9. Speech coding

Introduction to speech coding

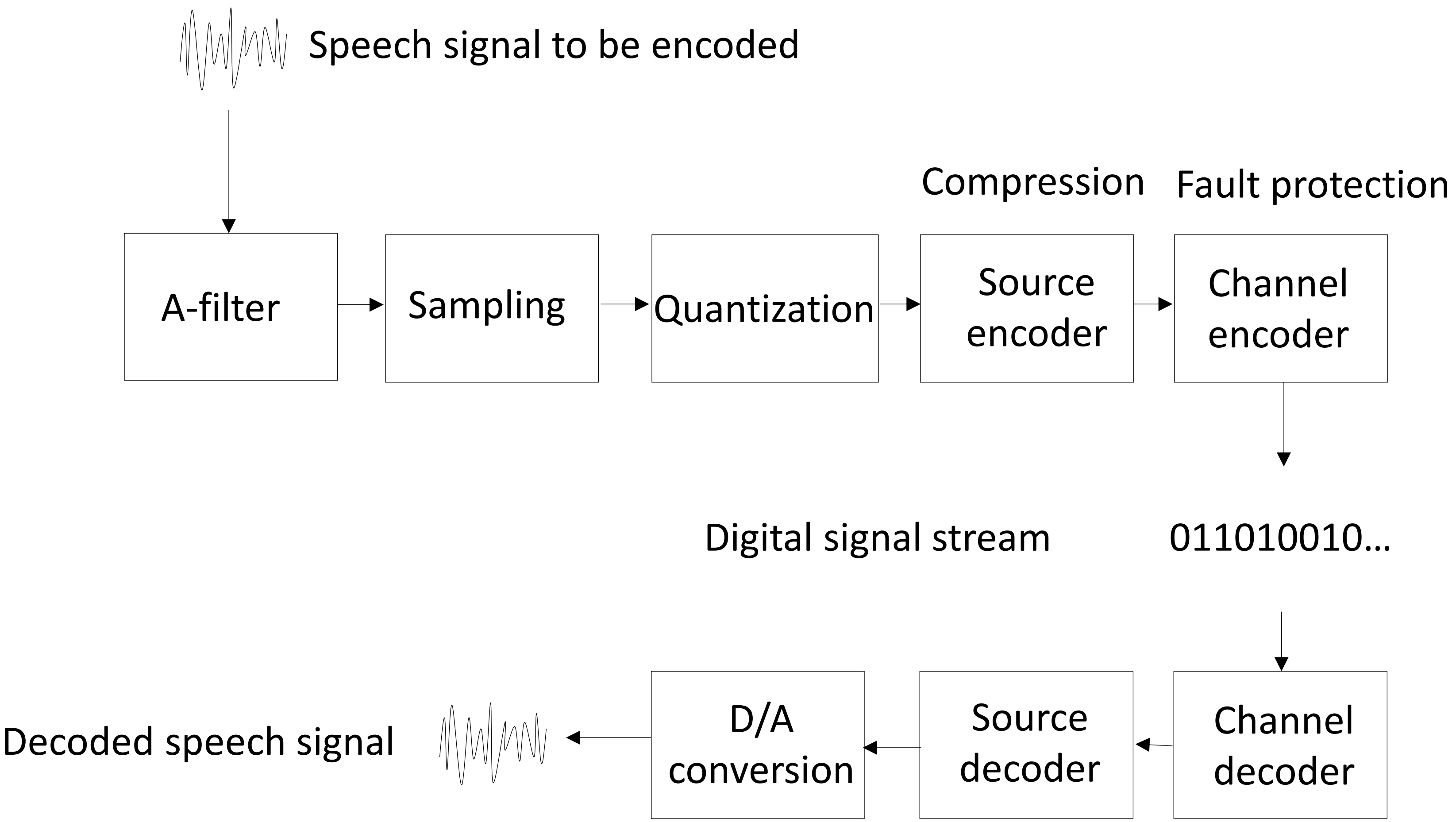
- **Introduction**

  - Speech is an analog signal
  - We use digital devices
- Sampling
- Quantization
- Coding

Time /Amplitude	Continuous	Discrete
Continuous	Analog signal	Asynchronous digital
Discrete	Sampled analog (CCD)	Synchronous digital

