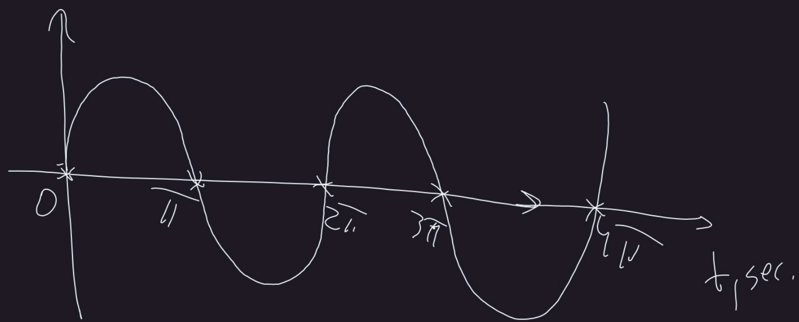
NNs for audio



2π sec. → 1 osc.

1 sec. → $\frac{1}{2\pi}$ osc.

Human perception:

16 Hz → 20 kHz

Wav-file:

bit rate 44.1 kHz

data width: 8-bit or 16-bit

number of channels (mono or stereo)

$$B = \frac{1}{2\pi} \text{ Hz}$$

44 kHz → 16 kHz

$$\frac{1}{2B} = \frac{2\pi}{2} = \pi$$

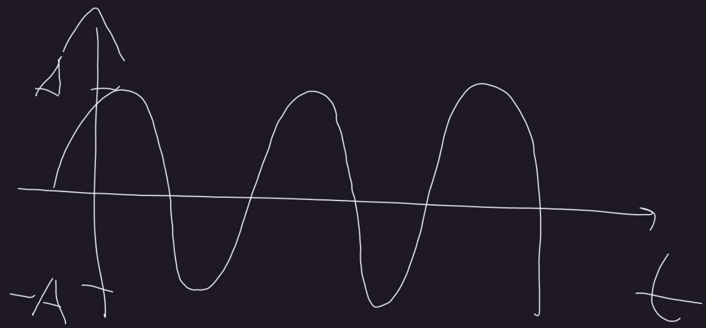
Th (Nyquist-Shannon)

$f(t)$ has largest freq. B Hz $\Rightarrow$

$f(t)$ can be restored from sampled signal,

if it is sampled $< \frac{1}{2B}$ samples per sec.

$$f(t) = A\sin(\omega t + \varphi)$$



A – amplitude

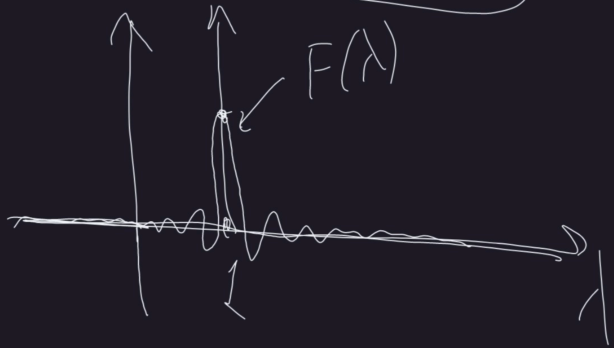
ω – freq.

φ – phase

$$f(t) \longrightarrow A, \omega, \varphi$$

$$\boxed{f(t) = \sin(t)}$$

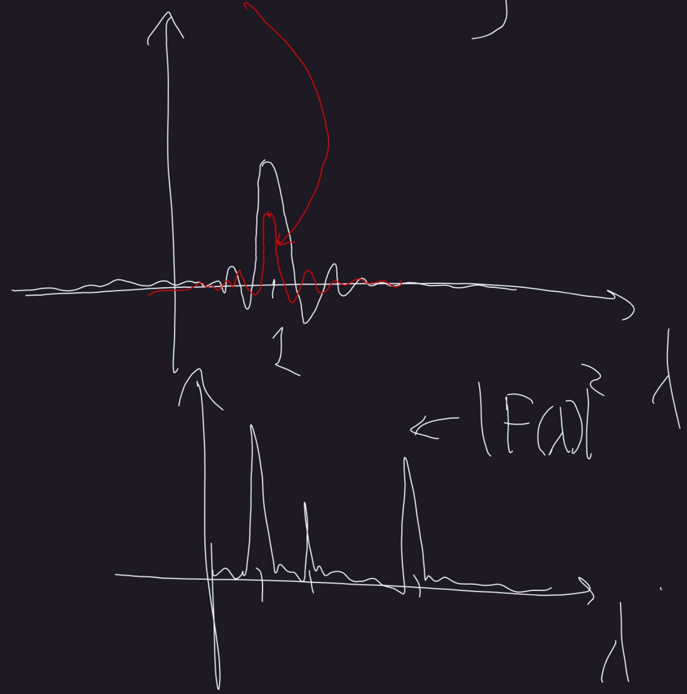
$$F(\lambda) := \int_{-\infty}^{+\infty} f(t)\sin(\lambda t)\,dt = \delta(\lambda - 1)$$



