), & q_i(l) = 0, \\ X_i^q(l), & \text{otherwise}, \end{cases} \tag{3.59}$$

$$\text{folder: } \tilde{X}_i^q(l) = \begin{cases} X_i^q(l) + \bar{X}_i^q(D-l) \cdot \min(t_i(b), 10), & q_i(l) = 0, \\ X_i^q(l), & \text{otherwise}, \end{cases} \tag{3.60}$$

with (3.50)–(3.53) as above, i. e., the pre-flattened LF tile contribution does not *replace* the zero-quantized (not noise filled) HF samples but is *added* to the (noise pre-filled) HF samples. Since the LF tile and HF noise-fill values are, by definition, uncorrelated, their per-b powers can be assumed to add up, so the technique of (3.59), (3.60) is reasonable.

3.5 Transform-Domain Semi-Parametric Stereo Filling

The previous section presented a TDA-domain HF gap filling scheme for high-quality, LC, and LD extension of the codable bandwidth at medium and low bit-rates. Emphasis was put on an overview of the fundamental algorithmic procedures in a single-channel perceptual coding infrastructure. In the following, an enhancement of the IGF principle for two-channel semi-parametric stereo coding, as a low-cost alternative to the pseudo-QMF-domain MPS [ISO07, ISO12] and Advanced Coupling [ETSI14] tools, is described. It is realized in an open-loop way that, unlike closed-loop methods [Elfitr11], avoids computationally expensive analysis-by-synthesis processes in the parameter acquisition.

Section 2.5, introducing the basic operation of a parametric stereo and multichannel audio codec, noted that the essential difference to — and advantage over — the simpler intensity stereo coding lies in a subjectively acceptable resynthesis of the spatial width (i. e., ICC) of the input signal, achieved by way of an appropriate derivation and usage of a decorrelation signal. The current section will, therefore, focus on very simple LC TDA-domain decorrelation for an optimal complexity-quality tradeoff in the given context.

Figure 3.24 illustrates the location of the IGF encoder and decoder modules in a one-channel transform codec. The basic idea behind the proposed HFR stereo extension is to

- move the IGF encoding (i. e., parameter extraction) and decoding (i. e., gap filling for HF and ICC regeneration) into the core coder's downmix/residual JS domain,
- extend the IGF to lower frequencies in the residual channel by performing, upon gap filling, a *copy-over* instead of a *copy-up* of said LF residual (subsection 3.5.3),

in a two-channel codec (or in a CPE of a multichannel codec). The signal flow of such a codec, augmented by the proposed unified semi-parametric HFR and JS technique to be referred to as spectral band substitution (SBS), is shown in Figures 3.25 and 3.26. These will serve as guides through the processing chain outlined in the following subsections.

The remainder of this section, whose content has been (co-)published by the author in [Hel15b, Schu16], is organized like section 2.5. Subsection 3.5.1 examines the ICC cue extraction and parametrization missing in conventional M/S or intensity stereo coding, and subsection 3.5.2 discusses a FD active downmix-residual calculation compensating for the lack of access to a reliable per-band IPD measure in the real-valued TDA-domain spatial synthesis. Moving to the decoder side, subsection 3.5.3 addresses the transform-domain generation and spectrotemporal shaping of a decorrelated source signal for the copy-over, and subsection 3.5.4 concludes the study with a brief overview of the upmix process reconstructing the original channel configuration from the gap-filled downmix-residual representation. The performance of the SBS principle is evaluated in Chapter 4.

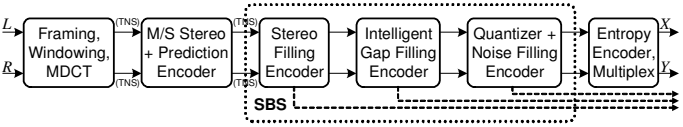


Figure 3.25. Detailed encoder of Fig. 3.23 with SBS in the JS domain. (- -) parameter data. (TNS) is optional.

3.5.1 Transform-Domain Extraction and Coding of Spatial Parameters

State-of-the-art PS codecs, as discussed in section 2.5, acquire several cross-channel spatial cues before downmixing the two input channels into a monophonic signal above a given PS start frequency (i. e., residual bandwidth) f_{rc}. The spatial cues comprise parameter-band-wise ILD, ICC, and, optionally, IPD and/or OPD values, which are quantized and entropy coded for low-rate transmission to the decoder. Moreover, the bandlimited residual signal below f_{rc}, if permitted by the RD target configuration, is waveform coded and multiplexed into the final bit-stream. The objective behind the proposed transform-domain SBS scheme is to maintain the extraction and coding of all these PS parameters or, if necessary, to construct equivalent parameters which are more readily accessible.

Regarding the ILD information, it is worth revisiting the JS approaches already in use in the FD core for "discrete" waveform preserving TDA-domain coding. Subsection 2.2.3 introduced three M/S-based intensity-stereo compatible joint-channel representations:

- normalized M/S stereo, as adopted in [VanS08, IETF12, Valin13], where the sum-difference matrix is applied on the band-wise RMS-normalized channel spectra. Here, the pair-wise ratios between the per-channel RMS data represent the ILDs.
- predictive M/S coding, with real- or complex-valued stereo prediction employed between the mid and side spectral coefficients [Helm11, ISO12, Neue13, ETSI14]. In this case, the ILDs are modeled by the real part ρ_i of the predictor coefficients.
- stereo rotation, equaling a size-2 KLT across the channel spectra, as in [Vand91]. If utilized in a band-wise manner, the rotation angles α_i indicate the bands' ILDs.

In other words, the above core-coder JS tools already provide useful parameterizations of the band-wise inter-channel intensity differences, i. e., spatial panning. The complex stereo prediction and KS tools, in addition, give access to a type of indirect IPD measure through the complex-valued predictor coefficient [Neue13] and channel-wise kernel selection, respectively, although their accuracy (two-frame R-to-I procedure in the stereo predictor) and resolution (no band-wise KS, only globally per frame in steps of 90°) are limited. Finally, the residual side spectra — or error spectra, in case of rotation-based JS coding — can be transmitted simply by avoiding a quantization to zero below f_{rc}. Hence, only the ICCs remain to be conveyed to the receiver to complete a UniSte-like data set.

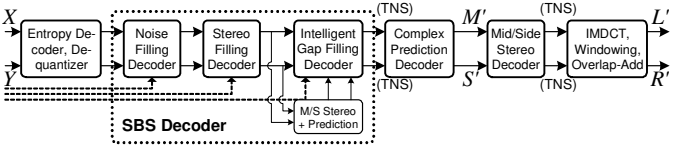


Figure 3.26. Detailed SBS enhanced USAC decoder of Fig. 3.23 associated with the encoder of Fig. 3.25.

In standalone and USAC MPS, residual coding and mixing-in of a decorrelation signal for adequate reconstruction of the inter-channel coherence are mutually exclusive: the decorrelation signal acts as a replacement for the residual signal. This substitution can, therefore, be regarded as a special form of NF (which, as noted earlier, is also, by design, uncorrelated with the zero-quantized FD coefficients at and for which it is applied). In the core NF, described in subsection 2.3.3, a frame-global or SFB-wise (in combination with scale factors) noise level nl_i controls the magnitude of the inserted pseudo-random values so as to prevent excessive noise substitution. A similar value can be determined for the desired magnitude of the decorrelation signal in the PS upmix process by simply measuring the band-wise RMS of the residual coefficients at and above f_{rc} prior to their quantization to zero (or, in general, their exclusion from transmission). More precisely, after the JS pre-processing yielding M_i and S_i or $\acute{S}_i$ (subsection 2.2.3), the residual RMS

$$er_i(b) = \left\lfloor 4 \cdot \log_2\left(s \cdot ER_i(b)\right)\right\rceil,\ ER_i(b) = \sqrt{\frac{1}{c_i(b)} \cdot \sum_{l \in p_i(b)} \left\{S_i(l) \text{ or } \acute{S}_i(l)\right\}^2}, \qquad (3.61)$$

is obtained for every band $b_{\mathrm{rc}} \leq b < b_s$, just like the IGF target envelope $e_i(b), b \geq b_s$ of (3.49) or, as will be clarified later, an equivalent of the complex-valued e_i' of (3.58). This *residual envelope* parameterizes the SFB-wise ratio between the energies of the, ideally, uncorrelated downmix (which, for the range $b_{\mathrm{rc}} \leq b < b_s$, is waveform coded and transmitted to the decoder) and residual to be quantized, i. e., the desired normalized ICC.

In the spatial upmix process of PS and MPS, the downmix contribution is maximized in order to mix in as little of the decorrelation signal (replacing the residual) as possible [Bree05, Bree07] since this results in the best subjective performance. In the context of transform-domain SBS, this design choice translates to the requirement of a maximally channel-compacting JS representation, i. e., a L/R-to-downmix/residual mapping where the (band-wise) ratio between the downmix and residual powers is as large as possible. Put differently, the JS analysis mapping of two input signals should yield a downmix and a residual signal which are uncorrelated (to a large extent) regardless of the amount of ICC between said input signals. The KLT, as is well known, is a maximally decorrelating transform, so its usage as a JS rotator according to [Vand91] and subsection 2.2.3 forms the ideal candidate for such a mapping. The normalized and predictive M/S stereo techniques, however, were found to decorrelate almost equally well (see also the Appendix).

3.5.2 IPD-Aware Calculation of JS Downmix and Residual Spectra

Given that the SBS proposal operates in — and, thus, relies upon — the JS downmix-residual domain of the transform coder, no special pre-processing steps are required in principle, and the joint-channel analysis (i. e., downmix) algorithm may operate as usual when no transform-domain parametric coding is employed. The presence of strong non-trivial IPDs in combination with PS-like "artificial" spatial reconstruction, however, can be problematic when, as in the given case, the IPD coding capabilities of the TDAC core codec are limited (e. g., to steps of 90° or to relatively low phase accuracy, see page 106). The coding and transmission of the residual spectra compensates for this shortcoming in the IPD estimation upon JS analysis channel mixing (the residual's energy increases), thereby allowing for waveform-preserving reconstruction of the spatial image in the JS upmix. If the residual is not transmitted, though, but substituted in the decoder with a decorrelation signal, scaled according to the ICC value derived as in the last subsection, a "phasy" signal will likely sound spatially wider after coding than the encoder input. In addition, destructive interference (spectral cancellation) may appear in the downmix.

Another view on this issue is that, in the presence of a distinct IPD in a spectral band, an IPD-"blind" JS analysis pre-processor creates a downmix and a residual signal which are insufficiently decorrelated, i. e., with suboptimal channel compaction into one of the two. This can be addressed by encoder-side pre-treatment of the input channel signals, with the goal of increasing the TDA-domain ICC between the input spectra in bands for which problematic IPDs are detected. In other words, on spatially directional (narrow) but phasy input, the encoder may apply intensity-stereo pre-treatment which maximizes the transform-domain ICC and, thereby, the channel compaction of the subsequent JS analysis pre-processor. An IPD-agnostic measure of spatial directionality is obtained by taking, in the encoder, the previously described complex-valued MCLT-like transform representation and by computing therewith the following normalized correlation value:

$$nc_i(b) = \frac{\left(\sum_l A_i(l)B_i(l) + \sum_l A'_i(l)B'_i(l)\right)^2 + \left(\sum_l A_i(l)B'_i(l) - \sum_l A'_i(l)B_i(l)\right)^2}{\left(\sum_l A_i(l)^2 + \sum_l A'_i(l)^2\right) \cdot \left(\sum_l B_i(l)^2 + \sum_l B'_i(l)^2\right)}, \quad (3.62)$$

with A_i and B_i denoting the channel spectra for frame i after KS and TNS coding, prime $'$ indicating the imaginary counterpart of the real-valued A and B, $l \in p_i(b)$ as previously, and $b_{\mathrm{rc}} \leq b < B$. If $nc_i(b)$ lies above a predefined threshold ($\frac{2}{3}$ was found to work well), the band's input can be considered narrow and directional, and coefficient-wise "active" stereo pre-downmixing can be performed to equalize the phases of each coefficient pair:

$$\ddot{A}_i^{(\prime)}(l) = \dot{A}_i^{(\prime)}(l) \cdot \sqrt{\frac{A_i(l)^2 + A'_i(l)^2}{\dot{A}_i(l)^2 + \dot{A}'_i(l)^2 + \varepsilon}}, \; \dot{A}_i^{(\prime)}(l) = (1-\gamma) \cdot A_i^{(\prime)}(l) + \gamma \cdot B_i^{(\prime)}(l), \quad (3.63)$$

$$\ddot{B}_i^{(\prime)}(l) = \dot{B}_i^{(\prime)}(l) \cdot \sqrt{\frac{B_i(l)^2 + B_i'(l)^2}{\dot{B}_i(l)^2 + \dot{B}_i'(l)^2 + \varepsilon}},\ \dot{B}_i^{(\prime)}(l) = \gamma \cdot A_i^{(\prime)}(l) + (1-\gamma) \cdot B_i^{(\prime)}(l), \quad (3.64)$$

where $0 \le \gamma \le 0.5$ is an equalization strength and ε, l are defined as ealier. With $\gamma = 0.5$, the IPDs between the sample pairs in b are fully removed, but the per-sample ILDs (i. e., channel-wise complex-domain magnitudes) are preserved for every γ. Note that this so-called *active downmix* approach is very similar to the process applied in intensity stereo coding. The difference is that, in the latter case, only the *band*-wise—but not *coefficient*-wise — ILDs (and, thereby, channel intensities) are maintained because the square-root factor is averaged over all $l \in p_i(b)$, and γ is fixed to 0.5. Given that, in the present work, a decorrelation signal is available as a substitute for the JS residual spectrum, the per-band intensity pre-downmixing gives little advantage over the perceptually more subtle sample-wise method of (3.63, 3.64) and may actually degrade the overall coding quality.

3.5.3 TDA-Domain Generation and Shaping of Decorrelation Signal

After executing (3.62) and, conditionally, (3.63, 3.64), conventional JS coding can be applied to $\ddot{A}_i$ and $\ddot{B}_i$, and a reliable measure of the band-wise ICC — namely, the residual RMS index vector er_i — can be easily acquired via (3.61). After entropy (de)coding of all $er_i(b)$ and reconstructive scaling, or "de-quantization", identically to the process carried out on $e_i(b)$ in (3.50), a residual RMS *target* value $ER_i^q(b)$ is obtained for each SFB. This parameter conveys the original energy of the residual spectrum in b to be synthesized.

Two quite sophisticated solutions for the MDCT-domain synthesis of a decorrelation signal, following the approach of, e. g., [Engd04], have recently been presented [Sure09, Sure12, Melk14]. Preliminary informal evaluation, however, indicated that the usage of the previous frame's reconstructed spectral downmix M_{i-1}^q as substitute for the current frame's zero-quantized residual S_i^q (or $\acute{S}_i^q$, in case of predictive JS coding) — a *copy-over* from downmix to residual — works equally well in the SBS context and is much simpler algorithmically. This observation confirms the conclusion in [Engd04] that, in the higher frequencies, such "delayed-downmix" decorrelators represent the best quality-complexity tradeoff (at low frequencies, residual coding as in UniSte MPS is used in SBS). In the USAC decoder, the delay operation, denoted by $D(M_i^q) = M_{i-1}^q$ hereafter, is already in use in the R-to-I process [Helm11] and, thus, comes at no additional cost. This makes the approach even more convenient and efficient in the given context. Moreover, for the case of plain M/S decoding (without prediction) and certain input signal configurations, the JS upmix of M_i^q and $D(M_i^q)$ yields a Lauridsen-type decorrelator [Laur54, Gribb14], as the M/S matrixing of a signal with a delayed version of itself is equivalent to a pair of

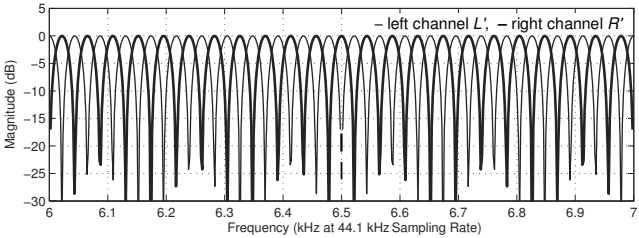


Figure 3.27. Per-channel frequency response of the Lauridsen decorrelator for an SBS-like configuration (frame length $N = 1024$ TD samples, sampling rate $f_s = 44.1$ kHz, i. e., a downmix delay of 23 ms).

complementary FIR comb filters. [Irwa02, Hel15b] examine this aspect in further detail. A visualization of the decorrelator's transfer function in SBS is provided in Figure 3.27.

Note that, as in IGF, the copy-over (instead of copy-up) spectrum could, optionally, be flattened in spectral direction (via multiplicative whitening) or temporal direction (via TNS-like predictive analysis filtering, see subsection 3.4.3). Such a pre-flattening design is likely to decorrelate M_i^q and $D(M_i^q)$ further as a perceptually useful side-effect, but it was not implemented in the course of this work. An evaluation of its necessity or benefit especially in very-low-rate coding is left to the reader as a topic of further investigation.

