OPEN ACCESS Freely available online
J. Herre, S. Dick, A. Biswas, M. Bosi, P. M. Delgado, G. Schuller, and M. Erne, "'Perceptual Audio Coders What to Listen For': The Free Web Edition," *J. Audio Eng. Soc.*, vol. 74, no. 9, pp. 576–593 (2026 Sep.). https://doi.org/10.17743/jaes.2026.0299.

"Perceptual Audio Coders – What to Listen For": The Free Web Edition

JÜRGEN HERRE,[1,*] *AES Fellow*, **SASCHA DICK,**[2,*] *AES Fellow*,
(juergen.herre@audiolabs-erlangen.de) (sascha.dick@iis.fraunhofer.de)

ARIJIT BISWAS,[3] *AES Member*, **MARINA BOSI,**[4] *AES Fellow*,
(arijit.biswas@dolby.com) (mab@ccrma.stanford.edu)

PABLO M. DELGADO,[2] *AES Member*,
(pablo.delgado@iis.fraunhofer.de)

GERALD SCHULLER,[2,5] *AES Member* **AND MARKUS ERNE**[†]
(gerald.schuller@tu-ilmenau.de)

[1]*International Audio Laboratories Erlangen, a joint institution of Fraunhofer Institute for Integrated Circuits and Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[2]*Fraunhofer Institute for Integrated Circuits, Erlangen, Germany*
[3]*Dolby Germany GmbH, Nürnberg, Germany*
[4]*Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA, USA*
[5]*Technische Universität Ilmenau, Ilmenau, Germany*

In 2001, the AES Technical Committee on Coding of Audio Signals produced the legendary educational CD-ROM "Perceptual Audio Coders – What To Listen For." It contains a taxonomy of common types of codec artifacts and tutorial information on the background of each one. Example audio signals with different degrees of impairment illustrate the nature of the artifacts and help in training test listeners' expertise. Since its initial release, several generations of CD-ROMs have been sold, and the material has found worldwide use for the education of the public. This paper presents the comprehensive second edition of the tutorial, based on the Technical Committee's efforts in producing a next-generation educational package. It covers an extended range of coding artifact types that can be typically experienced with classic and advanced audio coder processing (e.g., bandwidth extension or parametric stereo). The paper provides the theoretical background for coding artifacts and educates readers on subjective quality evaluation. The enhanced format of the tutorial allows for seamless display and playback on PCs, tablets and mobile phones and includes interactive graphical elements. This attractive educational package is available freely on the web.

Keywords: Artifacts, Listener Education, Perceptual Audio Coding

1 INTRODUCTION

Compared to other "classic" disciplines in audio engineering, the field of perceptual audio coding is a comparably young technology, combining elements from digital signal processing, coding theory, and psychoacoustics. By using sophisticated signal processing techniques and exploiting psychoacoustic phenomena, perceptual audio coding at very low bitrates—independent of specific implementations—can change the sound noticeably and can produce artifacts that are very different from traditional distortions in their nature. Such artifacts are frequently not obvious at all to the untrained listener. Moreover, the use of advanced psychoacoustics frequently led to questions from nonexperts ("how does that work?") and sometimes was even perceived as some kind of "black magic" within the coder ("I don't trust this psychoacoustics stuff").

Since many of the members of the AES Technical Committee on Coding of Audio Signals (TC-CAS) have been confronted with such questions or comments, the idea of an educational publication on this topic soon found broad support within the group once its concept was formulated. Implementing this idea eventually took several years within

*To whom correspondence should be addressed, email: juergen.herre@audiolabs-erlangen.de; sascha.dick@iis.fraunhofer.de
†Deceased 17 Aug. 2024.

the committee and resulted in the production of a first-generation tutorial "Perceptual Audio Coders – What To Listen For" [1] in 2001.

The tutorial was distributed as a hybrid CD-ROM that could be used both as an audio CD on conventional audio equipment and as a CD-ROM for computer use, containing the tutorial in an HTML-based style for web browsers. The tutorial provided a taxonomy of common perceptual audio coding artifact types and explained each of them in dedicated sections. Especially, it also included audio examples with different degrees of distortion, varying from "subtle" up to "obvious" in order to enable easy listener training. Several categories of coding artifacts were explained and demonstrated using audio examples including: pre-echo artifacts, speech reverberation artifacts, binaural masking level difference (BMLD) artifacts, loss of stereo image artifacts, high-frequency limitation artifacts, aliasing artifacts, artifacts due to tandem coding, and artifacts resulting from coding music with a speech codec. This resource covered all artifact types that were practically relevant for coding systems at that point in time. However, the demonstrated signals were generated independently of any specific commercially available coding system to be more generalized and still valid for more modern coding systems.

After the successful sale of several print generations of the original CD-ROM by the AES, the TC-CAS felt that it was time for an update of the tutorial material by discussing artifacts as they might be produced by new coding techniques that have become popular after the first-generation CD-ROM release. Specifically, this includes advanced (semi)parametric coding techniques like bandwidth extension or parametric stereo/multichannel coding that have become successful and widely deployed after 2000. In addition, the second generation material features enhancements in presentation style to further enhance its attractiveness.

This paper, targeted toward audio professionals, accompanies the enhanced second edition of the educational material called the "Free Web Edition," which addresses a broader audience and is freely available as a resource to anyone. The tutorial was created to be deliberately agnostic of any specific audio coding technology, in order to provide general knowledge to the readers while avoiding any potential commercial bias. Hence, the discussion throughout the paper focuses on general principles and perceptual effects, rather than on any specific coding standard or implementation. While numerous papers are cited with particular technologies, they merely serve as examples for the general viewpoint of the paper and tutorial.

This paper is organized as follows. SEC. 2 describes the Free Web Edition's features and access. SEC. 3 provides background information on perceptual audio coding, and SEC. 4 provides a tutorial on possible artifact types. SEC. 5 discusses issues of subjective evaluation and listener training. Finally, the paper closes with a summary and acknowledges the contributions to the Free Web Edition.

The provided overview reflects the most recently published version of the educational material at the time of writing. However, another advantage of the online format is that it can be easily updated and extended to cover new developments in the field of audio coding. For instance, material on neural codecs may be incorporated in the future when sufficient consolidated know-how has become available.

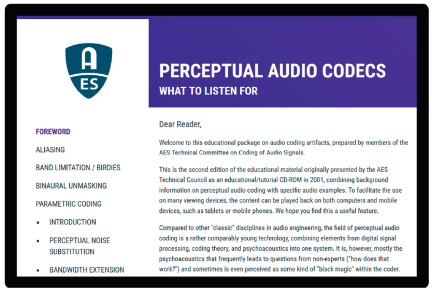


Fig. 1. Desktop device view.

2 ABOUT THE FREE WEB EDITION

The Free Web Edition extends the coverage of coding techniques and type of artifacts they might introduce from the original CD-ROM edition, and it is envisioned to be further extended in the future. Also, the tutorial's section about subjective quality evaluation of audio has been enhanced with interactive audio examples together with associated scores from subjective listening tests to enable listeners to compare their perception of quality to proven experimental results. For further reading, the tutorial contains an updated selection of articles, books, and links to other websites with interesting relevant tutorials and audio demonstrations. The Free Web Edition also features many enhancements in the presentation to increase the attractiveness of the educational material.

2.1 Access

As a gift to the community, the tutorial is freely accessible to the public (not limited to AES Members) at https://www.audiolabs-erlangen.de/content/resources/aesCodingTutorial/index.html.

2.2 Presentation Features

The educational descriptions, graphics, and audio examples of the Free Web Edition tutorial have been enriched with interactive audio playback and graphics:

- Web/download format: The new edition is delivered from a website either for use with arbitrary web browsers over the Internet or as a .ZIP download package for offline reading and listening.
- Mobile-friendly: To facilitate the use on many viewing devices, the content is formatted in a "mobile-friendly" way [i.e., it will display well not only on computers (Fig. 1) but also on tablets and smartphones (Fig. 2)].
- Interactive graphics: Many tutorial sections now make use of interactive graphics for visualizing and comparing the technical impact of an artifact. By repeatedly clicking the graphics, the display toggles

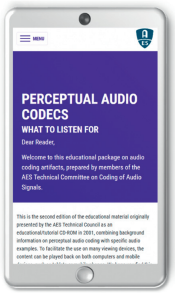


Fig. 2. Mobile device view.

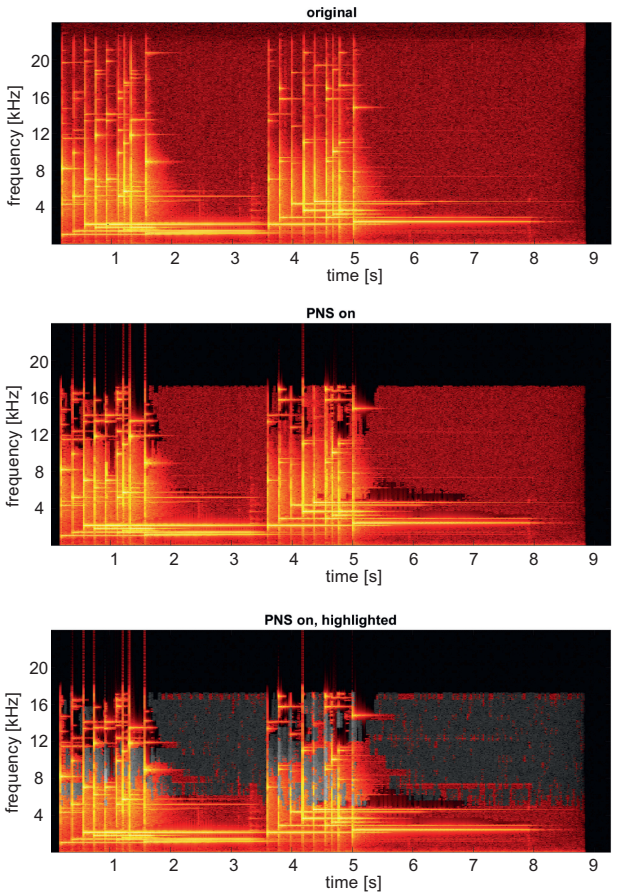


Fig. 3. Example visualization of spectral differences—interactive switching between views enables easy visual comparison (see "Perceptual Noise Substitution" section of the online tutorial).

between several "overlaid" views on the signals (e.g., spectrograms; Fig. 3).

- Integrated audio player: Audio playback using HTML 5 provides seamless playback on all devices without the need for an external player application in different views (examples are shown in Figs. 4 and 5).

