Digital Signal Processing – bonus slides

Markus Kuhn

Computer Laboratory, University of Cambridge

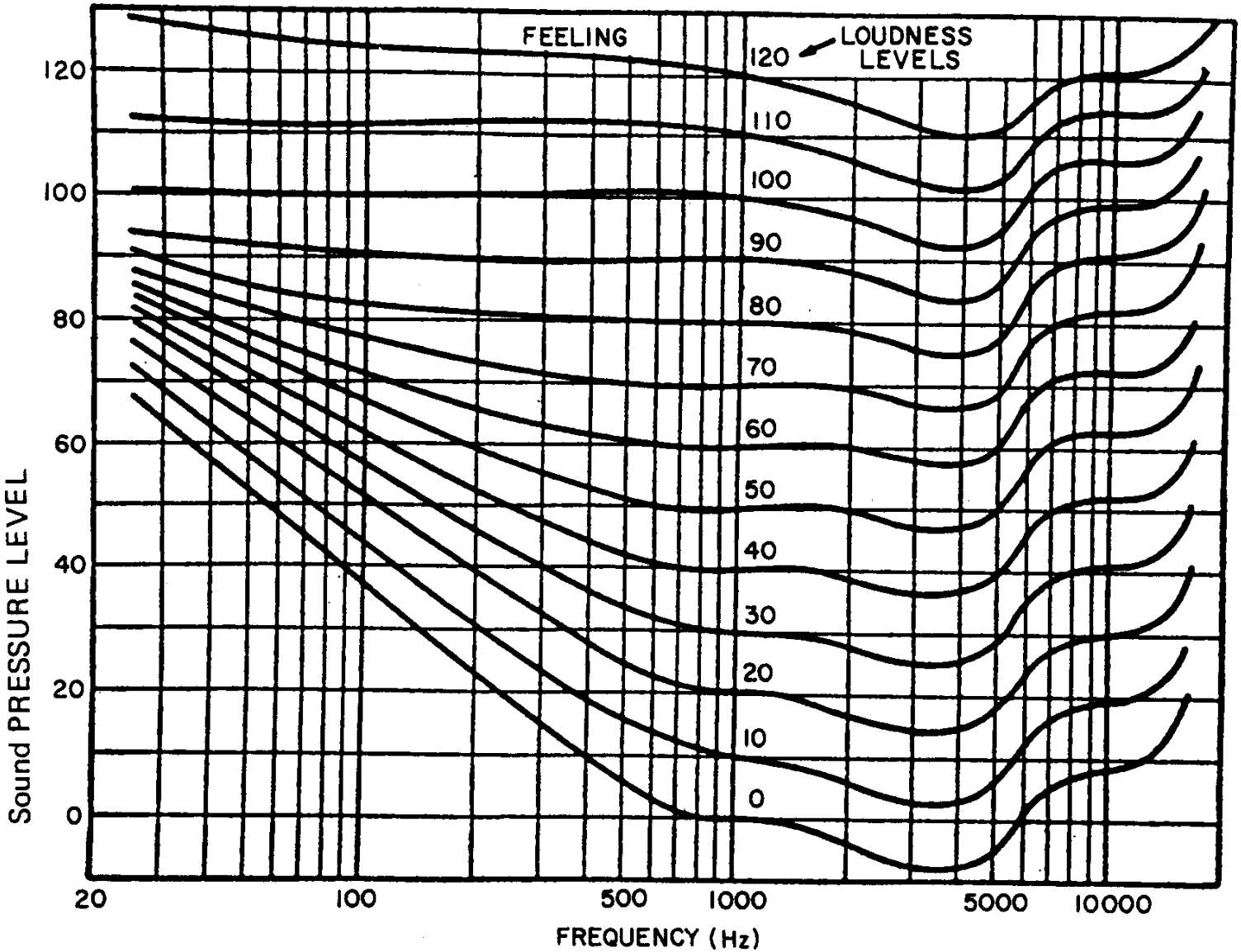
https://www.cl.cam.ac.uk/teaching/current/DSP/