3.5.4 Transform-Domain Stereo Filling, Spatial Upmixing via JS Module

The decorrelation spectrum $D(M_i^q)$, generated according to the previous subsection, represents the *tile* (or source) signal in the copy-over procedure. Its energy is given by

$$TR_i(b) = \sum_{l \in z_i(b)} \left(D(M_i^q(l))\right)^2 = \sum_{l \in z_i(b)} \left(M_{i-1}^q(l)\right)^2, \quad b_{\mathrm{rc}} \le b < b_s, \tag{3.65}$$

separately for each parameter band (here, SFB) b, in the absence of the abovementioned spectrotemporal flattening, with $z_i(b)$, b_{rc}, and b_s defined as in IGF. After reconstructive scaling and core NF of M_i^q and S_i^q or $\acute{S}_i^q$, the *survived* residual energies can be computed:

$$SR_i(b) = \sum_{l \in p_i(b)} \left(S_i^q(l)\right)^2 \text{ or } \sum_{l \in p_i(b)} \left(\acute{S}_i^q(l)\right)^2, \quad b_{\mathrm{rc}} \le b < b_s. \tag{3.66}$$

Given $ER_i^q(b)$ as the target RMS and the fact that, for $l \in z_i(b)$, $D(M_i^q(l))$ and $S_i^q(l)$ resp. $\acute{S}_i^q(l)$ are uncorrelated (since the latter are either zero or pseudo-randomly noise filled),

$$MR_i(b) = c_i(b) \cdot \left(ER_i^q(b)\right)^2 - SR_i(b), \quad b_{\mathrm{rc}} \le b < b_s, \tag{3.67}$$

yields the *missing* residual power in b, and gain $tr_i(b) = \sqrt{\max(MR_i(b), 0)/(TR_i(b) + \varepsilon)}$ equivalently to IGF. With $tr_i(b)$, the copy-fill operation can, finally, be applied to S_i^q or $\acute{S}_i^q$:

$$\text{w/o prediction: } \tilde{S}_i^q(l) = \begin{cases} S_i^q(l) + \overline{D}(M_i^q(l)) \cdot \min(tr_i(b), 10), & q_i(l) = 0, \\ S_i^q(l), & \text{otherwise,} \end{cases} \tag{3.68}$$

$$\text{with prediction: } \tilde{S}_i^q(l) = \begin{cases} \hat{S}_i^q(l) + \overline{D}(M_i^q(l)) \cdot \min(tr_i(b), 10), & q_i(l) = 0, \\ \hat{S}_i^q(l), & \text{otherwise,} \end{cases} \tag{3.69}$$

where, as in section 3.4, the bar accent on the *D* function indicates possible spectrotemporal pre-flattening of the decorrelation signal — which then has to be accounted for in (3.65)—and $l \in p_i(b)$. Due to its similarity to the HF gap filling operation, the process of applying (3.65)–(3.69) is called Stereo Filling (SF) in this work. M_i^q and $\tilde{S}_i^q$, representing the outputs of the SBS tool chain (NF, SF, and IGF, executed in that order as shown in Fig. 3.26), can now be subjected to the other FD algorithms in the core decoder. In xHE-AAC, these are the M/S upmixing, with or without prediction, and the TNS synthesis filtering, the order of which, as indicated in Figs. 3.25 and 3.26, is determined and signaled by the encoder (see also subsection 2.2.4). The JS upmix, as noted earlier, is the equivalent of the spatial upmix matrix in PS/MPS. The FDP proposed in section 3.3, being a JS-domain LF tool, is assumed to operate only in the range below f_{rc}, where residual coding is used and NF is not applied. As such, the FDP synthesis filtering can be performed either after or before the semi-parametric SBS decoding, assuming that said LF range is not utilized as a copy-up source in the IGF decoding. Some further conceptual details in the context of an implementation of SBS into USAC, including the role of the *M/S Stereo + Prediction* module in Fig. 3.26 and typical choices of b_{rc} (f_{rc}) and b_s, are discussed in [Hel15b].

The SF technique, as indicated by the above algorithmic description, is closely linked with the NF tool and, thus, shares with the latter the application start frequency f_{rc} and, by way of the noise level nl_i, allows for relatively fine mixing of decorrelated-downmix and pseudo-random spectral content, similarly to the HFR related noise blending in SPX [Field04] and SBR [Wolt03]. Furthermore, it is worth noting that M_i^q and $\tilde{S}_i^q$ are located in the TNS residual domain. The subsequent IIR filtering by the TNS decoder, thus, acts as a temporal shaper, rendering further sharpening methods like the "sub-band domain temporal processor" or "guided envelope shaper" in MPS (cf. subsection 2.5.3) obsolete. In addition, MPS's "ducker" post-processor for transient input portions is also implicitly contained in the transform coding functionality: SF, like IGF, is executed separately per *short*-block group, enabling sufficiently high time resolution in case of transiens. After a block length switch (e. g., in a *short* after a *start* frame), M_{i-1}^q is not available, so only IGF and NF are used. This further minimizes audible reverberation due to the decorrelator.

Generally, the residual signal will be fully quantized to zero in the SF spectral region. The possible transmission of isolated *surviving* residual coefficients in that region, however, allows for a highly flexible semi-parametric stereo design, in which the encoder can

- vary the residual bandwidth, in units of b, between b_{rc} and B from frame to frame,
- set the frame-wise residual bandwidth based on the load on the FD entropy coder,
- determine the suriving lines depending on psychoacoustic criteria, i. e., based on a measure of how well the parametric SBS coding is able to regenerate each FD line.

The last aspect has been implemented in the SF encoder as a means to compensate for destructive interferences (i. e., cancelation and, thereby, energy loss) in M_i^q. When these occur, Lauridsen decorrelation does not work, and SBS resorts to NF or discrete coding.

In completion of this subsection, two remaining aspects shall be reviewed. First, for multichannel input with more than two channels, the same parallelized approach as in standalone MPS, illustrated in Fig. 2.16(b), could be employed with SBS. An alternative (and more flexible) tree-like architecture, in which arbitrary JS-coded channel pairs can be constructd in a time-variant fashion and cascaded decorrelation can be avoided, has recently been presented by Schuh, Dick, *et al.* [Schu16]. Unlike legacy JS coding tools in MPEG, this architecture also supports KLT-like coding. For this publication, the present author contributed a description of how SF may be integrated into the proposed multichannel coding tool (MCT) and, in particular, the necessary modifications to $D(M_i^q)$, i. e., the derivation of the last frame's spectral downmix, given the time-variant operation.

Second, the TDA-compensating energy-preserving calculation of the parametric envelope, described in subsection 3.4.4 in the context of IGF, can also be adopted for SF. To be specific, in SFBs identified as being tonal, the imaginary counterparts M_i' and S_i' or $\acute{S}_i'$ of the real-valued downmix and residual spectra are obtained via (3.63), (3.64), and the employed JS analysis matrix for the given frame i. Then, for each band b, the RMS value

$$ER_i'(b) = \sqrt{\frac{1}{c_i(b)} \cdot \sum_{l \in p_i(b)} \left| S_i(l) + jS_i'(l) \right|^2} \quad \text{or} \quad \sqrt{\frac{1}{c_i(b)} \cdot \sum_{l \in p_i(b)} \left| \acute{S}_i(l) + j\acute{S}_i'(l) \right|^2}, \tag{3.70}$$

with $b_{\mathrm{rc}} \leq b < b_s$ as usual, represents the complex-domain SF target value, and utilizing

$$rr_i(b) = \sqrt{\sum_{l \in z_i(b)} \left(D(M_i(l)) \right)^2 \Big/ \sum_{l \in z_i(b)} \left| D(M_i(l)) + jD(M_i'(l)) \right|^2}, \tag{3.71}$$

with decorrelation filter D (or, if applicable, $\overline{D}$) as a real-to-complex source ratio similar to the one of (3.57), the compensated transform-domain envelope index can be derived:

$$er_i'(b) = \left\lfloor 4 \cdot \log_2\left(s \cdot rr_i(b) \cdot ER_i'(b) \right) \right\rceil. \tag{3.72}$$

Transmitting er_i' as SF envelope notably reduces the modulation on some tonal material. At the same time, it allows to sustain the fully real-valued decoding of the SF proposal.

3.6 Entropy Coding of Spectral Coefficients and Scale Factors

The previous discussions indicate that, among all signals or parameters contained in a codec bit-stream (see page 33), the quantized spectral coefficient values q_i as well as the SFB-wise scale factor data r_i and SBS envelopes e_i (or e_i') and er_i (or er_i'), all for each frame/group i, constitute the major share of the total bit consumption. To conclude this chapter, a brief review of unpublished and recently presented studies for more efficient entropy coding of said bit-stream components is, therefore, provided in the following.

3.6.1 Coarse and Fine Spectral Envelope Models for Coefficient Coding

Concerning lossless coding of the per-frame q_i vector, subsection 2.3.4 demonstrated the improved compression achieved by USAC's context-adaptive arithmetic coding over that of AAC's section-wise Huffman coding. In fact, experiments conducted by the present author revealed that, for highly stationary tonal signals (e. g., harpsichord or pitch pipe recordings, see Chapter 4), the overall bit-rate is reduced by up to 30% compared to the case where sectioned Huffman coding — with or without a joint optimization like that of [Baue06] — is utilized. The primary element behind this performance gain is the temporal per-tuple adaptation of the probability context, which works particularly well for said input. In case of transform length changes (*long* to *short* blocks or vice versa) or error resilient coding (prohibiting inter-frame dependencies), though, temporal context adaptation may not be feasible, so the arithmetic coder loses most of its advantage. For such situations, two improvements to the context-based design were devised recently:

- for very-low-rate coding, a LPC-based *coarse* spectral envelope model reflecting, e.g., the formants in speech and primary resonances of musical signals [Bäck15],
- for arbitrary bit-rates, a *fine* spectral model increasing the accuracy (efficiency) of the frequency-direction context adaptation for harmonic waveforms [Mori15].

For both studies and publications, the author assisted in the development and testing of the algorithm as well as the writing of the manuscript. Note that the harmonic context model follows the same principle as the FDP proposal herein; it is an attempt to further reduce the intra-transform redundancy before (or during) the entropy coding stage by exploiting a fractional FD spacing parameter similar to s_0 in section 3.3. This parameter, which conveys the mean transform-domain harmonic interval, controls the probability context adaptation in both the entropy encoder and decoder and, as such, is quantized and transmitted using up to 8 bit per channel and frame, just like s_0^q. It, therefore, seems useful to combine the LF FDP with a HF harmonic-model enhanced entropy coder and, if beneficial, let both tools share one instance of s_0. This, however, has not been studied.

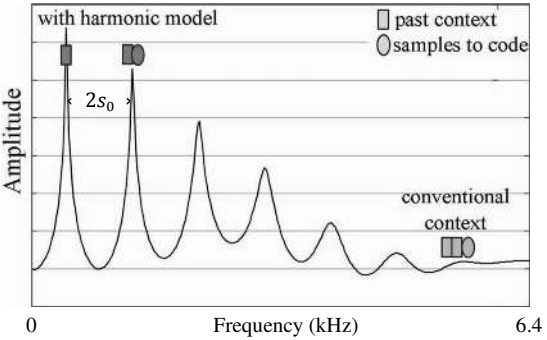


Figure 3.28. Context-adaptive arithmetic coder of Fig. 2.11(a), selectively enhanced by a harmonic context model. The latter is guided by an s_0-like interval parameter in a manner similar to the predictor in the FDP proposal. The harmonic context is only applied to samples on or near the harmonic grid [Mori15].

Figure 3.28 visualizes the harmonic model enhanced context adaptation, guided by a s_0^q-like interval parameter, in a manner consistent with Fig. 2.11(a) in subsection 2.3.4.

3.6.2 Context-Adaptive Arithmetic Coding of Scale Factors and SBS Envelopes

In MPEG audio coding, the scale factors r_i, representing the core "spectral envelope", are converted into a DPCM vector per group and frame, which is Huffman coded using a code alphabet, or book, with the per-symbol words listed in Table 3.2 [ISO97]. One can observe that the Huffman table is roughly symmetric about zero, and that the most frequently occurring small differential values are assigned accordingly short code words (value zero is associated with the shortest possible word length). On average, however, the words are still quite long, which explains why, as reported by Sreenivas and Dietz in [Sree98], the scale factor information consumes up to 20% of the total coding bit-rate in some low-rate configurations. As a countermeasure, said authors investigated three VQ methods for lossy coding of the scale factor vector (in unquantized form and, optionally, subjected to a DCT beforehand). Although rate savings of up to 50% were achieved, it is worth mentioning that this approach causes an interdependency between the spectral coefficient quantization and the scale factor coding, which greatly complicates both the encoding algorithm — especially the RD component — and the psychoacoustic model.

A more promising alternative to VQ coding is to apply the context adapted arithmetic coder also to the DPCM scale factor vector (appropriate retraining of all probability distribution tables assumed). This technique, in fact, has been adopted in the coding of the IGF envelope e_i (or e_i'), whose shape is very similar to that of the core envelope and even the SF envelope er_i/er_i', particularly in their differential forms. For better efficiency, the MPEG-H 3D Audio variant of the IGF tool [ISO15a] employs two-dimensional frequency-

time instead of one-dimensional frequency-only DPCM [Hel15a]. Utilizing envelope contexts along both frequency and time is analogous to image coding, where contexts along the horizontal and vertical direction of an image are derived. In [Wein99], a fixed linear predictor is used for plane fitting and basic edge detection, and the prediction residual is coded. In the IGF coding scheme, similar logarithmic-domain linear prediction allows to accurately model spectrotemporally constant and fading in/out energy areas. The actual prediction errors are encoded with a retrained version of USAC's arithmetic coder, assisted by an escape coding mechanism for the values beyond the distribution center.

Evaluations made during the development of the IGF tool show that, as with context adaptive arithmetic coding of the q_i, compression gains of about 30% over an optimized AAC-like Huffman coder design can be accomplished. This is consistent with the results obtained by the author in an experiment where the IGF arithmetic envelope coder was applied to the quantized scale factors r_i and SF envelope er_i/er_i' (after retraining, about 11.6 bit per frame and channel, or 0.5 kbit/s of the average per-channel bit-rate at 44.1 kHz sample rate, were saved in comparison to the Huffman delta coding in each case).

3.6.3 Biased Length-Limited Interleaved Golomb (BiLLIG) Coding of Scale Factors

In some devices requiring minimum software complexity (e. g., miniature low-power mobile hardware), Huffman-like variable-length coding (VLC) may be preferred over the more resource intensive arithmetic scheme. In the remainder of this chapter, an almost equally efficient and universal alternative to Huffman coding, which is even slightly less complex algorithmically, is devised. To the author's knowledge, the design, called biased length-limited interleaved Golomb (BiLLIG) coding, has not been published previously.

The basic motivation behind the BiLLIG code, a parametric prefix code based on the interleaved use of multiple differently configured Golomb-*m* codes [Salo07, sec. 2.23], is

- the construction of a code book which, in terms of the targeted probability distribution, better fits the given input data (here, DPCM scale factor or SBS vectors) than a single Golomb, Rice [Rice79], or subexponential [How94] book instance,
- the restriction of the maximum code word length to a predefined value in order to "rescue" the worst-case compression performance on primarily unlikely input,
- a respective decoding process that, on average, is faster than a Huffman decoder.

The Golomb-*m* codes are known to be appropriate for compressing data items that are distributed geometrically [Salo07]. Using, as a relevant example, AAC's or USAC's DPCM scale factor data, whose distribution is *not* entirely geometric, any Golomb book is, thus, considerably inferior to the dedicated Huffman table defined in [ISO97] and Tab. 3.2.

However, for certain partitions of the input value range, a respectively optimally chosen Golomb-m code set potentially performs equivalently to the Huffman table, i. e., exhibits identically long code words. Over the value range [–60, 60] of the DPCM scale factors r_i',

- $r_i' \in [0]$ can be treated as the first, most frequently occurring partition, for which $m = 0$ (yielding the unary code "0" of length one) is the most appropriate choice,
- $|r_i'| \in [1, 3]$ represents the second partition, with the positive and negative input values interleaved in order of their decreasing probability and $m = 3$ for best fit,
- $|r_i'| \in [4, 28]$ is the third partition, where alternated interleaving of the positive and negative input data and $m = 4$ (creating a Rice code) lead to the best results,
- $|r_i'| \in [29, 60]$ denotes the fourth, least frequent "outer" partition, with the same interleaving as in the third partition and a high $m = 64$ for limited word lengths.

Combining the above four Golomb instances into a single code book, with proper input biasing (to cover all data values) and avoidance of duplicate prefixes (to get unique code words), results, for instance, in the BiLLIG code table enumerated in the second column of Tab. 3.2. To reach a smooth progression of the word lengths towards decreasing symbol probability, the prefix for the fourth partition was manually selected from one of the words of the third partition. Put differently, part of one word of the third partition (the 111111111101.. originally intended for the input value ± 19) acts as an escape sequence into the range $|r_i'| \in [29, 60]$, and the remaining upper neighbors at $|r_i'| \geq 19$ are biased accordingly. Note that, for $|r_i'| \in [27, 28]$, reading and writing of the prefix-terminating zero may be skipped by limiting the reading of the leading ones to a maximum count of 15. For said four input values, this saves one additional bit in the BiLLIG word length. A source-code listing of the BiLLIG read and write algorithms is given in the Appendix.