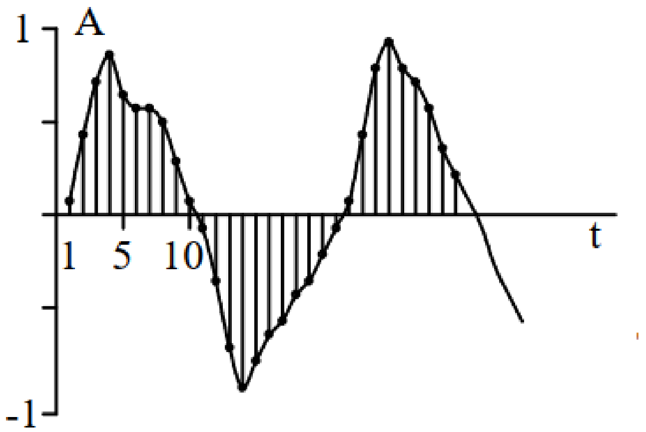
The process of encoding speech

<u>Source and channel encoding</u>

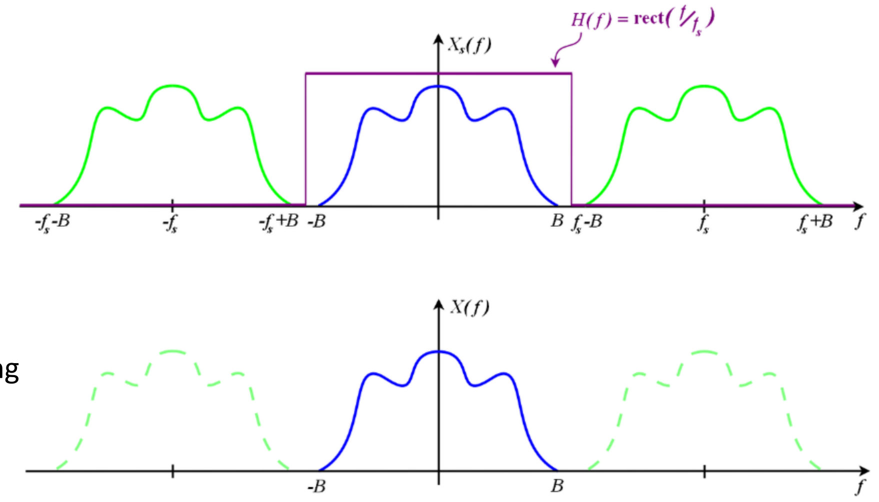


Sampling

- discretization



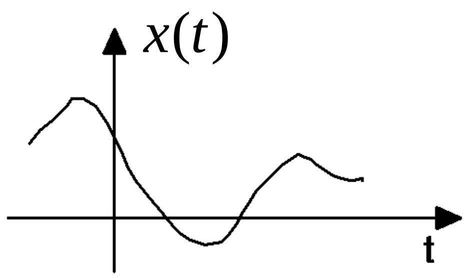
- Shannon's theorem: A signal limited to band B can be uniquely reconstructed from samples taken with a density of fs ≥ 2B.



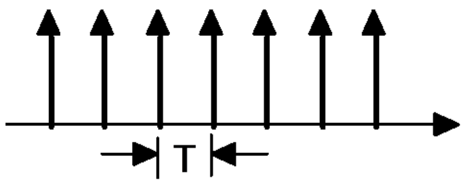
Source:
https://en.wikipedia.org/wiki/Nyquist%E2%80%93Shannon_sampling_theorem#/media/File:ReconstructFilter.png

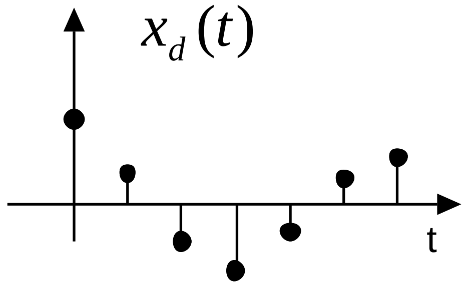
The sampling process

Time range



$$p(t) = \sum_{k=-\infty}^{\infty} \delta(t - kT)$$

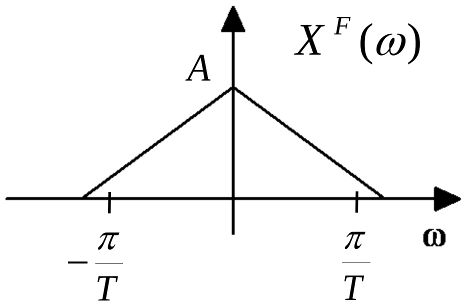




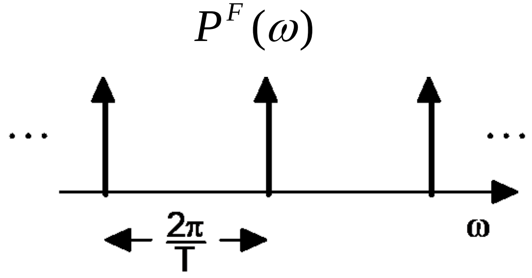
Frequency range

$$P^f(\omega) = \frac{2\pi}{T} \sum_{k=-\infty}^{\infty} \delta\left(\omega - k\frac{2\pi}{T}\right)$$

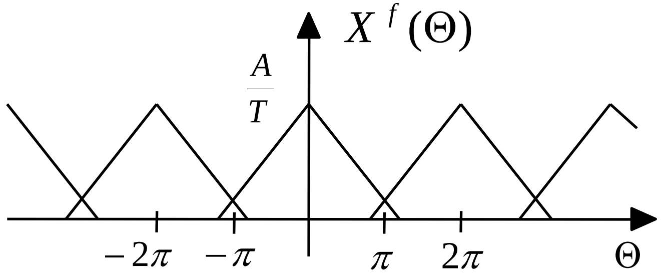
$$\boxed{X^f(\Theta) = \frac{1}{T} \sum_{k=-\infty}^{\infty} X^F\left(\frac{\Theta - 2\pi k}{T}\right)}$$



source signal spectrum



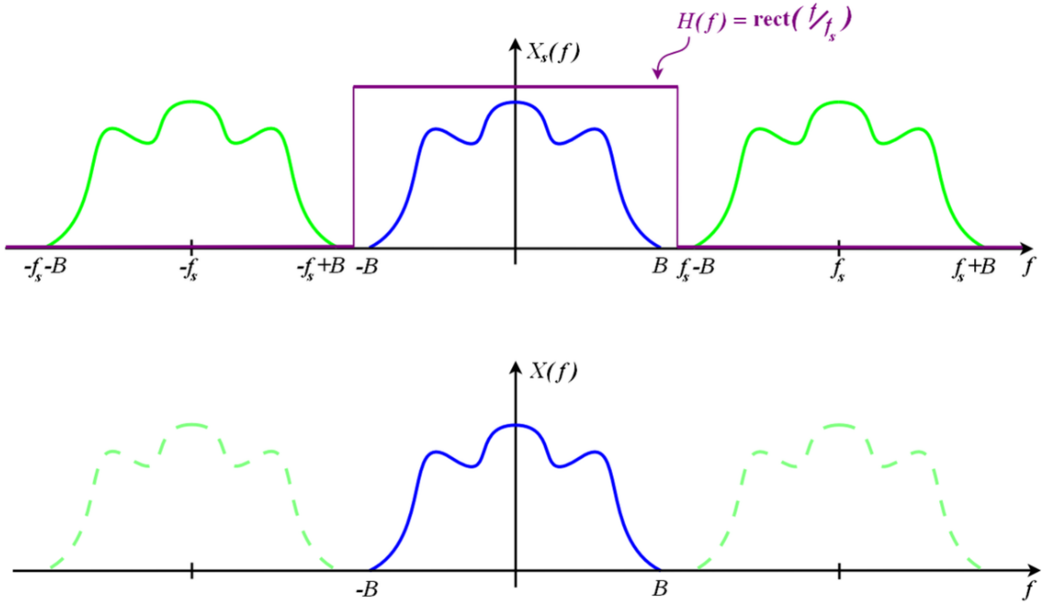
sampling function



sampled signal spectrum
(cumulative spectrum)