3 BACKGROUND

3.1 Psychoacoustic Foundations

Before discussing audio coding artifacts, a short review of the psychoacoustic foundations of perceptual audio coding is provided subsequently. From a signal processing perspective, the cochlea can be viewed as a bank of overlapping band-pass filters, leading to the notion of *critical bands* [2–5] to describe the inherent frequency selectivity of human hearing. The most commonly adopted critical band distribution in perceptual audio coding is based on empirical data reported in [5]. As can be observed in Fig. 6, the band-splitting is nonuniform with respect to frequency, showing constant-bandwidth filters up to around 500 Hz and almost constant-Q filters above 500 Hz. The so-called "Bark scale" is an absolute perceptual frequency scale that is constructed by concatenating critical bands over the audible frequency range as illustrated in Fig. 7. Typically, masking models used in perceptual audio coding perform their calculations in the Bark scale domain [6].

Simultaneous or spectral (frequency) masking is a phenomenon in which a low-level signal (the maskee) becomes inaudible (masked) due to the presence of a simultaneously occurring stronger signal (the masker), provided that the masker and maskee are close enough in frequency (see Fig. 8) [5].



Fig. 4. Integrated audio player (list view) allowing for easy comparison of differently coded versions (here: different processing crossover frequencies) (see "Perceptual Noise Substitution" section of the online tutorial).

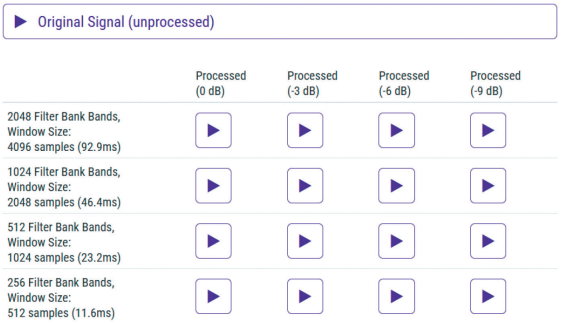


Fig. 5. Integrated audio player (matrix view) demonstrating the impact of two parameters (here: level and window size) (see "Speech Reverberation" section of the online tutorial).

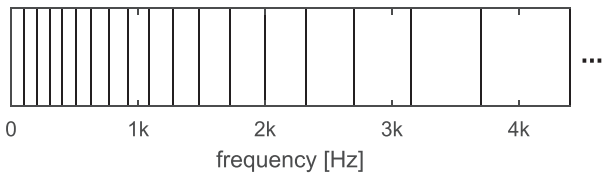


Fig. 6. Critical band distribution for frequencies up to 4 kHz based on data from [5].

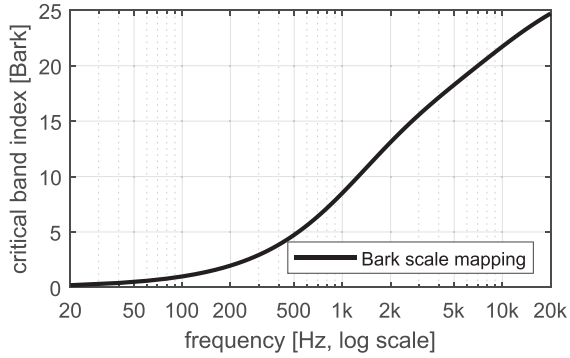


Fig. 7. Extended critical band index as a function of frequency, based on data from [5].

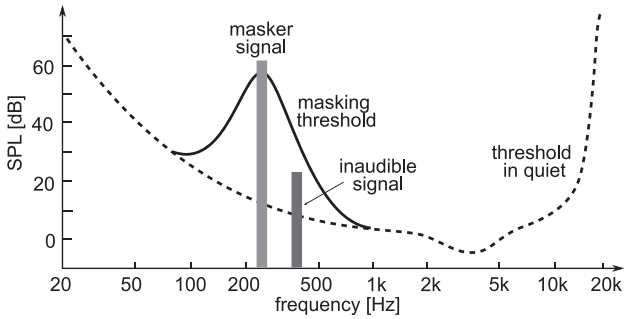


Fig. 8. Spectral masking.

Such masking is most pronounced in the critical band where the masker is located and is less effective in neighboring bands. The masking threshold relative to a masker defines the level below which a low-level signal becomes inaudible (masked). This masked signal may include low-level signal components, quantization noise, aliasing distortion, and/or transmission errors. The masking threshold, also known as threshold of just noticeable distortion [7], in the context of source coding, combines contributions of multiple maskers and the threshold in quiet [5] and varies over both frequency and time. Individual masker contributions depend on the SPL of the masker L_M and on the characteristics of both the masker and the maskee. In general, the slope of the masking threshold is steeper toward lower frequencies with respect to the center frequency of the masker, and it presents a gentler slope at higher frequencies (i.e., which means higher-frequency maskee components are more easily masked). Based on data from [5] for narrowband noise-masking-tone and from [8], one can approximate these slopes as 27 dB/Bark for frequencies below the center frequency of the masker and as

$$-27 \text{ dB/Bark} + 0.367 \max(L_M - 40,\ 0) \tag{1}$$

for frequencies above the center frequency of the masker. Additionally, there is an asymmetry of masking [9], where the level difference between the masker and the masking threshold is smaller in noise-masking-tone experiments than in tone-masking-noise experiments, which means noise is a more effective masker than a tone.

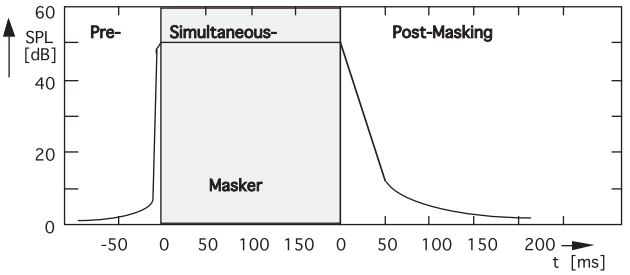


Fig. 9. Temporal masking adapted from [5].

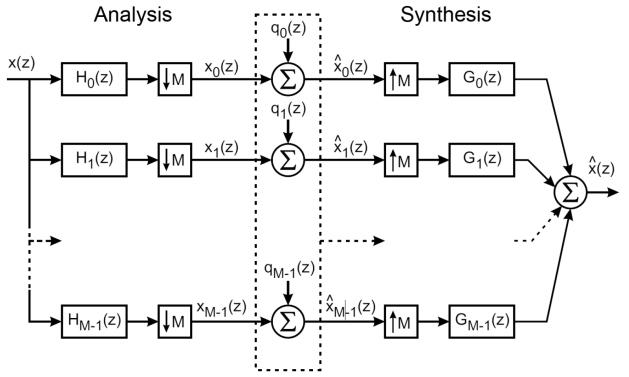


Fig. 10. Conceptual diagram for a sub-band coding system with quantizers based on [6]. Left half: analysis. Right half: synthesis.

In addition to spectral masking, the phenomenon of temporal masking (i.e., masking that occurs before and after the onset of the masker) plays an important role in auditory perception. Due to the finite integration time in the human hearing mechanism and depending on the masker SPL, the stronger sound may mask the weaker one, even if the maskee precedes the masker (premasking) as shown in Fig. 9. Premasking and postmasking can be exploited in perceptual coding algorithms and become important when discussing block artifacts, such as the spreading of quantization noise in time (pre-echoes) in coding systems.

The primary goal of any perceptual coding scheme is to dynamically and adaptively shape the coding error in a way that it always remains inaudible (or least disturbing). Ideally, if sufficient bitrate is available, the coding error lies below the masking threshold. Note that audibility of errors in the vicinity of the masking threshold is governed by a psychometric function (a probability of detection), rather than a hard boundary where a signal instantly goes from completely imperceptible to perfectly audible.

3.2 Principles of Perceptual Audio Coding

A common design paradigm in perceptual audio coding is to adaptively map the signal from the time domain into the frequency domain, following a filter bank–based coding paradigm. The signal spectrum is split into nearly uncorrelated spectral sub-band components that can then be quantized and entropy-coded independently for each sub-band. In the encoder, an analysis filter bank consisting of M band-pass filters is used to split the source spectrum accordingly (analysis; see the left side of Fig. 10). In the decoder,

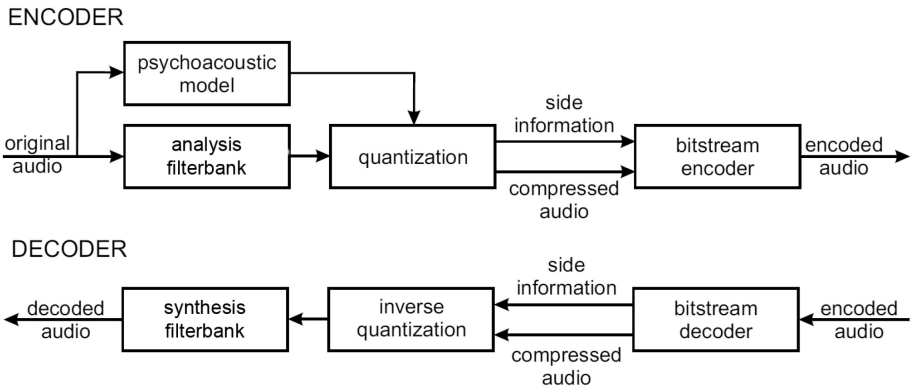


Fig. 11. Conceptual diagram for encoder and decoder of a perceptual coding system based on [6].

the set of sub-bands is recombined, using a synthesis filter bank (synthesis; see the right side of Fig. 10).

Each filter is critically sampled (i.e., sampled at twice the bandwidth of the corresponding band-pass filter). Since ideal "brick wall" band-pass filters do not exist, frequency domain aliasing cancellation techniques are used in such filter banks so that aliasing components between adjacent frequency bands will cancel. Pseudo/polyphase [10, 11] quadrature mirror filters are examples of sub-band filters originally utilized in perceptual audio coding, ensuring frequency-domain aliasing cancellation just between adjacent bands.

Another widely used example of a filter bank in audio coding is the modified discrete cosine transform (MDCT), which was derived from time-domain aliasing cancellation [12, 13]. MDCT filter banks are critically sampled and with 50% overlap between adjacent time domain blocks. They have perfect reconstruction after the overlap/add synthesis stage. The time-domain block sizes ideally are chosen as large as possible in order to code stationary signals most efficiently, achieving a good ratio between the amount of sub-band data and the required side information, such as bit allocation data, ancillary data, and scaling factors. These scaling factors are band-specific gain values that allow the encoder to shape the quantization noise so it remains below the threshold of hearing. Nonetheless, the block size should be small enough so that the audio signal within each block remains quasi-stationary, meaning its spectral characteristics stay relatively constant without abrupt temporal changes, thus avoiding pre-echo artifacts (see SEC. 4.2). To adapt to different audio signal characteristics, common audio coders can adaptively switch the block size, or switch the number of sub-bands, typically between 128 and 1,024 sub-bands [14].