In comparison to the AAC Huffman table in the first column of Tab. 3.2, the proposed BiLLIG alphabet exhibits some interesting properties, leading to some practical benefits:

- the code length for the quite likely input $r_i' = -7$ has been shortened by one bit,
- the largest code length (in the 4th partition) has been reduced from 19 to 18 bit,
- the per-symbol code lengths rarely increase, and if so, never by more than 3 bit,
- except for value $r_i' = -1$, the BiLLIG book is length-symmetric about zero input, and the last bit of each word holds the sign of the respective value (1: negative),
- unlike in a Huffman decoder, conditional progression down a probability tree is not required for each code bit, so the decoding process is a bit faster on average,
- the code book can easily accomodate a wider input range by extending the third partition beyond $|r_i'| = 28$ and/or adapting m of the fourth partition accordingly.

Since these advantages over the Huffman book are, however, negligible or irrelevant in a modern perceptual transform (de)coder, and an average compression gain of only one

bit per frame is achieved due to the partly shorter code lengths, it was decided to apply Huffman (or in case of the IGF, arithmetic) coding for the evaluation discussed hereafter.

input data	AAC Huffman word	code length	BiLLIG code word	code length
60	1111111111111110011	19	111111111101111110	**18**
38	1111111111111010010	19	111111111101010010	**18**
29	111111111110101	15	111111111101000000	18
28	1111111111110110	16	11111111111111$_{(0)}$10	17 $_{(18)}$
27	111111111110100	15	11111111111111$_{(0)}$00	17 $_{(18)}$
26	1111111111110000	16	11111111111111010	17
14	1111111000	10	11111111000	11
8	11110111	8	11111000	8
7	1111010	7	1111010	7
6	1111000	7	1111000	7
5	111011	6	111010	6
4	111001	6	111000	6
3	11011	5	11010	5
2	1100	4	1010	4
1	1010	4	1100	4
0	0	1	0	1
– 1	100	3	100	3
– 2	1011	4	1011	4
– 3	11010	5	11011	5
– 4	111000	6	111001	6
– 5	111010	6	111011	6
– 6	1111001	7	1111001	7
– 7	11110110	8	1111011	**7**
– 8	11111000	8	11111001	8
–14	1111111001	10	11111111001	11
–26	11111111110111	14	11111111111111011	17
–27	11111111110101	14	11111111111111$_{(0)}$01	17 $_{(18)}$
–28	11111111111001	14	11111111111111$_{(0)}$11	17 $_{(18)}$
–29	111111111110111	15	111111111101000001	18
–38	11111111111101111	17	111111111101010011	18
–60	111111111111101000	18	111111111101111111	18

Table 3.2. VLC alphabets for scale factor coding. Left: AAC (Huffman), right: BiLLIG proposal.

4 Objective and Subjective Performance Evaluation

The previous chapter presented various contributions to the design of a fully flexible perceptual transform codec, i. e., a coding system supporting both unconstrained and LD configurations as well as high and low bit-rates and channel counts with the same set of algorithmic tools. The two principal goals of this study, to repeat, are the simultaneous decrease of the computational codec complexity, especially at the decoding (receiving) end, and increase of the subjective reconstruction quality, particularly for tonal input.

To determine whether, and to what extent, the abovementioned goals were achieved, the codec extensions described in sections 3.1–3.6 were integrated into a floating-point C software implementation of the phase-1 MPEG-H 3D Audio [Herr14, Herr15, ISO15a] encoder and decoder, maintained by the Fraunhofer Institut für integrierte Schaltungen (IIS) in Erlangen, Germany. With the channel-based mono, 2.0 stereo, and 5.1 surround configurations investigated in this thesis, this first revision of the MPEG-H specification is almost identical to the MPEG-D USAC (xHE-AAC) standard [ISO12]. The only relevant differences are a newly introduced transform splitting (TS) tool for frame- and channel-wise intermediate spectrotemporal transform resolution between that of either *long* or *eight-short* blocks [Hel15d], and a quad-channel enhancement of USAC's MPS (UniSte) module for QMF-domain semi-parametric joint coding of four channels [Herr15]. Due to their IIR-like behavior, the FDP proposal and the context-adaptive arithmetic coders were, as a special case, implemented with only fixed-point integer arithmetic in order to ensure identical results on different computing platforms and/or operating systems.

Both objective and subjective evaluation was carried out, as described in respective sections hereafter. The objective assessment includes estimations and actual measurements of the computational complexity of the unchanged "reference" 3D Audio decoder implementation and the modified "proposed" variant thereof, as well as calculations of the total algorithmic delay of both versions of the codec software. For subjective testing, formal double-blind listening experiments, intended to assess both the "absolute" audio quality of the consolidated flexible codec proposal and the "relative" performance compared to the state of the art, were conducted for different signals and channel counts.

For reasons of brevity, not all listening tests performed in the course of this work are documented in detail herein. The interested reader may find the omitted test reports in

- [Helm14], where an early EVS proposal with the contributed LD block switching of section 3.1 and a tonality-adaptive NF design are evaluated at 48 kbit/s mono,
- [Fuch15], testing the final EVS codec with the same NF and IGF at 9.6–24 kbit/s,
- [Bäck15], comparing the envelope-based to USAC's arithmetic coding at 8 kbit/s,
- [Rämö15], presenting a comprehensive suite of independent EVS test results, and
- [Hel16c] with regard to an isolated assessment of the intra-channel FDP scheme.

Note, further, that, unless stated otherwise, both the objective and subjective evaluation are performed with all codecs under test operating in stereo at an input/output sample rate of 48 kHz and a mean overall bit-rate (including all side-information) of 48 kbit/s.

4.1 Objective Assessment of Delay and Decoder Complexity

According to [Alla99] and [Lutz04], the total algorithmic latency of the MPEG-2 AAC Low Complexity (LC) profile [ISO97] amounts to a minimum of $2N + 9N/16 = 2624$ TD samples, or 54.7 ms, when neglecting the buffering delay contributed by a bit reservoir. Based on this figure, it is easy to derive the respective delay values for FD-only versions (utilizing just the MDCT core and disabling the LPC-based speech core and QMF-domain parametric tools) of MPEG-D USAC [ISO12] and MPEG-H 3D Audio [ISO15a]. Since all of the added FD-core tools are delayless, or reuse existing delay sources such as the block switching component, both recent coding standards exhibit the same latency of 55 ms.

Among the author's algorithmic contributions of Chapter 3, comprising the LD block switching modification, KS, RS, and FDP functionality, as well as the semi-parametric IGF and SF extensions, only the LD block switching and RS techniques introduce additional codec latency in the form of extended encoder-side waveform lookahead. Specifically,

- the required LD block switching lookahead, as given by (3.2), equals 64 samples, i. e., 1.33 ms, which can be included in the 12 ms lookahead for the conventional block switching method [Bosi97, Alla99] when both are used in combination, or which replace said 12 ms of lookahead when only the LD proposal is employed,
- the RS scheme must extend the maximum windowing lookahead from $N = 1024$ to $L - N = 3072$ samples, i. e., from 21.3 to 64 ms, to accommodate the allowed steady-state MELT. Possible block-switch lookahead must be added to this value.

Thus, the total encoding-decoding delay of the proposed flexible transform codec architecture, avoiding any QMF-domain processing, ranges from 44 ms (LD) to 97.3 (RS) ms.

It is worth noting that, outside of any profile constraints, the basic syntax definitions of both the USAC and the 3D Audio standard include a *coreCoderFrameLength* element, which can be employed to signal, for the given audio stream, the use of a reduced frame size of $N' = 768$ samples [ISO12, ISO15a]. In doing so, all of the noted frame, transform, and lookahead lengths are multiplied by $768/1024 = 0.75$, which, e. g., allows the 44-ms configuration introduced on the previous page to be "compacted" into a 33-ms version [Hel15d]. This "LD 3D Audio" proposal, whose subjective performance is reported on in section 4.2), represents a very promising candidate for live broadcasting and communication applications since its delay of $2N' + N'/16 = 1584$ samples perfectly matches the target latency of approximately 33 ms established for such LD scenarios in Chapter 1.

To summarize, it can be stated that the delay-flexible MPEG-H based transform codec

- can be configured, during initialization of the audio stream, to exhibit a minimal encoding-decoding delay of 33 ms — including block switching functionality — at a sample rate of 48 kHz. This nearly equals the 32 ms used by 3GPP's EVS codec [Fuch15, ETSI16] and is lower than the 37.7 ms of AAC-ELD v2 [ISO09, LuVa10], both of which, like IETF's Opus [IETF12], represent dedicated state-of-the-art LD solutions for bidirectional Voice-over-IP (VoIP) telephony and teleconferencing.
- requires, under typical (unrestricted) operating conditions, a latency of 54.7 ms when not utilizing RS and 97.3 ms with full support for block switching, window overlap switching, KS, and RS. These values, of which the latter equals the maximum delay allocated by the proposed coding approach, compare favorably with the 129.3 ms used by HE-AAC [ISO09, Lutz04] or the at least 200 ms (owing to the MPS 2-1-2 pre-/post-processing) measured for USAC with UniSte coding [ISO12].

4.1.1 Estimation and Practical Measurement of Algorithmic Decoding Complexity

The specific computational complexity exhibited by a perceptual transform decoder depends not only on numerous design and realization details of the particular software implementation but also on the several aspects of the underlying hardware architecture on which the decoding algorithms are executed. A digital signal processor (DSP), which can generally be found in consumer and professional electronic equipment such as TVs or sound systems, for instance, performs some algorithmic instructions with a different clock count (i. e., relative speed) than a central processing unit (CPU) inside a personal computer. To enable a reasonably accurate assessment of the complexity, regardless of any hardware platform or software realization details, two measures, termed processor complexity unit (PCU) and RAM complexity unit (RCU), are utilized in MPEG standardization work [Bosi97, Neue13, ISO15b]. The former indicates the execution workload in million operations per second (MOPS), while the latter quantifies the memory usage.

Focusing, again for the sake of brevity, only on the PCU numbers, the overall and per-component complexities of the conventional decoders and the proposal are as follows:

- The MPEG-2 AAC decoder, for the LC profile, requires at most 47000 instructions per frame and channel, including 854/2 M/S stereo operations [Bosi97], which accumulate to an average of 2.2 MOPS per channel at a sampling rate of 48 kHz.
- The USAC decoder, in its Baseline profile, uses an average of 6 MOPS per channel for 2.0 stereo and 5.1 surround output [Neue13], i. e., 2.7 times the PCU count of MPEG-2 AAC-LC. The complex stereo prediction contributes 0.6 MOPS [Helm11].
- The 3D Audio core decoder, with a mean per-channel workload of 5.8 MOPS for 8- or 16-channel output signals [ISO15b], consumes nearly the same complexity as a USAC decoder. Further rendering tools more than double this value, though.

The increased workload of the USAC and 3D Audio decoders is primarily caused by the QMF-domain SBR and MPS 2-1-2 post-processors, which are responsible for worst-case complexity increments of 3.0 and 4.5 MOPS, respectively, for stereo signals [ISO15b]. If, as in the current proposal, these coding tools are deactivated and the core codec is run at the output sample rate of 48 kHz, the peak per-channel complexity of the two MPEG decoders can be reduced to about 3 MOPS each. Note that this number is quite close to that for the AAC-LC decoder. The slight increase is attributable to the use of arithmetic (de)coding, which is about 1.5 times more expensive than its Huffman counterpart, the complex prediction capability, with a mean cost of 0.3 MOPS per channel, the core filter bank extensions, and the support for a maximum TNS filter order of 15 instead of 12.

Taking the above MDCT-only 3D Audio core as the baseline codec and extending it by

- the contributed filter bank enhancements, including LD block switching, KS, and RS functionality with a combined cost of at most 0.6 MOPS per channel [Hel16b],
- support for JS-domain FDP decoding, with at most 0.27 MOPS/channel [Hel16c],
- TDA- instead of QMF-domain BWE using IGF, at 0.9 MOPS per channel [ISO15b],
- to complete the SBS approach, SF either per CPE or within the MCT, with a peak channel-averaged decoder-side complexity of 0.2/2 MOPS in each case [Schu16],

yields a total *worst-case* decoder-side workload of **4.87** MOPS/channel for the complete flexible transform codec proposal. When disabling support for RS, this value is reduced to roughly **4.3** MOPS, i. e., 195% of the MPEG-2 AAC-LC and 72% of the USAC/xHE-AAC decoder workload, representing the complexity and quality benchmarks, respectively.

To assess the *average* complexity, which is, usually, more relevant in practice, an ARM executable of the 3D Audio decoder was run on a Nexus 7 tablet. Even higher workload savings of 33–40% over the QMF-extended variant were observed thereon [ISO15b].

4.2 Subjective Evaluation of Overall Audio Coding Quality

The previous section revealed that both the average and peak algorithmic complexity of the developed flexible perceptual transform decoder reach only around two thirds of the respective measurements for the USAC and 3D Audio "reference" decoders on which the present proposal is based. In order to assess the subjective impact of this objective achievement in comparison with the uncoded PCM reference (i. e., input) signals as well as, e. g., the equivalent USAC (de)coded stimuli generated at the same average bit-rate, a number of formal double-blind listening tests were conducted. To prepare and conduct such blind evaluation in a consistent, controlled, and (largely) reproducible way, several test methodologies have been published. The most commonly known and utilized test procedures and/or protocols are the ITU-T recommendation P.800 [ITU96] for low-rate transmission quality particularly in mobile telecommunication applications, the ITU-R recommendation BS.1534 [ITU15b] for the comparative assessment of moderate quality degradations, and ITU-R recommendation BS.1116 [ITU15a] for the evaluation of small impairments, i. e., of high-quality audio systems. A comparison of the three methods is provided in [Raak12], noting, for instance, that P.800 testing is performed by naïve subjects and that, for the ITU-R-type tests, experienced participants are recommended. An almost identical variant of BS.1116, called ABC/HR (where the HR denotes the presence of a hidden reference, whose identity on the multi-slider user interface is concealed to the test subject), is widely employed for audio codec evaluation by the Web community [Hydr16, Soun12], even for comparisons of perceptual codecs at quite low bit-rates.

Targeting, in this work, the *good*-quality (i. e., slightly perceptible degradations)—but not *excellent*-quality (i. e., imperceptible degradations) — range of subjective judgments, the pursuance of the BS.1534 methodology, also known as MUlti-Stimulus with Hidden Reference and Anchor(s) (MUSHRA), was chosen for all listening experiments. In such a test, the subject is seated in an acoustically controlled room with low reverberation and environmental noise, and is presented with a multi-slider user interface, on which each slider is randomly associated with an uncoded (hidden reference or low-pass anchor) or coded (system under test) stimulus for every new session (i. e., input signal). A button for playback of the known uncoded reference is provided as well. Having played, looped through, and switched between all stimuli assigned to a slider, the subject then casts his or her vote on the "basic audio quality" [ITU15b] of each stimulus, on an opinion scale from 0 (lowest *bad* score) to 100 (highest *excellent* score), by way of the corresponding slider. The evaluation then continues with the next session, where the session order is randomized for every participant. Further details, including illustrations of the grading scale and the graphical user interface employed in this study, are provided in [ITU15b], and an exemplary user handout is located at www.ecodis.de/audio/guideline_high.html.

4.2.1 Performance Evaluation of General-Purpose and LD Stereo Configurations

Both the latency-unconstrained and LD configuration of the 3D Audio-based flexible codec proposal were evaluated in the course of this work. For a better overview, the 2.0 stereo (L/R) channel setting, assessed using STAX Lambda Pro electrostatic headphones with accompanying amplifier, will be discussed first. The 5.1 multichannel loudspeaker setup is addressed in subsection 4.2.2. A mean stereo target bit-rate of 48 kbit/s, i. e., 0.5 bit per time sample on average at a sampling rate of 48 kHz, was selected for testing.

Depending on the particular experiment, the proposed codec, which was configured for SF and IGF ranges of 3.75–9.0 and 9.0–16.5 kHz, respectively, was compared against

- Opus [IETF12] with encoder version 1.1 from www.opus-codec.org, which uses only its CELT core at this bit-rate and which was run at the default frame length of 20 ms and bandwidth of 20 kHz in constrained variable bit-rate (CVBR) mode,
- USAC [ISO12] or "phase 1" 3D Audio [ISO15a], being identical for stereo input at this bit-rate except for the added TS capability in the latter specification (which, however, has a negligible effect on the overall quality), operating in CVBR mode,
- QMF-less 3D Audio, as a LC benchmark similar to MPEG-2 AAC-LC, with only NF, predictive M/S, and TS, but no Enhanced SBR or UniSte MPS coding, also run in CVBR mode and, like the QMF-extended codecs, an audio bandwidth of 16.5 kHz.

For the LC codec proposal, the contributed LD block switching was disabled, and SBS's temporal tile flattening (page 99) and noise blending (page 104) were deactivated (i. e., core NF was applied only between 3.75 and 9 kHz) since their perceptual benefits were found to be too small at the given HFR start frequency to justify the complexity increase. Moreover, in the SF and IGF ranges, the predictive JS tool was limited to the real-valued design, bypassing the R-to-I algorithm. As described in subsection 3.2.1, a complex predictor is more counterproductive than advantageous at the given bit-rate, both in terms of quality and with regard to the additional JS side-information (see also section 5.1).