PAM type reconstruction

- Pulse Amplitude Modulation
- The signal can be restored from samples of a signal of bandwidth B using the PAM smoothing recunstruction filter if $f_s \geq 2B$ and $H(f) = M(f) * G(f)$ is constant in the range $-B \cdots +B$, and 0 in the vicinity of $k \cdot f_s$ B



Source:
https://en.wikipedia.org/wiki/Nyquist%E2%80%93Shannon_sampling_theorem#/media/File:ReconstructFilter.png

Spectrum of sampled signal

- There are infinite number of signals with identical samples
- It is valid for the time series of the signals: $x_i(kT_0) = x_j(kT_0), \forall k$
- What about spectra?

$$\sum_{m=-\infty}^{\infty} X_i(\omega - m\omega_0) = \sum_{l=-\infty}^{\infty} X_j(\omega - l\omega_0), \qquad \omega = \frac{2\pi}{T_0}$$

- We also suppose that

$$\sum_{n=-\infty}^{\infty} x^2(nT_0) < \infty$$

Nyquist equivalent

- Nyquist equivalent: the $-f_s/2 \leq f_s/2$ range of the sampled spectrum
- The sampled spectrum is periodic

$$\sum_{k=-\infty}^{\infty} X_i(\omega - k\omega_0) = T_0\left[\sum_{n=-\infty}^{\infty} x(nT_0)e^{(-j\omega nT_0)}\right]$$

- If the signal is band-limited to B and the sampling frequency is more than 2B then the Nyquist equivalent of the spectrum is itself

PAM type smoothing reconstructed output signal

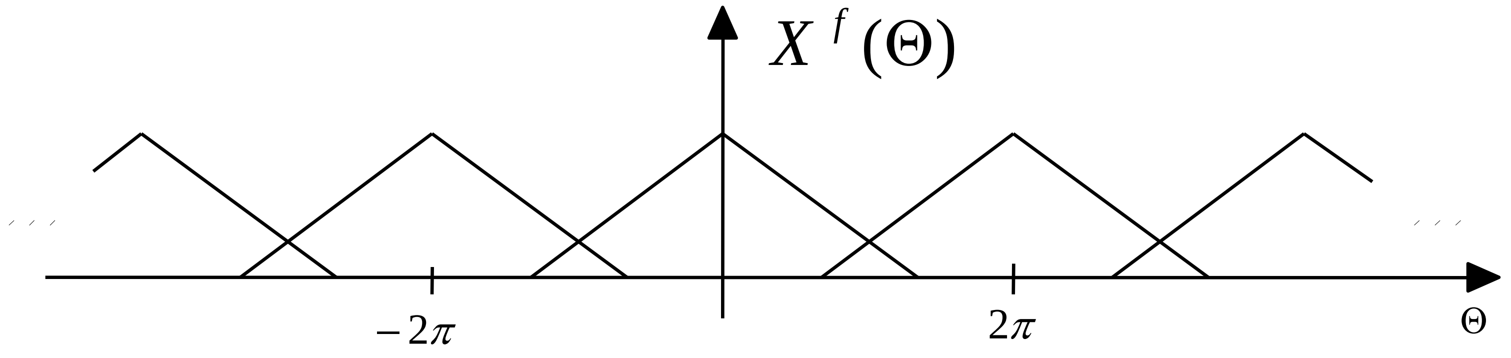
$$X(\omega) = T_0 \left[\sum_{n=-\infty}^{\infty} x(nT_0) e^{(j\omega n T_0)} \right] M(\omega) G(\omega) =$$

$$\left[\sum_{k=-\infty}^{\infty} X_i(\omega - k\omega_0) \right] M(\omega) G(\omega)$$

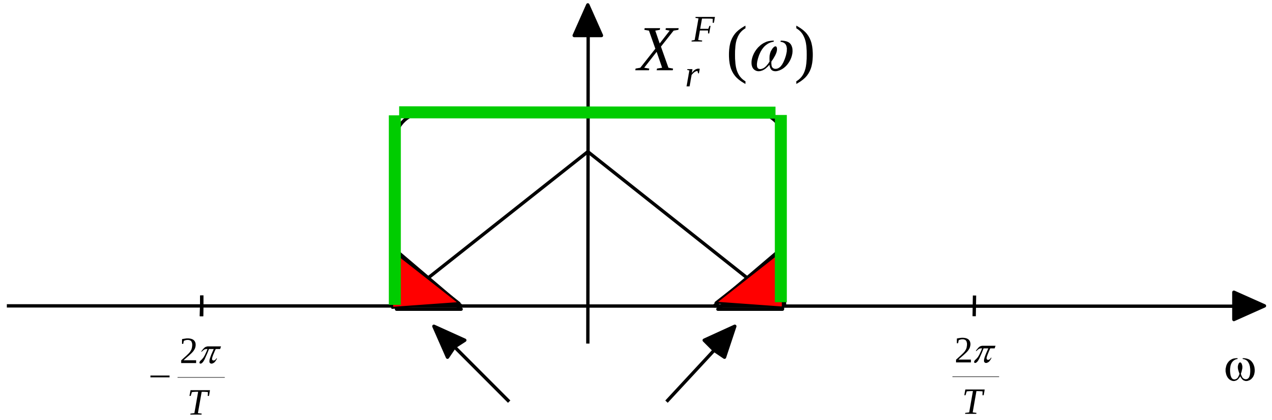
- If the band-limit is not kept, overlapping will occur
- The original signal cannot be reconstructed in case of spectral overlap

