INVITED PAPER

Perceptual Coding of High-Quality Digital Audio

This paper discusses how human perception aspects are integrated in the design of modern audio coding systems.

By Karlheinz Brandenburg, *Fellow IEEE*, Christof Faller, Juergen Herre, *Senior Member IEEE*, James D. Johnston, *Fellow IEEE*, and W. Bastiaan Kleijn, *Fellow IEEE*

ABSTRACT | This paper introduces high-quality audio coding using psychoacoustic models. This technology is now abundant, with gadgets named after a standard (mp3 players) and the ability to play high-quality audio from literally billions of devices. The usual paradigm for these systems is based on filterbanks, followed by quantization and coding, controlled by a model of human hearing. The paper describes the basic technology, theoretical framework to apply to check for optimality, and the most prominent standards built on the basic ideas and newer work.

KEYWORDS | Audio coding

I. INTRODUCTION

Many people consider low-bit-rate, high-quality audio coding as one of the most prominent examples of disruptive technologies. Some people even claim that audio coding has been the cause of death of the music industry as we have known it. Audio coding did speed up the transformation of the music industry from selling physical media to selling songs via the Internet. Economically and as a technology, audio coding is one of the most successful digital signal processing technologies of the last decades. Decoders for mp3, Dolby Digital, Advanced Audio Coding (AAC) and other advanced low-bit-rate, high-quality audio coding formats number in many billions, built into portable media players, mobile phones, TV sets, and other types of devices. High-quality audio coding at low bit rates is only possible because we have learned how to exploit human perception and omit information that is not relevant to the listener. The magic of audio coding lies in the combination of signal processing algorithms like advanced filterbanks, quantization and coding, and consideration of knowledge about human hearing, i.e., psychoacoustics. This paper explains the basic ideas and the psychoacoustic models used for audio coding.

The first ideas about low-bit-rate coding of high-quality digital audio go back to the mid-1970s. In [1], Blauert proposed to build an audio coding system based on analog filterbanks to be able to use psychoacoustics to control the quantization noise of later stages. This was never implemented. The first actual software for low-bit-rate perceptual coding of audio probably was the critical band coder proposed by Krasner at the MIT Lincoln Labs (Cambridge, MA, USA) [2].

The topic received much more attention in the mid 1980s, when digital coding of speech signals [3] was already a well-established field of research (and even applications) and, with the compact disc (CD), digital storage of music became a commercial reality. Research on audio coding started at several places in parallel. Sometimes only a few years later the proponents, giving their first public presentations on the new technology, became aware of other very similar systems. References [4]–[10] list a number of these early proposals.

Manuscript received September 9, 2012; revised March 6, 2013 and April 29, 2013; accepted May 2, 2013. Date of publication July 11, 2013; date of current version August 16, 2013.

K. Brandenburg is with the Fraunhofer-Institut für Digitale Medientechnologie (Fraunhofer IDMT), 98693 Ilmenau, Germany and also with the Ilmenau Technical University, 98693 Ilmenau, Germany.
C. Faller is with Illusonic GmbH, Uster 8610, Switzerland and also with the École Polytechnique Fédérale de Lausanne (EPFL), Lausanne 1015, Switzerland.
J. Herre is with International Audio Laboratories Erlangen, 91058 Erlangen, Germany.
J. D. Johnston is retired.
W. B. Kleijn is with the School of Engineering and Computers Science, Victoria University of Wellington, Wellington 6140, New Zealand and also with the Department of Intelligent Systems, Delft University of Technology, Delft 2628, The Netherlands.

Digital Object Identifier: 10.1109/JPROC.2013.2263371

Triggered by both, textbooks on psychoacoustics [11]–[14] and early papers explaining the usage of masking for improving speech coding (see, e.g., [15]), the notion of psychoacoustics-based coding of high-quality audio became a hot topic.

Several more general ideas triggered even more research and helped to establish research consortia with the necessary resources to further advance the art of high-quality audio coding.

- In 1987, the European Digital Audio Broadcasting (DAB) project needed a low-bit-rate representation of high-quality audio to pack enough channels of audio into its digital transmission system.
- In 1988, the Moving Picture Experts Group (MPEG) of the International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) issued a call for audio coding algorithms to accompany the planned video coding in a way that movies would fit on CD-ROM.
- Soon afterwards, the Advanced Television Standards Committee in the United States was looking for the sound coding part of the future high-definition (HD) and standard-definition (SD) digital TV system for the United States.

The audio codecs developed for these activities form the first generation of widely used low-bit-rate audio coding. They are MPEG-Audio Layer II [16], MPEG-Audio Layer-3 (now called mp3) [16]–[18], and Dolby AC-3 (better known as Dolby Digital) [19], [20]. From 1992 on, these systems went from standardization to wide spread adoption.

This paper explains the basics of high-quality audio coding in an abbreviated way. It introduces some simple models of human hearing, explains ideas about their usage for audio coding, and then introduces some audio coding systems according to the timeline of their introduction. This includes the original MPEG–Audio standards and newer work, all covering the last 20 years. For more in-depth coverage, the reader can find overviews in [21]–[25].

II. PSYCHOACOUSTICS

Unlike in speech coding, the knowledge about the human auditory system (HAS) is clearly a decisive factor needed to build audio coders capable of delivering high subjective quality for all possible input signals, including all kinds of music and speech signals. There is no low-bit-rate high-quality audio coding without build-in models of human hearing.

A. Auditory Masking

The performance of the HAS has been studied for many years. There are still many open questions regarding the mechanism of hearing, biology, and cognition, especially regarding spatial hearing. We do have a clear understanding of some ways the human hearing functions.

- The HAS starts in the cochlea with a highly redundant time/frequency analysis.

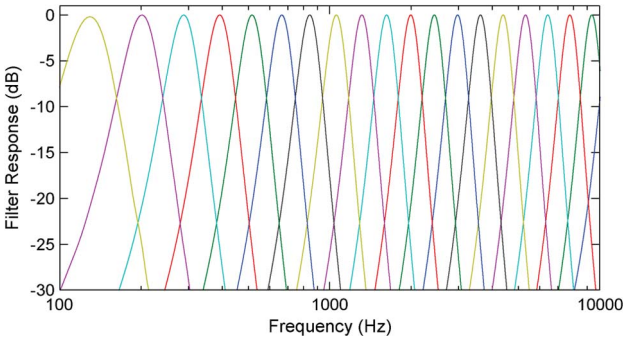


Fig. 1. ***Modeled auditory filterbank frequency response with band center frequencies arranged on an ERB scale.***

- There is an absolute threshold of hearing that varies with frequency.
- The inner hair cells, often viewed as detectors, have about a 30-dB signal-to-noise ratio (SNR).
- Outer hair cells via changing stiffness help to provide a much larger dynamic range of the HAS.
- In the presence of one sound, some other sounds become inaudible, denoted masking.
- Masking relates to the time/frequency resolution of the cochlea.

In any discussion of the HAS, the time/frequency analysis is a primary consideration. Several scales of associated filter bandwidths exist, including the "Bark" scale [11] and the equivalent rectangular bandwidth (ERB) scale [26]–[28]. A good estimate of the bandwidth of the auditory filter is 70-Hz bandwidth starting at 20 Hz, and converting to about one quarter octave when that exceeds 70 Hz.

Fig. 1 shows an example of modeled auditory filters [29] with center frequencies following an ERB scale. These filters implement a model of the level response as a function of space (position on the cochlea) [30] before the corresponding signal reaches the hair cells. In addition to the aforementioned bandwidth increase with frequency, the auditory filters are not symmetric in frequency, i.e., as a function of space, they decay more slowly toward higher than toward lower frequencies. For audio coding, this implies that quantization noise at lower frequencies compared to the masker is more audible than quantization noise at higher frequencies.

In addition to simultaneous masking, also temporal masking must be considered (i.e., when a masker and a masked signal do not appear simultaneously). Premasking time is in the 5 to under 1-ms range; postmasking for loud signals persists strongly for at least 10 or so ms, and has some effect out to 100–200 ms. A filterbank impulse response much larger than the premasking time may lead to the pre-echo artifacts described above. Premasking and postmasking are illustrated in Fig. 2.

Many researchers originally reached the obvious conclusion to split an audio signal into bands corresponding

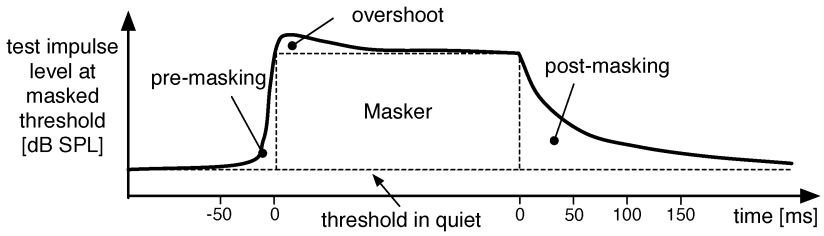


Fig. 2. *Basic idea of premasking and postmasking for a switched masker. Masking extends to before and after the actual masker duration. At the onset, there is even an increased amount of masking relative to the steady-state case (overshoot).*

to those of human hearing to obtain good audio coding performance; see, e.g., [31]. In this system, removal of irrelevancy (by considering masking) is assured, but removal of signal redundancy is not very effective. Because of this view (see Section III), usually filterbanks with higher frequency resolution and uniform bands are used, as illustrated in the top panel of Fig. 3. Masking is considered in groups of bands, following an ERB scale, illustrated in the bottom panel of Fig. 3.