Outline

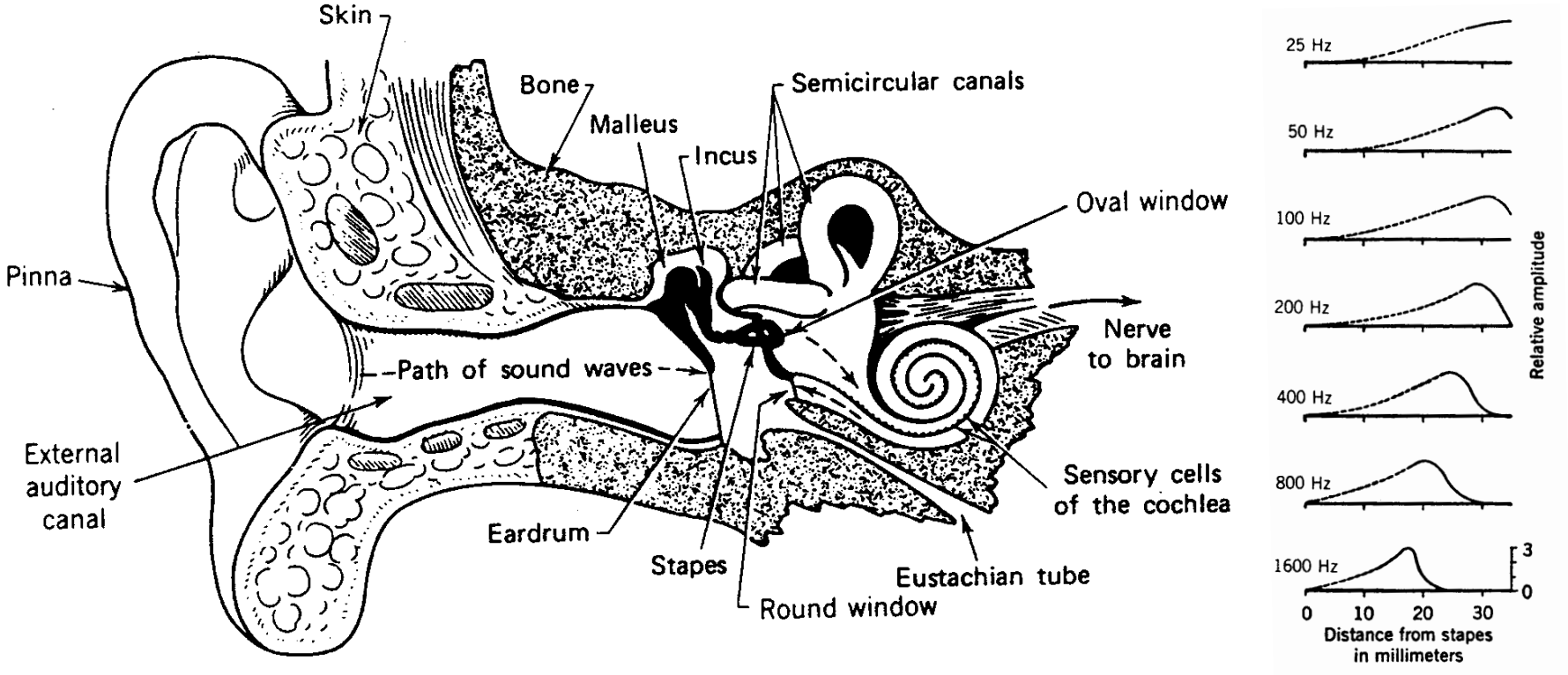
The human auditory system

- frequency range 20–16000 Hz (babies: 20 kHz)
- sound pressure range 0–140 $\mathrm{dB_{SPL}}$ (about 10^{-5}–10^{2} pascal)
- mechanical filter bank (cochlea) splits input into frequency components, physiological equivalent of Fourier transform
- most signal processing happens in the frequency domain where phase information is lost
- some time-domain processing below 500 Hz and for directional hearing
- sensitivity and difference limit are frequency dependent

Equiloudness curves and the unit "phon"



Each curve represents a loudness level in "phon". At 1 kHz, the loudness unit phon is identical to dB_{SPL} and 0 phon is the sensation limit.



Sound waves cause vibration in the eardrum. The three smallest human bones in the middle ear (malleus, incus, stapes) provide an "impedance match" between air and liquid and conduct the sound via a second membrane, the oval window, to the cochlea. Its three chambers are rolled up into a spiral. The basilar membrane that separates the two main chambers decreases in stiffness along the spiral, such that the end near the stapes vibrates best at the highest frequencies, whereas for lower frequencies that amplitude peak moves to the far end.

Frequency discrimination and critical bands

A pair of pure tones (sine functions) cannot be distinguished as two separate frequencies if both are in the same frequency group ("critical band"). Their loudness adds up, and both are perceived with their average frequency.

The human ear has about 24 critical bands whose width grows non-linearly with the center frequency.

Each audible frequency can be expressed on the "Bark scale" with values in the range 0–24. A good closed-form approximation is

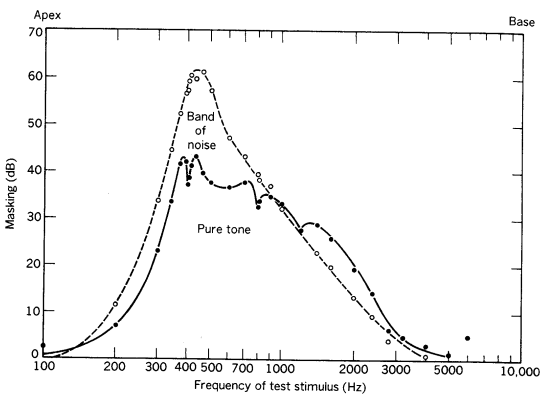
$$b \approx \frac{26.81}{1 + \frac{1960\ \text{Hz}}{f}} - 0.53$$

where f is the frequency and b the corresponding point on the Bark scale.