Fig. 11 provides a block diagram of a generic (monophonic) perceptual audio encoder and decoder. The input signal is decomposed into subsampled spectral coefficients by the analysis filter bank. The spectral coefficients are then quantized according to perceptual and rate requirements, encoded (e.g., entropy encoded) and assembled into a bitstream. To consider perceptual issues, a perceptual model analyses the input signal by, for example, determining its time-dependent and frequency-dependent masking threshold. This information is provided to the quantization or coding process in order to optimize perceptual quality. In the decoder, the bitstream is disassembled, and the quantized spectral coefficients are rescaled to their original range and mapped back from spectral to time domain by the synthesis filter bank.

Many of the intricacies behind the implementation of a practical perceptual encoder are related to the sophistication of the perception-guided strategy that adaptively drives the allocation of resources for different blocks. Generally, the amount of quantization noise typically decreases by approximately 6 dB for each additional bit allocated to the quantizer. A straightforward bit allocation strategy aims at keeping the quantization noise level below the masking threshold. Generally, bit allocation in an audio coding system must balance data rate and perceptual quality through an optimization process. To satisfy the data rate constraint, it is formulated as a rate-distortion optimization problem, where the constraint determines the maximum number of bits that can be allocated (see [6] and [15]).

3.3 Parametric Audio Coding

Modern perceptual audio coding approaches—across a variety of implementations—often incorporate additional tools to parameterize the audio signal beyond the conventional techniques used in waveform coding. Instead of relying solely on time–frequency representations, part of the signal is expressed through a sparse set of abstract parameters, such as spectral envelopes for high-frequency reconstruction [16]. The goal of these tools is to reproduce the perceptual properties of the signal, rather than restoring its original waveform, therefore ensuring good quality even at low bitrates (e.g., below 64 kb/s). One drawback of these techniques is that they often fail to achieve transparent (i.e., perceptually indistinguishable) quality at high data rates. Also, improper application of these tools can introduce characteristic artifacts, even at high data rates. Typical artifacts introduced by these techniques will be reviewed in SEC. 4.9.

4 AUDIO CODING ARTIFACTS

4.1 Overview

The following subsections introduce and provide a theoretical background of the most prominent types of audio coding artifacts, as they are contained in the artifacts tutorial together with listening examples and other related information. References to audio examples that are available in the tutorial are *highlighted in italics*.

Different stages of the perceptual coding system introduce distinct types of audible impairments, the character of which often reveals the underlying technical cause. If, for instance, quantization errors are introduced before the filter bank synthesis stage, the alias cancellation process is no longer perfect (see SEC. 4.4).

When the filter bank uses large blocks, and the signal within this block is statistically nonstationary (e.g., contains transients), a distortion due to temporal spreading of the quantization error over the block is introduced. This is known as pre-echo and can cause disturbing artifacts that will be discussed in SEC. 4.2.

High-frequency resolution filter banks, such as those using large block-length transforms, enhance the overall efficiency of the coder and offer a better framework for psychoacoustic models for stationary signals. However, high-frequency resolution comes at the cost of poor temporal resolution, making these filter banks less effective at handling nonstationary signals. Especially for speech signals, where voiced and unvoiced segments (including plosives) alternate, the use of a time-to-frequency mapping with high frequency selectivity may introduce some "echoiness" into the coded speech signal. This artifact is known as "speech reverberation" and will be discussed in SEC. 4.3. Most modern coders address these issues by utilizing switchable filter banks [17, 18], where the time–frequency resolution of the filter bank is adjusted (switched) upon detecting transients.

To meet data-rate constraints, the perceptual model may temporarily demand a bit allocation that quantizes certain sub-bands, especially at high frequencies, to zero for low data rates and complex signals. In this way, a high-frequency component of a sound may momentarily disappear and become audible again later, when the bit-allocation of a following block has changed. This artifact is called "birdies" and will be explained in SEC. 4.5.

Another interesting surprising phenomenon can be observed at low frequencies where the masking threshold of the ear can sometimes be considerably lowered when listening with two ears rather than with one due to differences in various spatial properties between sounds (e.g., audio signal and coding error). Binaural phase differences may reduce masking significantly with respect to the level estimated on monophonic signal modeling, especially for headphone listening. When this effect (known as BMLD [19, 20]) is not considered appropriately in the audio coder, artifacts will be introduced for certain signals as discussed in SEC. 4.6.

Generally speaking, spatial auditory perception and spatial masking add another level of processing on top of the monaural phenomena discussed previously. Spatial hearing and localization of sound are based on so-called spatial cues, some of which are described in the literature [20]. The use of mid/side (M/S) stereo coding, as one representative technique, is a simple method [21] that helps to prevent spatial unmasking for stereo signals, most prominently for near-monophonic ones. Other techniques, such as Intensity Stereo coding [22, 23], provide a high potential for saving bits but may create stereo imaging artifacts, which are discussed in detail in SEC. 4.9.

Since most processing in a typical audio distribution chain (e.g., postprocessing, editing, or remixing of streamed content) operates on uncompressed Pulse Code Modulation audio rather than a compressed representation, using audio coders for storage or transmission requires an additional decoding and re-encoding cycle. If several encoding/decoding steps accumulate, the audio quality will usually degrade with each cycle; this problem is discussed in SEC. 4.7.

For more details on the principles of perceptual coding, the reader is referred to [6, 14], and [24–26].

4.2 Pre-Echo Artifacts

Typical block sizes used in audio coders range from 128 to 2,048 samples per block. Generally, using long blocks to compute the filter bank output is preferred because it minimizes the relative side-information overhead and allows for efficient encoding of slowly varying, near-stationary signals.

4.2.1 The Problem of Block-Processing

A special situation arises if a time-domain transient occurs (e.g., a castanet signal) during the audio block to be coded. Since there is wide-band signal energy in the attack, a typical perceptual model may allocate only a limited number of bits to the quantizers in each sub-band, assuming enough spectral and temporal masking. In a corresponding decoding stage, the sub-band samples are dequantized, and the quantization noise, which was expected to be fully masked, spreads out in time over the complete block. Because this spread corresponds to the length of the filters, the noise can actually precede the time-domain transient. Specifically, the preceding quantization noise portion will cause audible time domain artifacts when it occurs more than approximately 2 ms before the transient onset, when it exceeds the temporal backward masking threshold. With usual block sizes for audio coding (20–40 ms), this is easily the case, hence the quantization error in the sub-bands is not as localized as the signal itself. The temporal spread of the quantization noise can be up to the length of the impulse response of the synthesis filters.

Fig. 12 illustrates this situation: A castanet transient occurs in the middle of the current block. Therefore its quantization error can be noticed prior to the signal attack as a so-called "pre-echo," which is shown in Fig. 12(c). It can easily be noticed that the quantization error preceding the attack may cause audible artifacts whereas the quantization error during the attack will be masked by the energy of the signal attack.

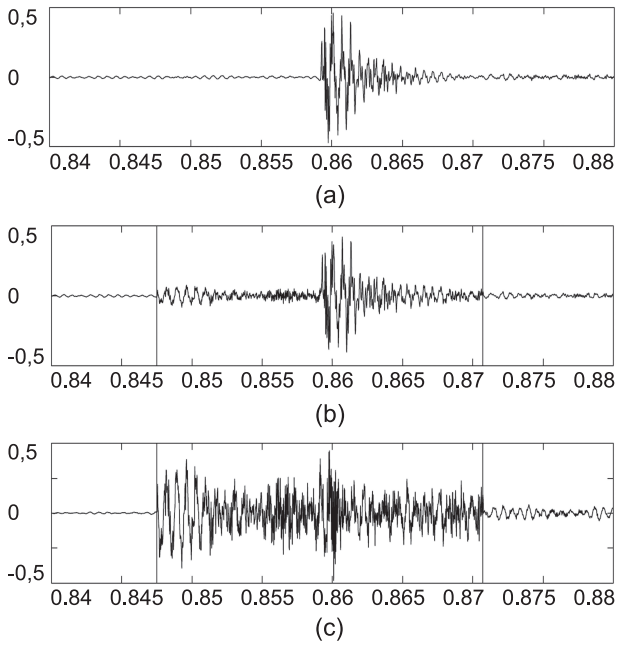


Fig. 12. Castanet signal. (a) Original, (b) encoded/decoded signal, and (c) difference (i.e., quantization error) between the two signals.

The tutorial provides an example with castanets as a particularly critical signal for these pre-echo artifacts because of its very sharp (transient) onsets. Coding this signal with a block size of 2,048 samples results in pronounced pre-echo artifacts.

4.2.2 How Can Pre-Echoes Be Avoided?

As discussed earlier, the amount of temporal spreading of quantization noise depends on the block size (filter length of the synthesis filter bank). Shorter block sizes create less temporal spread than long ones. However, using a short block size for coding all signals creates a large overhead for the side-information and does not allow to take advantage of long-term stationarity in slowly varying signals. Therefore, advanced audio coders make use of perceptual coding approaches such as adaptive block switching [17]. A look-ahead of the energy build-up in the upcoming block allows to detect transients and to switch to a shorter block size during the duration of the transient. For almost stationary signals, which represent the typical operating condition for most audio content, long block sizes are used (provided that the maximum delay permits). This strategy provides both optimal coding efficiency and avoids pre-echo artifacts. Still, for signals with densely spaced transients (like applause), it is difficult to avoid such artifacts entirely because running the coding process constantly with a short block size is inefficient.