Note that the absolute threshold of hearing, shown in Fig. 4, is usually not directly considered in audio coding. The problem is that audio reproduction systems have a volume control, and any considerations on absolute level will be made obsolete by the user changing listening level. Psychoacoustic models used in commercial encoders just set some very low threshold near the quantization noise of the input pulse code modulation (PCM) signal.

Aside from the time-domain aspects of pre-echo, there is another asymmetry to consider, that of tone–masking–noise versus that of noise–masking–tone, as described in an early paper by Hellman [32]. This asymmetry originates beyond the cochlea, in the central nervous system. A tone may require an SNR (inside of an ERB) close to 30 dB to sound the same as the original. Narrowband noise masking a tone may happen at 5.5 dB. In modern practice, the models used in audio coding assume, e.g., tone–masking–noise in an ERB to be circa 30 dB, and noise–masking–tone (or better call it noise–masking–noise) about 6 dB, for monaural signals.

In summary, a suitable masking model has to consider:

- pre-echo;
- ERB-scale frequency analysis;
- auditory filter shape;
- tone–masking–noise;
- noise–masking–noise.

B. Stereo Coding Issues

Stereo coding presents both additional problems and additional possibilities, for further bit-rate reduction. For example, in early audio coding work, we coded a nearly monophonic song, "Tom's Diner" by Suzanne Vega, in good audio quality using a mono audio coder. The same song coded at twice the bit rate with a stereo coder based on two mono coders yielded much lower quality. The signal is not completely monaural. The two independent mono coders created an output signal with uncorrelated noise on the left and right channels. The localization of uncorrelated noise is wide and not focused in the middle with the voice of Suzanne Vega singing. This noise is audible even if it is masked in a single-channel coder. This problem results from the ability of the auditory system to process time-domain

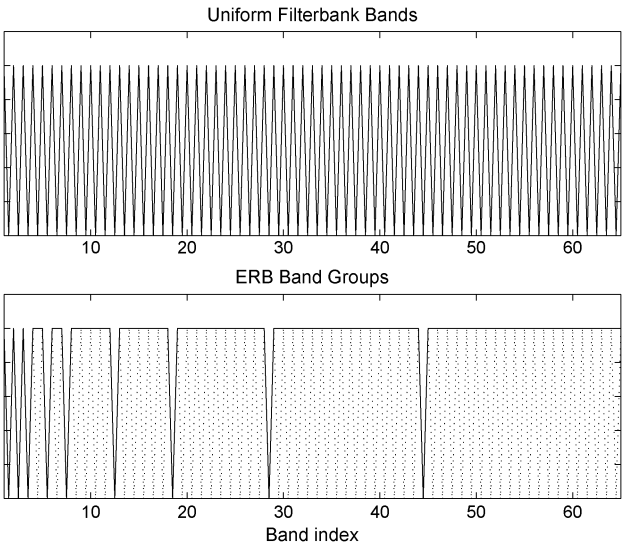


Fig. 3. *Bands of a uniform filterbank (top) and groups of bands following an ERB scale (bottom).*

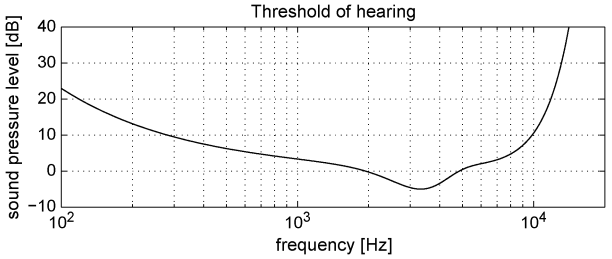


Fig. 4. *Absolute threshold of hearing.*

cues in a binaural signal, in particular, by localizing different parts of the signal at different positions. This leads to a lowering of the masked threshold, known as the binaural masking level difference (BMLD) [14].

Mid/side (M/S) stereo coding [33], also often referred to as sum-difference coding, addresses the mentioned issues. M/S coding switches between coding the left and right channels directly, and coding of the left–right sum and difference signals, whatever is more efficient. Other methods (i.e., intensity stereo coding [34]) fall under the general principles of spatial audio coding, as described in Section VI-C.

C. Further Work Regarding Psychoacoustic Models in Audio Coding

In recent years, the work regarding psychoacoustic models for use in audio coding went in two opposite directions.

On the one hand, there was work to use more detailed models to enhance the quality of audio coding systems, especially at lower bit rates. One example of such work is the Ph.D. dissertation of Baumgarte [35], who introduced nonlinear functions into the estimation of actual masked thresholds.

On the other hand, for commercially used encoders, fast computation became more important than the last little detail in the perceptual model. Some commercial systems today use psychoacoustic models more looking like the early work reported in [7].

The use of spatial cues (see Section VI) is another field where more accurate psychoacoustic models are needed. Currently, we are learning a lot about the interaction between hearing and cognition leading to auditory illusion.

III. TIME/FREQUENCY-BASED AUDIO CODING

Perceptual audio coding systems generally use a similar structure, which is shown in the block diagram in Fig. 5. The basic processing blocks are as follows.

- Analysis filterbank: A filterbank is used to decompose the input signal into a time sequence of vectors with as many elements as spectral components. Together with the corresponding filterbank in the decoder it forms an analysis/synthesis system.
- Perceptual model: Using either the time-domain input signal and/or the output of the analysis filterbank, an estimate of the actual (time- and frequency-dependent) masking threshold is computed using rules derived from psychoacoustics.
- Quantization and coding: To facilitate transmission, the signal must be represented with finite precision; it must be quantized. The general aim is to keep the resulting error signal (usually referred to as quantization noise) below the masking threshold.
- Bit stream multiplex: The bit stream typically consists of the quantized and coded spectral coefficients and some side information, such as bit allocation or information about the quantizers.

Before exploring the filterbank-based structure further, we note that coding applications that require short delay, such as mobile communications, are generally based on prediction, e.g., [36]–[38]. In contrast to filterbank-based architectures, prediction-based architectures can provide a significant coding gain, even at zero delay. Recent results [39] provide insight in the performance bounds on predictive coding.

A. Filterbanks

The objective of audio coding is to reduce the rate by not coding information multiple times, and by not coding what you cannot hear. That is, the aim is to remove redundancy and irrelevancy. The filterbank used in audio coding facilitates both aims [40].

The analysis filterbank enables subsequent processing to affect just a certain range of frequencies. Note that the filter responses have a finite length. Each coefficient of the resulting representation contributes to the description of a finite time/frequency region. This makes it straightforward

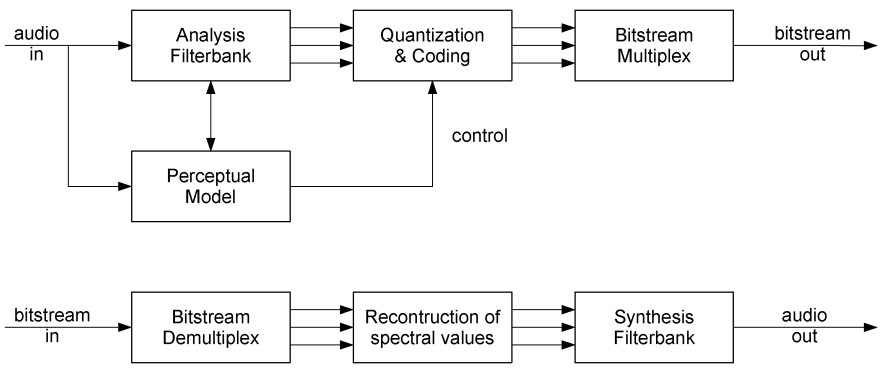


Fig. 5. *Block diagram of a perceptual encoding/decoding system.*

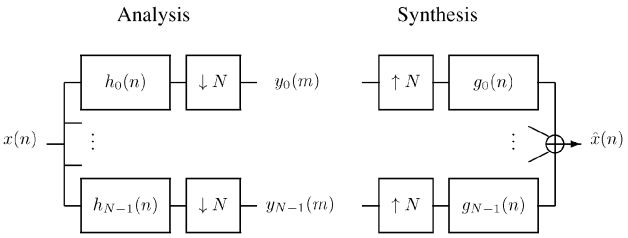


Fig. 6. *An N-band filterbank with critical sampling. Perfect reconstruction means that $\hat{x}(n) = x(n - n_d)$, where n_d is the system delay. $\downarrow N$ means downsampling by N, meaning only every Nth sample is kept, hence reducing the sampling rate accordingly. $\uparrow N$ means upsampling by N to obtain the original sampling rate. This is done by inserting $N - 1$ zeros after each sample.*

to base further processing on the psychoacoustic model by using masking-dependent quantizers. The result is that (most) irrelevant information is not coded.