$$F(\lambda) = \int_0^T f(t)\sin(\lambda t)\,dt \approx \frac{1}{T}\sum_{n=0}^{T^2} f\left(\frac{n}{T}\right)\sin\left(\lambda \frac{n}{T}\right)$$

$$f(t) = \sin(t + \varphi)$$

$$F(\lambda) = \int f(t) \sin(\lambda t)\, dt$$



$$\underset{\in \mathbb{C}}{F(\lambda)} = \int f(t) \exp(-i\lambda t)\, dt$$

$$A = |F(\lambda)|^2$$

$$\varphi = \operatorname{Arg} F(\lambda)$$

$$\begin{cases} F_1(\lambda) = \int f(t) \sin(\lambda t)\, dt \\ F_2(\lambda) = \int f(t) \cos(\lambda t)\, dt \end{cases}$$

$$A = F_1^2(\lambda) + F_2^2(\lambda)$$

doesn't depend on φ

$$\varphi = \arctan\left(\frac{F_1(\lambda)}{F_2(\lambda)}\right)$$

Discrete Fourier Transform:

$$x_0, x_1, \ldots, x_{N-1} \xrightarrow{DFT} \underset{\overset{\uparrow}{\mathbb{C}}}{X_0}, \underset{\overset{\uparrow}{\mathbb{C}}}{X_1}, \ldots, \underset{\overset{\uparrow}{\mathbb{C}}}{X_{N-1}}$$

$$X_k = \sum_{n=0}^{N-1} x_n \exp\left(-i\, 2\pi \frac{k}{N} n\right) \quad \forall k = 0, 1, \ldots, N-1$$

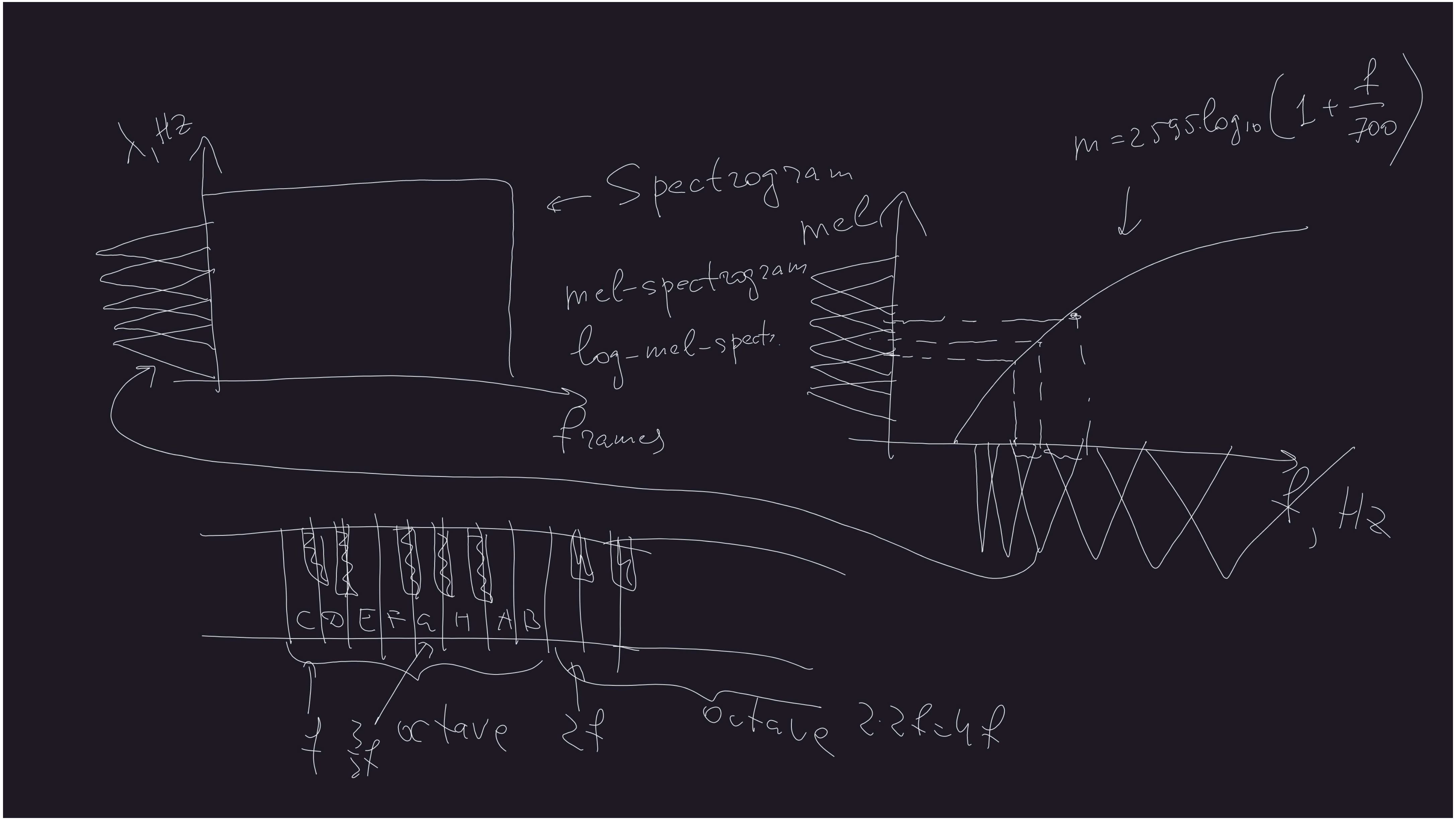
$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k \exp\left(i\, 2\pi \frac{k}{N} n\right) \quad \forall n = 0, 1, \ldots, N-1$$

