quick study

The perceptual basis for audio compression

Armin Kohlrausch

Because the human ear ignores small amounts of noise that accompany a strong signal, much of the information in audio files can be thrown away with little loss of fidelity.

Armin Kohlrausch is a research fellow at Philips Research Europe in Eindhoven, the Netherlands, and professor of auditory and multisensory perception at the Eindhoven University of Technology.

Have you ever wondered how it is possible to store thousands of songs—a quantity of music that would require several hundred compact disks—on a small flash player and still have a sound quality rivaling that of CDs? The secret behind such impressive audio data compression lies in a highly successful cooperative effort between two scientific disciplines that are usually found in different departments on campus, if they are found at all: psychology of hearing, or psychoacoustics, and digital signal processing. During the past 20 years, scientists and engineers from those disciplines have worked together to create perceptual-coding algorithms, or "perceptual audio coders." Their efforts have enabled the use of the internet for music distribution and sharing and have inspired a new segment in the consumer electronics industry. If you listen to music on a modern audio player during your daily commute, the compression algorithms are based on the principles in this Quick Study.

Digital sound representation

To understand perceptual coders, it is useful to first consider some aspects of digital audio in the context of a CD. On a standard CD, audio signals are sampled at a rate of 44.1 kHz, and each sample value is represented with a resolution of 16 bits. Taking into account that a CD has two audio channels, 1 second of music corresponds to $2 \times 44\,100 \times 16$ b, or 1.4 Mb of audio data.

The sampling rate and bit resolution are motivated by properties of human hearing. Theoretically, the useful bandwidth reaches up to half the sampling frequency, but in real systems, filters have finite slopes. As a result, the highest frequency for a CD is somewhat lower than the theoretical limit, 18–20 kHz, just about the highest frequencies that can be heard. The 16-bit amplitude resolution limits the dynamic range. For the CD, the quantization of the analog voltage amplitude at the output of a microphone into 2^{16} or 65 536 equal steps introduces an error. For most practical purposes, the error can be treated as an additive noise signal.

Sound engineers quantify that noise with the decibel, a logarithmic measure used to specify either a ratio between sound pressures or absolute values of the pressure amplitude—in which case one refers to the dB SPL, for sound pressure level. The ratio of a CD signal with maximum amplitude to the quantization noise level corresponds to nearly 100 dB, and it will increase by 6 dB for every extra bit used to represent the amplitude. The goal in CD amplitude quantization is to keep the error signal below the absolute threshold of hearing.

An engineer can diminish the overall bit rate in a CD coding scheme by reducing the sampling rate or by using fewer bits per sample. Both affect the perceived quality in a drastic way. A reduced sampling rate decreases the bandwidth of the signal and leads to the elimination of audible high-frequency components. Using fewer bits per sample increases the audibility of the quantization noise by 6 dB for every bit. The online version of this Quick Study links to sound files to which one or the other technique has been applied. As you'll readily hear, neither possibility preserves the sound quality of the CD.

Perceptual coding

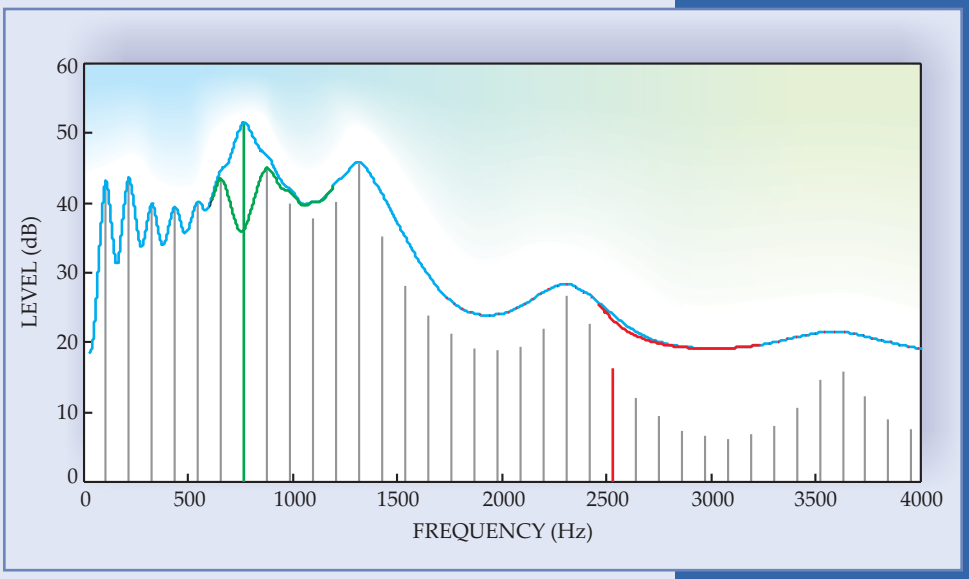
You've probably seen commercials for sweet-smelling products that "mask unwanted odors." Indeed, masking exists in all sensory modalities. That is, a signal easily perceived in isolation can be imperceptible in the presence of a different and stronger signal. Perceptual coders exploit masking in the quantizing of audio samples.

Audio masking is a reflection of auditory perception's limited resolution. If one begins by listening to a narrowband noise signal and then adds a music signal with the same frequency range, the noise becomes inaudible once it is about 6 dB lower than the music. For a person to enjoy the music, the signal-to-noise ratio need not be as large as is possible for a CD—a ratio of 10 dB is more than sufficient.