The filterbank also facilitates coding because it removes redundancy. In general, a vector can be decorrelated using the Karhunen–Loève transform (KLT). Such decorrelation also manifests itself as energy compaction in the coefficients. Decorrelation generally implies that the coefficients are more independent, and independence means that the scalar quantizers operating on the different coefficient do not encode the same information multiple times. Thus, Huang and Schultheiss, in 1963, proposed to perform KLT-based decorrelation on a block-by-block basis for the purpose of coding [41]. The computation of the KLT can be avoided if it is realized that, for a segment of a stationary signal, the KLT approaches a transform to the frequency domain with increasing segment length, e.g., [42] and [43]. That is, a Fourier transform can be used to decorrelate the signal samples. This operation can be approximated with a filterbank. Indeed, for most real-world audio signals, it can be shown that an increase in frequency resolution leads to more energy compaction, which corresponds to lower redundancy and, therefore, improved coding efficiency, if a scalar quantizer is used. (Naturally, this only holds if the resolution is not increased beyond the point where the stationarity assumption breaks down.)

Filterbanks can be interpreted both from a filtering viewpoint or a transform viewpoint. Historically, these views led to two separate starting points for audio coding systems: subband coders and transform coders. It is simple to show the equivalence of both concepts. Consider the critically sampled filterbank shown in the block diagram of Fig. 6. By time reversing the impulse responses h_i and g_i (and applying suitable time displacements), we obtain the corresponding basis functions of the forward and backward transforms. If the transform is orthonormal, the analysis and synthesis responses are identical. The equivalence to transforms also exists for oversampled filterbanks, where the transforms can be based on frame theory, e.g., [44] and [45]. For a simple transform that can be decomposed in a fixed window and modulation functions, the window corresponds to the prototype window of an equal bandwidth filterbank [46].

As discussed above, the KLT for a stationary signal corresponds to a frequency transform. For audio coding, we would like to use a sequence of overlapping windowed discrete Fourier transforms (DFTs), arranged to provide perfect reconstruction. The smoothness of the windows would reduce aliasing (which affects the spread of quantization errors) in a practical application. However, the Balian–Low theorem [45] shows that a perfect-reconstruction filterbank with smooth windowing based on the DFT requires oversampling: the number of transform coefficients is larger than the number of time samples. Obviously, oversampling is undesirable for audio coding as it reintroduces a redundancy problem. Surprisingly, the Balian–Low theorem does not carry over to the discrete cosine transform (DCT): critically sampled, smoothly windowed, perfect reconstruction filterbanks are possible with the DCT. As a result, the so-called modified discrete cosine transform (MDCT) has become a standard component in audio coding. As it is a frequency transform, it displays good energy compaction behavior (effective decorrelation) and can be implemented efficiently. The MDCT is known as well as cosine modulated filterbank or lapped orthogonal filterbank [46]–[51]. The MDCT was originally derived by considering aliasing explicitly. In an MDCT, the aliasing components introduced in analysis cancel each other during synthesis. This phenomenon is known as time domain aliasing cancellation (TDAC) [47]. For perfect reconstruction, the windows of the MDCT must be power complementary. The most common choices are cosine window and the so-called Kaiser–Bessel-derived (KBD) window [19].

A practical example for the analysis/synthesis filterbank in audio coding is the one used in MPEG AAC [52]. The prototype filter (analysis window) is 2048 samples in length and the number of transform frequencies is 1024. This corresponds to an analysis frame of 21.3 ms at 48-kHz sampling frequency. The window may straddle a transition in the signal, which means the stationarity condition is violated, reducing coding efficiency. More importantly, masking is not constant in time in such signal segments: as discussed in Section II, 20 ms is well beyond the so-called premasking time. Note that any quantization error is spread in time over a region of duration equal to the length of g_i in Fig. 5. That is, the duration of the error is equal to the window length. This leads to so-called pre-echo artifacts. Fig. 7 shows a well-known example of this effect: First, there is the time-domain plot of hitting a castanet. The next two plots show the effect of the spread of the quantization error over time, following the synthesis window of an audio coder (in this example, mp3). In the first output signal, the spread of the quantization noise extends well beyond the limit of premasking; see Section II-A. These errors are often clearly audible. If

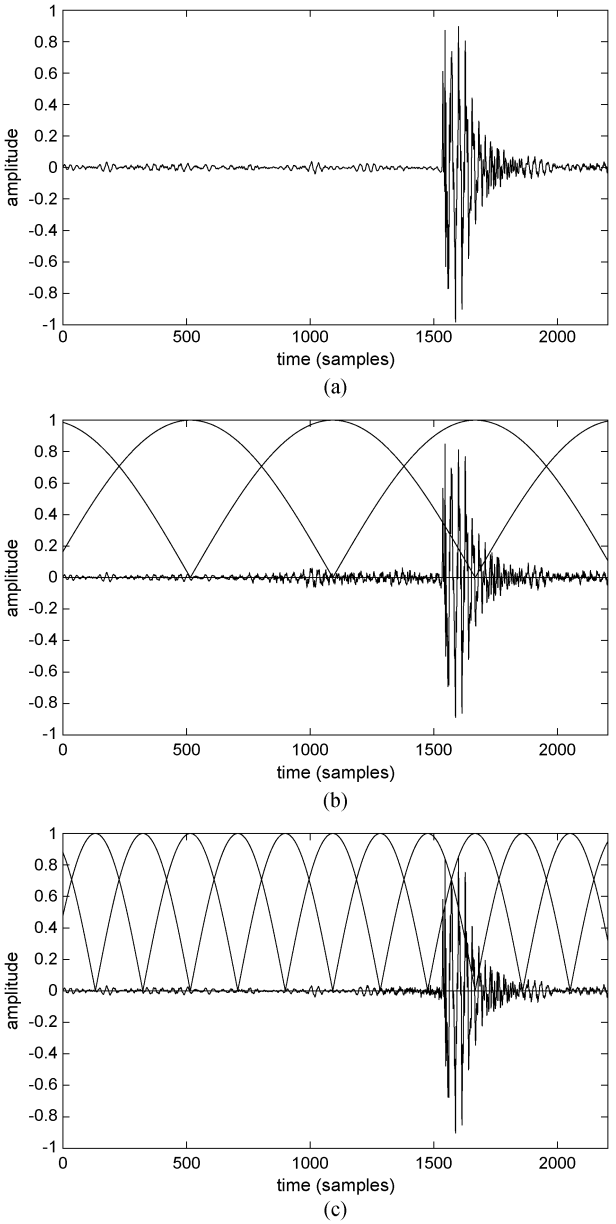


Fig. 7. *Example of a pre-echo: (a) Castanet attack, time-domain signal; (b) castanet coded with the synthesis windows shown, long blocks; and (c) castanet coded with the synthesis windows shown, short blocks.*

we use shorter windows, we can keep the noise within the bounds of premasking. To avoid pre-echo problems, window switching is used to change the time/frequency resolution properties of the filterbank from one time segment to the next. Fig. 8 shows a typical sequence of window switching [8], [53], [54] as used in AAC.

B. Quantization and Coding

The quantization and coding of the quantized values constitutes the main data reduction step in a perceptual

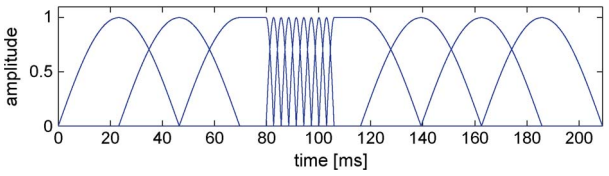


Fig. 8. *Block switching example for AAC: long blocks, start block, 8 short blocks, stop block, and long blocks.*

audio coder. Essentially all audio coders use scalar quantization to have a low level of computational complexity. In general, quantization can be performed under a fixed-rate constraint (resolution-constrained quantization) or under an average-rate constraint (entropy-constrained quantization). In the case of constrained-resolution coding, the quantization index itself is the code that is transmitted, whereas in the case of constrained-entropy coding, the index is encoded with an entropy coder. For both cases, the quantizer can be optimized either using high-rate theory or by using a Lloyd algorithm, e.g., [55] and [56].

In the case of entropy-constrained quantization, an entropy coder is needed to remove the redundancy from the quantization indices. Entropy coding can be performed with arithmetic coding [57] (or, equivalently, a range coder [58]), which can essentially reach optimal performance since it can code any number of indices together. An alternative is Huffman coding [59], which encodes all indices separately (resulting in lower performance), but generally is easier to implement and of lower computational complexity.

Fixed rate transmission is often required in practical applications. It is possible to obtain fixed rate even if a variable rate coder is used if delay is allowed (e.g., in broadcast applications). In this case, buffer control can be used to get a fixed rate across the transmission channel.

In practical audio coders, two methods for quantization and coding are commonly used.

- Block companding/block floating point: The values to be encoded are first normalized. The maximum of the absolute values serves as a factor to scale all the values (called scale factor), e.g., to a maximum of 1. The quantizer divides the interval into a uniform set of intervals and describes each interval with a label. The rate allocation (step size) of the frequency components is set by a bit allocation algorithm driven by the psychoacoustic model. In general, the quantization indices do not have a uniform probability distribution. However, the entropy coder is commonly omitted.
- Nonuniform quantization combined with Huffman coding: With nonuniform quantization, similar to the log-PCM of G.711 speech coding [60], it is possible to have the quantization error increase

with the magnitude of the spectral sample. Huffman coding of the variables is commonly used to reduce the bit rate.

IV. EXAMPLES OF STANDARDIZED CODECS

With the explanations above, it is easy to characterize some of the most used perceptual audio coding systems.