Out of almost 100 short test signals gathered by the author from the EBU SQAM and 5.1 surround sets [EBU08, EBU07], multiple archives accessible via the HydrogenAudio and SoundExpert web sites [Hydr16, Soun12], and collections of MPEG standardization material [Herr96, Bosi97, Ojan99, Bree05, Bree07, LuVa10, Song11, Helm11], about 23 highly codec-critical sequences per channel configuration (2.0 and 5.1), as described in Appendix A.4, were selected in a pre-audition procedure. These signals were resampled to 48 kHz/24 bit, concatenated, coded, decoded, and extracted (split) again for the tests. Whenever necessary, the decoded waveforms were synchronized (delay aligned) to the uncoded concatenated 48-kHz references. All stimuli exhibited a word length of 24 bit per sample, whose direct playback was supported by the computer's sound interface.

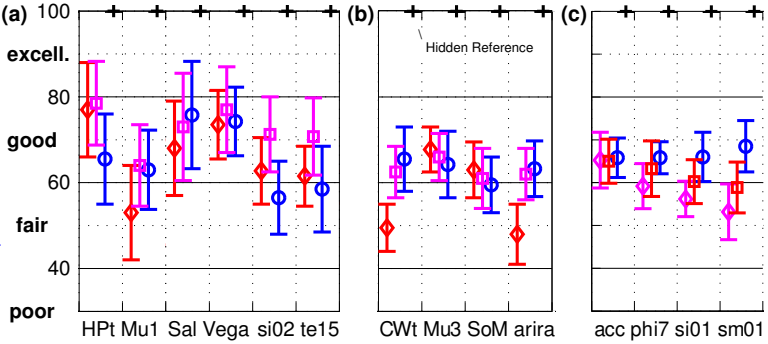


Figure 3.29. Outcome of the 3D Audio comparative listening test set with (red, magenta) LC FD-core-only and (blue, circles) UniSte+ SBR setups, normal delay.

Figure 3.29 depicts, for the general-purpose stereo configuration, the mean opinion score (MOS) for each sequence and stimulus, divided into the categories "conventional stereo" (a) [Hel15b], "IPD-critical stereo" (b) [Hel15c], and "stationary tonal stereo" (c) [Hel16a], along with the associated per-stimulus 95% confidence intervals (based on a Student-t distribution since a normal distribution cannot be assumed). A total number of 12 listeners (maximum age 39, incl. one or two females depending on the test) took part in each of the three subtests, with post-screening rules as in [Neue13] and statistical analysis as described in [ITU15b] applied. Owing to the extensive experience of the participants, no subject had to be post-screened, i. e., no scores were excluded. For best visibility, the quality axis is zoomed to the range 30–100, and the scores for the 3.5-kHz anchor stimuli (which are below 30 for all sequences) have been omitted in Fig. 3.29.

From Figs. 3.29(a) and (b) it can be observed that the flexible transform codec, with the IGF, SF, and KS proposals activated (although KS is not triggered on the first 6 items), achieves a MOS of at least 60, i. e., a *good* quality verdict, on every sequence tested. Also,

- compared to the QMF-less 3D Audio reference using only NF and IGF but none of the other contributions devised in Chapter 3, significant quality gains ($p < 0.05$) of 10 or more MOS points are reached for the noisy and transient Mu1, si02, and te15 as well as the "phasy" CWt and arira (with IPDs near ±90° on the latter two),
- the quality of QMF-enhanced 3D Audio (i. e., USAC with TS) is matched — or even exceeded — for all signals. Given that this *quality reference* needs 40–50% more decoding complexity and four times more delay, this result is highly satisfactory.

Naturally, on the tonal material of Fig. 3.29(c), the core-downsampled UniSte reference outperforms the QMF-less LC versions, but the RS and FDP enhancements (red squares) allow the codec proposal of the first two subtests (pink diamonds) to catch up halfway. Similar gains of a few score points due to the FDP were observed on item CWt [Hel16c].

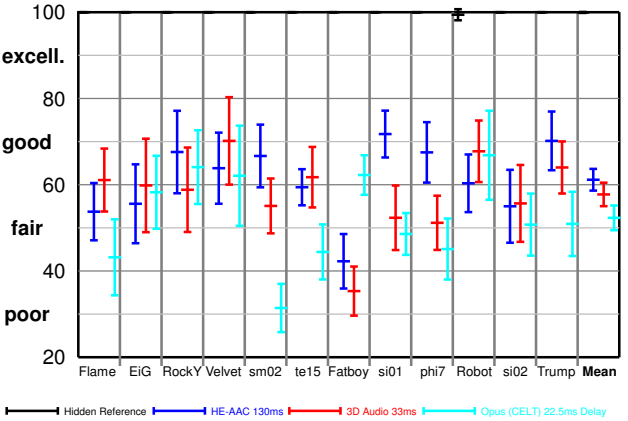


Figure 3.30. Results of the lower-delay stereo listening test comparing (red) the LD 3D Audio proposal against (blue) HE-AAC v1 and (cyan) Opus, using only CELT, again at 48 kbit/s. All codecs operated in a CVBR mode [Hel15d]. A 3.5-kHz lowpass anchor was included in the test but, for clarity, its scores are not displayed here.

For the subjective evaluation of the proposed LD configuration, another listening test with slightly different signals (because the tests participants were mostly the same) but including the most critical sequences from the general-purpose test of Fig. 3.29 (except for sm01, which was forgotten), was carried out. The 33-ms 3D Audio tuning, with the LD block switching proposal, KS, FDP, SBS (i. e., IGF and SF) but, for obvious reasons, no RS, was compared against Opus, an earlier version of which had been found to perform quite well in a previous MUSHRA test [Helm14]. The LD 3D Audio proposal saves 40% in algorithmic latency over the default 55-ms configuration, so to keep the assessment somewhat fair, HE-AAC was used as the quality reference instead of the QMF-enhanced 3D Audio (the former, in the v1 setting, requires roughly 40% less delay than the latter with its UniSte MPS 2-1-2). The number of listeners, test environment, equipment, and statistical analysis methods were the same as in the previously described experiments.

The overall and per-stimulus results of this MUHSRA test, visualized in Figure 3.30, indicate that on average, the LD 3D Audio variant shows a level of audio quality which is comparable to that of HE-AAC at this bit-rate since the 95% confidence intervals of the two conditions overlap. However, unlike in the unrestricted-delay tests discussed previously, the subjective difference between the QMF-less and the QMF-extended MPEG codec versions is statistically significant here ($p < 0.05$). The main reason for the quality advantage of the HE-AAC condition is, as expected given its four times greater delay, the much better performance on the highly tonal harpsichord (si01), pitch pipe (phi7), and Glockenspiel (sm02) recordings. Nonetheless, the LD 3D Audio condition outperforms the Opus codec with high significance ($p < 0.001$). This can also be expected: CELT's 10 ms lower latency (23 ms) is attained at the cost of reduced spectral selectivity [Helm14].

4.2.2 Subjective Evaluation of General-Purpose and LD Surround Configurations

The stereo MUSHRA tests, whose results are illustrated in Figs. 3.29 and 3.30, were repeated with identical codec configurations but on 5.1 multichannel sequences. To this end, the execution of the test sessions was moved to one of Fraunhofer IIS's loudspeaker equipped sound labs [Silzle09], and playback was carried out over five Dynaudio BM6A active transducers and a corresponding subwoofer. To facilitate the comparative quality assessment and grading procedure (which tends to be more difficult for surround than for stereo sound due to the increased channel number) to the participants, the general-purpose "regular delay" experiment was divided into 3 successive tests. In the first, the QMF-less 3D Audio proposal (55 ms delay), augmented by FDP and joint-channel coding via the MCT of [Schu16] and subsection 3.5.4, was evaluated against the QMF-extended 3D Audio variant utilizing the SBR and quad-channel MPS 2-1-2 enhancements (200 ms delay). The second test compared the QMF-less version of the first test with an identical configuration having, in addition, the SF tool integrated into the MCT module. The third assessment, finally, intended to quantify the combined perceptual benefit of the KS (i. e., MDST coding) and RS (MELT coding) tools when added on top of the MCT+SF design.

According to Table 1.1, the 5.1-surround equivalent bit-rate of the 48 kbit/s used in the 2.0-stereo evaluations can be estimated as 96 kbit/s. Since, however, a preliminary informal evaluation indicated that the multichannel test sequences tend to be a bit less codec-critical than their stereo counterparts, a slightly lower mean bit-rate of 80 kbit/s was chosen for all three tests. The best sounding encoder tuning at this operating point ended up being one where the core codec runs at a downsampled 32 kHz [Hel16b], with SF and IGF employed in the frequency ranges 2.5–8 and 8–16 kHz, respectively. The 3D Audio encoder itself was responsible for resampling the 48 kHz concatenated waveform input to 32 kHz, while the decoder output was manually upsampled (with high quality) to 48 kHz/24 bit, using McGill's ResampAudio, for playback in the blind test sessions.

The outcome of the first test, performed by 19 listeners of which one had to be post-screened due to an inability to consistently identify and/or rank the anchor conditions, is shown in Figure 3.31 (again with 95% confidence intervals based on the appropriate Student-t distribution). The results demonstrate that, overall and for each sequence, the subjective performance of the QMF-based 3D Audio reference is matched—or, in case of the applaus item, even exceeded by a highly significant margin — at a much lower delay and decoder complexity. Extending the "discrete" MCT to a semi-parametric solution by means of the SF method, as indicated by the scores of the second test depicted in Figure 3.32, leads to a substantial further improvement of the coding quality achievable via the QMF-less scheme, at least at the given bit-rate. For four of the 13 signals and the signal average, the audio quality improves significantly due to the application of SF [Schu16].

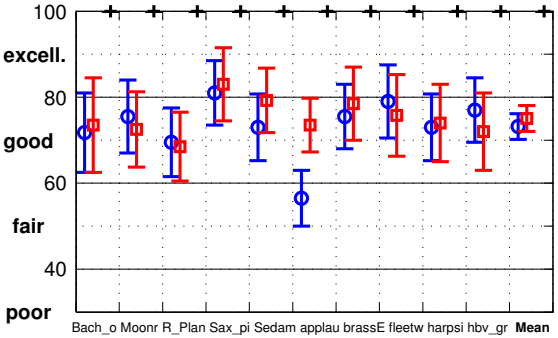


Figure 3.31. Pooled results of Fraunhofer IIS and Aalto University for the first of the three 5.1 multichannel blind tests at 80 kbit/s, regular delay. (blue, circles) 3D Audio with SBR and MPS, (red, squares) LC 3D Audio design with FDP, IGF, and MCT [ISO15b].

The results of the third listening experiment, illustrated in Figure 3.33, provide similar evidence as the stereo tests of Fig. 3.29 with regard to the perceptual benefit of KS and, in particular, RS (the former proposal is rarely triggered on the employed multichannel material). For six of the eight highly tonal and/or transient signals tested, and averaged over all subjects (10 instead of 19 in this evaluation), a quality gain is attributable to the integration of switchable MDST and MELT coding. Further analysis reveals that for two items (Harpsi and brassEx) as well as the overall mean score, the quality improvement is statistically significant. Nevertheless, the MOS increases are small, as for stereo, and QMF-enhanced 3D Audio is expected to still outperform the proposal on these signals.

To assess, as in the two-channel case, the subjective performance of the 33-ms delay-reduced 3D Audio configuration, with active LD block switching, FDP, and SBS in CPEs, but no KS and MCT coding (since stable implementations of the latter two could not be completed in time), a respective blind test was carried out. Again, the Opus codec (in an unconstrained VBR instead of CVBR mode this time) serves as a LD reference, whereas instead of HE-AAC another undisclosable but structurally and qualitatively comparable medium-latency codec is included. Given the relatively low overall quality observed in the stereo LD experiment, a somewhat higher bit-rate than in the previous 5.1 sessions, namely, 128 kbit/s, was chosen. The per-stimulus and overall MOS values and 95% confidence intervals acquired from this evaluation (based on the gradings of 7 experts), as depicted in Figure 3.34, confirm the audio quality advantage of the devised LD 3D Audio codec over the undisclosed and the IETF condition. Averaged over all 10 sequences, the proposal is found to outperform the other codecs with statistical significance, although the distance to Opus (here, CELT) is smaller than in the stereo case. The two reasons, to repeat, are the unrestricted VBR configuration of CELT, allowing more than 128 kbit/s on "difficult" input (even over a longer period than a few frames) and the lack of flexible MCT+SF coding in the 3D Audio variant (which is likely beneficial on, e. g., item applau).

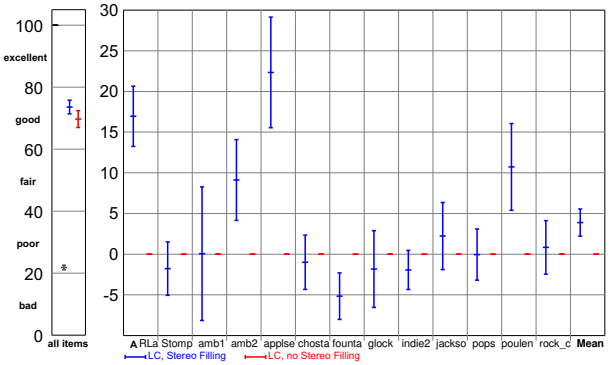
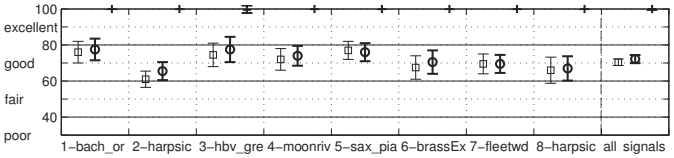


Figure 3.32. Results of the second 5.1 multichannel test at 80 kbit/s. Left: mean score across all 13 signals, right: mean differential scores for each sequence, relative to (red) the LC 3D Audio codec condition without activated Stereo Filling. Both configurations employ the MCT for joint-channel coding [Schu16].



(■) 3D Audio without proposal, (●) 3D Audio with proposal, (+) hidden reference

Figure 3.33. Results of the third 5.1 multichannel test, regular delay. LC 3D Audio (–,□) without KS,RS, (–,○) with combined KS and RS [Hel16b].

Among all MUSHRA test results presented so far, the Fatboy signal of the stereo test of Fig. 3.30 ends up being the most demanding waveform sequence for the MPEG audio codecs studied, including the QMF-less 3D Audio proposal. Surprisingly, it also appears to be one of the least critical items for the Opus (CELT) coder, using on such pulse-heavy vocoder waveforms, its sub-band merging and traditional intensity stereo downmixing quite extensively (in virtually every frame and starting at relatively low frequencies). To assess whether *good* quality can also be reached with the general-purpose QMF-less 3D Audio codec at 48 kbit/s on the Fatboy sequence, the encoder was modified as follows:

- In a transient frame with pitch pulses, as indicated by the usage of an *eight-short* block sequence, a SFB-wise active intensity downmix to panned mono (i. e., fully ILD-preserving and IPD-removing), following the sample-wise method of (3.63), (3.64) in subsection 3.5.2, is applied in the SBS range to minimize the JS residual.
- In case of multiple *eight-short* blocks in a row, the frames after the first transient one are restricted to at most two transform groups to limit the side information. Note that this tuning is rarely triggered by the other signals, even transient ones.

A formal listening test comparing the modified encoder against the LC 3D Audio system without the above two tunings indicates that, for the 12 listeners participating, a score very close to 60 (i. e., a MOS like that for item sm01 in Fig. 3.29) is, indeed, achievable.

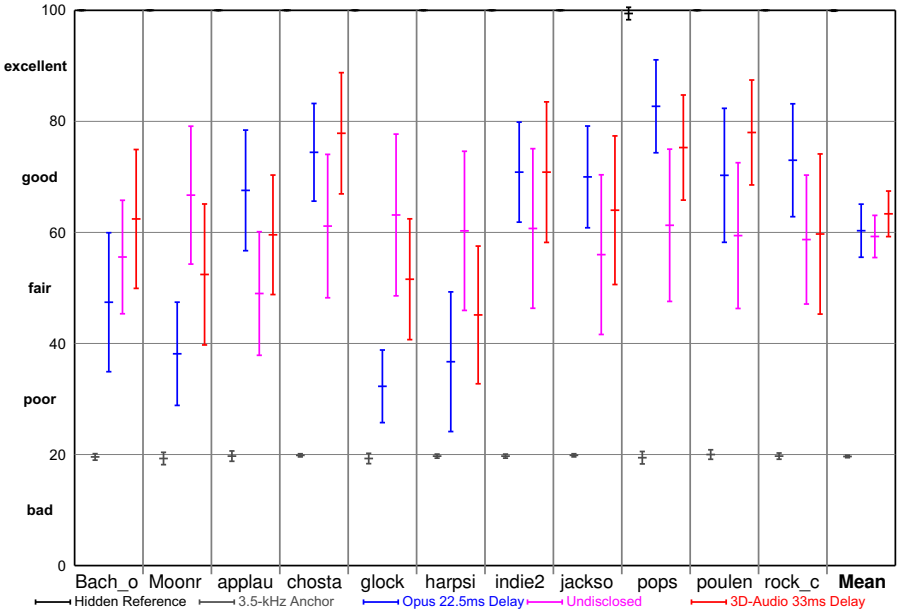


Figure 3.34. Results of the low-delay 5.1 multichannel MUSHRA test with 7 expert participants.