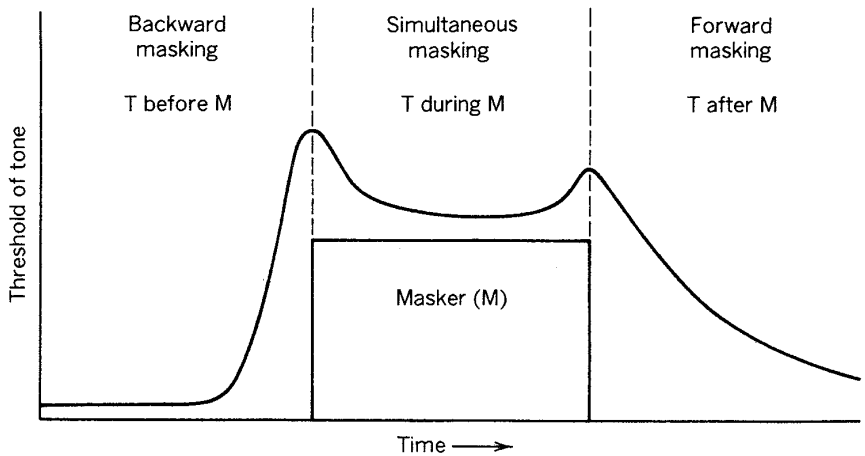
Two frequencies are in the same critical band if their distance is below 1 bark.

Masking

- Louder tones increase the sensation limit for nearby frequencies and suppress the perception of quieter tones.
- This increase is not symmetric. It extends about 3 barks to lower frequencies and 8 barks to higher ones.
- The sensation limit is increased less for pure tones of nearby frequencies, as these can still be perceived via their beat frequency. For the study of masking effects, pure tones therefore need to be distinguished from narrowband noise.
- Temporal masking: SL rises shortly before and after a masker.



Level of masking in decibels for test tones of various frequencies presented with a masker of either an 80-dB SPL 400 Hz tone or an 80-dB SPL narrow-band noise (90 Hz) with a center frequency of 410 Hz. From Egan and Hake (11).



Audio demo: loudness and masking

loudness.wav

Two sequences of tones with frequencies 40, 63, 100, 160, 250, 400, 630, 1000, 1600, 2500, 4000, 6300, 10000, and 16000 Hz.

- Sequence 1: tones have equal amplitude
- Sequence 2: tones have roughly equal perceived loudness
  Amplitude adjusted to IEC 60651 "A" weighting curve for soundlevel meters.