4.3 Speech Reverberation Artifacts

If a coder lacks adaptability in its time–frequency resolution, speech signals that naturally include transient events and other nonstationary segments can suffer from an artifact

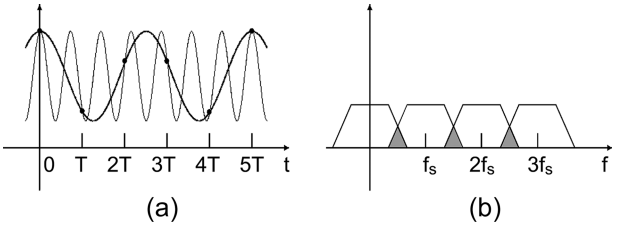


Fig. 13. Aliasing in the time and frequency domains. (a) Time domain: undersampling at interval T causes the actual signal (thin line) to be misrepresented as a lower-frequency alias (thick line). (b) Frequency domain: undersampling causes periodic spectral copies at multiples of f_s to overlap. Shaded regions denote high-frequency fold-back distortion.

known as the "speech reverberation" effect. On a technical level, the cause of this artifact is closely related to previously discussed pre-echo artifacts. However, the resulting speech reverberation artifacts produces a different perceptual effect: for a sharp transient, quantization noise that is temporally smeared before the actual signal onset can be perceived as an extra onset, resembling a (pre-)echo effect. In contrast, speech signals have a temporal fine-structure that consists of multiple periodic transients. Here, temporal smearing of quantization noise can be perceived as a diffusion of the signal's temporal fine-structure, resembling a reverberation effect. *The tutorial offers listening examples of speech reverberation with varying degrees of temporal smearing and artifact severity, as shown above in Fig. 5.*

To overcome the problem of speech reverberation, Temporal Noise Shaping can be applied [15] to flexibly adapt the time–frequency resolution trade-off of the filter bank to the signal also in cases where switching to short transform or filter bank windows would be uneconomic.

4.4 Aliasing Artifacts

Aliasing is well known from the sampling theorem, which states that the sampling frequency must be at least twice the bandwidth of the signal to be sampled [6]. Otherwise, aliasing in the time or frequency domain can occur as illustrated in Fig. 13, which can result in additional signal components that may sound like additional inharmonic overtones or resonances. *The example in the tutorial uses the "Quartet" test signal (i.e., a small choir that makes aliasing artifacts easy to recognize because of familiarity with human voices).*

Since sub-band coders require a set of band-pass filters and these filters are nonideal FIR filters, their passband-ripple and their stopband attenuation will be limited. The MDCT filter bank [13] is a popular example. Although these filters have aliasing cancellation, the cancellation does not apply when the sub-band signals are modified (e.g., by applying quantization or attenuation between the analysis and the synthesis part of the filter bank). Thus, aliasing distortions may occur from the combination of analysis or synthesis filter banks and coding of the spectral coefficients or sub-band samples. In practice, however, these artifacts are

hardly audible for most signal types due to the good selectivity of the filters. Hence, in applications where sub-bands may be strongly processed (like in parametric surround sound or bandwidth extension), pseudo/polyphase quadrature mirror filter banks [11] are used. They are derived from a single prototype filter that is then cosine modulated to form a filter bank of equally spaced band-pass filters. Its prototype filter is usually much longer than for the MDCT; hence, it can get a much better stopband attenuation. This is also needed to minimize aliasing, since only aliasing from directly neighboring sub-bands is canceled. However, the high stopband attenuation also makes it more robust against modifications of sub-band signals that are not immediate neighbors.

4.5 Time-Varying Spectral Artifacts ("Birdies")

A perceptual audio coder includes, by definition, an algorithmic model of the human auditory system [6]. Most perceptual coders estimate masking thresholds to determine, for each frequency band, the maximum noise level that can be introduced while remaining imperceptible.

In addition to such "low-level auditory processing," higher-level mechanisms in the brainstem segment the incoming sound into individual items (auditory objects) and organize them into several streams [27]. For example, a tune with a violin and a flute playing simultaneously contains multiple auditory objects (individual notes) grouped into two streams: one for the violin and one for the flute.

Although there has been considerable research on "Auditory Scene Analysis" [27], researchers are still far from fully understanding how the auditory system identifies instruments from mixtures potentially containing hundreds of instruments, based on only one or two sensor (ear) signals. Basically, the auditory streaming process assembles subsets of the time–frequency components into particular auditory objects, guided by specific time–frequency relations between the time–frequency atoms. Notably, isolated time–frequency components can create a new timbre and new perceived objects can be created "simply" by changing timbre. *To illustrate this effect, the tutorial provides an audio example in which modification (switching on or off) of a harmonic overtone leads to the perception of an additional auditory object.*

These psychoacoustic mechanisms directly explain why certain coder operations can produce unnatural distortions. In perceptual coders, the bit-allocation algorithm selects the best quantizer step size to be used under the constraint that the quantization noise remains inaudible. One possible quantizer is the "zero-quantizer," which sets selected transform coefficients to zero. As the bit assignment varies from block to block, spectral coefficients may temporarily appear or disappear. Because the human auditory system isolates these rapidly fluctuating high-frequency components and streams them into new auditory objects, the listener perceives distinct timbral changes and high-frequency energy fluctuations, commonly known as "birdie" artifacts [18]. Fig. 14 illustrates this effect.

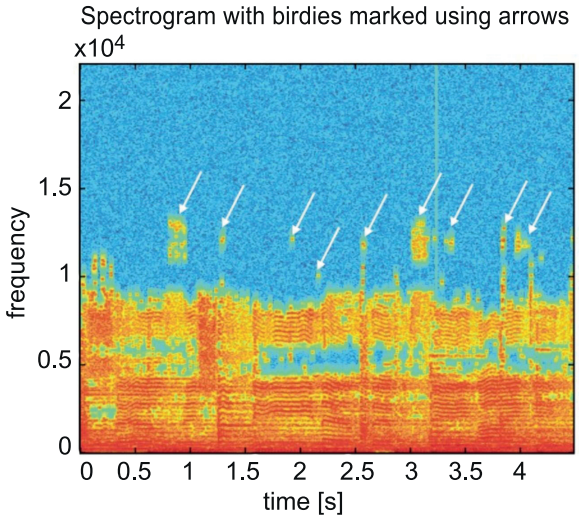


Fig. 14. Spectrogram of a sound excerpt, indicating the "birdies" artifact.

One solution to avoid these types of artifacts is to band-limit the audio signal prior to coding. With increasing band limitation, fewer bits are allocated to high frequencies, and allocation instabilities (birdies) are reduced or eliminated, at the expense of perceived audio bandwidth. *The tutorial discusses the process of auditory streaming and birdies artifacts in detail with several listening examples.*

4.6 Binaural Unmasking Artifacts

To hear spatial imaging artifacts, listeners must stop focusing on the often easier recognizable timbral artifacts. Instead, they should try to allow the stereo signal to construct an overall spatial sound impression. By doing this, they can listen to artifacts and the position of elements within the stereo field, noting both omissions and commissions—new additions to the spatial image that are not present in the reference.

Differences in the spatial arrangement of the elements can cause unmasking. A prominent example of this is the previously mentioned BMLD phenomenon that describes how masking effects can be reduced due to binaural hearing [20]. Sounds that would be masked for monaural hearing can still be detected when additional binaural cues are available. The BMLD phenomenon is observed at low frequencies when a masker and probe have specific time relationships, for example, in the following classic experiment: A narrowband noise masker of one critical bandwidth is presented identically at both ears (i.e., the same time signal), while the phase of the masked sinusoidal probe is alternated (i.e., presented in phase and then out of phase) to the two ears. For the full effect to be noticed, this must be done at or below 500 Hz center frequency, although some effect has been noted up to between 2 and 3 kHz.

When both masker and probe are the same in both ears, a masking threshold very similar to the single-ear masking threshold is observed. In an example provided in the tutorial, the tone probe is inaudible and below the threshold of masking when presented in phase with the masker, because

its level falls below the masking threshold, which is approximately 5.5 dB below the masker level when presented in phase with the masker. However, in the case where only the probe is out of phase, a difference between the signal with and without the probe becomes easily audible at this level. The experiment can also be run with the masker being applied in and out of phase, in which case, a release of the masking threshold of up to approximately 20 dB may be observed in specific cases. *The tutorial provides a number of sound examples with in-phase and out-of-phase probe signals of different levels to demonstrate the effect and enable listener training.*

The BMLD effect has a significant impact on perceptual audio coding of stereo signals because the coding errors of the left and right channels have to be hidden below the binaural masking thresholds that in turn depend on the phase relationships between the stereo channels. Without further measures, separate encoding of left and right channels will produce independent (uncorrelated) error components in both channels that can easily be unmasked for near monophonic input signals [28]. Simply put, two good audio coders that are applied independently to the stereo channel signals are not enough to form a good stereo coder.

For coding of high-quality stereophonic (or multichannel) audio signals at low bitrates, joint coding techniques have proven to be extremely valuable. On one hand, they provide mechanisms to account for binaural psychoacoustic effects; on the other hand, the required bitrate for the stereophonic signals may be reduced significantly below the bitrate for separate coding of the input channels. Common joint stereo coding techniques are M/S stereo coding [21, 28] and Intensity Stereo coding [22, 23].

4.7 Tandem Coding Artifacts

In each perceptual coder, the spectral representation of the input signal is altered slightly by the quantization process. The lower the bitrate, the coarser quantization is required to represent the signal within this given target bitrate. In this way, distortion is introduced that can be thought of as the addition of quantization noise ("coding noise"), which is shaped according to perceptual criteria.

With the increasing deployment of low-bitrate audio coding, audio compression can be used at various stages of media production and distribution (contribution to production/studio, distribution between production facilities, emission/transmission of content to consumers, etc.). Since processing/transmission of audio is still mostly done in uncompressed or even analog representation, this may lead to repeated cycles of decoding, processing, and re-encoding of the audio content. Furthermore, a change of audio coding formats and bitrates usually requires a decoding/re-encoding cycle for format conversion. This effect is called "tandem coding" and can lead to increasing loss in audio quality with each subsequent encoding/decoding cycle ("generation").

To demonstrate this effect, the tutorial includes a sound excerpt in its original version plus encoded/decoded versions up to the tenth tandem coding generation. It can be noticed that—although a single encoding/decoding step is almost perceptually transparent—repeating this process builds up more distortion from generation to generation.

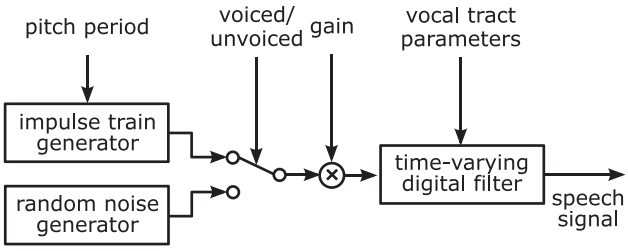


Fig. 15. Speech synthesis model based on linear predictive coding.

There are a few possible solutions. The simplest solution to avoid tandem coding distortions is to avoid re-encoding of content altogether [e.g., by always keeping an uncompressed (or losslessly compressed) version of the original content]. Alternatively, necessary signal manipulations in a content distribution chain (such as level changes or equalization) can sometimes be carried out in the compressed domain, thus avoiding an extra generation of decoding and re-encoding.