- *MPEG Layers I and II:* These codecs use a 32-band polyphase filterbank. The frame length is 384 (Layer I), respectively 1152 (Layer II), samples. All 12 samples in time-domain direction (succeeding outputs of the filterbank at a given frequency) are coded using block companding. A perceptual model driven from the output of a separate DFT-based filterbank (to get higher frequency resolution) together with the scale factors drive a bit allocation algorithm. MPEG Audio Layers I and II support two-channel stereo with intensity stereo coding. MPEG Audio is defined in [16] and [61]; see the original paper in [17] and [62].
- *MPEG Audio Layer-3 (mp3):* The filterbank consists of the polyphase filterbank, as used in Layers I and II, cascaded with an 18- or (at short windows mode) 6-band MDCT with optional window switching. Quantization is done with a power law (quantization step size increases with $x^{0.75}$). The decoder reconstructs the values according to $v = i^{4/3} * s$, where v is the reconstructed value, i is the index which was transmitted, and s is a scale factor. The scale factors modify the quantization step size according to the output of a psychoacoustic model. Actual encoding is done using (limited) adaptive Huffman coding. The tables are not generated on the fly, but there are a number of predefined tables. One solution is selected for a locally minimum bit rate. The quantization and coding is usually done using an iterative procedure which modifies the scale factors using the analysis by synthesis. MPEG Audio Layer-3 supports two-channel stereo with either full-band meter per second switching or intensity stereo.
- *AC-3:* The default audio coder for DVD and digital TV (in many countries) is also known as Dolby Digital. The filterbank uses some modification of the MDCT, as described above, with different window switching and 256 frequency lines. Quantization and coding is done by the block floating point. Different frequency lines (number adapted) go into the block companding step. One specialty of AC-3 is the possible combination of backwards adaptive bit allocation (just calculated from the exponents of the block floating point representation), needing no additional side information for the bit allocation information, and forward adaptive bit allocation. AC-3 supports a number of stereo modes, including 5.1 channel stereo.
- *AAC:* MPEG-2 AAC [63], [64] is a successor of mp3 that incorporates several improvements, such as, e.g., multichannel capability, and is used widely for, e.g., music distribution over the Internet and portable music playback. It follows the basic structure of mp3 (high-resolution filterbank, however with a 1028/128 spectral lines MDCT, nonuniform quantization, and entropy coding) with several enhancements regarding handling of transient and tonal signals, joint coding of several channels, and general efficiency. AAC supports a number of stereo modes, including 5.1 channel stereo using frequency-band-based meter per second coding [19], [20].

The bit stream of these audio coders contains all the necessary data to allow decoding just from the audio bit stream. Usually, it is organized into frames of, e.g., 1152 samples for MPEG Audio Layers II or III or 1024 samples for MPEG Audio AAC. Each frame contains header information for standalone decoding (without additional data), actual coded audio, and some provisioning for nonaudio-related metadata.

V. SOME NEWLY STANDARDIZED AUDIO CODECS SINCE ABOUT 2000

Starting around the year 2000, many new developments started which extended the traditional architecture in novel ways and, e.g., left the paradigm of waveform coding, i.e., perfectly reconstructing the input signal if infinite bit rate was available.

While there have been many more successful codec developments [like sinusoidal coding or coders developed within the International Telecommunication Union—Telecommunication Standardization Sector (ITU-T)], this paper takes the coders developed by the ISO/MPEG standardization group as examples to illustrate the more recent evolution of perceptual audio coding.

A. Toward Lower Bit Rates: The AAC Codec Family

Starting from the MPEG-2 AAC [64], a number of extensions were developed subsequently which established a family of backwards compatible coders.

The first key idea extending the traditional audio coder concept was bandwidth extension which provided a substantial gain in audio quality for coding at low bit rates (say, below 48 kb/s per channel). Up to this point, all audio coders that used scalar quantizers showed the tendency to introduce annoying artifacts if run at very low bit rates. To avoid these artifacts, frequently, a low-pass filter was introduced, leading to a very perceptible loss of high frequencies. As explained in more detail below, bandwidth extension processors bring back the high-frequency (HF)

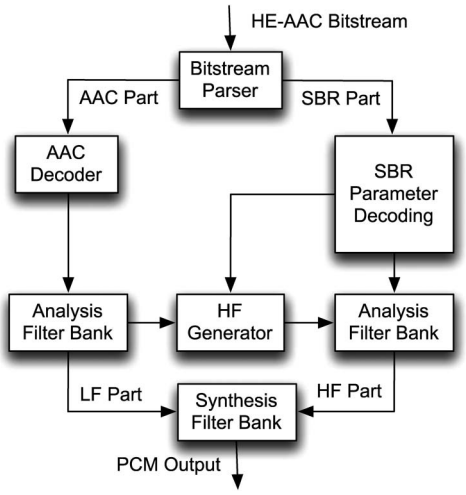


Fig. 9. ***The HE–AAC decoding process.***

content by resynthesizing it from the transmitted low-frequency (LF) part. The characteristics of the original HF part (spectral envelope and other aspects) are extracted by the encoder and sent as very compact side information (few kilobits per second) to the decoder along with the main LF bitstream.

While schemes for bandwidth extension have been around for some time for speech signals [65], [66], the first significant scheme for bandwidth extension in a general audio coder was the spectral band replication (SBR) scheme introduced around 2000 [67]. The combination of SBR with the MPEG-2/4 AAC coder is called high-efficiency AAC (HE–AAC) [68], [69] and allows good quality stereo at bit rates as low as 48 kb/s. Fig. 9 shows the basic signal flow of HE–AAC decoding. The bit stream is decomposed into an AAC part and an SBR part. The former is decoded to PCM by an AAC decoder and fed into an analysis filter bank, forming the LF part of the final output signal. From this, the HF generator generates the HF part of the output, which is subsequently adjusted in its time/frequency envelope and other characteristics, based on the transmitted SBR parameters. Both LF and HF parts are put together and converted back to time domain by a synthesis filterbank. While the operation for SBR HF generation is just a simple translation (copy-up) operation, more recent schemes for bandwidth extension also make use of other more elaborated techniques; see, e.g., [70].

As a second substantial step ahead, the HE–AAC codec was subsequently combined with a so-called parametric stereo [71] tool that allows efficient parametric coding of stereo signals at very low bit rates. The principle is that, instead of transmitting two discrete audio channels, a single (mono) audio channel is encoded together with some compact spatial side information; see Section VI-C for a general explanation of the underlying concept of spatial audio coding. The combination of HE–AAC and parametric stereo emerged in 2004 and is known as HE–AAC v2 [68].

A third important type of progress can be seen in the recent convergence between the worlds of perceptual audio coding and speech coding. While the former did not make use of specific source models and focused on exploiting the receiver characteristics (auditory perception), the latter extensively relies on models of human speech production and some basic perceptual weighting. As a consequence, when looking at very low bit rates, perceptual audio coders exhibit better performance for music signals whereas speech coders have a distinct advantage for speech. In a quest to establish a truly universal coder that "performs comparable to or better than the best coding technology that might be tailored specifically to coding of either speech or general audio content," both technology approaches were integrated in the so-called MPEG Universal Speech and Audio Coding (USAC) scheme [70]. The architecture features an enhanced bandwidth extension module, an improved parametric stereo scheme [72], and a coding core which uses one of two coding modes alternatively on a frame-by-frame basis. As a first mode, an enhanced AAC-type filterbank-based audio coder is available for coding of general audio signals, while an algebraic code-excited linear prediction (ACELP)-type coding mode provides optimum performance for speech signals. Both coding modes have been integrated to share some processing functions, such as the LPC-based noise shaping. Published in early 2012, the integrated USAC coder [73] indeed consistently outperforms [74] both HE–AAC v2 and AMR–WB+ [75].

To summarize, Fig. 10 illustrates the overall structure of a modern audio codec with both stereo and HF extension pre/postprocessing wrapped around the actual core coders.

B. Audio Coding for High-Quality Telecommunication

In parallel to the evolution of "regular" perceptual audio coding (as was discussed previously for the example of the AAC codec family), the same concepts were also applied to the problem of communication coding, i.e., encoding/decoding with a low overall delay which forms an essential requirement for two-way communication,

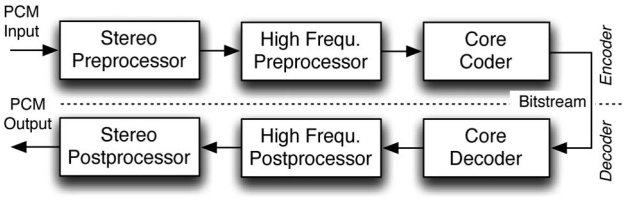


Fig. 10. ***Structure of modern audio codecs, including stereo and HF processing.***

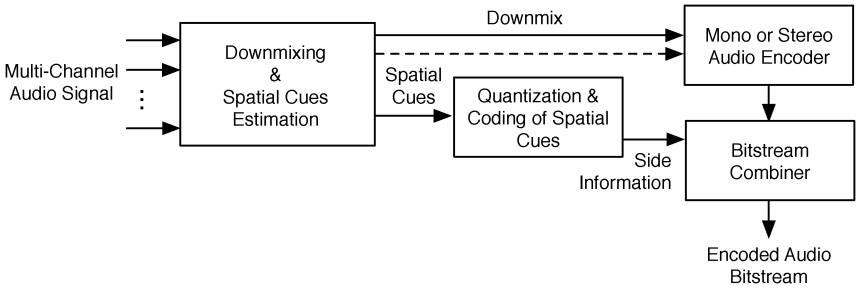


Fig. 11. *SAC encoder: Spatial cues between the input audio channels are estimated and encoded, the input signal is downmixed and encoded with a conventional audio coder, and the encoded spatial cues and downmix are combined to a bit stream.*

such as telephony, A/V teleconferencing, or telepresence. While usual perceptual audio coders are able to deliver good compression and high audio quality (full audio bandwidth, low perceived distortion) for a wide range of input signals, the delay inherent in their algorithms substantially exceeds the desired target of less than 30 ms.