To complete this chapter, it is noted that, in the 5.1 multichannel experiments, none of the contributed tools were employed in the LFE channel. The FDP could, in theory, be advantageous on the often quite stationary LF waveform in this channel but, in practice, the overall subjective benefit is too small to justify the needed signaling and processing.

5 Summary and Conclusion

In this thesis, the persistent demand for efficient perceptual audio transform coding solutions in various applications was addressed. In particular, Chapter 1 noted the need for low-rate but high-quality codecs in order to maintain a high QoS even when several consumers must share a single network path, such as a radio cell, and/or when current-generation network infrastructure is not available and a fallback to slower transmission rates becomes necessary. This was mentioned to be especially relevant in case of a high number of input/output audio channels, as in 5.1 surround or 7.1+4 immersive applications which, compared to the traditional stereo configuration, are growing in popularity. For use cases like on-site acquisition of live-broadcast material and bidirectional voice communication or teleconferencing across the Internet, i. e., applying the VoIP principle, the additional requirement of low overall algorithmic coding and decoding latency was identified. Given the prevalent usage of battery equipped mobile devices such as smart watches, phones, and tablets for the consumption of digital media, as well as the trend towards an Internet of things (IoT) potentially allowing for the same (at least regarding audio), a necessity for low computational (de)coding complexity was also introduced.

Unifying the requirements of low coding bit-rate, high audio reconstruction quality and, in some applications, channel counts, as well as a low algorithmic codec delay and complexity, the design and development of a very flexible and efficient audio transform codec, satisfying all of the previously described use-case specific needs, was established as the goal of this work. To this end, the USAC/xHE-AAC specification, standardized by the ISO/IEC MPEG in [ISO12] and, as illustrated in section 2.6, representing the state of the art in perceptual stereo audio coding, was chosen as the starting point or *baseline*.

Having introduced the reader to the purpose and design of the essential components of a modern audio transform codec like USAC, AC-4 [Kjör16], or Opus [IETF12], namely,

- overlapped T/F mapping by way of a block switching TDAC filter bank (sec. 2.1),
- transform-domain spectrotemporal pre/post-processing and JS coding (sec. 2.2),
- frequency band partitioning, scaling, quantization, and entropy coding (sec. 2.3),
- spectral coefficient substitution (sec. 2.3), complex-QMF-domain HFR (sec. 2.4),
- hybrid pseudo-QMF-domain parametric stereo or multichannel coding (sec. 2.5),

in Chapter 2, some drawbacks of said components, with regard to the set of application specific requirements, were identified in section 2.6. Most importantly, a fundamental conflict was described: at low bit-rates, highest subjective audio quality is only achieved using the auxiliary parametric tools operating around the core transform codec, but for best possible perceptual performance (using, e. g., phase coding techniques), these tools require additional instances of a complex-QMF bank, which increase the overall codec complexity and latency considerably. The conventional block length switching approach was also noted to necessitate further waveform lookahead (i. e., delay) in the encoder.

To address these disadvantages of the state of the art, Chapter 3 proposed a modified codec design employing only algorithmic tools operating directly in the TDA transform domain. Specifically, six contributions for realizing the codec proposal were presented:

- a low-delay block switching method, requiring only 1–2 ms of extra lookahead at the encoder side while preserving the TDAC and, thereby, PR property (sec. 3.1),
- a signal-adaptive transform kernel switching approach, enabling better channel compaction (and JS coding efficiency) on two-channel input with IPDs near $\pm 90°$ by transitioning from MDCT to MDST-IV coding in one of the channels (sec. 3.2),
- overlap ratio switching via transitions from the MDCT/MDST to so-called MELT coding, i. e., from 50% to 75% inter-transform overlap or vice versa, to partially compensate for the lack of downsampled coding when not using SBR (sec. 3.2),
- a closed-loop frequency-domain long-term predictor, operating on the JS coded transform coefficients, i. e., very close to the spectral quantizer, thereby allowing for much lower algorithmic complexity than, e. g., the predictor in MPEG-2 AAC; this delayless FDP proposal intends to assist the MELT in the coding of harmonic quasi-stationary input, or may substitute the latter in LD applications (sec. 3.3),
- an extended transform-domain coefficient substitution technique, also known as IGF, which unifies the principles of PNS-like core noise filling and SBR-like HFR into an efficient low-complexity algorithm and which, moreover, supports mixed parametric and waveform-preserving coding even at high frequencies (sec. 3.4),
- a TDA-domain semi-parametric Stereo Filling design extending the IGF principle towards lower frequencies in the JS residual spectra by performing a coefficient *copy-over* (from the last frame's downmix) instead of a *copy-up* (from LF regions of the same spectrum); the combination of IGF and SF was termed SBS (sec. 3.5).

All of these contributions were implemented into the MPEG-H 3D Audio codec [ISO15a], whose transform core specification is virtually identical to that of USAC. Then, the "LC" 3D Audio version, with the proposed tools activated and all complex-QMF components disabled, was evaluated against the QMF-enhanced "baseline" 3D Audio variant, both in terms of the required decoding complexity and with regard to subjective coding quality.

The objective assessment, described in section 4.1, demonstrated that the proposed modified 3D Audio codec, thanks to its restriction to only transform-domain algorithms,

- exhibits, at 48 kHz input/output sampling rate, a total encoding-decoding delay between 97.3 ms (including all contributed tools and the default frame length of $N = 1024$ samples) and 33 ms (LD instead of traditional block switching, no ratio switching, and a reduced frame length of $N = 768$). The 97.3 ms undermatch the delays of the QMF-enhanced general-purpose reference codecs, (x)HE-AAC and USAC or 3D Audio, with latencies between 129 and at least 200 ms, respectively. The delay of 33 ms matches that of state-of-the-art communication codecs like 3GPP EVS and MPEG-4 ELD, requiring between 32 [ETSI16] and 39 ms [LuVa10], and is only 6–10 ms higher than that of the very-low-delay Opus codec [Valin13].
- consumes an average decoding complexity which is less than two thirds of that necessitated by a USAC/xHE-AAC or baseline 3D Audio decoder. The worst-case decoder workload, in PCU units, was found to be 4.3 MOPS, or about 72% of the USAC/3D Audio complexity, when disabling ratio switching (i. e., MELT) coding. It is worth noting that the 4.3 MOPS include PCU estimates for super-wideband (SWB) complex predictive JS coding, but in the subjective tests summarized on the next page, the complex-valued stereo prediction was limited to the spectral region below the noise/stereo filling range, i. e., a bandwidth of 3.75 kHz. When restricting the complex prediction support to this NB width, and allowing only real-valued predictive JS at or above 3.75 kHz, the worst-case complexity of the predictive JS decoder can be cut in half. This, in turn, further reduces the overall worst-case decoder workload to about 4.0 MOPS, i.e., two thirds of the reference decoder complexity, just as in case of the average measurement described above.

Section 1.1 established the objective that the flexible codec proposal, in its unrestricted-delay configuration, shall not exhibit a higher decoding complexity than HE-AAC v1 (no Parametric Stereo). To determine whether this goal has been reached by the developed QMF-less 3D Audio design, a look into [ISO15b] again serves well. Therein, the decoder complexities of (A) MPEG-2 AAC-LC at 128 kbit/s and (B) dual-rate MPEG-4 HE-AAC at 56 kbit/s, measured on Cortex A9 hardware for 2.0 stereo output, are tabulated in MHz. Comparison of the two values reveals a complexity ratio of 1.78, i. e., decoder (B) needs 78% additional processing workload than decoder (A) in the given scenario. Applying this factor to the worst-case MPEG-2 AAC-LC complexity of 2.2 MOPS noted in subsection 4.1.1—which was also obtained for 128 kbit/s stereo [Bosi97]—an, arguably, quite accurate estimate of the worst-case HE-AAC decoder complexity at 56 kbit/s stereo, in this case 3.9 MOPS, is obtained. Assuming the abovementioned NB complex-prediction limit, the QMF-less proposal uses 4.0 MOPS, i. e., nearly the same, so the goal is reached.

Naturally, the ultimate objective of any perceptual codec is maximum reconstruction quality for any input signal. To address this aspect, two goals were set out in section 1.1:

- At unrestricted latency, the proposed flexible coding framework should outperform HE-AAC, with the latter utilizing the best encoder available. On [Soun12], a blind listening test was performed at 64 and 96 kbit/s stereo. In both cases, the Winamp 5.63 HE-AAC encoder scored highest by statistically significant margins, and it is a very similar encoder that was employed in the USAC verification tests [Neue13]. The high-rate USAC evaluation depicted in Fig. 2.17, on the one hand, revealed that, for the relevant bit-rate range of 32–64 kbit/s stereo, the MPEG-D codec, utilizing the QMF-based SBR and UniSte MPS extensions inherited by the 3D Audio system, clearly outperforms HE-AAC. Averaging, on the other hand, the per-stimulus MOS values for 3D Audio with SBR and UniSte MPS (blue) and the QMF-less 3D Audio proposal (pink) across Figs. 3.29(a)–(c) yields a mean score of 65 for the former and 67, or 65.5 without MELT coding, for the latter. In other words, the proposed LC 3D Audio codec achieves, on the given item set, a higher overall MOS than the QMF-extended 3D Audio baseline (although no conclusion on statistical significance can be drawn). In view of these observations, it is safe to conclude that, on average, the proposed LC 3D Audio system, at the very least, matches the reconstruction quality of the USAC/baseline 3D Audio codecs and, by inference, subjectively outperforms the HE-AAC codec at the tested bit-rate of 48 kbit/s stereo. Hence, it can be stated that the initial goal has been reached.

- For restricted-delay, i. e., LD, applications, the objective was to at least match the perceptual quality of the general-purpose HE-AAC standard (whose latency is, of course, not constrained) and to exceed that of AAC-ELD or Opus. Unfortunately, this goal has been attained only partially. As illustrated in Figs. 3.30 and 3.34, the proposed 33-ms 3D Audio configuration does outperform Opus (here, just CELT) which, in [Helm14], was shown to achieve at least the same audio quality as ELD. In comparison with HE-AAC, however, the proposal sounds significantly worse, at least at 48 kbit/s stereo. A further 5.1-surround listening test including LD 3D Audio and HE-AAC should be conducted at 80 or 96 kbit/s in order to allow for conclusions with respect to multichannel use cases like VoIP teleconferencing.

In summary, though, it can be concluded that the transform-domain SBS approach is a highly useful semi-parametric substitute for the QMF-based SBR and UniSte MPS pre- and post-processors. Moreover, the kernel switching design successfully assists the SF method on "phasy" input with IPDs near $\pm 90°$, and the IGF scheme, supporting complex-valued envelope calculation, allows for substantial BWE of the coded signal at low rates, without exhibiting issues such as a high HFR complexity or energy loss effects [HsuL11].

In [Kjör16], it is stated that AC-4 circumvents one of the fundamental limitations of earlier systems, namely, the inability to accurately reconstruct important transient and tonal HF components. Thanks to its semi-parametric design, the SBS technique devised herein serves the same purpose and, unlike A-SPX in AC-4, operates directly in the core transform domain, thus bearing the discussed delay and complexity advantage. It shall also be noted in this context that, even when employing fully parametric HFR coding on the highly transient signals of the test sets at hand, as done for the 48-kbit/s evaluation, the QMF-less SBS-extended 3D Audio codec already outclasses the QMF-based reference by about 10 MOS points (see items si02, te15 in Fig. 3.29 and Flame, Robot in Fig. 3.30).

5.1 Considerations for Future Research and Development

Section 4.2 illustrated that, at 48 kbit/s stereo and 80 kbit/s 5.1 multichannel, *good* basic audio quality can be achieved using the codec proposal on all but the most critical input sequences, even when assessed by expert listeners as in all tests of this study. The most critical sequences, for which MOS results close to — but still below — 60 MUSHRA points were obtained with the present encoder tuning, are Fatboy (vocoder), sm01 (bag pipes), and, when not using MELT coding, si01 (harpsichord). All of these audio signals are quasi-stationary, at least over the course of a few frames, with a highly tonal and, in case of the Fatboy Slim excerpt, pitch pulse-like character. For this type of material, the author encourages further research and encoder tuning in the following two directions:

- TD pre-/post-processing with low complexity. Phase-2 3D Audio coding [ISO16] already supports two appropriate tools: a temporal long-term post-filter (LTPF), applied exclusively at the decoder side for fine harmonic shaping of the decoded waveform similar to the speech post-filter of Chen and Gersho [Chen95, Fuch15], and a high-resolution envelope processor based on Vaupel's signal companding [Vaup91] and MPEG-2 AAC's gain control approach [Bosi97], intended primarily for applause-like input. The latter can be regarded as the dual paradigm to TNS, with very efficient context based arithmetic parameter coding if used repeatedly in multiple successive frames. To allow scaling towards perceptually transparent coding, the LTPF can be turned into an open-loop pre-/post-processor (like the TD compander) by applying the corresponding inverse filter before the analysis transforms at the encoder side. It then behaves equivalently to CELT's pitch filter [Valin13], which appears to be highly beneficial on, e. g., the Fatboy waveform. It is expected that the combination of these two recent 3D Audio extensions allows to ensure *good*-range coding quality even on the abovenoted four critical signals when the encoder-side parameter selectors and coders are tuned appropriately.

- CVBR coding for speech-like input. One further property of the Fatboy excerpt is its speech-like character, exhibiting considerable frame-energy fluctuations over time in dependence on the vocal activity pattern. A bit-reservoir governed CVBR encoder as used in MPEG audio may not fully exploit low-level (and, thus, mostly inaudible) pauses between active speech segments, thereby often "wasting bits" in the affected frames that would be more useful in high-energy frames. In other words, said CVBR scheme is closer to a constant bit-rate (CBR) design than to an unconstrained VBR system like Opus. For instance, for the proposed transform codec, a speech-tuned CVBR encoder may be devised which uses only a fraction of the target bit-rate on speech (and, possibly, music) pauses. At the same time, and guided by a respective psychoacoustic model, it could exceed the target rate by at most 20% over a short- or medium-term period — i. e., much less than the Opus encoder [Rämö15] — during "difficult" input passages. At the mean rate of 48 kbit/s on which this work focused, the maximum instantaneous consumption would total 57.6 kbit/s, which still fits into the MCS-9 rate of 58.4 kbit/s used in EDGE. This would turn the QMF-less codec proposal into an interesting solution for, e. g., connection and quality reliable VoIP or Internet radio streaming.

Having addressed the above two aspects, and to complete the current study, a direct comparative MUSHRA test between the QMF-less 3D Audio proposal, in its unrestricted delay setup, and, e. g., Winamp's HE-AAC codec at 48 kbit/s stereo is worth performing. With regard to further longer-term research, the author is under the impression that the objective as well as subjective performance of frame-based perceptual audio coding has saturated during the last decade for both general-purpose and low-delay configurations. Particularly in the *good*-to-*excellent* quality range, all recently developed codec designs provide, at equivalent bit-rates (see also Tab. 1.1), only quite subtle overall audio quality improvements over (x)HE-AAC, as can be concluded from Figs. 2.17, 3.29, and 3.31. In addition, the author observed that, at low bit-rates, the parameters transmitted by the developed coding tools, although consuming only a small part of the total rate, seem to already affect the overall bit-budget so much that the possible coding gains are partially canceled. Transmitting a considerable amount of further side-information should, thus, be avoided. In the author's experience and opinion, remaining potential for coding gain may lie in lower-rate long-term backward-adaptive or bidirectional predictive designs (i. e., using predictions from past and/or future reconstructed frames), as extensions of the work described in [Paras95].

A Appendices

A.1 Comparative Evaluation of Joint-Stereo Coding Algorithms

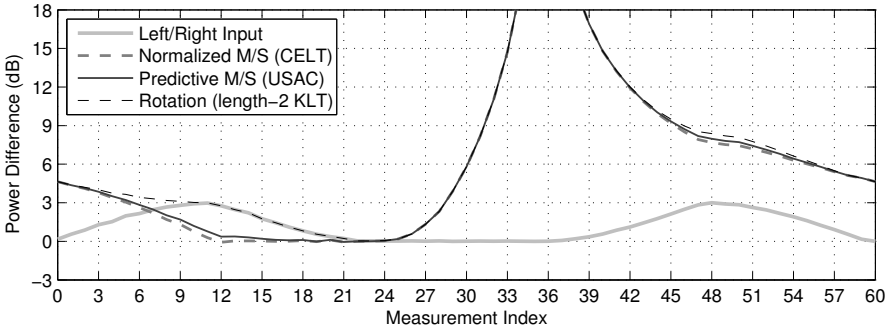


Figure A.1. Channel compaction performances, measured as the frame-wise energy difference between the downmix and residual spectrum, for the three joint-stereo coding methods discussed in subsections 2.2.3 and 3.5.1. The two-channel input waveform is mixed from three independent white pseudo-random noise sources panned fully to the left, right, and center, respectively, and is available at www.ecodis.de/noise.wav.