Example of overlap

Spectrum of a sampled (discrete) signal

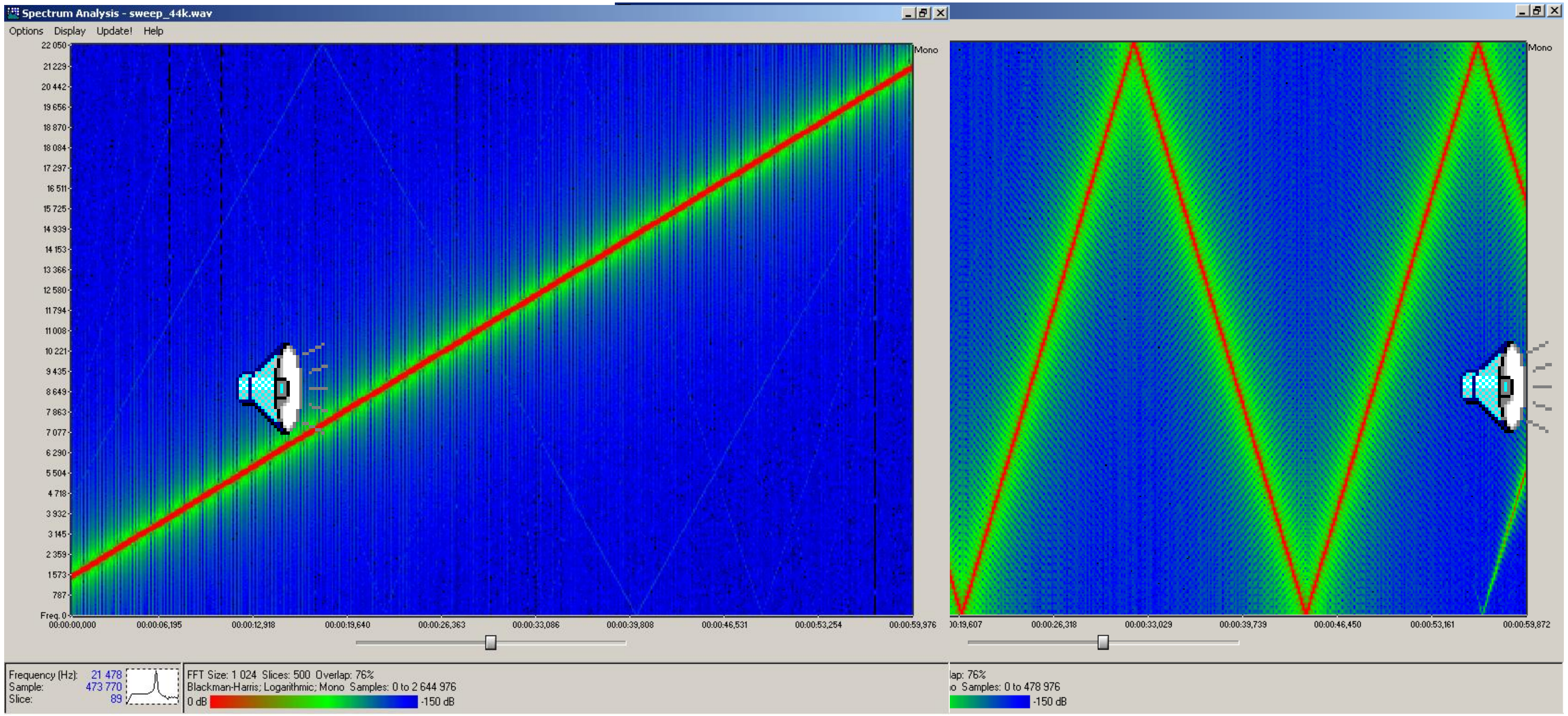


Spectrum of a restored (filtered, interpolated) continuous signal



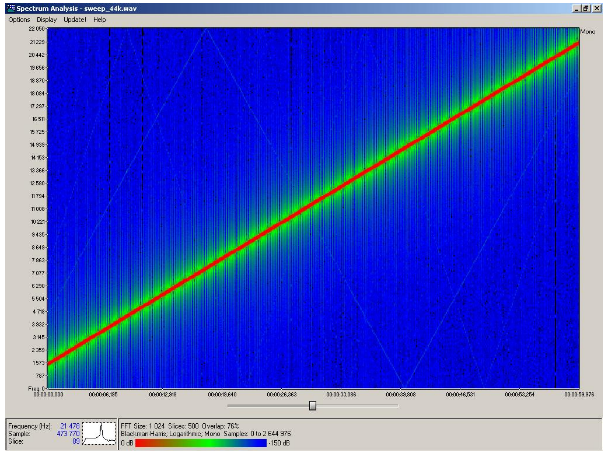
"Aliasing"

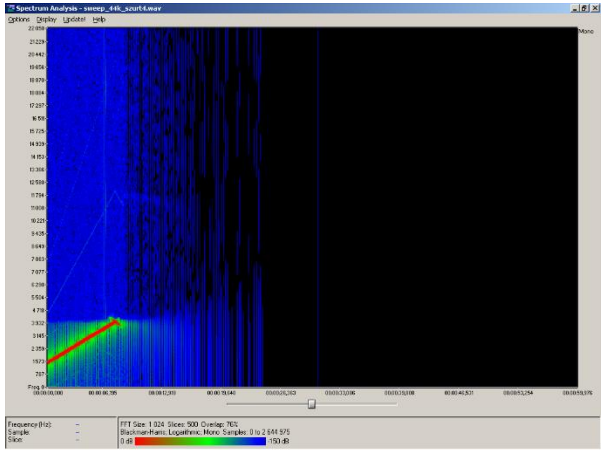
Sinusoidal example of overlap



Protection against overlap

We put a low-pass filter in front of the sampler , which cuts at half the sampling frequency → we ensure that the condition of the sampling law is met





If it is not there: unpleasant sound + cannot be filtered out afterwards!