Thus, if one could shape quantization noise in such a way that its level was always 10 dB below the spectrum of the audio signal, the noise should be inaudible. Such a quantization would require only 2 bits per sample, a great reduction relative to the CD. This key perceptual concept of audio coding is relatively simple, but those who build coders face a significant challenge: Audio signals have dynamic spectral and temporal qualities that you may have seen visualized if, for example, you used the "scope" utility while playing music with a Winamp media player. As a consequence, designers of audio coders must know the spectral and temporal resolution with which the audio signal needs to be analyzed to give the correct shape to the quantization noise. To that end, psychoacoustic models of masking come into play.

Consider first the spectral variations. In transforming a sound wave into neural activity, the inner ear works in a way similar to a bank of bandpass filters. In other words, the input sound spectrum is transformed into a spectrally smoothed internal excitation; as an example, the figure shows how a particular vowel sound is smoothed. The masking effect of a narrowband signal will be largest in the spectral range covered by the signal, and less for lower and higher frequencies.

Based on the spectrum of a short audio signal segment,



Individual spectral lines show the power spectrum of the vowel "a" as sounded in "father." The continuous line shows, on a comparable decibel scale, the resulting internal excitation based on psychoacoustic models. The removal of a strong spectral component from the audio signal—for example the green element—leads to a significant change in the excitation pattern, as indicated by the pronounced green dip. When a weak spectral component (red) is removed, change in the excitation pattern is imperceptible. Conversely, if quantization noise added to an audio pattern does not lead to a significant change in the excitation pattern, the noise will be inaudible, or masked. That phenomenon is the basis for audio coders.

the encoding algorithm estimates the spectral masking in the form of a so-called masking curve. The next step is to distribute the available bits in such a way that the resulting quantization noise spectrum remains below the predicted masking curve. Such a shaping requires that the quantization of the audio signal be done separately in different spectral regions. That is, the quantization noise level is chosen separately for each region to more or less follow the masking curve.

The psychophysical phenomena behind the masking curve are unique, but translating them into real-time signal processing algorithms is a challenge. Thus one finds a great number of different algorithms—either standardized or proprietary—for audio coding. The perceptual quality of an audio coder depends to a large extent on the quality of the psychoacoustical model and on the repertoire of signal processing algorithms for shaping quantization noise according to the model.

Temporal aspects of masking are much more difficult to handle in models and coding applications. A person's hearing system has a remarkable temporal resolution: In continuous noise, gaps as short as 2–3 ms can be perceived. Periodic amplitude modulations of a high-frequency audio signal can be detected for modulation rates of up to 500 Hz.

On the other hand, the auditory system can also accumulate temporal information over durations of several hundred milliseconds so as to improve the audibility of weak signals. That time scale is a full two orders of magnitude longer than the minimal duration of a gap that can be detected in continuous noise.

Thus, to capture the high temporal resolution, masking estimates should be done with audio signals cut into very short time windows. But, in view of temporal accumulation, one also needs to look over time periods of hundreds of milliseconds. The solution in audio coders is pragmatic: The audio signals are analyzed with a fixed temporal window length, typically from 20 to 30 ms. For critical situations, like the very fast rise times typically found in the attacks of percussion instruments, specific tools increase temporal resolution.

State of the art

The first generations of perceptual audio coders, developed in the 1980s, mainly exploited spectral masking. Such algorithms form the bulk of audio coders in use today. The most efficient present-day audio coders apply improved psychoacoustic models and signal processing techniques. They give stereo signals with CD quality in the great majority of conditions, but at a rate of 92 kb/s—a nearly 15-fold compression ratio. Should you expect a major revolution that will bring the compression ratio up to 100 or more?

The greatest potential for further compression is expected, not from improving masking models but from exploiting the perceptual similarity of the various channels of a stereo or multichannel signal. In the first audio coders, relatively little gain was achieved from inter-channel relations, and the bit rate for a stereo signal was close to the sum of the bit rates needed to code the right and left channels separately. Over the past 5 years or so, very efficient representations of the spatial aspects of sound fields have been realized and internationally standardized by using knowledge about spatial perception to model the perceptual similarity between channels. Such coding algorithms will, in my opinion, be commonly used in the near future.

I thank Jeroen Breebaart, Bert den Brinker, Ralph van Dinther, and Steven van de Par for feedback on an earlier version of this article.

Additional resources

▶ Information about MPEG (Moving Picture Experts Group) standards is available at http://www.chiariglione.org/mpeg.

▶ J. L. Hall, in *The Digital Signal Processing Handbook*, V. K. Madisetti, D. B. Williams, eds., CRC Press, Boca Raton, FL (1998), chap. 39.

▶ J. Breebaart, S. van de Par, A. Kohlrausch, E. Schuijers, *EURASIP J. Appl. Signal Process.* **9**, 1305 (2005), available online at http://www.jeroenbreebaart.com/papers/jasp/jasp2005.pdf.

▶ S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, S. H. Jensen, *EURASIP J. Appl. Signal Process.* **9**, 1292 (2005), available online at http://www.hindawi.com/getarticle.aspx?doi=10.1155/asp.2005.1292.