A.2 Scale Factor Band Offsets and Widths since MPEG-2 AAC

```
sfb_offsets_long = [0, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 48, 56, 64, 72, 80, 88, 96, 108, ...
                    120, 132, 144, 160, 176, 196, 216, 240, 264, 292, 320, 352, 384, ...
                    416, 448, 480, 512, 544, 576, 608, 640, 672, 704, 736, 768, 800, ...
                    832, 864, 896, 928, 960, 992, 1024];

if (fs > 32) sfb_offsets_long(end−2:end) = 1024; end   % sample rate (fs) is 44.1 or 48 kHz

sfb_offsets_short = [0, 4, 8, 12, 16, 20, 28, 36, 44, 56, 68, 80, 96, 112, 128];

sfb_widths_long = diff(sfb_offsets_long);     sfb_widths_short = diff(sfb_offsets_short);
```

A.3 Pseudo-Code for BiLLIG Encoding and Decoding Routines

numBitsWritten = **encodeData**(bitStr, value)

```
// look-up table for code words for values 2...28
cw = [10, 26, 56, 58, 120, 122, 248, 250, 504,
      506, 1016, 1018, 2040, 2042, 4088, 4090,
      8184, 16376, 16378, 32760, 32762, 65528,
      65530, 131064, 131066, 131068, 131070];
// look-up table for code lengths for values 2...28
cln = [4, 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 11, 11, 12,
      12, 13, 14, 14, 15, 15, 16, 16, 17, 17, 17, 17];

if (value == 0) {
  numBitsWritten = bitStr.setBits(0, 1);   // 0
} else if (value == –1) {
  numBitsWritten = bitStr.setBits(4, 3);   // 100
} else if (value == 1) {
  numBitsWritten = bitStr.setBits(12, 4); // 1100
} else {
  signBit = 0;
  if (value < 0) {
    signBit = 1;
  }
  absValue = |value|;
  if (absValue < 29) {
    absValue = absValue – 2;
    numBitsWritten = bitStr.setBits(cw[absValue]
                     + signBit, cln[absValue]);
  } else {
    absValue = absValue * 2 – 58;
    numBitsWritten = bitStr.setBits(261952
                     + absValue + signBit, 18);
  }
}
return numBitsWritten;
```

value = **decodeData**(bitStr)

```
bit = 0, count = –1;
do {
  bit = bitStr.getBits(1);
  count = count + 1;
} while (bit ≠ 0 && count < 14);
if (count == 0) {
  value = 0;   // 0
} else if (count ≤ 2) {
  if (bitStr.getBits(1) ≠ 0) {
    value = count + 1;
    if (bitStr.getBits(1) ≠ 0) {
      value = value * –1;
    }
  } else value = (–1)^count;   // 1(1)00
} else if (count == 10) {
  if (bitStr.getBits(1) ≠ 0) {
    value = bitStr.getBits(5) + 29;
    if (bitStr.getBits(1) ≠ 0) {
      value = value * –1;
    }
  } else {
    value = 18;
    if (bitStr.getBits(1) ≠ 0) {
      value = value * –1;
    }
  }
} else {
  count = count + bit;
  value = bitStr.getBits(1) – 2 + count*2;
  if (count > 10) {
    value = value – 1;
  }
  if (bitStr.getBits(1) ≠ 0) {
    value = value * –1;
  }
}
return value;
```

Table A.2. BiLLIG algorithms as pseudo-code. setBits/getBits(…, *n*) write/read an *n*-bit integer.

A.4 Stereo and 5.1 Material Used for the Subjective Evaluation

2.0 Item	Description	Source	**5.1** Item	Description	Source
acc	Accordion (MPEG te16 item)	AAC st	amb1	ambience, city station/hall	MPS st
arira	Korean speech (Arirang)	USAC st	amb2	ambience, Wimbledon, live	MPS st
CWt	Max Mutzke – Can't Wait		applau	applause in rear channels	EBU07
	Until Tonight (Wurlitzer, B)	Hydr16	applse	other applause in all chans.	MPS st
EiG	Abfahrt Hinwil – Everything		ARL_a	ARL applause, dense claps	MPS st
	is Green (B)	Hydr16	Bach_o	church organ with stop-outs	EBU07
Fatboy	Fatboy Slim – Kalifornia (B)	Hydr16	brassE	Exodus, brass instruments	EBU07
Flame	Flamenco guitar, castanets	Hydr16	chosta	Chostakovitch, orch./strings	MPS st
HPt	Harry Potter (audio book,		fleetw	Fleetwood Mac, guitar, male	EBU07
	male English narrator, E)	USAC st	founta	fountain music, piano, birds	MPS st
Mu1	Astral Doors – Of the Son		glock	Glockenspiel and Timpani	MPS st
	and the Father (E)	USAC st	Harpsi	harpsichord, isolated notes	EBU07
Mu3	The Beatles – I Feel Fine (B)	USAC st	harpsi	Bach, harpsichord concert	EBU07
phi7	Pitch pipe (shortened si03)	USAC st	hbv_gr	small choir, Gregorian chant	EBU07
Robot	Kraftwerk – die Roboter (E)	Hydr16	indie2	Indiana Jones, movie scene	MPS st
RockY	Jennifer Warnes – Rock You	CD, un-	jackso	Jackson gospel singers, live	MPS st
	Gently (CD The Hunter, B)	known	Moonr	Mancini, mouth org., strings	EBU07
Sal	Salvation (monch choir)	USAC st	pops	Japanese pop song, female	MPS st
si01	Harpsichord (CD track 40)	EBU08	poulen	Poulenc, orchestra/organ	MPS st
si02	Castanets (CD track 27, B)	EBU08	R_Plan	rock, Robert Plant – Whole	
sm01	Bag pipes	AAC st		Lotta Love, live, applause	EBU07
sm02	Glockenspiel (CD tr. 35, E)	EBU08	rock_c	rock concert with clapping	MPS st
SoM	Speech over music (TV ad)	USAC st	Sax_pi	Jazz, saxophone and piano	EBU07
te15	Bizet – Carmen (Aragon. IV)	AAC st	Sedam	Sedambonjou, atmospheric	
Trump	recording of single trumpet	Hydr16		Latin-American Salsa music	EBU07
Vega/$_{es01}$	Suzanne Vega – Tom's Diner	AAC st	Stomp	Stomp, percussion, live	MPS st
Velvet	Green Velvet – Coitus (B)	Hydr16			

Table A.3. Selected sequences for listening tests. B: beginning, E: excerpt, st: standardization.

Acknowledgments

In early 2000 I had the opportunity to hear an MP3 file, created from an audio CD, for the first time. Astonished by its sound quality despite its small size, I decided to embark upon what turned out to be a 16-year journey through the fields of digital audio signal processing and coding. This journey led me to numerous places and fellow researchers across the world and culminated in the writing of this thesis, whose completion would not have been possible without the expertise, guidance and support of said researchers.

The first eight years of my journey I devoted to the familiarization with and understanding of the fundamental concepts of audio processing and coding in order to be able to answer the question: "**How does MP3 work?**" I sincerely thank Prof. Dr. Udo Zölzer for, during this period, introducing me to the academic field of audio coding, for showing me how to approach scientific study in this area in an organized, systematic, and chronological way (before meeting him in 2006 I was a coding autodidact) and for giving me the opportunity to craft and present my first own contribution to the audio engineering community. Further gratitude goes to Dr. Martin Holters for allowing me to pursue both my project and my Master's thesis on psychoacoustically optimized quantization noise shaping under his supervision. His diligence greatly influenced the way I work today.

The second eight years of my journey I spent on the meticulous study of the details of state-of-the-art audio codec designs with the ultimate objective of being able to answer the question: "**How can you do better than MP3?**" My research towards this goal was conducted almost exclusively at the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany. Of the countless colleagues I met — and had the honor of working with — there over the years, I would like to especially thank for the fruitful discussion, helpful advice, constructive criticism, careful listening, debugging, and/or other support in many different ways (in alphabetical order): Andreas Niedermeier, Axel Horndasch, Benjamin Schubert, the CDK team, Christian Neukam, Christopher Oates, Dr. Emmanuel Ravelli, Florian Schuh, Dr. Florin Ghido, Dr. Frederik Nagel, Prof. Dr. Gerald Schuller, Goran Marković, Dr. Guillaume Fuchs, Jérémie Lecomte, Johannes Hilpert, Julien Robilliard, María Luis Valero, Markus Multrus, Max Neuendorf, Michael Fischer, Nikolaus Rettelbach, Dr. Ralf Geiger, Richard Füg, Sascha Dick, Dr. Sascha Disch, Dr. Stefan Bayer, Dr. Takehiro Moriya, Tobias Schwegler, and last but not least, Prof. Dr. Tom Bäckström. The folks at the well moderated hydrogenaud.io discussion forums, with their highly useful collection of codec test samples and experience in the preparation, execution, statistical analysis, and extensive documentation of blind listening tests, deserve gratitude as well.

Finally, I gratefully acknowledge Prof. Dr. Walter Kellermann's and Prof. Dr. Elmar Nöth's participation in my doctorate defense and the careful review of my thesis by Prof. Dr. Bernd Edler, Prof. Dr. Rudolf Rabenstein, Florian Schuh, and Goran Marković. Moreover, I'd like to thank the professors, students, and staff of the International Audio Laboratories Erlangen, a joint institution of Fraunhofer IIS and the Friedrich-Alexander University (FAU) of Erlangen-Nürnberg which since 2013 has been my place of work in pursuance of this Dr.-Ing. dissertation. I extend particular appreciation and gratitude to Dr. Bernhard Grill as Fraunhofer IIS representative for granting scientific and financial support of my doctorate position at the Audio Labs, Prof. Dr. Jürgen Herre as FAU representative for his experience and expertise extended to me in several fruitful discussions and, of course, my own group at the Audio Labs. This group, which in April 2016 comprised Dr. Armin Taghipour, Esther Feichtner, Fabian-Robert Stöter, Konstantin Schmidt, and Nils Werner, is headed by Prof. Dr. Bernd Edler, who always offered an open ear and door to me when I needed it. His knowledge and advice was essential in the completion of this work, and, as is the case with Prof. Bäckström, his sense for detail (identifying all algorithmic parameters) and exhaustive study (trying all possibilities) is truly inspiring.

"*The prize is the pleasure of finding the thing out,*
the kick in the discovery,
the observation that other people use it."

Richard Feynman (1918–1988)

References

[Alla99] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, "MPEG-4 Low Delay Audio Coding Based on the AAC Codec," in *Proc. AES 106th Convention*, Munich, no. 4929, May 1999.

[Anna06] R. Annadana, E. V. Harinarayanan, A. J. Ferreira, and D. Sinha, "New Results in Low Bit Rate Speech Coding and Bandwidth Extension," in *Proc. AES 121st Convention*, San Francisco, no. 6876, Oct. 2006.

[ATSC12] Advanced Television Systems Committee, "ATSC Standard: Digital Audio Compression (AC-3, E-AC-3)," doc. A/52:2015, Dec. 2012. atsc.org/standards/

[Bäck15] T. Bäckström and C. R. Helmrich, "Arithmetic Coding of Speech and Audio Spectra Using TCX Based on Linear Predictive Spectral Envelopes," in *Proc. IEEE ICASSP*, Brisbane, pp. 5127–5131, Apr. 2015.

[Baue06] C. Bauer and M. Vinton, "Joint Optimization of Scale Factors and Huffman Code Books for MPEG-4 AAC," *IEEE Trans. Signal Processing*, vol. 54, no. 1, pp. 177–189, Jan. 2006.

[Baum03] F. Baumgarte and C. Faller, "Binaural Cue Coding—Part I: Psychoacoustic Fundamentals and Design Principles," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 509–519, Nov. 2003.

[Blau96] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, revised edition, MIT Press, 1996.

[Bosi97] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Y. Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding," *J. Audio Eng. Soc.*, vol. 45, no. 10, pp. 789–814, Oct. 1997.

[Bran97] K. Brandenburg and M. Bosi, "Overview of MPEG Audio: Current and Future Standards for Low-Bitrate Audio Coding," *J. Audio Eng. Soc.*, vol. 45, no. 1/2, pp. 4–21, Jan. 1997.

[Bran13] K. Brandenburg, C. Faller, J. Herre, J. D. Johnston, and W. B. Kleijn, "Perceptual Coding of High-Quality Digital Audio," *Proc. of IEEE*, vol. 101, no. 9, pp. 1905–1919, Sep. 2013.

[Bree05] J. Breebaart, S. van de Par, A. Kohlrausch, and E. Schuijers, "Parametric Coding of Stereo Audio," *EURASIP J. Applied Signal Processing*, vol. 2005, no. 9, pp. 1305–1322, Sep. 2005.

[Bree07] J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, and S. van de Par, "Background, Concept, and Architecture for the Recent MPEG Surround Standard on Multichannel Audio Compression," *J. Audio Eng. Soc.*, vol. 55, no. 5, pp. 331–351, May 2007.

[Brita03] V. Britanak, "A Unified Fast Computation of the Evenly and Oddly Stacked MDCT/MDST," in *Proc. 4th EURASIP Conf. VIP-MC*, Zagreb, pp. 233–238, 2003.

[Chen95] J.-H. Chen and A. Gersho, "Adaptive Postfiltering for Quality Enhancement of Coded Speech," *IEEE Trans. Speech & Audio Processing*, vol. 3, no. 1, Jan. 1995.

[Chen04] C. Cheng, "Method for Estimating Magnitude and Phase in the MDCT Domain," in *Proc. AES 116th Convention*, Berlin, no. 6091, May 2004.

[DenB09] A. C. den Brinker, J. Breebaart, P. Ekstrand, J. Engdegård, F. Henn, K. Kjörling, W. Oomen, and H. Purnhagen, "An Overview of the Coding Standard MPEG-4 Audio Amendments 1 and 2: HE-AAC, SSC, and HE-AAC v2," *EURASIP J. Audio, Speech, and Music Processing*, vol. 2009, article ID 468971, June 2009.

[Disch15] S. Disch, C. Neukam, and K. Schmidt, "Temporal Tile Shaping for Spectral Gap Filling in Audio Transform Coding in EVS," in *Proc. IEEE ICASSP*, Brisbane, pp. 5873–5877, Apr. 2015.

[EBU07] European Broadcasting Union, technical report 3324, "EBU Evaluations of Multichannel Audio Codecs," Sep. 2007. tech.ebu.ch/docs/tech/tech3324.pdf

[EBU08] European Broadcasting Union, technical report 3253, "Sound Quality Assessment Material recordings for subjective tests (Users' handbook)," Sep. 2008. tech.ebu.ch/docs/tech/tech3253.pdf and tech.ebu.ch/publications/sqamcd/

[Edler89] B. Edler, "Codierung von Audiosignalen mit überlappender Transformation und adaptiven Fensterfunktionen," *Frequenz*, vol. 43, pp. 252–256, Sep. 1989.

[Elfitr11] I. Elfitri, B. Günel, and A. M. Kondoz, "Multichannel Audio Coding Based on Analysis by Synthesis," *Proc. of IEEE*, vol. 99, no. 4, pp. 657–670, Apr. 2011.

[Engd04] J. Engdegård, H. Purnhagen, J. Rödén, and L. Liljeryd, "Synthetic Ambience in Parametric Stereo Coding," in *Proc. AES 116th Convention*, Berlin, no. 6074, May 2004.

[ETSI12] ETSI/GSM, techn. specification TS 145 001, "Digital Cellular Telecommunications System (Phase 2+); Physical layer on the radio path; General description (3GPP TS 45.001 version 10.1.0)," v10.1.0, Jan. 2012. www.etsi.org/standards-search#search=145001

[ETSI13] ETSI/EBU, standard ES 201 980, "Digital Radio Mondiale (DRM); System Specification," v4.1.1, Jan. 2014. www.etsi.org/standards-search#search=201980

[ETSI14] ETSI/EBU, techn. specification TS 103 190, "Digital Audio Compression (AC-4) Standard," v1.1.1, Apr. 2014. www.etsi.org/standards-search#search=103190

[ETSI15] ETSI/EBU, techn. specification TS 103 190-2, "Digital Audio Compression (AC-4) Standard, Part 2: Immersive and personalized audio," v1.1.1, Sep. 2015.

[ETSI16] ETSI/EBU, techn. specification TS 126 445, "Universal Mobile Telecommunications System (UMTS); LTE; Codec for Enhanced Voice Services (EVS); Detailed algorithmic description (3GPP TS 26.445 version 13.0.0)," v13.0.0, Feb. 2016.

[Falle03] C. Faller and F. Baumgarte, "Binaural Cue Coding — Part II: Schemes and Applications," *IEEE Trans. Speech & Audio Processing*, vol. 11, no. 6, pp. 520–531, Nov. 2003.

[Fastl07] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd, Springer, 2007.

[Ferre05] A. J. Ferreira and D. Sinha, "Accurate Spectral Replacement," in *Proc. AES 118th Convention*, Barcelona, no. 6383, May 2005.

[Field89] L. D. Fielder, "Low-Complexity Transform Coder for Music Applications," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, session 2-1, Oct. 1989.

[Field96] L. D. Fielder, M. Bosi, G. Davidson, M. Davis, C. Todd, and S. Vernon, "AC-2 and AC-3: Low-Complexity Transform-Based Audio Coding," *AES Collected Papers on Digital Audio Bit-Rate Reduction*, pp. 54–72, 1996.