The first standardized generic low delay audio coder was the low delay AAC (AAC–LD) codec which was derived from AAC by using a shorter transform size, avoiding the use of the bit reservoir and replacing adaptive window switching by a technique which does not require look-ahead and, thus, does not incur additional delay in the encoder [76]. The codec was standardized by MPEG in 2000, has a minimum algorithmic delay of 20 ms, and performs comparable to regular, not delay-constrained, AAC with an average penalty in coding efficiency of ca. 8 kb/s per channel. AAC–LD has become widely deployed as part of many hardware or desktop-computer-based systems for videoconferencing.

In order to make the benefits of bandwidth extension available for communication coding, a low delay version of spectral band replication (SBR–LD) was developed and combined with an enhanced version of AAC–LD [77]. Both for the core audio coder and for the low delay SBR scheme, dedicated novel low delay filter banks were designed that reduce the delay while keeping a good frequency resolution. The resulting coder was standardized as enhanced low delay AAC (AAC–ELD) in 2008 and provides a marked increase in compression performance at low bit rates, owing to the additional use of bandwidth extension.

C. Toward Highest Quality: (Near) Lossless Audio Coding

While the main thrust of perceptual audio coding research certainly has focused on enhancing subjective quality at very low bit rates, also the opposite direction has been pursued, i.e., coding at very high quality, which is desired for archival applications and studio and production purposes. A number of codecs have been developed for coding audio at "better than perceptually transparent" quality (i.e., with some headroom toward audibility), in a near-lossless or fully lossless fashion, where the latter exclusively aims at reducing the redundancy present in the encoded audio signals [78]–[81]. Some of these also support coding of high-resolution audio signals, such as 96-kHz/24-b content.

VI. SPATIAL AUDIO CODING

A. SAC Encoding and Decoding

Spatial audio coding (SAC) enables higher compression ratios for two-channel stereo and multichannel surround audio signals. This is achieved by waveform coding only a downmix of the input audio signal. The downmix contains all signal components (disregarding spatial aspects) which are present in the original two-channel or multichannel audio signal. In addition, parameters describing "perceptually relevant differences" (in terms of spatial hearing) between the original audio channels are estimated. These parameters, denoted spatial cues, contain about two orders of magnitude less information than the waveforms themselves. Thus, the bit rate is significantly reduced by transmitting them as opposed to transmitting all audio channels. In the decoder, the downmix is processed such that the spatial cues between the synthesized channels approximate those of the original audio channels.

SAC-based audio coders initially used a mono downmix [71], [82]–[90]. Later versions, such as MP3 Surround [91], [92] and MPEG Surround [93]–[95] use also a stereo downmix, enabling stereo backwards compatibility and higher audio quality. Fig. 11 shows a block diagram of a SAC encoder. A corresponding SAC decoder is illustrated in Fig. 12.

B. Spatial Hearing and Cues

Similarly to the way humans perceive a visual image, humans are also able to perceive an auditory spatial image. The different objects which are part of the auditory spatial image are denoted auditory events. When two-channel or multichannel audio signals are played back over headphones or loudspeakers, they evoke an auditory spatial

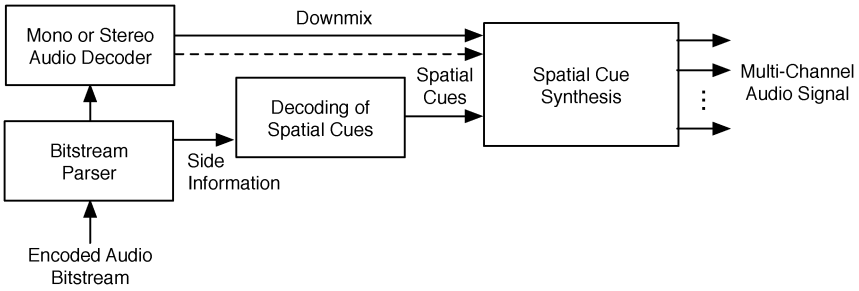


Fig. 12. *SAC decoder: The bitstream is separated into bitstreams for the encoded spatial cues and downmix. The downmix is decoded with a conventional audio coder and the output signal is synthesized from the downmix using the decoded spatial cues.*

image in the listener. To a large degree, the perceived auditory spatial image relates to the binaural cues [14] between the signals at the left and right ear entrances: interaural level difference (ILD), interaural time difference (ITD), and interaural coherence (IC).

Fig. 13 illustrates perceptions of a listener when exposed to left and right ear signals with different binaural cues. ILD and ITD relate to the location of the perceived auditory object, while IC relates to the size of the object.

When the same signals are played back over a stereo loudspeaker system, similar perceptions are evoked in the listener, but this time the localization is not in his head but between the two loudspeakers, as illustrated in Fig. 14. Because of this similarity in perception, the same signals (e.g., from a CD) are suitable for headphone and loudspeaker playback. For loudspeaker playback, the interchannel cues between the channels are related to, but not the same as the cues between the left and right ear entrance signals. Thus, in the following, we use the term interchannel cues as opposed to binaural cues. A thorough discussion about the relation between interchannel cues and auditory spatial image perception can be found in [94].

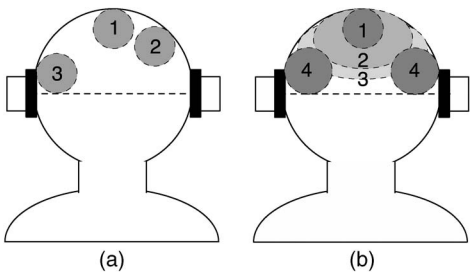


Fig. 13. *(a) ILD and ITD between a pair of headphone signals determine the location of the auditory event which appears in the frontal section of the upper head. (b) The width of the auditory event increases (1–3) as the IC between the left and right headphone signals decreases, until two distinct auditory events appear at the sides (4).*

C. Spatial Synthesis

First, stereo synthesis from a mono downmix is considered, illustrated in Fig. 15. The downmix is converted by a filterbank to short-time spectra. A decorrelation filter is applied to obtain spectra of two independent audio signals (the downmix and the decorrelated signal). For each frequency band, the mixing gains a_1, a_2, b_1, and b_2 are computed such that the mixed two channels have an interchannel level difference (ICLDs) and interchannel coherence (ICC), as specified by the corresponding ICLD and ICC from the spatial cues bit stream. The so-obtained spectra of the left and right channels are converted to the time domain, using an inverse filterbank.

Not to use ICTDs here is a simplification compared to the theoretically correct solution which has been chosen for MPEG surround and, in this system, seems to work well enough. This is especially true for mono-compatible (intensity only) two-channel stereo mixes.

Similar principles are applied for multichannel synthesis, where the interchannel cues are defined between different channel pairs [91] or two-channel synthesis blocks are cascaded [94]. MPEG Surround spatial synthesis, illustrated in Fig. 16, applies the latter principle. Additionally, the center channel is synthesized by means of prediction. Spatial audio coding principles have also

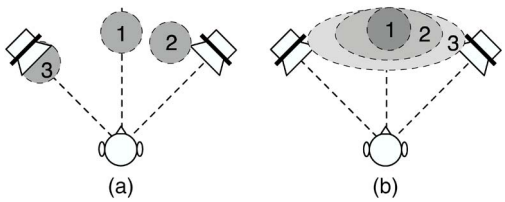


Fig. 14. *(a) ICTDs and ICLDs between a pair of coherent source signals determine the location of the auditory event which appears between the two sources. (b) The width of the auditory event increases (1–3) as the ICC between left and right source signals decreases.*

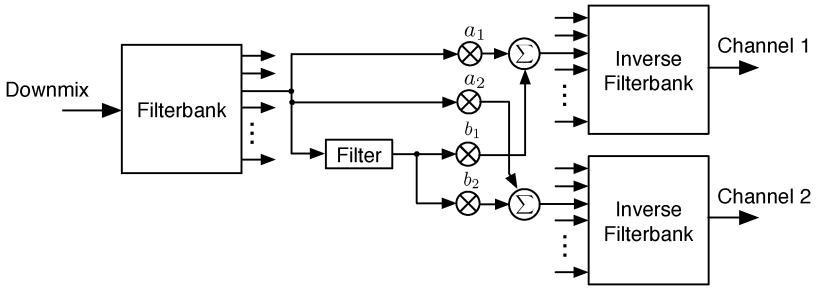


Fig. 15. *Synthesis of a stereo signal from the mono downmix.*

been applied for object-based coding with mixing or remixing capability at the decoder [82], [96]–[98].