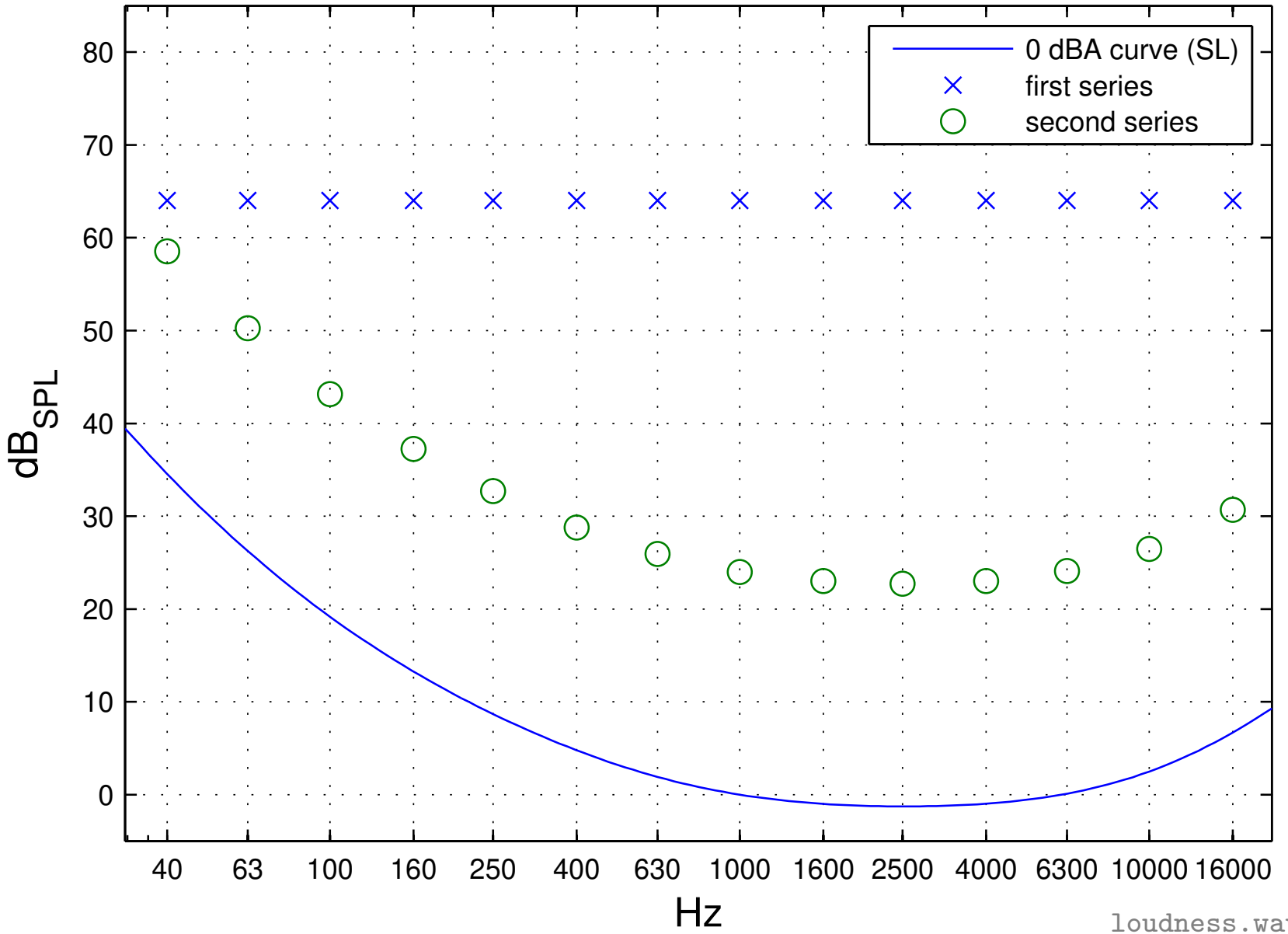
masking.wav

Twelve sequences, each with twelve probe-tone pulses and a 1200 Hz masking tone during pulses 5 to 8.

Probing tone frequency and relative masking tone amplitude:

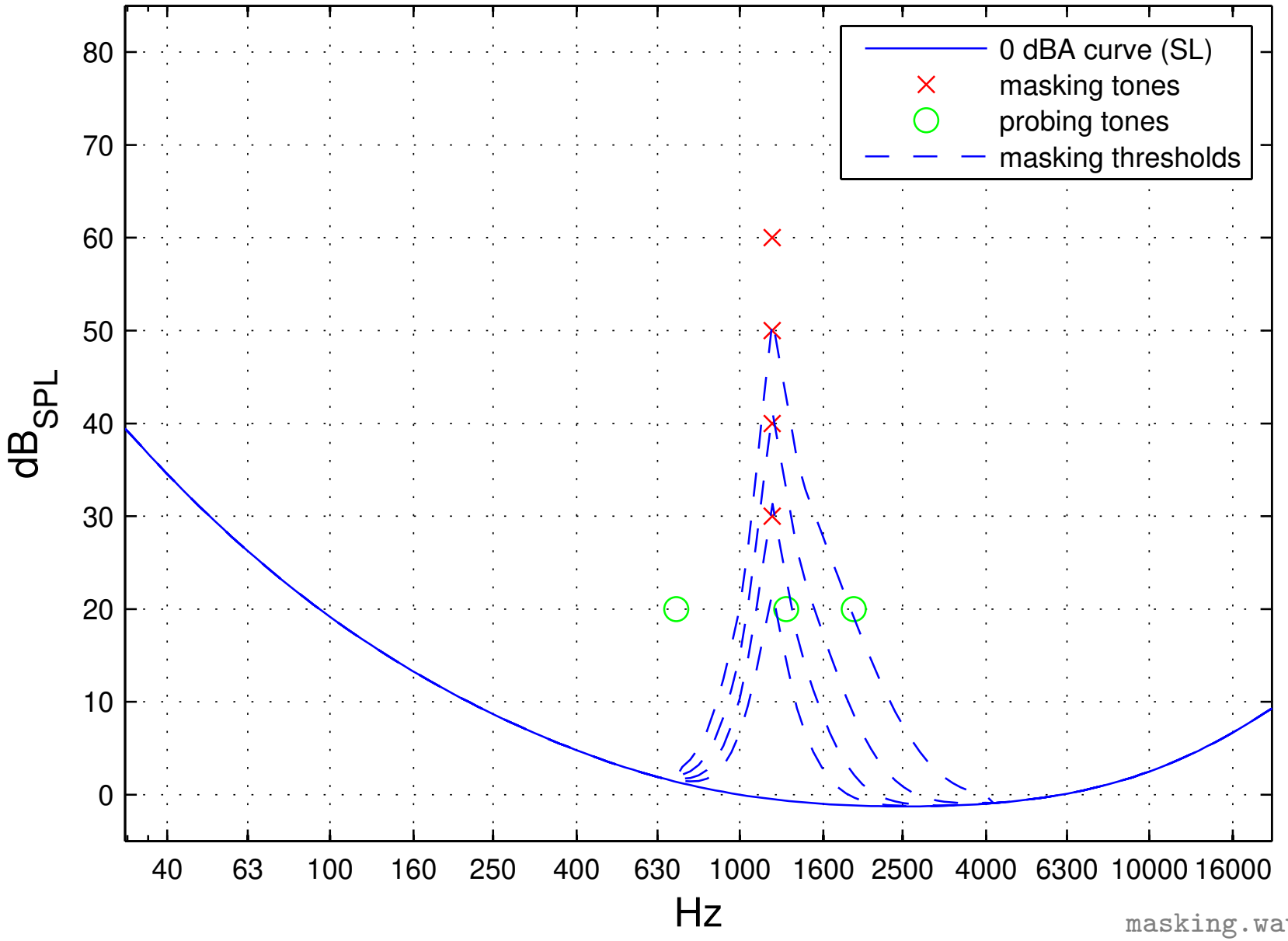
	10 dB	20 dB	30 dB	40 dB
1300 Hz				
1900 Hz				
700 Hz				