Music (44.1kHz)

Music (16kHz)

Music + aliasing (16kHz)

Speech (8kHz)

Speech+aliasing (8kHz)

Important practical considerations

Comment:

The above applies to both analog signal sampling and digital signal re-sampling (e.g. 44.1kHz → 8kHz).

An example occurred at a Hungarian telephone company:

“The studio recording wasn’t good – it sounded really bad at 8kHz… We re-recorded it all over the phone.”

How to choose a sampling frequency?

Choosing a sampling frequency: depends on the frequency range we want to restore.

Application	Frequency range	Sampling frequency
Phone	300-3400Hz	8 kHz
Wideband speech	50-7000Hz	16 kHz
Music	20-20,000Hz	44.1kHz

Speech: telephone 8 kHz 16 kHz 44.1 kHz

Music: 8 kHz 16 kHz 44.1 kHz

Telephone: goal is intelligibility; speaker not always identifiable, some sounds are hard to distinguish (sz-f)

Wideband speech: sense of presence, (e.g. for speakerphone, conference calls)

Music: higher quality expectations (e.g. CD digital audio)

Basic characteristics of quantization

- Discretization in the amplitude value set
- Quantization is irreversible
- Quantization steps may change
- Dead-zone quantization around zero (keep or change)

Linear quantization

- Quantization: amplitude discretization
- Difference between the real and quantized signal: additive white noise -> quantization noise
- Quantization steps may change

$$x[n] = Q\{x[n]\} + \varepsilon[n]$$

- Dead-zone quantization around zero (keep or change)
- Uniform quantization quantization step: Δ
- $\varepsilon[n]$ uniform distribution, with 0 mean value
- Noise power:

$$P_\varepsilon = M(\varepsilon^2) = \int_{-\infty}^{\infty} \varepsilon^2 f_\varepsilon(\varepsilon) d\varepsilon = \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \varepsilon^2 \frac{1}{\Delta} d\varepsilon = \left| \frac{1}{\Delta} \frac{\varepsilon^3}{3} \right|_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} = \frac{\Delta^2}{12}$$

- Sinewave with amplitude A: $2A = N\Delta = 2^n \Delta$ $\quad P_{\sin} = \frac{A^2}{2}$

$$SNR = \frac{P_{\sin}}{P_\varepsilon} = 6\frac{A^2}{\Delta^2} = 6\frac{(\frac{N\Delta}{2})^2}{\Delta^2} = \frac{3}{2} 2^{2n} \quad SNR^{[dB]} = 10 \cdot \lg(SNR) = 1.76dB + n \cdot 6.02dB$$