VII. CONCLUSION

Perceptual coding of high-quality audio has gone a long way since the first papers appeared, showing the possibility of substantial bit-rate reduction for near-CD quality audio. While most of the applications today are based on the standards devised in the 1990s, there is still active research leading to new standards. Over the last 15 years, the inclusion of coding tools based on parametric techniques, like SBR or spatial-cue-based coding of stereo (including multichannel) signals has considerably advanced the possibilities of coding at low bit rates. While these systems are in widespread use, other, even more ambitious ideas like fully parametric coding, have gained much less acceptance in the market. If we look at the current usage of high-quality audio coding, we can even find a tendency to higher bit rates for applications like storage on portable media or Internet streaming. At the same time, the psychoacoustic models implemented into consumer devices capable of encoding (not just play back) got simpler instead of more advanced and more accurate. Research in the last years and currently is focusing on two major areas: For some applications, even lower bit rates are necessary while a high quality needs to be maintained for both speech and audio signals. As one example, USAC addresses this area of applications. The second area of research is the goal to reach better acoustic illusion for audio playback. While this research on upmix techniques, object-based audio storage and advanced playback algorithms does not fall into the scope of this paper, clearly some of the very same ideas from psychoacoustics apply.

The first author has to acknowledge that current systems have advanced the state of the art to a point he did

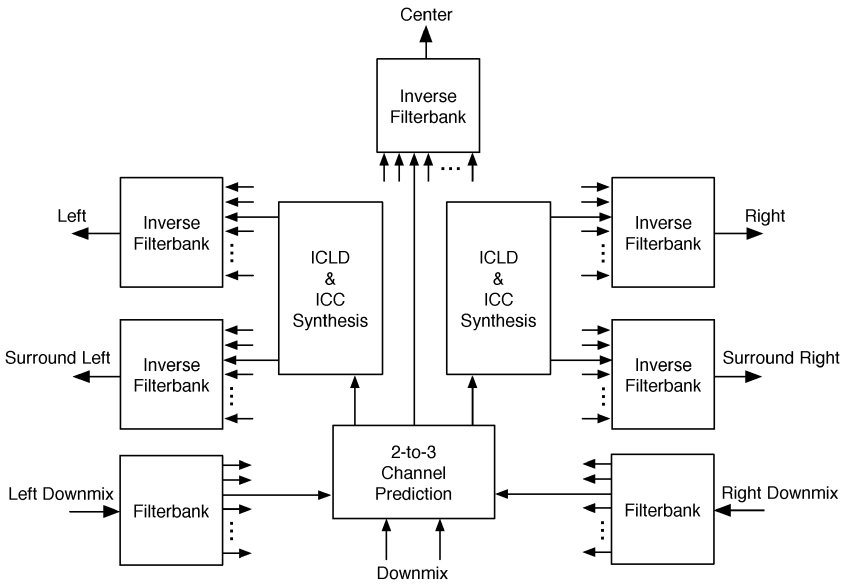


Fig. 16. *A MPEG Surround decoder uses one two-to-three channel prediction process and two one-to-two channel spatial cue synthesis processes for 5.1 surround synthesis.*

not think to be possible 20 years ago. Audio coding based on perceptual models enables the availability of music on the move and, via streaming services wherever we go, it finally enables communication approaching live-like quality. While research seemed to have reached (unknown) theoretical limits ten years ago, we still see a lively research community leading to new applications made possible. ■

Acknowledgment

The authors would like to thank H. Lukashevich and S. Turowski for help in editing, and G. Schuller and B. Edler for providing figures. The work described here covers the work of many contributors at many labs over more than two decades, so a special thanks is due to all who helped to advance the state of the art of these technologies.

REFERENCES

[1] J. Blauert and P. Tritthart, "Ausnutzung von Verdeckungseffekten bei der Sprachcodierung," (in German), in *Proc. Deutsche Jahrestagung für Akustik (DAGA)*, 1975, pp. 377–380.

[2] M. Krasner, "The critical band coder-digital encoding of speech signals based on the perceptual requirements of the auditory system," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1980, vol. 5, pp. 327–331.

[3] W. B. Kleijn and K. K. Paliwal, *An Introduction to Speech Coding*. Amsterdam, The Netherlands: Elsevier, 1995.

[4] S. E. Forshay, "An economical digital audio system for consumer delivery," *IEEE Trans. Consumer Electron.*, vol. CE-31, no. 3, pp. 269–277, Aug. 1985.

[5] J. D. Johnston and R. E. Crochiere, "An all-digital 'commentary grade' subband coder," *J. Audio Eng. Soc.*, vol. 27, no. 11, pp. 855–865, 1979.

[6] K. Brandenburg, G. G. Langenbucher, H. Schramm, and D. Seitzer, "A digital signal processor for real time adaptive transform coding of audio signal up to 20 kHz bandwidth," in *Proc. IEEE Int. Conf. Circuits Comput.*, 1982, pp. 474–477.

[7] K. Brandenburg, "OCF—A new coding algorithm for high quality sound signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1987, vol. 12, pp. 141–144.

[8] A. Sugiyama, F. Hazu, M. Iwadare, and T. Nishitani, "Adaptive transform coding with an adaptive block size (ATC-ABS)," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1990, pp. 1093–1096.

[9] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 314–323, Feb. 1988.

[10] G. Stoll, M. Link, and G. Theile, "Masking-pattern adapted subband coding: Use of the dynamic bit-rate margin," in *Proc. 84th Conv. Audio Eng. Soc.*, Mar. 1988, paper 2585.

[11] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. New York, NY, USA: Springer-Verlag, 1999.

[12] B. C. Moore, *Introduction to the Psychology of Hearing*. Baltimore, MD, USA: Univ. Park Press, 1977.

[13] B. C. Moore, "Hearing," in *Handbook of Perception and Cognition*, 2nd ed. San Diego, CA, USA: Academic, 1995.

[14] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. Cambridge, MA, USA: MIT Press, 1997.

[15] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *J. Acoust. Soc. Amer.*, vol. 66, no. 6, pp. 1647–1652, Dec. 1979.

[16] *MPEG-1: Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mbit/s-Part 3: Audio*, ISO/IEC 11172-3 International Standard, ISO/IEC, 1993, JTC1/SC29/WG11.

[17] K. Brandenburg and G. Stoll, "ISO-MPEG-1 audio: A generic standard for coding of high-quality digital audio," *J. Audio Eng. Soc.*, vol. 42, no. 10, pp. 780–792, Oct. 1994.

[18] K. Brandenburg, "MP3 and AAC explained," in *Proc. AES 17th Int. Conf. High Quality Audio Coding*, 1999, paper 17-009.

[19] 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," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. New York, NY, USA: AES, 1996, pp. 54–72.

[20] *Digital Audio Compression Standard*, AC-3, U.S. Television Systems Committee (ATSC), Dec. 1995.

[21] D. Pan, "A tutorial on MPEG/audio compression," *IEEE Multimedia*, vol. 2, no. 2, pp. 60–74, Summer, 1995.

[22] P. Noll, "MPEG digital audio coding," *IEEE Signal Process. Mag.*, vol. 14, no. 5, pp. 59–81, Sep. 1997.

[23] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, Apr. 2000.

[24] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards*, Boston, MA, USA: Kluwer, 2003, vol. 721, ser. *Engineering and Computer Science*.

[25] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*. Hoboken, NJ, USA: Wiley, 2007.

[26] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Res.*, vol. 47, pp. 103–138, 1990.

[27] B. C. J. Moore and B. R. Glasberg, "Suggested formula for calculating auditory-filter bandwidth and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, pp. 750–753, 1983.

[28] J. O. Smith and J. S. Abel, "Bark and ERB bilinear transform," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 697–708, Nov. 1999.

[29] R. D. Patterson, M. H. Allerhand, and C. Giguére, "Time-domain modeling of peripheral auditory processing: A modular architecture and software platform," *J. Acoust. Soc. Amer.*, vol. 98, no. 4, pp. 1890–1894, Oct. 1995.

[30] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed. Berlin, Germany: Springer-Verlag, 2007, vol. 22, ser. *Information Sciences*.

[31] M. Erne, G. Moschytz, and C. Faller, "Best wavelet-packet bases for audio coding using perceptual and rate-distortion criteria," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 1999, vol. 2, pp. 909–912.

[32] R. P. Hellman, "Asymmetry of masking between noise and tone," *Percept. Psychophys.*, vol. 11, no. 3, pp. 241–246, 1972.

[33] J. D. Johnston and A. J. Ferreira, "Sum-difference stereo transform coding," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1992, pp. 569–572.

[34] R. G. v. d. Waal and R. N. J. Veldhuis, "Subband coding of stereophonic digital audio signals," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1991, pp. 3601–3604.

[35] F. Baumgarte, "Ein psychophysiologisches gehörmodell zur nachbildung von wahrnehmungsschwellen für die audiocodierung," Ph.D. dissertation, Fakultät für Elektrotechnik und Informatik, Universität Hannover, Hanover, Germany, 2000.

[36] International Telecommunication Union—Telecommunication Standardization Sector (ITU-T), *Wideband coding of speech at around 16 kbit/s using adaptive multi-rate wideband (AMR-WB)*, Recommendation G.722.2—G.722.2, Jul. 2003.