Audio demo: loudness.wav



loudness.wav

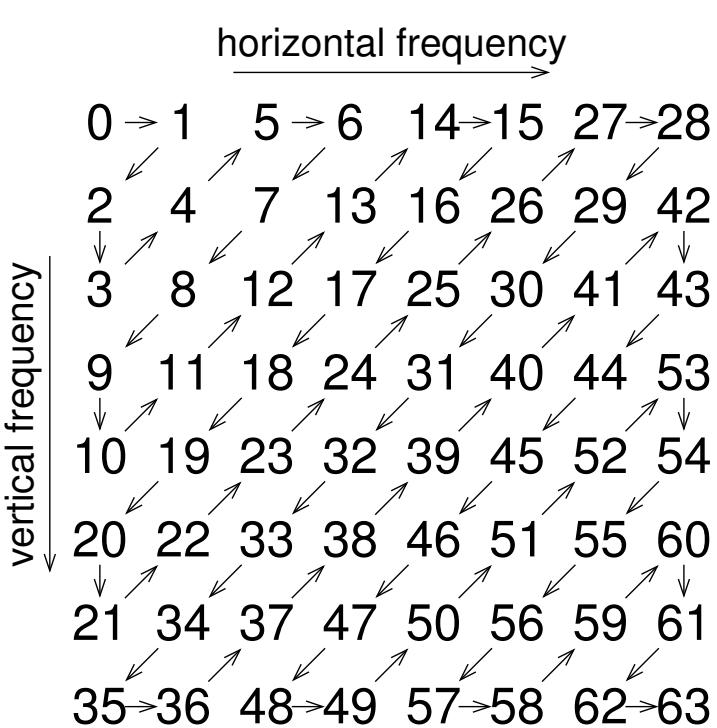
Audio demo: masking.wav

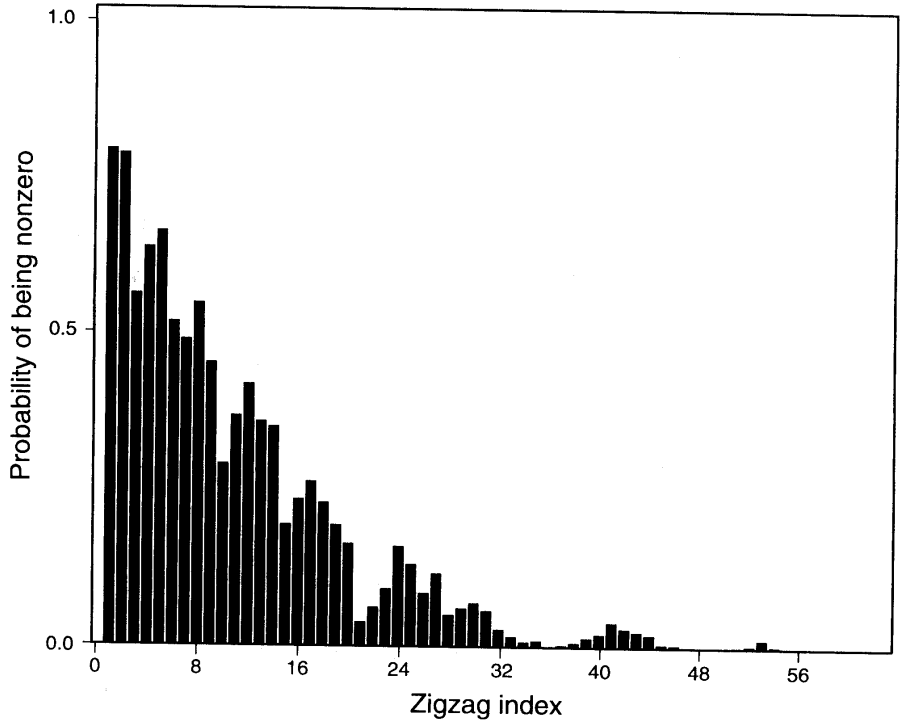


masking.wav

Outline

Storing DCT coefficients in zigzag order





After the 8×8 coefficients produced by the discrete cosine transform have been quantized, the values are processed in the above zigzag order by a run-length encoding step.