Examples of quantization

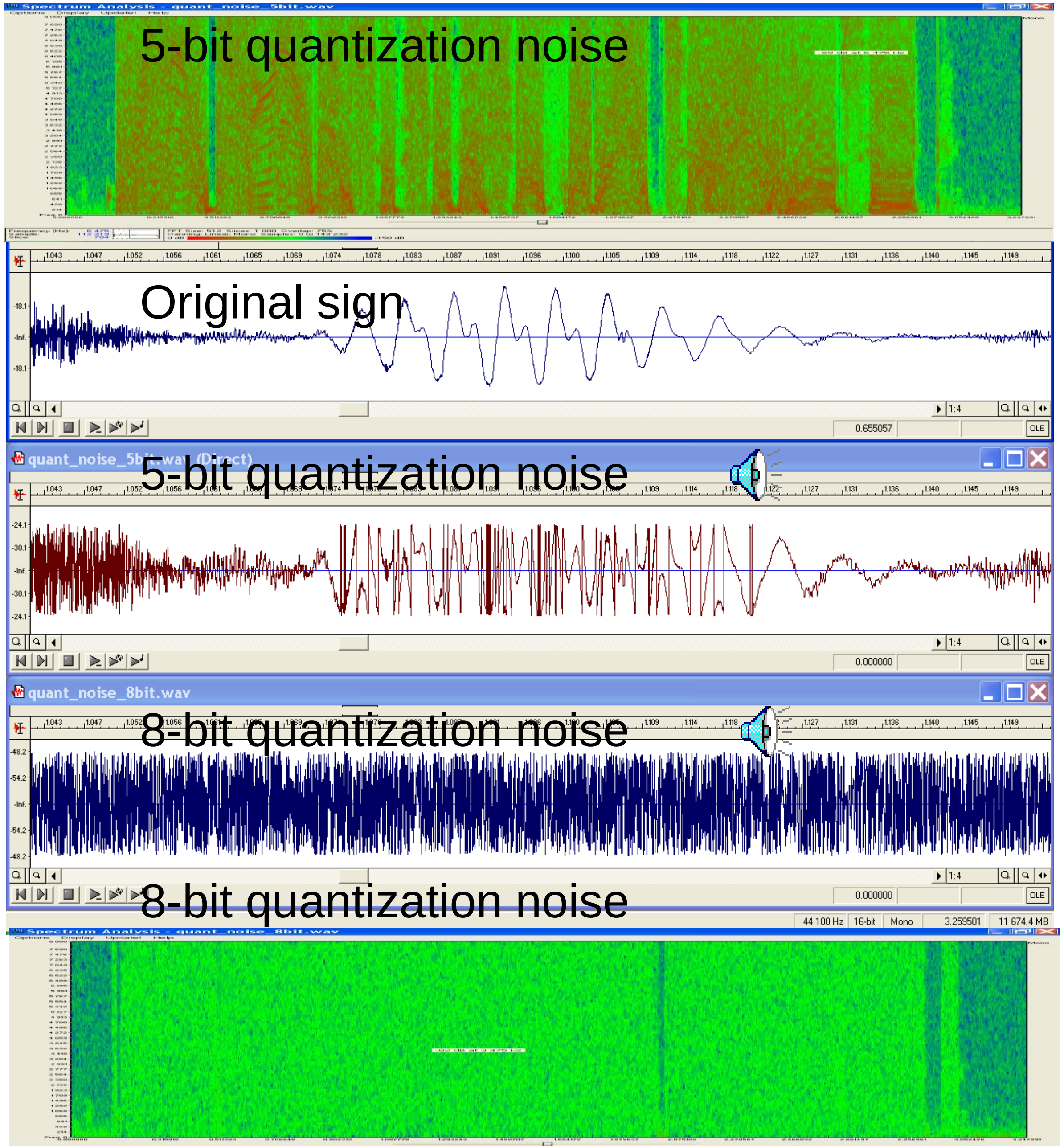
5-bit quantization

(not white noise!)

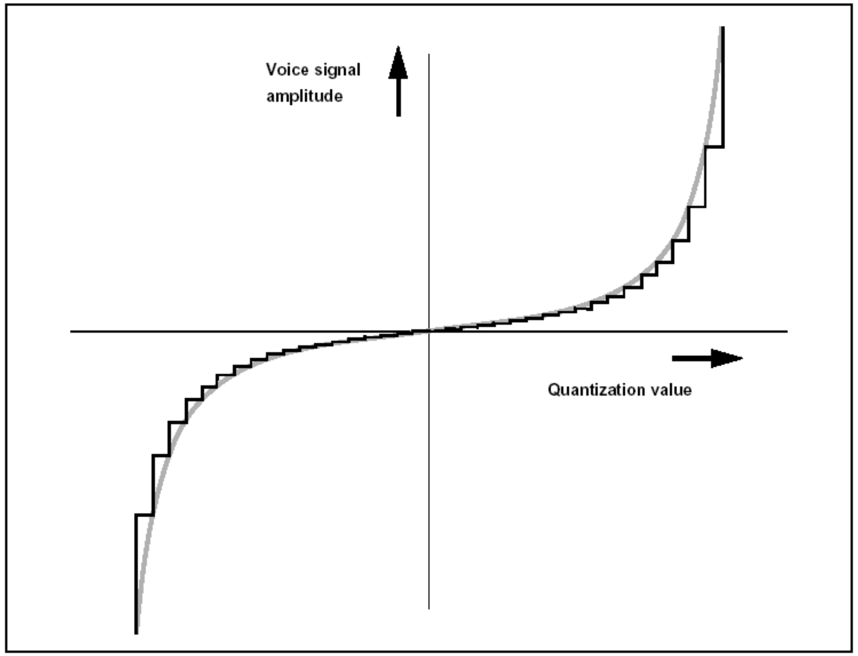
8-bit quantization

(now white noise)

16-bit quantization



Logarithmic quantization



- smaller amplitude values are finely quantized
- larger roughly → quantization noise increases

A-law (Europe):

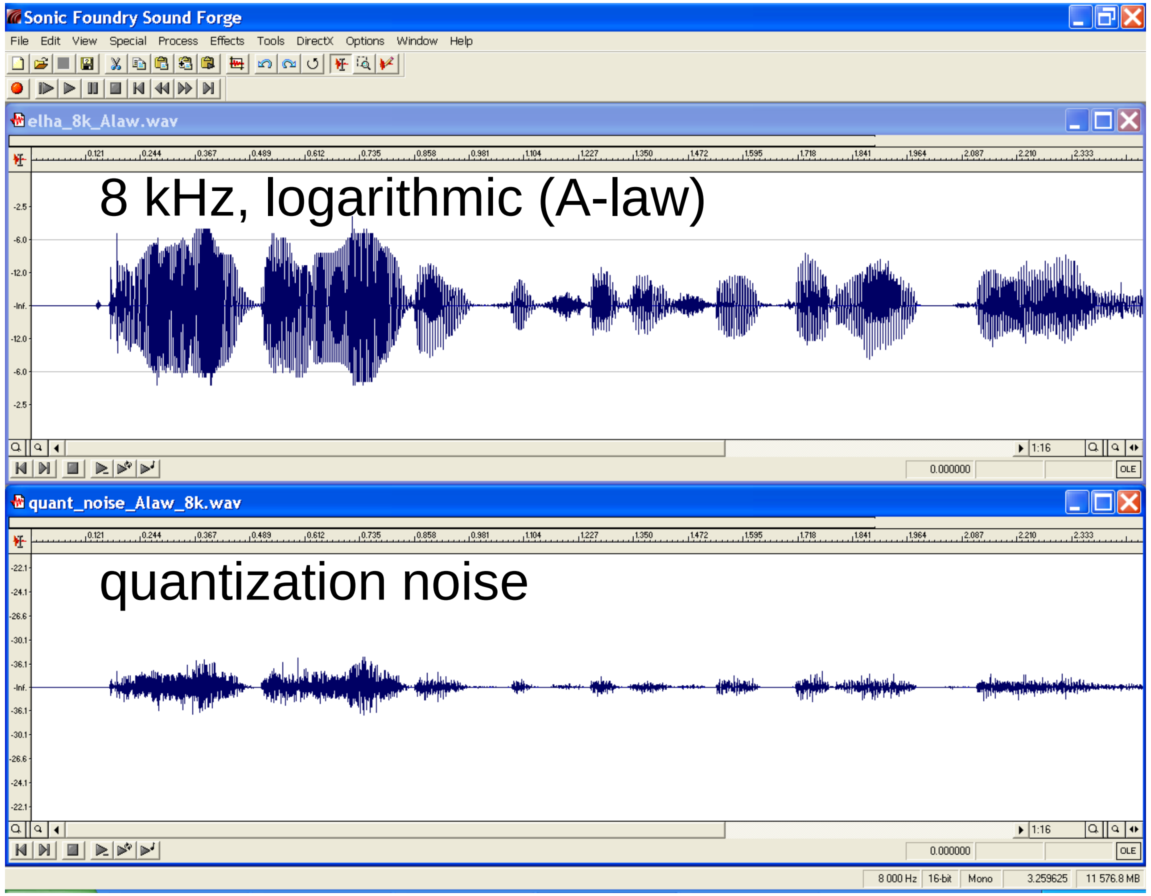
$$y = \begin{cases} \dfrac{Ax}{1+\log A}, & 0 \le |x| \le 1/A \\ \dfrac{1+\log Ax}{1+\log A}, & 1/A \le |x| \le 1 \end{cases}$$