If tandem coding is unavoidable in a distribution chain, it is important to plan for sufficient safety margin of the introduced coding distortions toward audibility in each encoding step by using higher bitrates than would be required for one step of perceptually transparent coding. More specifically, in order for the final coded audio to meet a desired target quality, the coding quality at intermediate coding steps must be significantly better than the target quality. In this way, quality losses due to tandem coding can be compensated by increased coding quality (and required bitrate) in intermediate coding steps. Additional concepts for avoiding tandem coding loss include the use of Inverse Decoders [29] instead of conventional re-encoding and the use of MOLE-data and ancillary data to convey previous encoder parameters in digital interfaces [30].

4.8 Artifacts From Audio Over Speech Codecs

Although sub-band coding techniques have also been applied to speech coding, many traditional speech coders are based on a model of the speech generation process of the human vocal tract. A simple vocoder based on linear predictive coding [31] as shown in Fig. 15 makes use of a time-varying all-pole filter with either an impulse train (tonal) or a random noise excitation. Since this system cannot synthesize tonal and noise-like components simultaneously, it will sound unnatural when applied to broadband generic audio signals.

Code Excited Linear Prediction [32] is another commonly used speech coding system based on an excitation codebook, a long-term prediction, and a time-varying postfilter. In general, speech coders offer a rather limited quality for generic wide-band audio signals (music), due to the following reasons:

- Their underlying model is based on the generation of a single pitch period at a time and thus cannot represent polyphonic music accurately. Also, reproduction of a single pure sinusoid (like from a solo musical instrument) is beyond the model's capabilities.
- Shaping of the quantization noise is based on overly simple perceptual criteria (i.e., the weighted mean square error).
- All coding parameters are optimized for speech.

To illustrate the perceptual characteristics of these effects, the tutorial provides a range of audio signals (speech, singing, music) in their original and narrow-band speech coded versions. Most prominently, the noise-like distortion of tonal monophonic or polyphonic sound sources can be noticed.

4.9 Parametric Coding Artifacts

Modern perceptual audio coders that target low bitrates are usually equipped with a set of coding tools that enable an efficient parametric (and semiparametric) representation of the signal (i.e., they describe the signal by a sparse set of abstract parameters rather than using a time or frequency representation). The idea behind these tools is to recreate the signal's perceptual properties rather than to restore its original waveform. In this way, good quality can be achieved even at very low bitrates (for instance below 64 kb/s). However, due to their non–waveform-preserving nature, these tools typically do not scale well toward transparent quality. Therefore, improper use of such tools can lead to characteristic artifacts even at higher bitrates. *To showcase such effects, the tutorial provides examples where signal components have been reconstructed to various extent by parametric coding techniques.*

4.9.1 Artifacts From Perceptual Noise Substitution

Perceptual Noise Substitution [33, 34], as used in MPEG-4 Advanced Audio Coding (AAC), makes use of the human hearing's inability to distinguish between noise-like signals that have similar spectro-temporal envelopes. Therefore, noise-like signal components can be resynthesized in the decoder, based only on a small amount of side information with a substantially lower bitrate than that required for transmitting the quantized spectral components. However, when this technique is improperly applied to substitute tonal signal components by noise, audible signal degradations can occur.

4.9.2 Bandwidth Extension Artifacts

Bandwidth extension [16] is a semiparametric audio coding technique that has become an essential ingredient of all modern perceptual audio coders that operate at lower bitrates (for instance below 32 kb/s). It provides full audio bandwidth even at those low data rates by parametrically reconstructing the high-frequency portion of the output signal from the transmitted low-frequency part and some transmitted compact parametric side information describing the properties of the original high-frequency signal as illustrated in Fig. 16. However, when an improper parametrization or reconstruction method is employed, an audible mismatch of the harmonic or temporal structure of the reconstructed high-frequency signal portion can occur, as illustrated in Fig. 17.

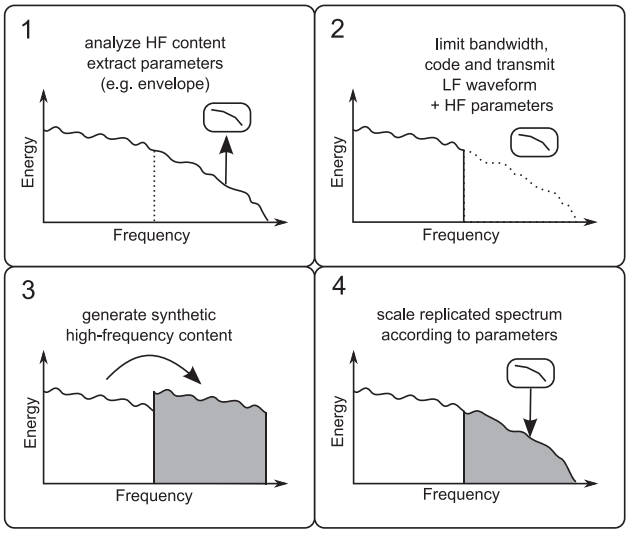


Fig. 16. Illustration of working principle of bandwidth extension.

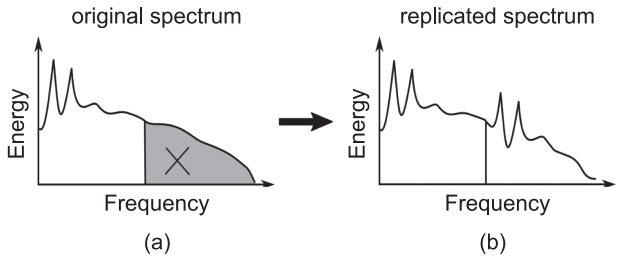


Fig. 17. Illustration of bandwidth extension resulting in tonality mismatch artifacts. (a) Original spectrum, where the cross (X) indicates the smooth, noise-like high-frequency content that is discarded by the encoder to save bitrate. (b) Replicated spectrum at the decoder. Because the tonal low-frequency components were naively copied into the high-frequency band, the reconstructed signal contains artificial tonal peaks that mismatch the original noise-like high frequencies.

4.9.3 Parametric Stereo and Multi-Channel Coding

Parametric spatial coding methods (such as Intensity Stereo [21, 23] and Parametric Stereo [35, 36]) and parametric multichannel coding (such as MPEG-Surround [37]) have become another essential ingredient of all modern perceptual audio coders that carry stereo or multichannel signals at very low bitrates. They provide spatial output from one (or a few) transmitted audio channels and some transmitted compact parametric side information that describes the spatial properties of the original signal. However, due

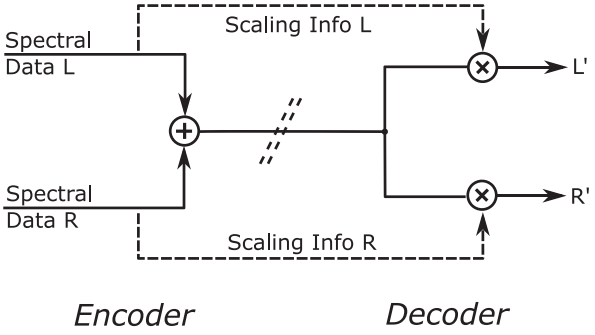


Fig. 18. Principle of Intensity Stereo coding.

to the parametric reconstruction of the stereo signal, audible changes in the stereo image and timbral artifacts may be introduced. In the following, such effects are first illustrated for the classic case of Intensity Stereo Coding between two channels and then discussed for more recent advanced parametric spatial audio coding techniques that generalize from two channels to multichannel reproduction on loudspeakers.

4.9.3.1 Intensity Stereo and Loss of Stereo Image. A straightforward approach for parametric stereo coding is Intensity Stereo (see Fig. 18) [21]. It exploits the fact that the perception of high-frequency sound components (e.g., above 4 kHz) mainly relies on the analysis of their energy-time envelopes [22] rather than the waveform itself. Thus, it is assumed to be sufficient to code the envelope of such a signal instead of its waveform. This is done by transmitting only one down-mixed mono signal instead of the full stereo signal. In the decoder, this is up-mixed again by scaling the signal individually for each signal channel to match its original average envelope (or signal energy) for the respective coder frame. The scaling information is calculated and transmitted in a lower resolution for a group of spectral coefficients (scale factor bands). Effectively, the stereo image is recreated at the decoder side by a pan-pot–like operation for each spectral coder band using amplitude panning, a process conceptually similar to a panoramic potentiometer, or "pan-pot," on a mixing console.

As a consequence of the Intensity Stereo coding process, all output signals are scaled versions of each other [i.e., they share the same envelope fine structure for the duration of the coded block (e.g., 10–20 ms)]. This does not present a major problem for stationary signals or for signals that have similar envelope fine structures in the Intensity Stereo–coded channels. For transient signals with dissimilar envelopes across different channels, however, the original distribution of the envelope onsets between the coded channels cannot be recovered. This effect can be observed for the example of an applause signal shown in Figs. 19 and 20, with arrows indicating affected transients. As a consequence, the perceived width of the sound stage decreases and image instabilities may occur for such critical signals.

4.9.3.2 Further Parametric Spatial Audio Coding Artifacts. In addition to the potential loss of stereo image width, spatial or multichannel coding techniques may introduce artifacts when reproducing complex 3D or surround material. Rapidly changing spatial parameters often require temporal smoothing in the rendering algorithm (e.g., first-order low-pass filters of 10–100 ms), leading to a trade-off between preserving rapid changes in the perceived location of auditory objects and maintaining spatial stability [38]. Failure to match fast signal variations (e.g., applause) can cause auditory images to migrate toward the center of the spatial field or to shift unpredictably, with noticeable effects in specific frequency bands.

Another potential artifact in parametric stereo or multichannel coding arises from the generation of diffuse decorrelated sound. Unlike in Intensity Stereo coding, decorrelation algorithms are employed to render a wide sound image (e.g., convolution with a short noise burst). This processing can distort the quality of fine-grained impulse-like applause signals if the time resolution is not sufficiently high [39].