[37] U. Krämer, G. Schuller, S. Wabnik, J. Klier, and J. Hirschfeld, "Ultra low delay audio coding with constant bit rate," in *Proc. 117th Audio Eng. Soc. Conv.*, San Francisco, CA, USA, 2004, pp. 22–33.

[38] O. A. Moussa, M. Li, and W. B. Kleijn, "Predictive audio coding using rate-distortion-optimal pre-and post-filtering," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2011, pp. 213–216.

[39] R. Zamir, Y. Kochman, and U. Erez, "Achieving the Gaussian rate-distortion function by prediction," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3354–3364, Jul. 2008.

[40] K. Brandenburg, E. Eberlein, J. Herre, and B. Edler, "Comparison of filterbanks for high quality audio coding," in *Proc. IEEE Int. Symp. Circuits Syst.*, 1992, vol. 3, pp. 1336–1339.

[41] J. Huang and P. Schultheiss, "Block quantization of correlated Gaussian random variables," *IEEE Trans. Commun. Syst.*, vol. 11, no. 3, pp. 289–296, Sep. 1963.

[42] R. M. Gray, "Toeplitz and circulant matrices: A review," Stanford Univ., Stanford, CA, USA, Tech. Rep., 1971.

[43] U. Grenander and G. Szego, *Toeplitz Forms and Their Applications*. New York, NY, USA: Chelsea, 1984.

[44] H. Bölcskei, F. Hlawatsch, and H. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans. Signal Process.*, vol. 46, no. 12, pp. 3256–3268, Dec. 1998.

[45] I. Daubechies, *Ten Lectures on Wavelets*, Philadelphia, PA, USA: SIAM, 1992, vol. 61, ser. *CBMS-SF Regional Conference Series in Applied Mathematics*.

[46] B. Edler, "Equivalence of transforms and filterbanks in source coding," (in German) Ph.D. dissertation, Fakultät für Maschinenwesen, Univ. Hanover, Hanover, Germany, 1995, Appendix A.

[47] J. P. Princen and A. B. Bradley, "Analysis and synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 34, no. 5, pp. 277–284, May 1986.

[48] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain alias cancellation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-34, no. 5, pp. 969–978, Oct. 1986.

[49] J. Princen, A. Johnson, and A. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1987, vol. 12, pp. 2161–2164.

[50] H. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, no. 6, pp. 969–978, Jun. 1990.

[51] T. Ramstad and J. Tanem, "Cosine-modulated analysis-synthesis filterbank with critical sampling and perfect reconstruction," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1991, pp. 1789–1792.

[52] 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.

[53] B. Edler, "Coding of audio signals with overlapping block transform and adaptive window functions,"(in German) *Frequenz*, vol. 43, no. 9, pp. 252–256, 1989.

[54] C. Todd, G. Davidson, M. Davis, L. Fielder, B. Link, and S. Vernon, "AC-3: Flexible perceptual coding for audio transmission and storage," in *Proc. 96th Conv. Aud. Eng. Soc.*, Feb. 1994, Paper 3796.

[55] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.

[56] P. Chou, T. Lookabough, and R. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, no. 1, pp. 31–42, Jan. 1989.

[57] J. J. Rissanen and G. Langdon, "Arithmetic coding," *IBM J. Res. Develop.*, vol. 23, no. 2, pp. 149–162, 1979.

[58] G. N. N. Martin, "Range encoding: An algorithm for removing redundancy from a digitised message," in *Proc. Video Data Recording Conf.*, Southampton, U.K., 1979, pp. 173–180.

[59] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proc. IRE*, vol. 40, no. 9, pp. 1098–1101, Sep. 1952.

[60] International Telecommunication Union (ITU), *Pulse code modulation (PCM) of voice frequencies*, Recommendation ITU-T G.711, 1993. [Online]. Available: http://www.itu.org

[61] *MPEG-2: Generic Coding of Moving Pictures and Associated Audio Information—Part 3: Audio*, ISO/IEC 13818-3 International Standard, ISO/IEC, 1995, JTC1/SC29/WG11.

[62] G. Stoll, "ISO-MPEG-2 audio: A generic standard for the coding of two-channel and multichannel sound," in *Collected Papers on Digital Audio Bit-Rate Reduction*, N. Gilchrist and C. Grewin, Eds. New York, NY, USA: Audio Engineering Society, 1996, pp. 43–53.

[63] *MPEG-2: Generic Coding of Moving Pictures and Associated Audio Information—Part 7: Advanced Audio Coding*, ISO/IEC 13818-7 International Standard, ISO/IEC, 1997 JTC1/SC29/WG11.

[64] 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, 1997.

[65] E. Larsen and R. Aarts, *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. New York, NY, USA: Wiley, 2004.

[66] B. Iser, W. Minker, and G. Schmidt, *Bandwidth Extension of Speech Signals*. New York, NY, USA: Springer-Verlag, 2008.

[67] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *Proc. 112th Conv. Aud. Eng. Soc.*, May 2002, Paper 5553.

[68] J. Herre and M. Dietz, "Standards in a nutshell: MPEG-4 high-efficiency AAC coding," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 137–142, May 2008.

[69] *MPEG-4: Coding of Audio-Visual Objects—Part 3: Audio*, ISO/IEC 14496-3 International Standard, ISO/IEC, 2010, JTC1/SC29/WG11.

[70] M. Neuendorf, M. Multrus, N. Rettelbach, G. Fuchs, J. Robilliard, J. Lecomte, S. Wilde, S. Bayer, S. Disch, C. 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, C. K. Seng, E. Oh, M. Kim, S. Quackenbush, and B. Grill, "MPEG unified speech and audio coding—The ISO/MPEG standard for high-efficiency audio coding of all content types," in *Proc. 132nd Conv. Aud. Eng. Soc.*, Apr. 2012, Paper 8654.

[71] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegard, "Low complexity parametric stereo coding," in *Proc. 117th Conv. Aud. Eng. Soc.*, May 2004, Paper 6073.

[72] J. Kim, E. Oh, and J. Robilliard, "Enhanced stereo coding with phase parameters for MPEG unified speech and audio coding," in *Proc. 127th Conv. Aud. Eng. Soc.*, Oct. 2009, Paper 7875.

[73] *MPEG-D: MPEG Audio Technologies—USAC*, ISO/IEC 23003-3 International Standard, ISO/IEC, 2012, JTC1/SC29/WG11.

[74] S. Quackenbush and R. Lefebvre, "Performance of MPEG unified speech and audio coding," in *Proc. 131st Conv. Aud. Eng. Soc.*, Oct. 2011, paper 8514.

[75] J. Makinen, 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 Int. Conf. Acoust. Speech Signal Process.*, 2005, vol. 2, pp. 1109–1112.

[76] E. Allamanche, R. Geiger, J. Herre, and T. Sporer, "MPEG-4 low delay audio coding based on the AAC codec," in *Proc. 106th Conv. Aud. Eng. Soc.*, May 1999, Paper 4929.

[77] M. Schnell, M. Schmidt, M. Jander, T. Albert, R. Geiger, V. Ruoppila, P. Ekstrand, and G. Bernhard, "MPEG-4 enhanced low delay AAC—A new standard for high quality communication," in *Proc. 125th Conv. Aud. Eng. Soc.*, Oct. 2008, Paper 7503.

[78] N. Harada, Y. Kamamoto, T. Liebchen, T. Moriya, and Y. A. Reznik, "The MPEG-4 audio lossless coding (ALS) standard—Technology and applications," in *Proc. 119th Conv. Aud. Eng. Soc.*, Oct. 2005, Paper 6589.

[79] R. Geiger, R. Yu, and J. Herre, "ISO/IEC MPEG-4 high-definition scalable Advanced Audio Coding," *J. Audio Eng. Soc.*, vol. 55, no. 1/2, pp. 27–43, Jan./Feb. 2007.

[80] M. Hans and R. W. Schafer, "Lossless compression of digital audio," *IEEE Signal Process. Mag.*, vol. 18, no. 4, pp. 21–32, Jul. 2001.

[81] L. Fielder, S. Lyman, S. Vernon, and C. Todd, "Professional audio coder optimized for use with video," in *Proc. 112th Conv. Aud. Eng. Soc.*, 1999, Paper 5033.

[82] C. Faller and F. Baumgarte, "Efficient representation of spatial audio using perceptual parametrization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2001, pp. 199–202.

[83] C. Faller and F. Baumgarte, "Binaural cue coding applied to stereo and multi-channel audio compression," in *Proc. 112th Conv. Aud. Eng. Soc.*, May 2002, Paper 5574.

[84] C. Faller and F. Baumgarte, "Binaural cue coding applied to audio compression with flexible rendering," in *Proc. 113th Conv. Aud. Eng. Soc.*, Oct. 2002, Paper 5686.

[85] F. Baumgarte and C. Faller, "Binaural cue coding—Part I: Psychoacoustic fundamentals and design principles," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 509–519, Nov. 2003.

[86] C. Faller and F. Baumgarte, "Binaural cue coding—Part II: Schemes and applications," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 520–531, Nov. 2003.

[87] E. Schuijers, W. Oomen, A. C. den Brinker, and A. J. Gerrits, "Advances parametric coding for high-quality audio," in *Proc. IEEE Benelux Workshop Model Based Process. Coding Audio*, Nov. 2002, pp. 73–79.