[Field04] L. D. Fielder, R. Andersen, B. Crockett, G. Davidson, M. Davis, S. Turner, M. Vinton, and P. Williams, "Introduction to Dolby Digital Plus, an Enhancement to the Dolby Digital Coding System," in *Proc. AES 117th Convention*, San Francisco, no. 6196, Oct. 2004.

[Fuch93] H. Fuchs, "Improving Joint Stereo Audio Coding by Adaptive Inter-Channel Prediction," in *Proc. IEEE WASPAA*, New Paltz, pp. 39–42, Oct. 1993.

[Fuch95] H. Fuchs, "Improving MPEG Audio Coding by Backward Adaptive Linear Stereo Prediction," in *Proc. AES 99th Convention*, New York, no. 4086, Oct. 1995.

[Fuch11] G. Fuchs, V. Subbaraman, and M. Multrus, "Efficient Context Adaptive Entropy Coding for Real-Time Applications," in *Proc. IEEE ICASSP*, Prague, pp. 493–496, May 2011.

[Fuch15] G. Fuchs, C. R. Helmrich, G. Marković, M. Neusinger, E. Ravelli, and T. Moriya, "Low Delay LPC and MDCT-Based Audio Coding in the EVS Codec," in *Proc. IEEE ICASSP*, Brisbane, pp. 5723–5727, Apr. 2015.

[Geig02] R. Geiger and G. D. T. Schuller, "Integer Low-Delay and MDCT Filter Banks," in *Proc. Asilomar Conf. on Signals, Syst., Comput.*, vol. 1, pp. 811–815, Sep. 2002.

[Gers94] A. Gersho, "Advances in Speech and Audio Compression," *Proc. of IEEE*, vol. 82, no. 6, pp. 900–918, June 1994.

[Gribb14] C. Gribben and H. Lee, "The Perceptual Effects of Horizontal and Vertical Interchannel Decorrelation, Using the Lauridsen Decorrelator," in *Proc. AES 136th Convention*, Berlin, no. 9027, Apr. 2014.

[Hame05] K. M. A. Hameed and E. Elias, "Extended Lapped Transforms with Linear Phase Basis Functions and Perfect Reconstruction," in *Proc. IEEE Int. Conf. on Electronics, Circuits, and Systems (ICECS)*, Gammarth, pp. 1–4, Dec. 2005.

[Hayk14] S. Haykin, *Adaptive Filter Theory*, 5th international edition, Pearson, 2014.

[Helm10] C. R. Helmrich, "On the Use of Sums of Sines in the Design of Signal Windows," in *Proc. DAFx-10*, Graz, Sep. 2010. http://dafx10.iem.at/proceedings/

[Helm11] C. R. Helmrich, P. Carlsson, S. Disch, B. Edler, J. Hilpert, M. Neusinger, H. Purnhagen, N. Rettelbach, J. Robilliard, and L. Villemoes, "Efficient Transform Coding of Two-Channel Audio Signals by Means of Complex-Valued Stereo Prediction," in *Proc. IEEE ICASSP*, Prague, pp. 497–500, May 2011.

[Helm14] C. R. Helmrich, G. Marković, and B. Edler, "Improved Low-Delay MDCT-Based Coding of Both Stationary and Transient Audio Signals," in *Proc. IEEE ICASSP*, Florence, pp. 6954–6958, May 2014.

[Hel15a] C. R. Helmrich, A. Niedermeier, S. Disch, and F. Ghido, "Spectral Envelope Reconstruction via IGF for Audio Transform Coding," in *Proc. IEEE ICASSP*, Brisbane, pp. 389–393, Apr. 2015.

[Hel15b] C. R. Helmrich, A. Niedermeier, S. Bayer, and B. Edler, "Low-Complexity Semi-Parametric Joint-Stereo Audio Transform Coding," in *Proc. EURASIP 23rd EUSIPCO*, Nice, pp. 794–798, Sep. 2015.

[Hel15c] C. R. Helmrich and B. Edler, "Signal-Adaptive Transform Kernel Switching for Stereo Audio Coding," in *Proc. IEEE WASPAA*, New Paltz, pp. 1–5, Oct. 2015.

[Hel15d] C. R. Helmrich and M. Fischer, "Low-Delay Transform Coding Using the MPEG-H 3D Audio Codec," in *Proc. AES 139th Convention*, New York, no. 9355, Oct. 2015.

[Hel16a] C. R. Helmrich and B. Edler, "Signal-Adaptive Switching of Overlap Ratio in Audio Transform Coding," *Proc. IEEE ICASSP*, Shanghai, pp. 639–643, Mar. 2016.

[Hel16b] C. R. Helmrich and B. Edler, "Audio Coding Using Overlap and Kernel Adaptation," *IEEE Signal Processing Letters*, vol. 23, no. 5, pp. 590–594, May 2016.

[Hel16c] C. R. Helmrich, R. Füg, and B. Edler, "Frequency-Domain Prediction for Audio Coding," submitted to *IEEE Signal Processing Letters*, 2016. www.ecodis.de

[Herr96] J. Herre and J. D. Johnston, "Enhancing the Performance of Perceptual Audio Coders by Using Temporal Noise Shaping (TNS)," in *Proc. AES 101st Convention*, Los Angeles, no. 4384, Nov. 1996.

[Her97a] J. Herre and J. D. Johnston, "Exploiting Both Time and Frequency Structure in a System That Uses an Analysis/Synthesis Filterbank with High Frequency Resolution," in *Proc. AES 103rd Convention*, New York, no. 4519, Sep. 1997.

[Her97b] J. Herre and J. D. Johnston, "Continuously Signal-Adaptive Filterbank for High-Quality Perceptual Audio Coding," in *Proc. IEEE WASPAA*, New Paltz, Oct. 1997.

[Herr98] J. Herre and D. Schulz, "Extending the MPEG-4 AAC Codec by Perceptual Noise Substitution," in *Proc. AES 104th Convention*, Amsterdam, no. 4720, May 1998.

[Herr08] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG Surround — The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," *J. Audio Eng. Soc.*, vol. 56, no. 11, pp. 932–955, Nov. 2008.

[Herr14] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio — The New Standard for Universal Spatial/3D Audio Coding," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830, Dec. 2014.

[Herr15] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D Audio — The New Standard for Coding of Immersive Spatial Audio," *IEEE J. Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779, Aug. 2015.

[Hoth08] G. Hotho, L. Villemoes, and J. Breebaart, "A Backward-Compatible Multichannel Audio Codec," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 83–93, Jan. 2008.

[How94] P. G. Howard and J. S. Vitter, "Fast Progressive Lossless Image Compression," in *Proc. SPIE*, no. 2186, pp. 98–109, Mar. 1994.

[HsuL11] H.-W. Hsu and C.-M. Liu, "Decimation-Whitening Filter in Spectral Band Replication," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2304–2313, Nov. 2011.

[Huff52] D. Huffman, "A Method for the Construction of Minimum-Redundancy Codes," *Proc. of I.R.E.*, pp. 1098–1101, Sep. 1952.

[Hydr16] HydrogenAudio, "listening tests". hydrogenaud.io/index.php/board,40.0.html and http://listening-tests.hydrogenaud.io/sebastian/

[Hyun12] D. Hyun, Y. Park, and D. H. Youn, "Estimation and Quantization of ICC-dependent Phase Parameters for Parametric Stereo Audio Coding," *EURASIP J. Audio, Speech, and Music Processing*, vol. 2012, no. 27, Nov. 2012.

[IETF12] Internet Engineering Task Force, J.-M. Valin, K. Vos, and T. Terriberry, Request for Comments (RFC) 6716, "Definition of the Opus Audio Codec," ISSN 2070-1721, Sep. 2012. https://tools.ietf.org/html/rfc6716

[Irwa02] R. Irwan and R. M. Aarts, "Two-to-Five Channel Sound Processing," *J. Audio Eng. Soc.*, vol. 50, no. 11, pp. 914–926, Nov. 2002.

[ISO93] ISO/IEC International Standard 11172-3, "Information technology — Coding of moving pictures and associated audio for digital storage media at up to 1.5 Mbit/s (MPEG-1) — Part 3: Audio," Geneva, Aug. 1993.

[ISO97] ISO/IEC International Standard 13818-7, "Information technology — Generic coding of moving pictures and associated audio information (MPEG-2) — Part 7: Advanced Audio Coding (AAC)," Geneva, Dec. 1997.

[ISO07] ISO/IEC International Standard 23003-1, "Information technology — MPEG audio technologies (MPEG-D) — Part 1: MPEG Surround," Geneva, Feb. 2007.

[ISO09] ISO/IEC International Standard 14496-3, "Information technology — Coding of audio-visual objects (MPEG-4) — Part 3: Audio," Geneva, Dec. 2001–2009.

[ISO12] ISO/IEC International Standard 23003-3, "Information technology — MPEG audio technologies — Part 3: Unified speech and audio coding," Geneva, 2012.

[ISO15a] ISO/IEC International Standard 23008-3, "Information technology—High efficiency coding and media delivery in heterogeneous environments (MPEG-H) — Part 3: 3D audio," Geneva, Oct. 2015.

[ISO15b] ISO/IEC Input document M37167, "Proposal for profiles and levels for 3D Audio," Geneva, Oct. 2015.

[ISO15c] ISO/IEC Input document M37238, "Proposal for the definition of a Ultra-Low Latency HEVC profile for content production applications," Geneva, Oct. 2015.

[ISO16] ISO/IEC Output document N16391, "Draft of ISO/IEC 23008-3:201x, MPEG-H 3D Audio, Second Edition," Chengdu, Oct. 2016.

[ITU96] International Telecommunication Union (ITU), Telecommunication Standard. Sector, Recommendation ITU-T P.800, "Methods for subjective determination of transmission quality," Geneva, Aug. 1996.

[ITU15a] International Telecommunication Union (ITU), Radiocommunication Sector, Recommendation ITU-R BS.1116-3, "Method for the subjective assessment of small impairments in audio systems," Geneva, Feb. 2015.

[ITU15b] International Telecommunication Union (ITU), Radiocommunication Sector, Recommendation ITU-R BS.1534-3, "Method for the subjective assessment of intermediate quality level of audio systems," Geneva, Oct. 2015.

[JaNo84] N. S. Jayant and P. Noll, *Digital Coding of Waveforms — Principles and Applications to Speech and Video*, 1st, Prentice-Hall Signal Processing Series, 1984.

[Jelín09] M. Jelínek, T. Vaillancourt, and J. Gibbs, "G.718: A New Embedded Speech and Audio Coding Standard with High Resilience to Error-Prone Transmission Channels," *IEEE Commun. Mag.*, vol. 47, no. 10, pp. 117–123, Oct. 2009.

[John88] J. D. Johnston, "Transform Coding of Audio Signals Using Perceptual Noise Criteria," *IEEE J. Selected Areas in Commun.*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[John89] J. D. Johnston, "Perceptual Transform Coding of Wideband Stereo Signals," in *Proc. IEEE ICASSP*, Glasgow, vol. 3, pp. 1993–1996, May 1989.

[John92] J. D. Johnston and A. J. Ferreira, "Sum-Difference Stereo Transform Coding," in *Proc. IEEE ICASSP*, San Francisco, vol. 2, pp. 569–572, Mar. 1992.

[Kjör16] K. Kjörling, J. Rödén, M. Wolters, J. Riedmiller, *et al.*, "AC-4 — The Next Generation Audio Codec," in *Proc. AES 140th Convention*, Paris, no. 9491, June 2016.

[Krüg08] H. Krüger and P. Vary, "A New Approach for Low-Delay Joint-Stereo Coding," in *Proc. ITG-Fachtagung Sprachkommunikation*, Aachen, pp. 1–4, VDE, Oct. 2008.

[KuoJ01] S.-S. Kuo and J. D. Johnston, "A Study of Why Cross Channel Prediction is Not Applicable to Perceptual Audio Coding," *IEEE Signal Processing Letters*, vol. 8, no. 9, pp. 245–247, Sep. 2001.

[Laak05] A. Laaksonen, "Bandwidth extension in high-quality audio coding," M. Sc. thesis, Helsinki Univ. of Technol., May 2005. http://urn.fi/urn:nbn:fi:tkk-007914

[Laur54] H. Lauridsen, "Experiments Concerning Different Kinds of Room-Acoustics Recording (in Danish)," *Ingenioren*, vol. 47, pp. 906–910, Dec. 1954.

[LeeC13] Y. H. Lee and S. H. Choi, "Superwideband Bandwidth Extension Using Normalized MDCT Coefficients for Scalable Speech and Audio Coding," *Advances in Multimedia*, vol. 2013, no. 909124, Hindawi, June 2013.

[Lieb02] T. Liebchen, "Lossless Audio Coding Using Adaptive Multichannel Prediction," in *Proc. AES 113th Convention*, Los Angeles, no. 5680, Oct. 2002.

[Lind05] J. Lindblom, J. H. Plasberg, and R. Vafin, "Flexible Sum-Difference Stereo Coding Based on Time-Aligned Signal Components," in *Proc. IEEE WASPAA*, New Paltz, pp. 255–258, Oct. 2005.

[Lutz04] M. Lutzky, G. Schuller, M. Gayer, U. Krämer, and S. Wabnik, "A Guideline to Audio Codec Delay," in *Proc. AES 116th Convention*, Berlin, no. 6062, May 2004.

[LuVa10] M. Luis Valero, A. Hölzer, M. Schnell, J. Hilpert, M. Lutzky, *et al.*, "A New Parametric Stereo and Multi-Channel Extension for MPEG-4 Enhanced Low Delay AAC (AAC-ELD)," in *Proc. AES 128th Convention*, London, no. 8099, May 2010.

[Mahi89] Y. Mahieux, J. P. Petit, and A. Charbonnier, "Transform Coding of Audio Signals using correlation between successive transform blocks," in *Proc. IEEE ICASSP*, Glasgow, vol. 3, pp. 2021–2024, May 1989.

[Makh79] J. Makhoul and M. Berouti, "High-Frequency Regeneration in Speech Coding Systems," in *Proc. IEEE ICASSP*, Washington, vol. 4, pp. 428–431, Apr. 1979.

[Mäki05] J. Mäkinen, B. Bessette, S. Bruhn, P. Ojala, R. Salami, and A. Taleb, "AMR-WB+: A New Audio Coding Standard for 3rd Generation Mobile Audio Services," in *Proc. IEEE ICASSP*, Philadelphia, vol. 2, pp. 1109–1112, Mar. 2005.

[Mal90a] H. S. Malvar, "Modulated QMF Filter Banks with Perfect Reconstruction," *Electronics Letters*, vol. 26, no. 13, pp. 906–907, June 1990.

[Mal90b] H. S. Malvar, "Lapped Transforms for Efficient Transform/Subband Coding," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, no. 6, pp. 969–978, June 1990.

[Mal92a] H. S. Malvar, "Extended Lapped Transforms: Properties, Applications, and Fast Algorithms," *IEEE Trans. Signal Processing*, vol. 40, no.11, pp. 2703–2714, Nov. 1992.

[Mal92b] H. S. Malvar, *Signal Processing with Lapped Transforms*, Artech House, 1992.

[Malv99] H. S. Malvar, "A Modulated Complex Lapped Transform and its Applications to Audio Processing," in *Proc. IEEE ICASSP*, Phoenix, pp. 1421–1424, Mar. 1999.

[Mart79] G. N. N. Martin, "Range Encoding: An Algorithm for Removing Redundancy from a Digitized Message," in *Proc. Inst. Electron. Radio Eng. Int. Conf. on Video and Data Recording*, Southampton, July 1979.

[Mau95] J. Mau, J. Valot, and D. Minaud, "Time-Varying Orthogonal Filter Banks without Transient Filters," in *Proc. IEEE ICASSP*, Detroit, vol. 2, pp. 1328–1331, May 1995.

[Mein05] N. Meine and B. Edler, "Improved Quantization and Lossless Coding for Subband Audio Coding," in *Proc. AES 118th Convention*, Barcelona, no. 6468, May 2005.

[Melk14] V. Melkote, K. Yen, M. Fellers, G. Davidson, and V. Kumar, "Transform-Domain Decorrelation in Dolby Digital Plus," in *Proc. IEEE ICASSP*, Florence, pp. 6949–6953, May 2014.

[Moor12] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 6th, Emerald, 2012.

[Mori96] T. Moriya, N. Iwakami, K. Ikeda, and S. Miki, "Extension and Complexity Reduction of TwinVQ Audio Coder," in *Proc. IEEE ICASSP*, Atlanta, vol. 2, pp. 1029–1032, May 1996.

[Mori15] T. Moriya, Y. Kamamoto, N. Harada, T. Bäckström, C. R. Helmrich, and G. Fuchs, "Harmonic Model for MDCT Based Audio Coding with LPC Envelope," in *Proc. EURASIP 23rd EUSIPCO*, Nice, pp. 789–793, Sep. 2015.

[Neue13] M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. R. Helmrich, R. Lefebvre, P. Gournay, B. Bessette, J. Lapierre, K. Kjörling, H. Purnhagen, L. Villemoes, W. Oomen, E. Schuijers, K. Kikuiri, T. Chinen, T. Norimatsu, K. Chong, E. Oh, M. Kim, S. Quackenbush, and B. Grill, "The ISO/MPEG Unified Speech and Audio Coding Standard — Consistent High Quality for All Content Types and at All Bit Rates," *J. Audio Eng. Soc.*, vol. 61, no. 12, pp. 956–977, Dec. 2013.