Finally, when listening to a coded/decoded stereo or multichannel signal outside the optimal "sweet spot," a listener

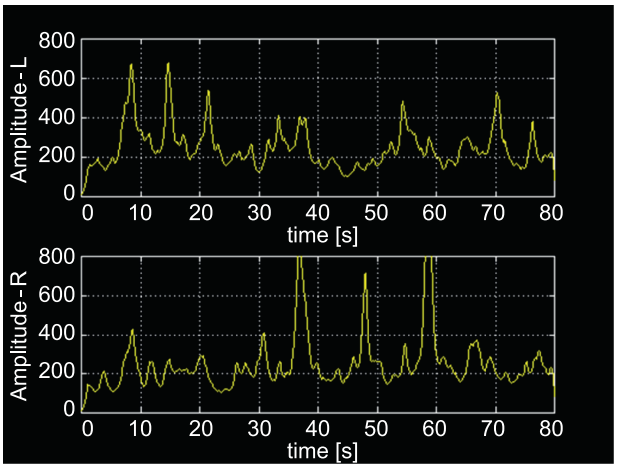


Fig. 19. Excerpt from an "applause" item, showing left (top panel) and right channel (bottom panel) high frequency signal envelopes.

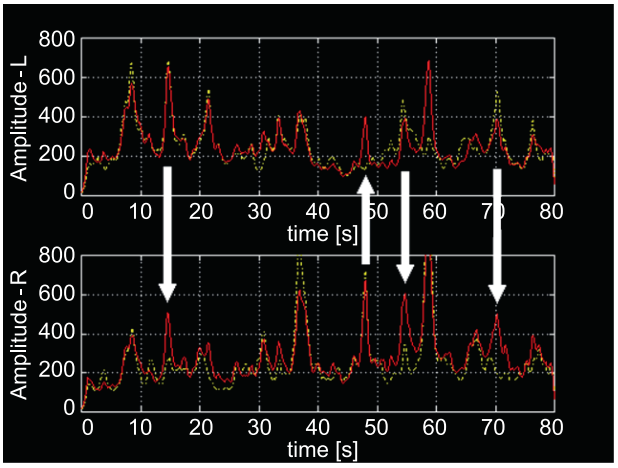


Fig. 20 Excerpt from "applause" item, showing left (top panel) and right channel (bottom panel) high frequency envelope after Intensity Stereo encoding/decoding (yellow: original; red: intensity coded).

might experience artifacts that result from the violation of masking assumptions made in the coder's perceptual model or processing (e.g., a certain amount of masking between loudspeaker signals). Consequently, processing artifacts may become revealed that are normally inaudible when listening at or near the sweet spot.

5 SUBJECTIVE EVALUATION AND LISTENER TRAINING

While the focus of the tutorial is on acquainting the reader with the background and perceptual qualities of various artifact types in perceptual audio coding, it is also important to understand the fundamentals of audio quality evaluation, as comprehensive testing must be undertaken in practice to assess a codec's fitness for an intended application (e.g., a wireless streaming service). As audio signals are intended primarily for humans, a subjective evaluation of audio quality remains the golden standard, carried out in the form of listening tests directly involving humans and the human auditory system. As a note, objective measurement technology [40, 41] can provide helpful tools but is not (yet) capable of fully substituting for a listening test. Different types of listening tests are used to answer critical perceptual questions, such as determining which codec provides superior quality, whether there are audible differences between the coded signal and the reference (original), and how natural a coding system sounds.

5.1 Fundamentals in Subjective Evaluation

There are two fundamentally different types of listening tests used to gather reliable subjective data:

- Preference tests involve direct comparisons of two or more stimuli, and participants are asked to identify which stimulus they prefer (i.e., to answer the question "which do you like better?").
- Auditory measurement tests assess how much a certain auditory attribute differs between two or more stimuli (i.e., they answer the question "how different is this?"). The auditory attribute can be either very broad (e.g., Basic Audio Quality, which includes both timbre and spatial aspects) or more specific (e.g., brightness or dullness).

In both cases, the tutorial presents different signals ("stimuli") to the listener ("subject"), and a qualitative scale or a quantitative measure is chosen to express the subject's perception.

It is important to understand the difference between preference testing and auditory measurement also in the context of audio coding. Usually, the quality of compressed audio is assessed by measuring how much the processed signal differs from the original reference signal, meaning that only the reference, or signals perceptually indistinguishable from it, receives the maximum grade. If a preference test is used, on the other hand, it may very well happen that a perceptually coded version of a music signal sometimes receives a higher preference than the original since it may sound smoother or less harsh than the original mix.

Another important aspect of well-designed listening test procedures is that they are designed to minimize the influence of undesired factors:

- A mental bias toward a certain direction (e.g., "coded audio always sounds bad") can be avoided by a blind presentation of the signals (i.e., the subject is not informed which signal is being played at a given time). Furthermore, a so-called "hidden reference" is sometimes introduced as one of the test signals, duplicating the reference signal. If a subject grades the hidden reference very low, it is clear that the subject is unreliable (e.g., because of mental bias or a lack of listening capability, and the results are discarded).
- Since it is known that the quality of the previously presented signal has some influence on the quality rating of the subsequent stimulus (good stimuli are usually rated even better after a previous bad stimulus), the order of presentation is usually randomized individually for each subject.
- It has also been found that the usage of the scale range depends on the presence/absence of stimuli with large degradations (i.e., when everything is of high quality, some listeners tend to become more critical). Thus, sometimes the so-called "anchor signals" are added that exhibit a well-defined degree of degradation and thus stabilize the use of the grading scale.

While some listening test methods prescribe the use of untrained ("naive") test listeners to better represent typical consumer perception, many test methods in the field of audio recommend or even mandate trained ("expert") listeners [42]. Having undergone many hours of listening experience, trained listeners excel in identifying nuanced differences and are essential for tasks requiring high precision or repeated measurements, such as product optimization [6, 42].

5.2 Methods for Evaluating Audio Quality

Historically, advancements in audio coding technology and the quest for higher quality at lower data rates have consistently involved the creation of suitable subjective quality evaluation methods. Numerous methods have been developed to cater to various application domains [43–45]. Most methods for testing the quality of perceptual audio codecs adhere to the 'auditory measurement test' paradigm. A well-established and popular example is the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) method, described in the ITU-R Recommendation BS.1534 [44]. Since 2001, the MUSHRA recommendation has seen several improvements and has become one of the most widely used approaches to evaluate audio quality [42].

The MUSHRA method aims to compare and rank the quality performance of different audio systems while ob-

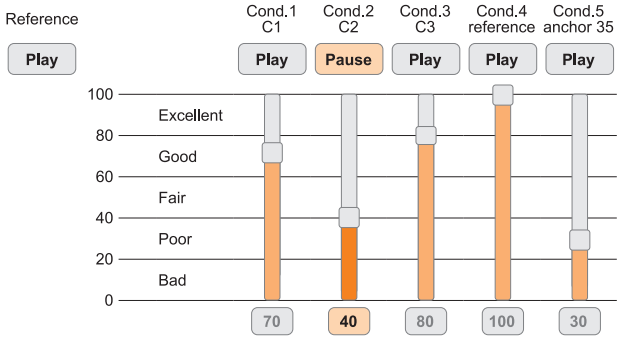


Fig. 21. Example of a MUSHRA test GUI. Source: https://www.audiolabs-erlangen.de/resources/webMUSHRA.

taining absolute quality scores using a 0–100 scale divided into five intervals: "Bad," "Poor," "Fair," "Good," and "Excellent." An open (known) reference establishes the expected maximum quality score at 100 points, and the overall quality grading scale is stabilized by predefined anchor conditions. During the test, listeners are presented with multiple unlabeled conditions corresponding to different treatments of the original signal (an "audio item"), stemming from different coding systems, in random order. An example of a MUSHRA presentation screen can be seen in Fig. 21. The method also specifies two possible anchor conditions: the original signal processed by a low-pass filter at 7 kHz (mid-quality anchor) and 3.5 kHz (low-quality anchor), and a hidden reference (i.e., the original signal) included among the conditions. Subjects can freely switch between all conditions and the open reference, comparing each to the open reference and to one another using the quality scale. If a listener rates the hidden reference below 90 points in more than 15% of trials, the listener data are deemed unreliable and are excluded from further statistical analysis through a process called postscreening.

5.3 Selection of Test Material

A coding algorithm for general audio is expected to deliver robust performance across a wide range of real-world audio signals. It is expected that, at a given bitrate, a codec will maintain a rather consistent overall level of quality. Still, performance may vary within a certain quality range across different signal characteristics, and there can be specific test signals that are especially challenging for an algorithm. Randomly selecting a limited set of test items will potentially lead to test results that do not accurately reflect the codec's performance for the intended application. This is generally addressed through experience-based manual selection of the material. Specifically, the purpose is to ensure that the evaluation comprehensively represents the targeted use cases and relevant audio content types. This usually involves the following:

- Coverage of diverse material, ensuring that the test items reflect a wide range of audio content, rather than, for example, only a specific music style, in order to evaluate the codec performance effectively across different scenarios or within a specific application.
- Including *challenging* ("critical") audio material to evaluate the worst-case performance of the system. Examples are transient sounds (e.g., castanets, applause) that are known to cause pre-echo artifacts and stationary tones (e.g., pitch pipe) or combinations of both (e.g., glockenspiel) and voice recordings with difficult utterances or speakers (e.g., a German male speech sample identified as critical in AAC core experiments [6, 46]).
- Inclusion of audio material that is *relevant* to and *typical* for the intended application (e.g., broadcasting). This may include speech and music with background noise, announcements, advertisements, and narration [47].

In the current era of AI-based codec development (i.e., codecs that make use of generative audio and speech deep-learning models [48]), it is important to emphasize that even for these newer codec types, the listening test material must adequately represent the diversity of audio signals, as outlined above. If the majority of excerpts in such a test set consist of noncritical audio material (which may be necessary for training the AI-based neural codec), it will not effectively stress-test the coding scheme and thus not allow one to draw conclusions about the codec's actual performance and shortcomings.

5.4 Statistical Analysis of Subject Responses

ITU-R Recommendation BS.1534 [44] outlines methods for the statistical analysis of subject responses, recommending parametric statistics when test results the residuals (i.e., the deviations of the individual listener scores from the fitted model) can be assumed to follow a Gaussian distribution. This and other basic assumptions are reasonably satisfied. Reporting of the results typically involves the calculation of mean scores, the standard deviation and a 95% confidence interval to assess the reliability of the results. The confidence interval depends on the number of listeners and decreases as the number of listeners increases.