The idea is to group all higher-frequency coefficients together at the end of the sequence. As many image blocks contain little high-frequency information, the bottom-right corner of the quantized DCT matrix is often entirely zero. The zigzag scan helps the run-length coder to make best use of this observation.

Huffman coding in JPEG

s	value range
0	0
1	$-1, 1$
2	$-3, -2, 2, 3$
3	$-7 \ldots -4, 4 \ldots 7$
4	$-15 \ldots -8, 8 \ldots 15$
5	$-31 \ldots -16, 16 \ldots 31$
6	$-63 \ldots -32, 32 \ldots 63$
$\ldots$	$\ldots$
i	$-(2^i - 1) \ldots -2^{i-1}, 2^{i-1} \ldots 2^i - 1$

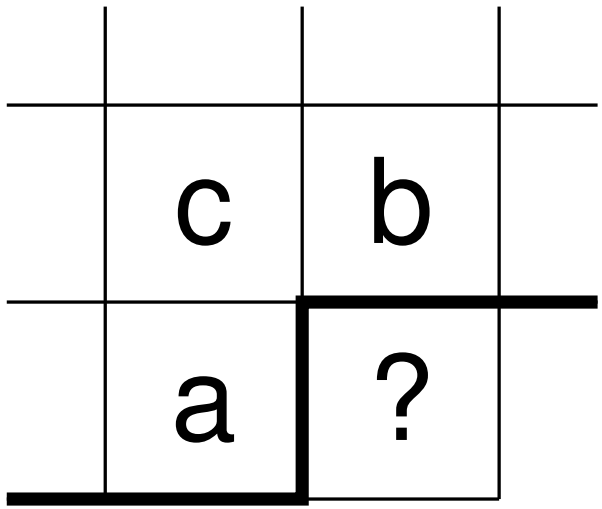
DCT coefficients have 11-bit resolution and would lead to huge Huffman tables (up to 2048 code words). JPEG therefore uses a Huffman table only to encode the magnitude category $s = \lceil \log_2(|v| + 1) \rceil$ of a DCT value v. A sign bit plus the $(s-1)$-bit binary value $|v| - 2^{s-1}$ are appended to each Huffman code word, to distinguish between the 2^s different values within magnitude category s.

When storing DCT coefficients in zigzag order, the symbols in the Huffman tree are actually tuples (r, s), where r is the number of zero coefficients preceding the coded value (run-length).

Lossless JPEG algorithm

In addition to the DCT-based lossy compression, JPEG also defines a lossless mode. It offers a selection of seven linear prediction mechanisms based on three previously coded neighbour pixels:

$$
\begin{aligned}
1: &\quad x = a \\
2: &\quad x = b \\
3: &\quad x = c \\
4: &\quad x = a + b - c \\
5: &\quad x = a + (b - c)/2 \\
6: &\quad x = b + (a - c)/2 \\
7: &\quad x = (a + b)/2
\end{aligned}
$$



Predictor 1 is used for the top row, predictor 2 for the left-most row. The predictor used for the rest of the image is chosen in a header. The difference between the predicted and actual value is fed into either a Huffman or arithmetic coder.

Advanced JPEG features

Beyond the baseline and lossless modes already discussed, JPEG provides these additional features:

- 8 or 12 bits per pixel input resolution for DCT modes
- 2–16 bits per pixel for lossless mode
- progressive mode permits the transmission of more-significant DCT bits or lower-frequency DCT coefficients first, such that a low-quality version of the image can be displayed early during a transmission
- the transmission order of colour components, lines, as well as DCT coefficients and their bits can be interleaved in many ways
- the hierarchical mode first transmits a low-resolution image, followed by a sequence of differential layers that code the difference to the next higher resolution

Not all of these features are widely used today. Several follow-on standards exist: JPEG XR uses a fully invertible DCT-like 4×4 block transform, JPEG 2000 uses a Cohen-Daubechies-Feauveau wavelet transform.

JPEG examples (baseline DCT)

1:5 (1.6 bit/pixel)

1:10 (0.8 bit/pixel)

JPEG examples (baseline DCT)

1:20 (0.4 bit/pixel)

1:50 (0.16 bit/pixel)

Better image quality at a compression ratio 1:50 can be achieved by applying DCT JPEG to a 50% scaled down version of the image (and then interpolate back to full resolution after decompression):

Outline

Moving Pictures Experts Group – MPEG

- MPEG-1: Coding of video and audio optimized for 1.5 Mbit/s (1× CD-ROM). ISO 11172 (1993).
- MPEG-2: Adds support for interlaced video scan, optimized for broadcast TV (2–8 Mbit/s) and HDTV, scalability options. Used by DVD and DVB. ISO 13818 (1995).
- MPEG-4: Adds advanced video codec (AVC) and advanced audio codec (AAC) for lower bitrate applications. ISO 14496 (2001).
- System layer multiplexes several audio and video streams, time stamp synchronization, buffer control.
- Standard defines decoder semantics.
- Asymmetric workload: Encoder needs significantly more computational power than decoder (for bit-rate adjustment, motion estimation, perceptual modeling, etc.)