μ -law (USA):

$$y = \frac{\log(1+\mu x)}{\log(1+x)} \ for \ \ x \ge 0$$

18

Logarithmic quantization example



- 8-bit logarithmic quantization has sound quality equivalent to 13-bit linear (uniform) quantization
- Suitable for speech sampled at 8kHz

- Louder quantization noise is masked by louder speech
- But only up to about 4kHz! (fs=8kHz)

Counterexample (fs=44.1kHz):

SmartLab
Intelligent Interactions
smartlab.tmit.bme.hu

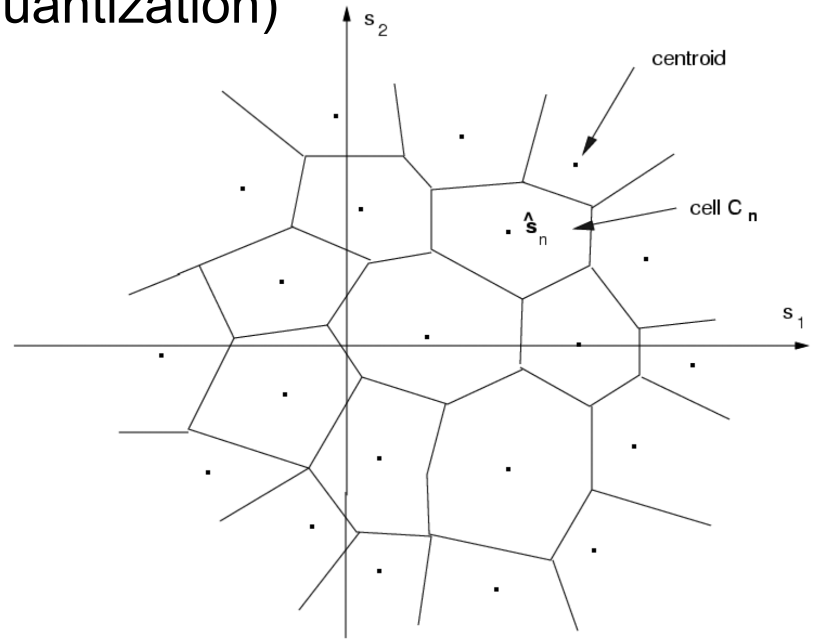
BME TMIT

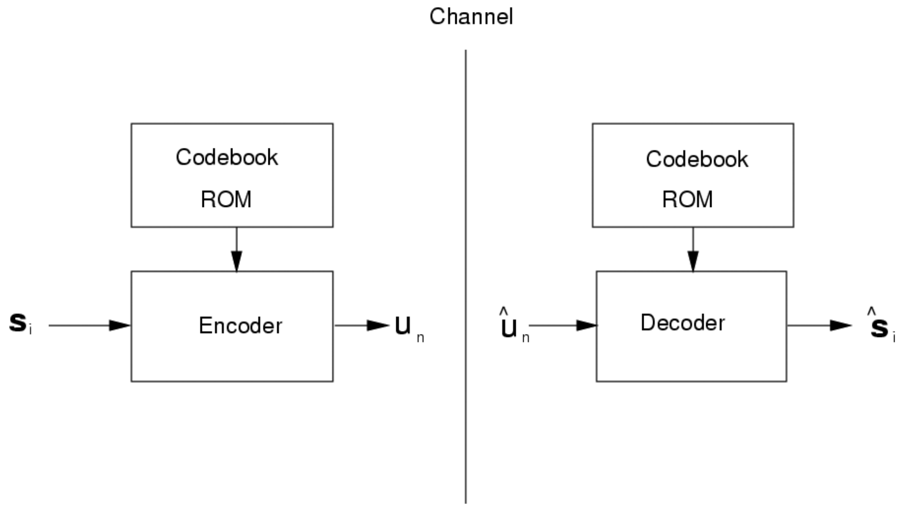
NVIDIA GPU EDUCATION CENTER

Vector quantization

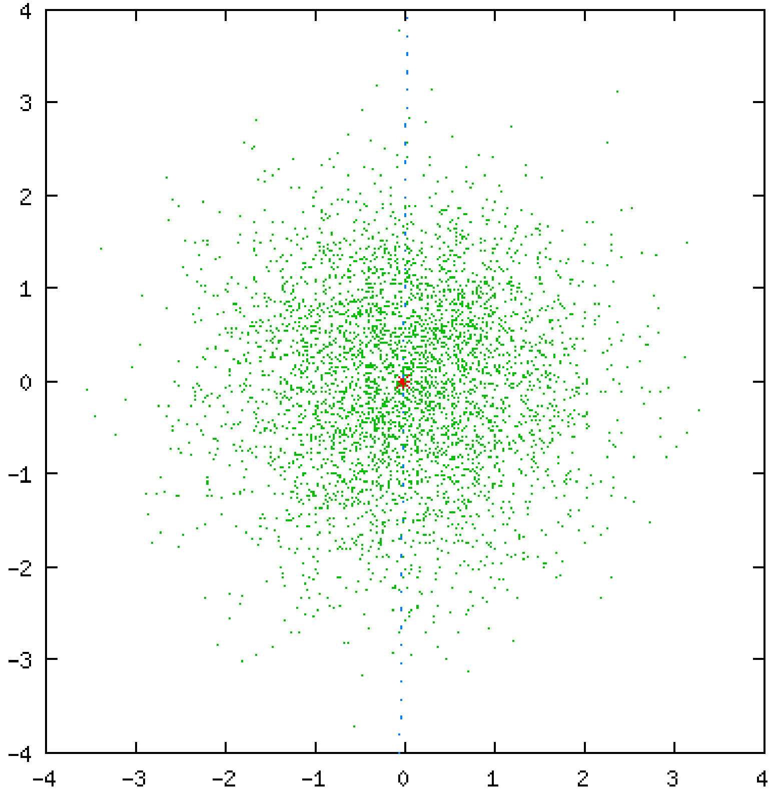
(VQ: Vector Quantization)

- We quantize
- Number of symbols coded together: dimension
- We transfer the cell index
- In a decoder, the centroid represents the vector
- Advantage: more efficient compression
- Disadvantage: higher computational requirements

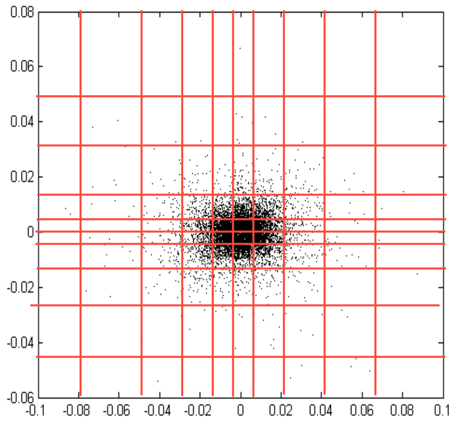




Codebook development with an iterative learning algorithm



- Pdf (distribution function) optimized (scalar can also do it, but not as much):