Complexity of DFT

Direct comp. $O(N^2)$

FFT $O(N \log N)$

ASR: Audio $\longrightarrow$ Spectrogram $\xrightarrow{NN}$ Text



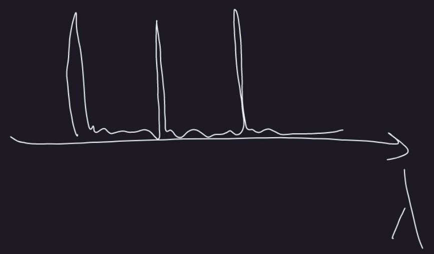
$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$
x, Hz
Spectrogram
mel
mel-spectrogram
log-mel-spectr.
frames
f, Hz
C D E F G H A B
f
$\frac{3}{2}f$
octave
2f
octave
2·2f=4f

MFCC spectrogram

mel-frequency-cepstrum-coefs

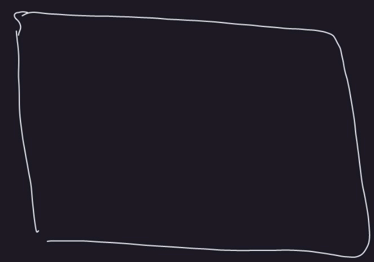
spectrum

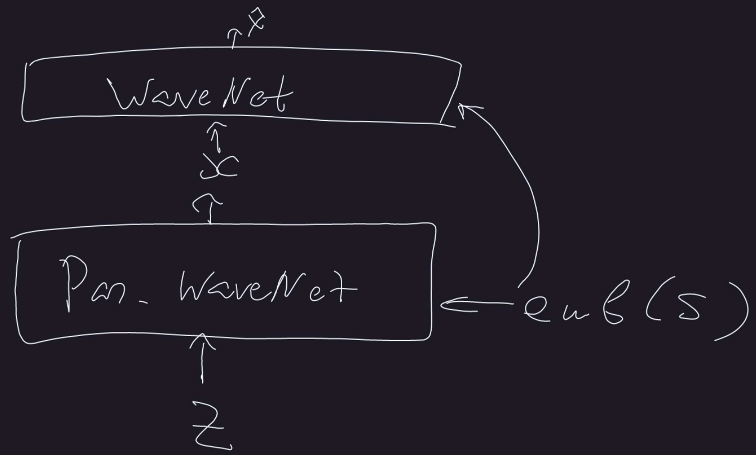
$$MFCC = \left| F\left(\log \left| F(f(t)) \right|^2 \right) \right|^2$$



spectr.

Audio →





$$x_t = z_t \cdot \sigma(z_{<t}) + \mu(z_{<t})$$

$$z \sim p(z) = \mathcal{N}(z \mid 0, I)$$

$$\log p(x) = \log p(z) - \log\left|\det \frac{\partial x}{\partial z}\right| = \sum_t \log p(z_t) - \log \prod_{t=1}^{T} \sigma(z_{<t}) =$$

$$= \sum_t \log p(z_t) - \sum_t \log \sigma(z_{<t})$$

$$KL\left(p_S(x|\theta)\,\|\,p_T(x)\right) \to \min_\theta$$

$$\parallel$$

$$\mathbb{E}_{p_S(x|\theta)} \log \frac{p_S(x|\theta)}{p_T(x)} = \mathbb{E}_{p_S(x|\theta)}\left(\sum_t \log p(z_t) - \sum_t \log \sigma(z_{<t},\theta)\right) -$$

$$- \mathbb{E}_{p_S(x|\theta)} \sum_t \log p_T(x_t | x_{<t}) = \left\{ \begin{array}{l} z \sim p(z),\ x = f(z) \\ \mathbb{E}_{p(x)} g(x) = \mathbb{E}_{p(z)} g(f(z)) \end{array} \right\} =$$

$$= \mathbb{E}_{p(z)}\left(\overset{\text{const}}{\cancel{\sum_t \log p(z_t)}} - \sum_t \log \sigma(z_{<t},\theta)\right) - \mathbb{E}_{p(z)} \sum_t \log p_T\Big(\underbrace{z_t \cdot \sigma(z_{<t}) + \mu(z_{<t})}_{x_t} \,\Big|\, \ldots\Big)$$