`https://mpeg.chiariglione.org/`

MPEG video coding

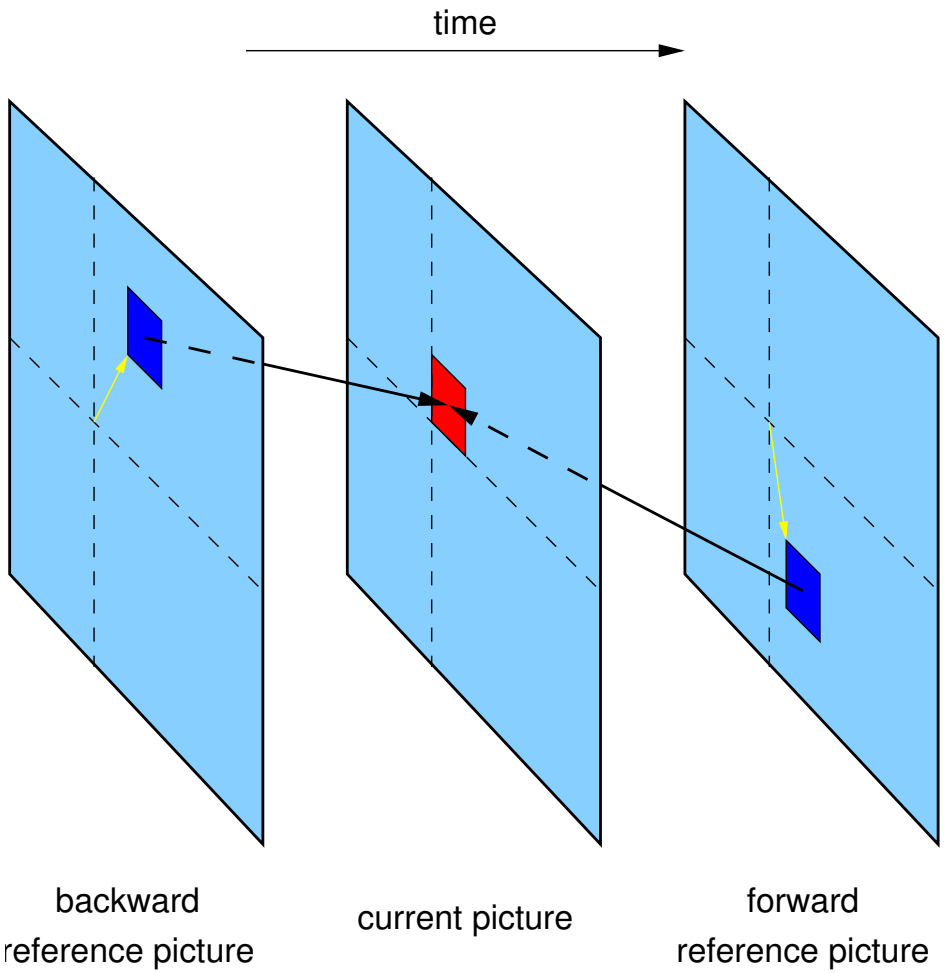
- Uses YCrCb colour transform, 8×8-pixel DCT, quantization, zigzag scan, run-length and Huffman encoding, similar to JPEG
- the zigzag scan pattern is adapted to handle interlaced fields
- Huffman coding with fixed code tables defined in the standard
  MPEG has no arithmetic coder option.
- adaptive quantization
- SNR and spatially scalable coding (enables separate transmission of a moderate-quality video signal and an enhancement signal to reduce noise or improve resolution)
- Predictive coding with motion compensation based on 16×16 macro blocks.

J. Mitchell, W. Pennebaker, Ch. Fogg, D. LeGall: MPEG video compression standard. ISBN 0412087715, 1997. (CL library: I.4.20)

B. Haskell et al.: Digital Video: Introduction to MPEG-2. Kluwer Academic, 1997. (CL library: I.4.27)

John Watkinson: The MPEG Handbook. Focal Press, 2001. (CL library: I.4.31)