An analysis of variance (ANOVA) [42, 49] procedure can identify significant factors such as the item, system, and listener impact on the quality ratings. However, it has been shown that MUSHRA data can violate several key ANOVA assumptions (interval scale, normality, homogeneity of variances, and independence), which can substantially increase the risk of type I errors (false positives) [50]. In particular, homogeneity of variances and normality can be strongly affected by ceiling effects for the hidden reference and floor effects for the anchor. In addition, clearly skewed score distributions at intermediate quality levels can influence the validity of the ANOVA procedure [44]. These deviations can occur for several reasons. For example, listeners may use the rating scale in a compressed or otherwise nonuniform way (contraction bias [51]), the tested stimuli may not cover the full quality range evenly (stimulus bias [51]), or different listeners may focus on different parts or

attributes of the sound (multimodality [50, 51]). Other contributing factors are also possible (see, e.g., [51]). For such data, nonparametric methods (e.g., Friedman or Wilcoxon tests) or resampling-based approaches are generally recommended as more robust alternatives [42, 50]. For a robust statistical analysis, between eight and 20 trained listeners are typically sufficient [44], and overlapping confidence intervals suggest that differences between systems are small relative to listener variability (or that the number of listeners is inadequate) [44, 52]. In light of these diverse risks to statistical validity, exploratory analysis of the data is indispensable to identify possible underlying problematic patterns and to validate whether the data satisfy the prerequisites for the intended statistical model.

5.5 Tutorial Section on Audio Content and Subjective Evaluation

The "What To Listen For" tutorial contains a section on "Audio Content and Evaluation" that provides listening test results of a MUSHRA test for a series of the tutorial's sound examples, in order to guide the reader on how professional expert listeners would grade such signals and in this way to help them better understand subjective evaluation. The listening test panel used in the tutorial's test included nine participants, comprising five females and four males, aged 21–39, with an average age of 26 years. Audio playback was carried out using a PC with a digital I/O card, external DACs, and STAX Lambda Pro headphones. The listening tests had been conducted for the original AES CD-ROM back in 2001. None of the listener results were deemed unreliable at the time, since the MUSHRA recommendation had not yet defined a postscreening process. When revisiting the results using current criteria in [44], it was found that several listeners' data would be rejected. Further details can be found in the corresponding tutorial section.

6 CONCLUSION

This paper introduces the freely available Free Web Edition of the "Perceptual Audio Coders – What To Listen For" educational package, prepared by the AES TC-CAS. The tutorial covers a plethora of potential artifact types, including their theoretical background and sound examples of different artifact strengths. The described artifacts are not specific to any particular coding scheme but also include recent advanced coding strategies and allow for a general understanding of possible effects. The tutorial comes with attractive presentation features and can be viewed online or downloaded as a package.

Although the tutorial's focus was chosen to be on artifacts, this by no means suggests that perceptual audio coding will always create audible artifacts. On the contrary, advances in time–frequency signal processing and perceptual modeling have led to coding schemes that, at sufficiently high bitrates, can achieve perceptual transparency (i.e., differences that are not detectable under controlled listening conditions as defined in standardized listening tests [6, 43]). It is, however, this flexibility of choosing the bitrate that is frequently used by service providers to squeeze as many audio streams as possible into their transmission channels while accepting mediocre quality. Naturally, this can lead to a bad reputation for perceptual audio coding among the public. It should be kept in mind that even the most advanced and established coding scheme (be it perceptual or not) can be driven to provide bad-sounding output when the bitrate is reduced progressively below its recommended range.

It was the ultimate goal of this tutorial project to provide universities, students, researchers, recording engineers, performing artists, musicians, publishers, journalists, and the interested reader with a way to easily enter the realm of audio coding. The project is certainly unique in that, so far, never in the history of audio coding has such a vast body of experience from the world's leading researchers been accumulated, resulting in a multimedia project that offers fundamental information, illustrations, and many unique audio demonstrations.

7 ACKNOWLEDGMENT

The tutorial was created by members of the AES TC-CAS, including (in alphabetical order) Antonio Pena, Arijit Biswas, Bernd Edler, Chris Lanciani, Christof Faller, Frederik Nagel, Gerald Schuller, Heiko Purnhagen, James D. Johnston, Jan Skoglund, Jürgen Herre, Karlheinz Brandenburg, Marina Bosi, Pablo Delgado, Sascha Dick, Schuyler Quackenbush, Scott Kramer, and Thomas Sporer. Special thanks go to Markus Erne as the editor of the first-generation tutorial, Sascha Dick as the editor of the second generation, and Schuyler Quackenbush and James D. Johnston as former co-Chairs of the AES TC-CAS.

8 REFERENCES

[1] M. Erne, "Perceptual Audio Coders – What to Listen For," in Proc. *111th Conv. Audio Eng. Soc.* paper 5489:1–10, New York, NY, USA (2001 Nov.).

[2] H. Fletcher, "Auditory Patterns," *Rev. Mod. Phys.*, vol. 12, pp. 47–55 (1940 Jan.), https://doi.org/10.1103/RevModPhys.12.47.

[3] D. Greenwood, "Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane," *J. Acous. Soc. Am.*, vol. 33, pp. 1344–1356 (1961 Oct.), https://doi.org/10.1121/1.1908437.

[4] B. Scharf, "Critical Bands," in J. V. Tobias (Ed.), *Foundations of Modern Auditory Theory*, pp. 159–202 (Academic Press, New York, NY, USA, 1970).

[5] E. Zwicker, *Psychoacoustics: Facts and Models* (Springer, Berlin, Germany, 1990), https://doi.org/10.1007/978-3-540-68888-4.

[6] M. Bosi and R. E. Goldberg, *Introduction to Digital Audio Coding and Standards* (Springer, Boston, MA, USA, 2003), https://doi.org/10.1007/978-1-4615-0327-9.

[7] N. Jayant, J. Johnston, and R. Safranek, "Signal Compression Based on Models of Human Perception," *Proc. IEEE*, vol. 81, no. 10, pp. 1385–1422 (1993 Oct.), https://doi.org/10.1109/5.241504.

[8] M. Bosi, "Overview of Perceptual Audio Coding," *IEEE Signal Process. Mag.*, vol. 14, no. 5, pp. 43–49 (1997 Sep.).

[9] R. Hellman, "Asymmetry of Masking Between Noise and Tone," *Percep. Psychphys.*, vol. 11, pp. 241–246 (1972 May), https://doi.org/10.3758/BF03206257.

[10] J. H. Rothweiler, "Polyphase Quadrature Filters – A New Subband Coding Technique," in Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1280–1283, Boston, MA, USA (1983 Apr.), https://doi.org/10.1109/ICASSP.1983.1172005.

[11] H. Nussbaumer, "Pseudo QMF Filterbanks," *IBM Tech. Discl. Bull.*, vol. 24, pp. 3081–3087 (1981 Nov.).

[12] J. P. Princen and A. B. Bradley, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 5, pp. 1153–1161 (1986 Oct.), https://doi.org/10.1109/TASSP.1986.1164954.

[13] J. P. Princen, A. W. Johnson, and A. B. Bradley, "Subband/Transform Coding Using Filter Bank Designs Based on Time Domain Aliasing Cancellation," in Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 2161–2164, Dallas, TX, USA (1987 Apr.), https://doi.org/10.1109/ICASSP.1987.1169405.

[14] G. Schuller, *Filter Banks and Audio Coding: Compressing Audio Signals Using Python* (Springer, Cham, Switzerland, 2020), https://doi.org/10.1007/978-3-030-51249-1.

[15] J. Herre, "Temporal Noise Shaping, Quantization and Coding Methods in Perceptual Audio Coding – A Tutorial Introduction," in Proc. *17th AES Int. Conf. High Quality Audio Coding*, pp. 312–325, Florence, Italy (1999 Sep.).

[16] P. Ekstrand, "Bandwidth Extension of Audio Signals by Spectral Band Replication," in Proc. *1st IEEE Benelux Workshop on Model Based Processing and Coding of Audio (MPCA)*, pp. 73–79, Leuven, Belgium (2002 Nov.).

[17] B. Edler, "Coding of Audio Signals With Overlapping Transform and Adaptive Window Functions," *Frequenz*, vol. 43, no. 9–10, pp. 252–256 (1989 Sep.), https://doi.org/10.15488/3111.

[18] M. Bosi and G. A. Davidson, "High-Quality, Low-Rate Audio Transform Coding for Transmission and Multimedia Applications," in Proc. *93rd Conv. Audio Eng. Soc.*, paper 3365:1–21, San Francisco, CA, USA (1992 Dec.).

[19] B. C. J. Moore, *An Introduction to the Psychology of Hearing* (Academic Press, Orlando, FL, USA, 1982).

[20] J. Blauert, *Spatial Hearing* (MIT Press, Cambridge, MA, USA, 1983).

[21] J. Johnston and A. Ferreira, "Sum-Difference Stereo Transform Coding," in Proc. *IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 569–571, San Francisco, CA (1992 Mar.), https://doi.org/10.1109/ICASSP.1992.225993.

[22] R. Veldhuis and R. van der Waal, "Subband Coding of Stereophonic Digital Audio Signals," in Proc. *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 3601–3604, Toronto, Canada (1991 Apr.), https://doi.org/10.1109/ICASSP.1991.151053.

[23] J. Herre, K. Brandenburg, and D. Lederer, "Intensity Stereo Coding," in Proc. *96th Conv. Audio Eng. Soc.* paper 3799:1–10, Amsterdam, the Netherlands (1994 Apr.).

[24] K. Brandenburg, "MP3 and AAC Explained," in Proc. *17th AES Int. Conf. High Quality Audio Coding*, pp. 99–110, Florence, Italy (1999 Sep.).

[25] M. Bosi, "Filter Banks in Perceptual Audio," in Proc. *17th AES Int. Conf. High Quality Audio Coding*, pp. 125–136, Florence, Italy (1999 Sep.).

[26] M. Erne, "Digital Audio Compression Algorithms," in Proc. *1st COST-G6 Workshop Digital Audio Effects*, pp. 99–110, Barcelona, Spain (1998 Nov.).

[27] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA, USA, 1990).

[28] J. Herre, E. Eberlein, and K. Brandenburg, "Combined Stereo Coding," in Proc. *93rd Conv. Audio Eng. Soc.*, paper 3369:1–18, San Francisco, CA, USA (1992 Oct.).

[29] J. Herre and M. Schug, "Analysis of Decompressed Audio – The Inverse Decoder," in Proc. *109th Conv. Audio Eng. Soc.*, paper 5256:1–23, Los Angeles, CA, USA (2000 Sep.).

[30] J. Fletcher, "ISO/MPEG Layer-2 – Optimum Re-Encoding of Decoded Audio Using a MOLE Signal," in Proc. *104th Conv. Audio Eng. Soc.*, paper 4706:1–20, Amsterdam, the Netherlands (1998 May).