[88] E. Schuijers, W. Oomen, B. den Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," in *Proc. 114th Conv. Aud. Eng. Soc.*, Mar. 2003, Paper 5852.

[89] H. Purnhagen, "Low complexity parametric stereo coding in MPEG-4," in *Proc. 7th Int. Conf. Audio Effects*, Naples, Italy, 2004, pp. 163–168.

[90] J. Engdegard, H. Purnhagen, J. Roden, and L. Liljeryd, "Synthetic ambience in parametric stereo coding," in *Proc. 117th Conv. Aud. Eng. Soc.*, May 2004, Paper 6074.

[91] C. Faller, "Coding of spatial audio compatible with different playback formats," in *Proc. 117th Conv. Aud. Eng. Soc.*, Oct. 2004, Paper 6187.

[92] J. Herre, C. Faller, C. Ertel, J. Hilpert, A. Hoelzer, and C. Spenger, "MP3 surround: Efficient and compatible coding of multi-channel audio," in *Proc. 116th Conv. Aud. Eng. Soc.*, May 2004, Paper 6049.

[93] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Röden, 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, 2008.

[94] J. Breebaart and C. Faller, *Spatial Audio Processing: MPEG Surround and Other Applications*. Hoboken, NJ, USA: Wiley, 2007.

[95] *MPEG-D: MPEG Audio Technologies—MPEG Surround*, ISO/IEC 23003-1 International Standard, ISO/IEC, 2012, JTC1/SC29/WG11.

[96] C. Faller, "Parametric joint-coding of audio sources," in *Proc. 120th Conv. Aud. Eng. Soc.*, May 2006, Paper 6752.

[97] J. Engdegard, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hoelzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, and W. Oomen, "Spatial audio object coding (SAOC): The upcoming MPEG standard on parametric object based audio coding," in *Proc. 124th Conv. Aud. Eng. Soc.*, May 2009, Paper 7377.

[98] H.-O. Oh, C. Faller, A. Favrot, and Y.-W. Jung, "Enhancing stereo audio with remix capability," in *Proc. 129th Conv. Aud. Eng. Soc.*, Nov. 2010, Paper 8290.

ABOUT THE AUTHORS

Karlheinz Brandenburg (Fellow, IEEE) received the Dipl.-Ing. and Dipl.-Math. degrees in electrical engineering and mathematics and the Dr.-Ing. degree in electrical engineering from the Friedrich-Alexander-Universität, Erlangen-Nürnberg, Erlangen, Germany, in 1980, 1982, and 1989, respectively.

From 1989 to 1990, he held a research position at AT&T Bell Laboratories, Murray Hill, NJ, USA. After returning to Friedrich-Alexander-Universität, Erlangen-Nürnberg, in 1990, he joined the Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany, as head of the Audio and Multimedia Department. Since 2000, he has been a Professor at the Institute of Media Technology, Ilmenau University of Technology, Ilmenau, Germany, and at the same time director of the Fraunhofer-Institut für Digitale Medientechnologie (Fraunhofer IDMT), Ilmenau, Germany. He is acclaimed for pioneering work in digital audio coding, perceptual measurement techniques, wave field synthesis, psychoacoustics, and analysis of audio and video signals. The research results of his dissertation are the basis of MPEG-1 Layer 3 (MP3), MPEG-2 Advanced Audio Coding (AAC), and most other modern audio compression schemes. He published numerous articles in scientific conferences, coauthored the book *Applications of Digital Signal Processing to Audio and Acoustics* (New York, NY, USA: Springer-Verlag, 1998), and holds about 100 patents.

Prof. Brandenburg is a Fellow of Audio Engineering Society (AES), a member of "Fernseh-und Kinotechnische Gesellschaft e. V." (FKTG), a member of "Deutsche Gesellschaft für Akustik" (DEGA), and has been elected as a member of "Saxonian Academy of Sciences in Leipzig." For many years, he served as a member of the Technical Committee on Audio and Acoustics of the IEEE Signal Processing Society and as General Chair of the 2002 IEEE International Symposium on Consumer Electronics (ISCE) in Ilmenau, Germany. His honors include the Cross of the Order of Merit of the Federal Republic of Germany, the IEEE Masaru Ibuka Consumer Electronics Award, the German Future Award (shared with his colleagues), the IEEE Engineering Excellence Award, and the Audio Engineering Society Silver Medal Award. Furthermore, he is a member of the Hall of Fame of the Consumer Electronics Association and of the International Electrotechnical Commission and received honorary doctorate degrees from the University Koblenz-Landau, Germany, and the Leuphana University of Lüneburg, Germany, for his outstanding research work in the field of audio coding.

Christof Faller received the Dipl. Ing. degree in electrical engineering from ETH Zurich, Zurich, Switzerland, in 2000 and the Ph.D. degree for his work on parametric multichannel audio coding from the École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2004.

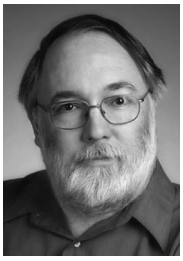
From 2000 to 2004, he worked in the Speech and Acoustics Research Department, Bell Labs Lucent and its spinoff, Agere Systems, where he worked on audio coding for satellite radio, MP3 surround, and the MPEG surround international standard. He is currently Managing Director at Illusonic GmbH, Uster, Switzerland, a company he founded in 2006, and teaches at EPFL.

Dr. Faller has won a number of awards for his contributions and inventions in spatial audio.

Juergen Herre (Senior Member, IEEE) received the Dipl.-Ing. and Dr.-Ing. degrees in electrical engineering from the Friedrich-Alexander-Universität, Erlangen-Nürnberg, Erlangen, Germany, in 1989 and 1995, respectively.

In 1989, he joined the Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany. In 1995, he joined Bell Laboratories for a Postdoc term working on the development of MPEG-2 Advanced Audio Coding (AAC). By the end of 1996, he went back to Fraunhofer to work on the development of more advanced multimedia technology, including MPEG-4, MPEG-7, and MPEG-D, currently as the Chief Scientist for the Audio/Multimedia activities at Fraunhofer IIS, Erlangen. In September 2011, he was appointed Professor at the University of Erlangen and the International Audio Laboratories Erlangen. He is an expert in low bit rate audio coding/perceptual audio coding, spatial audio coding, parametric audio object coding, perceptual signal processing, and semantic audio processing.

Prof. Herre is member of the IEEE Technical Committee on Audio and Acoustic Signal Processing and served as an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He is a Fellow of the Audio Engineering Society (AES), Cochair of the AES Technical Committee on Coding of Audio Signals, and Vice Chair of the AES Technical Council.



James D. Johnston (Fellow, IEEE) received the B.S.E.E. and M.S.E.E. degrees from Carnegie Mellon University, Pittsburgh, PA, USA, in 1975 and 1976, respectively.

He worked for AT&T Bell Labs and its heir organization AT&T Research in 1975 (summer) and from 1976 to 2002, starting as a casual lab tech, and retiring as a "Technology Leader" in the field of audio and auditory perception. After that, he was an Audio Architect at Microsoft in the Media Group until 2008, and Chief Scientist for Neural Audio and its purchaser until 2011, when he retired. He has published in the *Journal of the Audio Engineering Society* and its conventions and has been papers chair for several AES conventions. His current interests are in spatial audio perception, diffuse and direct acoustic sensation, simulation of acoustic environments (real or imaginary), and new loudspeaker technology.

Mr. Johnston is currently a member of the Signal Processing, Communications, and Broadcast Societies in the IEEE. He is also 25+ year member and Fellow of the Audio Engineering Society and a life member of the National Speleological Society. He has given the Heyser Lecture of the Audio Engineering Society. He is a corecipient of the Donald Fink award for tutorial paper, and is the 2006 James L. Flanagan Signal Processing awardee. He is a former member of the Technical Committee on Audio and Acoustics Signal Processing, a former Associate Editor of the IEEE TRANSACTIONS ON SPEECH, AUDIO, AND LANGUAGE PROCESSING, and has reviewed many papers for International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and other IEEE and other functions.

W. Bastiaan Kleijn (Fellow, IEEE) received the M.S. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1984, the M.S. degree in physics and the Ph.D. degree in soil science from the University of California Riverside, Riverside, CA, USA, both in 1981, and the Ph.D. degree in electrical engineering from the Delft University of Technology, Delft, The Netherlands, in 1991.

He has been a Professor at Victoria University of Wellington, Wellington, New Zealand, since 2010. He is also a Professor at Delft University of Technology (part-time, since 2011) and at KTH in Stockholm, where he was the Head of the Sound and Image Laboratory. Before joining KTH in 1996, he worked at AT&T Bell Laboratories (Research) on speech processing. He was a founder of Global IP Solutions, which developed voice and video processing technology, and was acquired by Google in 2010. He has authored or coauthored over 230 peer-reviewed papers and holds close to 40 U.S. patents.

He has served or is serving on the Editorial Boards of IEEE Signal Processing Letters, the IEEE Transactions on Speech and Audio, IEEE Signal Processing Magazine, and *Signal Processing*. He was Technical Chair of the 1999 International Conference on Acoustics, Speech, and Signal Processing (ICASSP) and the 2011 European Signal Processing Conference (EUSIPCO).