- Average distortion decreases
- But for rarer values, there is greater distortion! → may be disruptive with audio signals
- Cell index lookup is slow when encoding

Structured (e.g. lattice) codebook

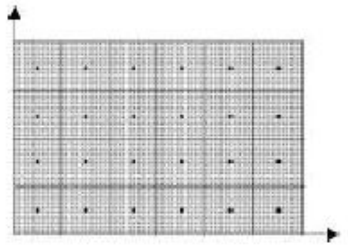
Uniform quantization for multiple dimensions

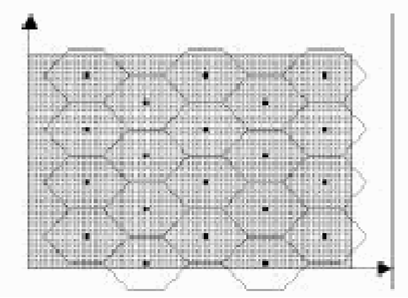
→ coverage with fewer cells (2D: hexa -mesh)

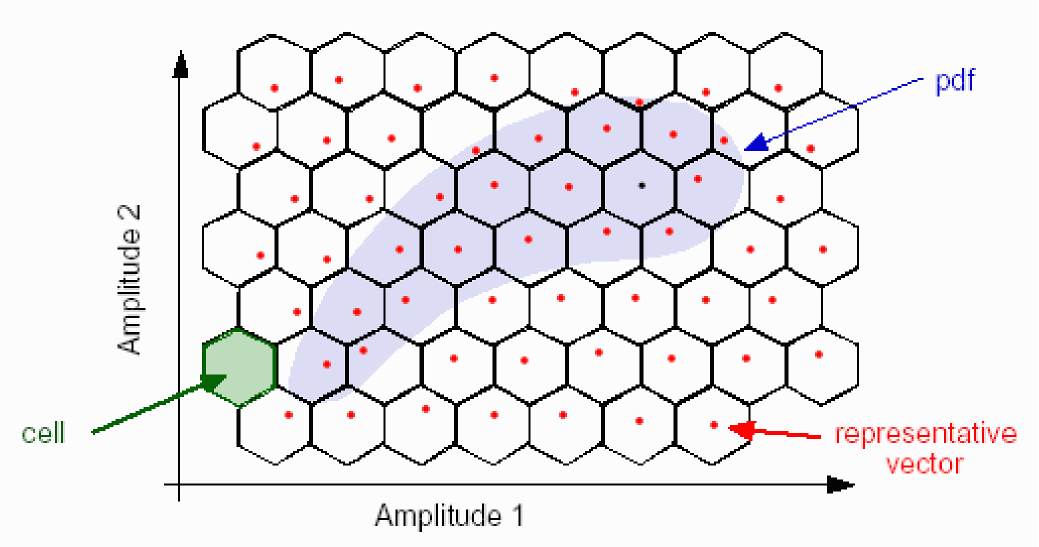
Entropy coding for indices (e.g. Huffman)

→more efficient compression than scalar quantization

→fast search due to structure







(Linear Prediction (LP)

Motivation

- In speech, it often happens that the next sample is not very different from the previous one
  - Estimation uses the values of previous samples, therefore ”prediction”

$$\widehat{x}(n) = \sum_{i=1}^{P} \alpha_i x(n-i)$$

- *x* (*n*) the signal
- α_i the prediction coefficients (L*PC*2)
- p is the prediction degree
- Linear Predictive Coding (*LPC* 1) <> (*LPC* 2) Linear Predictive Coefficients