[Neuk13] C. Neukam, F. Nagel, G. Schuller, and M. Schnabel, "A MDCT Based Harmonic Spectral Bandwidth Extension Method," in *Proc. IEEE ICASSP*, Vancouver, pp. 566–570, May 2013.

[Niam03] O. A. Niamut and R. Heusdens, "Subband Merging in Cosine-Modulated Filter Banks," *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 111–114, Apr. 2003.

[Nutt81] A. H. Nuttall, "Some Windows with Very Good Sidelobe Behavior," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, no. 1, pp. 84–91, Feb. 1981.

[Ojan99] J. Ojanperä, M. Väänänen, and L. Yin, "Long Term Predictor for Transform Domain Perceptual Audio Coding," in *Proc. AES 107th Convention*, New York, no. 5036, Sep. 1999.

[Padm92] M. Padmanabhan and K. Martin, "Some Further Results on Modulated/Extended Lapped Transforms," in *Proc. IEEE ICASSP*, vol. 4, pp. 265–268, Mar. 1992.

[Paras95] M. Paraskevas and J. Mourjopoulos, "A Differential Perceptual Audio Coding Method with Reduced Bitrate Requirements," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 6, pp. 490–503, Nov. 1995.

[Phili08] P. Philippe, D. Virette, and B. Kövesi, "Time-Varying Transform for High Quality Audio Communication Codecs," in *Proc. AES 124th Convention*, Amsterdam, no. 7333, May 2008.

[Prab85] K. M. M. Prabhu, "A Set of Sum-Cosine Window Functions," *Int. J. of Electronics*, vol. 58, no. 6, pp. 969–974, June 1985.

[Prin86] J. P. Princen and A. B. Bradley, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 5, pp. 1153–1161, Oct. 1986.

[Prin87] J. P. Princen, A. W. Johnson, and A. B. Bradley, "Subband/Transform Coding Using Filter Bank Design Based on Time Domain Aliasing Cancellation," in *Proc. IEEE ICASSP*, Dallas, vol. 12, pp. 2161–2164, Apr. 1987.

[Purn16] H. Purnhagen, T. Hirvonen, L. Villemoes, J. Samuelsson, and J. Klejsa, "Immersive Audio Delivery Using Joint Object Coding," in *Proc. AES 140th Convention*, Paris, no. 9587, June 2016.

[Raak12] A. Raake, M. Wältermann, U. Wüstenhagen, and B. Feiten, "How to Talk about Speech and Audio Quality with Speech and Audio People," *J. Audio Eng. Soc.*, vol. 60, no. 3, pp. 147–155, Mar. 2012.

[Rämö15] A. Rämö and H. Toukomaa, "Subjective Quality Evaluation of the 3GPP EVS Codec," in *Proc. IEEE ICASSP*, Brisbane, pp. 5157–5161, Apr. 2015.

[Ramp03] S. A. Ramprashad, "The Multimode Transform Predictive Coding Paradigm," *IEEE Trans. Speech & Audio Processing*, vol. 11, no. 2, pp. 117–129, Mar. 2003.

[Rice79] R. F. Rice, "Some Practical Universal Noiseless Coding Techniques," Jet Propulsion Laboratory, JPL publication 79-22, Pasadena, California, Mar. 1979.

[Roth83] J. H. Rothweiler, "Polyphase Quadrature Filters — A New Subband Coding Technique," in *Proc. IEEE ICASSP*, Boston, vol. 2, pp. 1280–1283, Apr. 1983.

[Rout69] E. R. Rout and A. H. Jones, "The Use of Pulse Code Modulation for Point-to-Point Music Transmission," *The Radio and Electronic Engineer*, pp. 199–207, Apr. 1969.

[Sala06] R. Salami, R. Lefebvre, A. Lakaniemi, K. Kontola, S. Bruhn, and A. Taleb, "Extended AMR-WB for High-Quality Audio on Mobile Devices," *IEEE Commun. Mag.*, vol. 44, no. 5, pp. 90–97, May 2006.

[Salo07] D. Salomon, *Variable-length Codes for Data Compression*, 1st, Springer, 2007.

[Schm16] K. Schmidt and C. Neukam, "Low Complexity Tonality Control in the Intelligent Gap Filling Tool," in *Proc. IEEE ICASSP*, Shanghai, pp. 644–648, Mar. 2016.

[Schn08] M. Schnell, M. Schmidt, M. Jander, T. Albert, R. Geiger, V. Ruoppila, P. Ekstrand, *et al.*, "MPEG-4 Enhanced Low Delay AAC — A New Standard for High Quality Communication," in *Proc. AES 125th Conv.*, San Francisco, no. 7503, Oct. 2008.

[Schn16] M. Schnell, W. Jaegers, P. Delgado, C. Benndorf, T. Albert, and M. Lutzky, "Delay Reduced Mode of MPEG-4 Enhanced Low Delay AAC (AAC-ELD)," in *Proc. AES 140th Convention*, Paris, no. 9488, June 2016.

[Schu16] F. Schuh, S. Dick, R. Füg, C. R. Helmrich, N. Rettelbach, and T. Schwegler, "Efficient Multichannel Audio Transform Coding with Low Delay and Complexity," in *Proc. AES 141st Convention*, Los Angeles, no. 9660, Oct. 2016.

[Schui04] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, "Low Complexity Parametric Stereo Coding," in *Proc. AES 116th Convention*, Berlin, no. 6073, May 2004.

[Schul00] G. D. T. Schuller and T. Karp, "Modulated Filter Banks with Arbitrary System Delay: Efficient Implementations and the Time-Varying Case," *IEEE Trans. Signal Processing*, vol. 48, no. 3, pp. 737–748, Mar. 2000.

[Schul96] D. Schulz, "Improving Audio Codecs by Noise Substitution," *J. Audio Eng. Soc.*, vol. 44, no. 7/8, pp. 593–598, July/Aug. 1996.

[ShiR14] D. Shi, H. Ruimin, W. Xiaochen, Y. Yuhong, and T. Weiping, "Expanded Three-Channel Mid/Side Coding for Three-Dimensional Multichannel Audio Systems," *EURASIP J. Audio, Speech, and Music Processing*, vol. 2014, no. 10, Mar. 2010.

[Shlie97] S. Shlien, "The Modulated Lapped Transform, Its Time-Varying Forms, and Its Applications to Audio Coding Standards," *IEEE Trans. Speech and Audio Processing*, vol. 5, no. 4, pp. 359–366, July 1997.

[Silzle09] A. Silzle, S. Geyersberger, G. Brohasga, D. Weninger, and M. Leistner, "Vision and Technique Behind the New Studios and Listening Rooms of the Fraunhofer IIS Audio Laboratory," in *Proc. AES 126th Convention*, Munich, no. 7672, May 2009.

[Sinha06] D. Sinha, A. J. S. Ferreira, and E. V. Harinarayanan, "A Novel Integrated Audio Bandwidth Extension Toolkit (ABET)," in *Proc. AES 120th Convention*, Paris, no. 6788, May 2006.

[Smar94] G. Smart and A. B. Bradley, "Filter Bank Design Based on Time Domain Aliasing Cancellation with Non-Identical Windows," in *Proc. IEEE ICASSP*, Adelaide, vol. 3, pp. 185–188, Apr. 1994.

[Smith11] J. O. Smith III, *Spectral Audio Signal Processing*, Center for Comp. Res. Music & Acoustics, Stanford University, 2015. https://ccrma.stanford.edu/~jos/sasp/

[Song10] J. Song, C.-H. Lee, H.-O Oh, and H.-G. Kang, "Harmonic Enhancement in Low Bitrate Audio Coding Using an Efficient Long-Term Predictor," *EURASIP J. Advances in Signal Processing*, vol. 2010, ID 939542, Aug. 2010.

[Song11] J. Song, H.-O. Oh, and H.-G. Kang, "Enhanced Long-Term Predictor for Unified Speech and Audio Coding," in *Proc. IEEE ICASSP*, Prague, pp. 505–508, May 2011.

[Soun12] SoundExpert, S. Smirnoff, "News & Articles: Opus, AAC and Vorbis in 64 kbit/s section," Nov. 2012. http://soundexpert.org/news/-/blogs/opus-aac-and-vorbis-in-64-kbit-s-section and http://soundexpert.org/encoders-64-kbps

[Span07] A. Spanias, T. Painter, V. Atti, *Audio Signal Processing and Coding*, Wiley, 2007.

[Sree98] T. V. Sreenivas and M. Dietz, "Vector Quantization of Scale Factors in Advanced Audio Coder (AAC)," in *Proc. IEEE ICASSP*, Seattle, pp. 3641–3644, May 1998.

[Sure09] K. Suresh and T. V. Sreenivas, "Parametric Stereo Coder with Only MDCT Domain Computations," in *Proc. IEEE Int. Symposium on Signal Processing and Information Technology (ISSPIT)*, Ajman, pp. 61–64, Dec. 2009.

[Sure12] K. Suresh and A. R. Raj, "MDCT Domain Parametric Stereo Audio Coding," in *Proc. IEEE Int. Conf. on Signal Processing and Communications (SPCOM)*, Bangalore, pp. 1–4, July 2012.

[Tam09] M. Tammi, L. Laaksonen, A. Rämö, and H. Toukomaa, "Scalable Superwideband Extension for Wideband Coding," in *Proc. IEEE ICASSP*, Taipei, pp. 161–164, April 2009.

[Teme93] M. Temerinac and B. Edler, "LINC: A Common Theory of Transform and Subband Coding," *IEEE Trans. Communic.*, vol. 41, no. 2, pp. 266–274, Feb. 1993.

[Teme95] M. Temerinac and B. Edler, "Overlapping Block Transform: Window Design, Fast Algorithm, and an Image Coding Experiment," *IEEE Trans. Communic.*, vol. 43, no. 9, pp. 2417–2425, Sep. 1995.

[Theil11] G. Theile and H. Wittek, "Principles in Surround Recordings with Height," in *Proc. AES 130th Convention*, London, no. 8403, May 2011.

[Tsuji09] K. Tsujino and K. Kikuiri, "Low-Complexity Bandwidth Extension in MDCT Domain for Low-Bitrate Speech Coding," in *Proc. IEEE ICASSP*, Taipei, pp. 4145–4148, April 2009.

[Vaill08] T. Vaillancourt, M. Jelínek, A. E. Ertan, J. Stachurski, A. Rämö, L. Laaksonen, J. Gibbs, U. Mittal, S. Bruhn, V. Grancharov, M. Oshikiri, H. Ehara, D. Zhang, F. Ma, D. Virette, and S. Ragot, "ITU-T EV-VBR: A Robust 8–32 kbit/s Scalable Coder for Error Prone Telecommunications Channels," in *Proc. EURASIP 16th EUSIPCO*, Lausanne, pp. 1–5, Aug. 2008.

[Valin10] J.-M. Valin, T. B. Terriberry, C. Montgomery, and G. Maxwell, "A High-Quality Speech and Audio Codec with Less Than 10-ms Delay," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 58–67, Jan. 2010.

[Valin13] J.-M. Valin, G. Maxwell, T. B. Terriberry, and K. Vos, "High-Quality, Low-Delay Music Coding in the Opus Codec," in *Proc. AES 135th Convention*, New York, no. 8942, Oct. 2013.

[Vand91] R. G. van der Waal and R. N. J. Veldhuis, "Subband Coding of Stereophonic Digital Audio Signals," in *Proc. IEEE ICASSP*, Toronto, pp. 3601–3604, Apr. 1991.

[VanO10] D. van Opdenbosch, "Entwurf und Implementierung einer MDCT-basierten Bandbreitenerweiterung (in German)," B. Sc. thesis, supervisors Dr. F. Nagel, C. R. Helmrich, Friedrich-Alexander Univ. of Erlangen-Nürnberg, July 2010.

[VanS08] N. H. van Schijndel, J. Bensa, M. G. Christensen, C. Colomes, B. Edler, R. Heusdens, J. Jensen, S. H. Jensen, W. B. Kleijn, V. Kot, B. Kövesi, J. Lindblom, D. Massaloux, O. A. Niamut, F. Nordén, J. H. Plasberg, R. Vafin, S. van de Par, D. Virette, and O. Wübbolt, "Adaptive RD Optimized Hybrid Sound Coding," *J. Audio Eng. Soc.*, vol. 56, no. 10, pp. 787–809, Oct. 2008.

[Vaup90] T. Vaupel, "Transform Coding with Multiple Overlapping Blocks and Time Domain Aliasing Cancellation (in German)," *Frequenz*, vol. 44, no. 11–12, pp. 291–298, Nov. 1990.

[Vaup91] T. Vaupel, "Ein Beitrag zur Transformationscodierung von Audiosignalen unter Verwendung der Methode der 'Time Domain Aliasing Cancellation (TDAC)' und einer Signalkompandierung im Zeitbereich (in German)," Ph. D. thesis, Univ. of Duisburg, Apr. 1991.

[Viret08] D. Virette, B. Kövesi, and P. Philippe, "Adaptive Time-Frequency Resolution in Modulated Transform at Reduced Delay," in *Proc. IEEE ICASSP*, Las Vegas, pp. 3781–3784, Apr. 2008.

[Wein99] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I Lossless Image Compression Algorithm: Principles and Standardization into JPEG-LS," HPL, 1999. http://www.hpl.hp.com/research/info_theory/loco/HPL-98-193R1.pdf

[Wies90] D. Wiese and G. Stoll, "Bitrate Reduction of High Quality Audio Signals by Modeling the Ears' Masking Thresholds," in *Proc. AES 89th Convention*, Los Angeles, no. 2970, Sep. 1990.

[Wolt03] M. Wolters, K. Kjörling, D. Homm, and H. Purnhagen, "A Closer Look Into MPEG-4 High Efficiency AAC," in *Proc. AES 115th Convention*, New York, no. 5871, Oct. 2003.

[Xiph15] Xiph.org Found., "Vorbis I specification," Feb. 2015. www.xiph.org/vorbis/doc

[Yama07] Yamaha Corp., "AV Receiver RX-Z11 — Owner's Manual," 2007. http://usa.yamaha.com/products/audio-visual/av-receivers-amps/rx/rx-z11_black__u/

[Yang00] D. T. Yang, H. Ai, C. Kyriakakis, and C.-C. J. Kuo, "An Inter-Channel Redundancy Removal Approach for High-Quality Multichannel Audio Compression," in *Proc. AES 109th Convention*, Los Angeles, no. 5238, Sep. 2000.

[Yang06] D. T. Yang, C. Kyriakakis, and C.-C. J. Kuo, *High-Fidelity Multichannel Audio Coding*, EURASIP Book Series on Signal Processing and Commun., Hindawi, 2006.

[YinS97] L. Yin, M. Suonio, and M. Väänänen, "A New Backward Predictor for MPEG Audio Coding," in *Proc. AES 103rd Convention*, New York, no. 4521, Sep. 1997.

[Yoko06] Y. Yokotani, R. Geiger, G. D. T. Schuller, S. Oraintara, and K. R. Rao, "Lossless

Audio Coding Using the IntMDCT and Rounding Error Shaping," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 6, pp. 2201–2211, Nov. 2006.

[Yoon06] B.-J. Yoon and H. S. Malvar, "A Practical Approach for the Design of Nonuniform Lapped Transforms," *IEEE Signal Processing Letters*, vol. 13, no. 8, pp. 469–472, Aug. 2006.

[Yoon08] B.-J. Yoon and H. S. Malvar, "Coding Overcomplete Representations of Audio Using the MCLT," in *Proc. IEEE Data Compression Conference (DCC)*, Snowbird, pp. 152–161, Mar. 2008.

[Zhan13] S. Zhang, W. Dou, and H. Yang, "MDCT Sinusoidal Analysis for Audio Signals Analysis and Processing," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1403–1414, July 2013.

[Zhon11] H. Zhong, L. Villemoes, P. Ekstrand, S. Disch, F. Nagel, S. Wilde, K.-S. Chong, and T. Norimatsu, "QMF Based Harmonic Spectral Band Replication," in *Proc. AES 131st Convention*, New York, no. 8517, Oct. 2011.

Index of Acronyms

About the Author

Christian R. Helmrich, born in 1981 in Cuxhaven, Germany, received his B. Sc. degree in computer science from Capitol Technology University (formerly Capitol College), Laurel, MD, USA, in 2005 and his M. Sc. degree in information and media technologies from Hamburg University of Technology (TUHH), Hamburg-Harburg, Germany, in 2008. Between 2008 and 2013 he worked on numerous speech and audio coding solutions at the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany, partly as a Senior Engineer. From 2013 to 2016 Mr. Helmrich continued this work as a research assistant at the International Audio Laboratories Erlangen, a joint institution of Fraunhofer IIS and Friedrich-Alexander University (FAU) of Erlangen-Nürnberg, where he completed his Dr.-Ing. doctorate degree in audio signal analysis and coding. In summer of 2016 he joined the Video Coding and Analytics (VCA) department of the Fraunhofer Heinrich Hertz Institute (HHI) in Berlin, Germany, as a video coding researcher and developer. His main research interests include audio and video coding, storage, and preservation as well as restoration from analog sources.