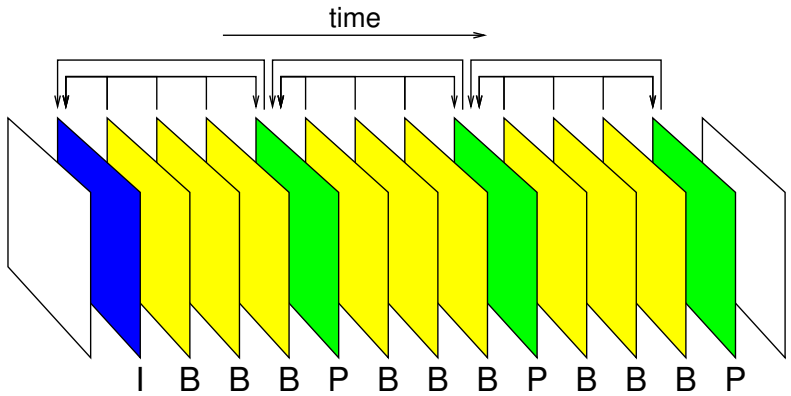
MPEG motion compensation



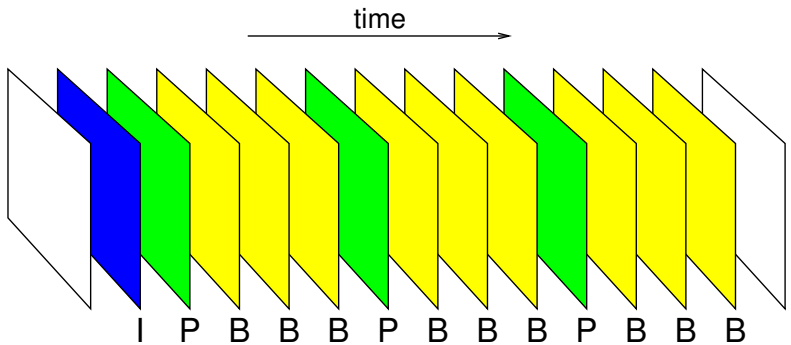
Each MPEG image is split into 16×16-pixel large *macroblocks*. The predictor forms a linear combination of the content of one or two other blocks of the same size in a preceding (and following) reference image. The relative positions of these reference blocks are encoded along with the differences.

MPEG reordering of reference images

Display order of frames:



Coding order:



MPEG distinguishes between I-frames that encode an image independent of any others, P-frames that encode differences to a previous P- or I-frame, and B-frames that interpolate between the two neighbouring B- and/or I-frames. A frame has to be transmitted before the first B-frame that makes a forward reference to it. This requires the coding order to differ from the display order.

MPEG audio coding

Three different algorithms are specified, each increasing the processing power required in the decoder.

Supported sampling frequencies: 32, 44.1 or 48 kHz.

Layer I

- Waveforms are split into segments of 384 samples each (8 ms at 48 kHz).
- Each segment is passed through an orthogonal filter bank that splits the signal into 32 subbands, each 750 Hz wide (for 48 kHz).
  This approximates the critical bands of human hearing.
- Each subband is then sampled at 1.5 kHz (for 48 kHz).
  12 samples per window $\rightarrow$ again 384 samples for all 32 bands
- This is followed by scaling, bit allocation and uniform quantization.
  Each subband gets a 6-bit scale factor (2 dB resolution, 120 dB range, like floating-point coding). Layer I uses a fixed bitrate without buffering. A bit allocation step uses the psychoacoustic model to distribute all available resolution bits across the 32 bands (0–15 bits for each sample). With a sufficient bit rate, the quantization noise will remain below the sensation limit.
- Encoded frame contains bit allocation, scale factors and sub-band samples.

Layer II

Uses better encoding of scale factors and bit allocation information.

Unless there is significant change, only one out of three scale factors is transmitted. Explicit zero code leads to odd numbers of quantization levels and wastes one codeword. Layer II combines several quantized values into a *granule* that is encoded via a lookup table (e.g., 3×5 levels: 125 values require 7 instead of 9 bits). Layer II is used in Digital Audio Broadcasting (DAB).

Layer III

- Modified DCT step decomposes subbands further into 18 or 6 frequencies
- dynamic switching between MDCT with 36-samples (28 ms, 576 freq.) and 12-samples (8 ms, 192 freq.)
  enables control of pre-echos before sharp percussive sounds (Heisenberg)
- non-uniform quantization
- Huffman entropy coding
- buffer with short-term variable bitrate
- joint stereo processing

MPEG audio layer III is the widely used "MP3" music compression format.

Psychoacoustic models

MPEG audio encoders use a psychoacoustic model to estimate the spectral and temporal masking that the human ear will apply. The subband quantization levels are selected such that the quantization noise remains below the masking threshold in each subband.

The masking model is not standardized and each encoder developer can chose a different one. The steps typically involved are:

- Fourier transform for spectral analysis
- Group the resulting frequencies into "critical bands" within which masking effects will not vary significantly
- Distinguish tonal and non-tonal (noise-like) components
- Apply masking function
- Calculate threshold per subband
- Calculate signal-to-mask ratio (SMR) for each subband

Masking is not linear and can be estimated accurately only if the actual sound pressure levels reaching the ear are known. Encoder operators usually cannot know the sound pressure level selected by the decoder user. Therefore the model must use worst-case SMRs.