[31] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580 (1975 Apr.), https://doi.org/10.1109/PROC.1975.9792.

[32] B. Atal and M. Schroeder, "Stochastic Coding of Speech Signals at Very Low Bit Rates," in Proc. *IEEE Int. Conf. Communications*, pp. 1610–1613, Amsterdam, the Netherlands (1984 May).

[33] J. Herre and D. Schulz, "Extending the MPEG-4 AAC Codec by Perceptual Noise Substitution," in Proc. *104th Conv. Audio Eng. Soc.*paper 4720:1–14, Amsterdam, the Netherlands (1998 May).

[34] J. Herre and S. Disch, "Perceptual Audio Coding," in R. Chellappa and S. Theodoridis (Eds.), *Academic Press Library in Signal Processing, Volume 4*, pp. 757–799 (Academic Press, Oxford, UK, 2013), https://doi.org/10.1016/B978-0-12-396501-1.00028-5.

[35] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, "Low-Complexity Parametric Stereo Coding," in Proc. *116th Conv. Audio Eng. Soc.*, paper 6072:1–11, Berlin, Germany (2004 May).

[36] J. Herre and M. Dietz, "MPEG-4 High-Efficiency AAC Coding [Standards in a Nutshell]," *IEEE Signal Process. Mag.*, vol. 25, no. 3, pp. 137–142 (2008 May), https://doi.org/10.1109/MSP.2008.918684.

[37] J. Herre, K. Kjörling, J. Breebaart, et al., "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 Nov.).

[38] M. Takanen, O. Santala, and V. Pulkki, "Binaural Assessment of Parametrically Coded Spatial Audio Signals," in J. Blauert (Ed.), *The Technology of Binaural Listening*, pp. 333–358. (Springer, Berlin, Germany, 2013), https://doi.org/10.1007/978-3-642-37762-4_13.

[39] S. Disch and A. Kuntz, "A Dedicated Decorrelator for Parametric Spatial Coding of Applause-Like Audio Signals," in A. Heuberger, G. Elst, and R. Hanke (Eds.), *Microelectronic Systems: Circuits, Systems and Applications*, pp. 363–371 (Springer, Berlin, Germany, 2011), https://doi.org/10.1007/978-3-642-23071-4_34.

[40] M. Chinen, F. A. Kelly, N. Waller, J. McGee, and J. Skoglund, "ViSQOL v3: An Open Source Production-Ready Objective Speech and Audio Metric," in Proc. *12th Int. Conf. Quality of Multimedia Experience (QoMEX)*, pp. 1–6, Athlone, Ireland (2020 May), https://doi.org/10.1109/QoMEX48832.2020.9123150.

[41] P. M. Delgado and J. Herre, "Towards Improved Objective Perceptual Audio Quality Assessment – Part 1: A Novel Data-Driven Cognitive Model," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 4661–4675 (2024 Oct.), https://doi.org/10.1109/TASLP.2024.3477291.

[42] S. Bech and N. Zacharov, *Perceptual Audio Evaluation: Theory, Method and Application* (John Wiley & Sons, Chichester, UK, 2007), https://doi.org/10.1002/9780470869253.

[43] ITU-R, *Recommendation BS.1116-3: Methods for the Subjective Assessment of Small Impairments in Audio Systems* (2015).

[44] ITU-R, *Recommendation BS.1534-3: Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems* (2015).

[45] ITU-T, *Recommendation P.800: Methods for Subjective Determination of Transmission Quality* (1996).

[46] EBU, "Sound Quality Assessment Material for Stereo Sound Systems (SQAM)," Tech. Rep. 3253 (2008 Sep.).

[47] ISO/IEC, "Report on the Verification Test of MPEG-4 Enhanced Low Delay AAC N10032," Tech. Rep. JTC 1/SC 29/WG 11 (2008 Jul.).

[48] M. Kim and J. Skoglund, "Neural Speech and Audio Coding: Modern AI Technology Meets Traditional Codecs," *IEEE Signal Process. Mag.*, vol. 41, no. 6, pp. 85–93 (2024 Nov.), https://doi.org/10.1109/MSP.2024.3444318.

[49] D. C. Montgomery, *Design and Analysis of Experiments* (John Wiley & Sons, Hoboken, NJ, USA, 2020).

[50] C. Mendonça and S. Delikaris-Manias, "Statistical Tests With MUSHRA Data," in Proc. *144th Conv. Audio Eng. Soc.*, pp. 23–26, Milan, Italy (2018 May).

[51] S. Zielinski, F. Rumsey, and S. Bech, "On Some Biases Encountered in Modern Audio Quality Listening Tests – A Review," *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451 (2008 Jun.).

[52] F. Nagel, T. Sporer, and P. Sedlmeier, "Toward a Statistically Well-Grounded Evaluation of Listening Tests – Avoiding Pitfalls, Misuse, and Misconceptions," in Proc. *128th Conv. Audio Eng. Soc.* paper 8146:1–10, London, UK (2010 May).

THE AUTHORS


Jürgen Herre


Sascha Dick


Arijit Biswas


Marina Bosi


Pablo M. Delgado


Gerald Schuller


Markus Erne

Jürgen Herre joined the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany, in 1989. Since then, he has been involved in the development of perceptual coding algorithms for high-quality audio, including the well-known ISO/MPEG-Audio Layer III coder ("MP3"). After his postdoctoral term at Bell Laboratories in 1995–1996, where he worked on MPEG-2 AAC, he returned to Fraunhofer IIS developing more advanced technology including MPEG-4, MPEG-7, MPEG-D, MPEG-H, and MPEG-I, currently as the Chief Executive Scientist for the Audio/Media Technologies division at Fraunhofer IIS. Since 2010, Dr. Herre has been a professor at the University of Erlangen and International Audio Laboratories Erlangen. He is a Fellow of the AES, Chair of the AES TC-CAS, has served as Chair and Vice Chair of the AES Technical Council for 25 years, and was an active member of the MPEG audio standardization group for almost 30 years.

•

Sascha Dick received his Dipl.-Ing. degree in Information and Communication Technologies from the Friedrich Alexander University (FAU) of Erlangen-Nuremberg, Germany, in 2011 with a thesis on an improved psychoacoustic model for spatial audio coding and joined the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen, Germany, as a Research Engineer in the same year. He has contributed to the development and standardization of audio coders such as MPEG-H 3D Audio. In 2017, he joined the International Audio Laboratories at FAU to research psychoacoustic effects for 3D audio and received a Ph.D. degree for his thesis on *Psychoacoustic Effects and Models for Processing and Coding of 3-Dimensional Audio* in 2023. He has since joined Fraunhofer IIS again to further pursue research and development of 3D-audio applications. His research interests include psychoacoustics, multichannel signal processing, and audio coding.

•

Arijit Biswas received a Ph.D. degree in audio coding from Technische Universiteit Eindhoven, the Netherlands, in 2007, following doctoral research at Philips Research Laboratories. He joined Coding Technologies in Nürnberg, Germany, later acquired by Dolby Laboratories, where he is currently a Senior Staff Researcher. He has contributed to audio coding and processing technologies including high-efficiency AAC, Dolby Digital Plus, Dolby AC-4, and Dolby Atmos. His research focuses on deep learning for audio coding, coder enhancement, and objective audio quality assessment. Dr. Biswas is a Senior Member of the IEEE and serves on the AES Technical Committees on Coding of Audio Signals and on Machine Learning and Artificial Intelligence. He received the Best Student Paper Award at the 121st AES Convention and was honored at the 2021 Primetime Emmy Awards for contributions to an Emmy-winning engineering achievement presented to Dolby Laboratories.

•

Marina Bosi, a pioneer in digital audio coding, is a founding director of Moving Picture, Audio and Data Coding by Artificial Intelligence (MPAI) and Chair of the IEEE Standards Association Context-based Audio Enhancement (CAE) Working Group. A past president of the AES, she has served as Chief Technology Officer of MPEG LA, Vice President of Technology at DTS, and was a member of the Dolby Laboratories research team that created Dolby Digital. She also led the development of MPEG-2 AAC, earning the ISO/IEC 1997 Editor Award. Dr. Bosi holds degrees in physics (University of Florence; dissertation at IRCAM, Paris) and from the Conservatory of Florence. A graduate of the Stanford Business School's Executive Program, she has held fiduciary roles on several boards. Dr. Bosi holds multiple patents, has authored influential publications, including *Introduction to Digital Audio Coding and Standards*, and has received numerous honors, including the AES Silver Medal.

•

Pablo Delgado studied electrical engineering at the University of Buenos Aires and the École nationale supérieure d'ingénieurs de Caen. He joined Fraunhofer Institute for Integrated Circuits (IIS) in 2014 to work in research and development of audio coding algorithms within numerous industry projects. From 2018 to 2022, he was a researcher at the International Audio Laboratories in Erlangen, Germany, where he received his doctoral degree (Dr.-Ing.) in 2023. Dr. Delgado currently works as a Senior Scientist at the Advanced Audio Research Group in Fraunhofer IIS, where he focuses on the intersection of psychoacoustics and machine learning, spatial audio, and virtual reality.

•

Gerald Schuller has been a full professor at the Institute for Media Technology, Technical University of Ilmenau, since 2008, and has been affiliated with the Fraunhofer Institute for Digital Media Technology since 2002, where he previously led the Audio Coding for Special Applications group. Before joining Fraunhofer, he was a member of technical staff at Bell Laboratories, Lucent Technologies, and Agere Systems (1998–2001). He received his Diplom degree in electrical engineering from the Technical University of Berlin and his Ph.D. (Dr.-Ing.) from the University of Hanover, and also studied at MIT and Georgia Tech. He served as Associate Editor for several IEEE Transactions and received the 2006 IEEE Best Paper Award in Audio and Electroacoustics. His research focuses on filter banks, audio coding, music signal processing, and deep learning for multimedia. He is best known for low-delay filter banks incorporated into the MPEG-4 Enhanced Low Delay AAC (ELD-AAC) standard.

•

Markus Erne studied electrical engineering and information technology at the ETH Zürich and earned his Ph.D. in 1991. He started his professional career as a development engineer at Studer Professional Audio GmbH and then joined Merging Technologies in 1991. He was also with Scopein Research for a considerable period of time. His main focus was on digital signal processing, audio technology, and low bitrate audio coding. He is the author of several books on the topic of Digital Audio. Dr. Erne also served as a Chairman of the AES Board of Governors and was the editor of the first edition of the "Perceptual Audio Coders – What To Listen For" tutorial of the AES TC-CAS. He died on 17 Aug. 2024.