Arab Republic of Egypt
Ministry of Communications and Information Technology

Information Technology Institute

EPITA
SCHOOL OF ENGINEERING & COMPUTER SCIENCE

A.I. IN AUDIO & SIGNAL PROCESSING

Session 1: Signal, Audio, Speech encoding

INTRODUCTION

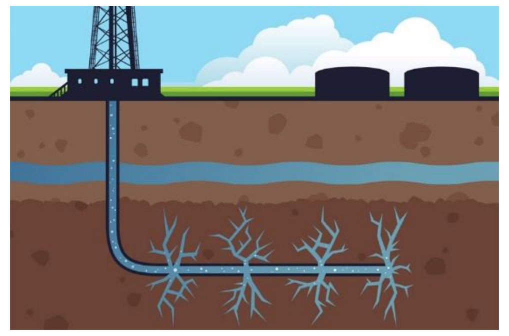
Gaël Laqueille

Parrot

move wireless

CentraleSupélec

(Supélec)



Schlumberger

faurecia

inspiring mobility

COURSE STRUCTURE

Quick Summary

Audio processing for AI

- Signal, audio, speech encoding (4h)
- Deep learning for audio processing (4h)

Automata for language modelling

- HMM for speech processing (4h)
- Automata and transducer (4h)

Towards speaking with an AI-bot

- Speech synthesis (4h)
- Automatic speech recognition (4h)
- Speaker and emotion recognition (4h)

SESSION 1: SIGNAL, AUDIO, SPEECH ENCODING

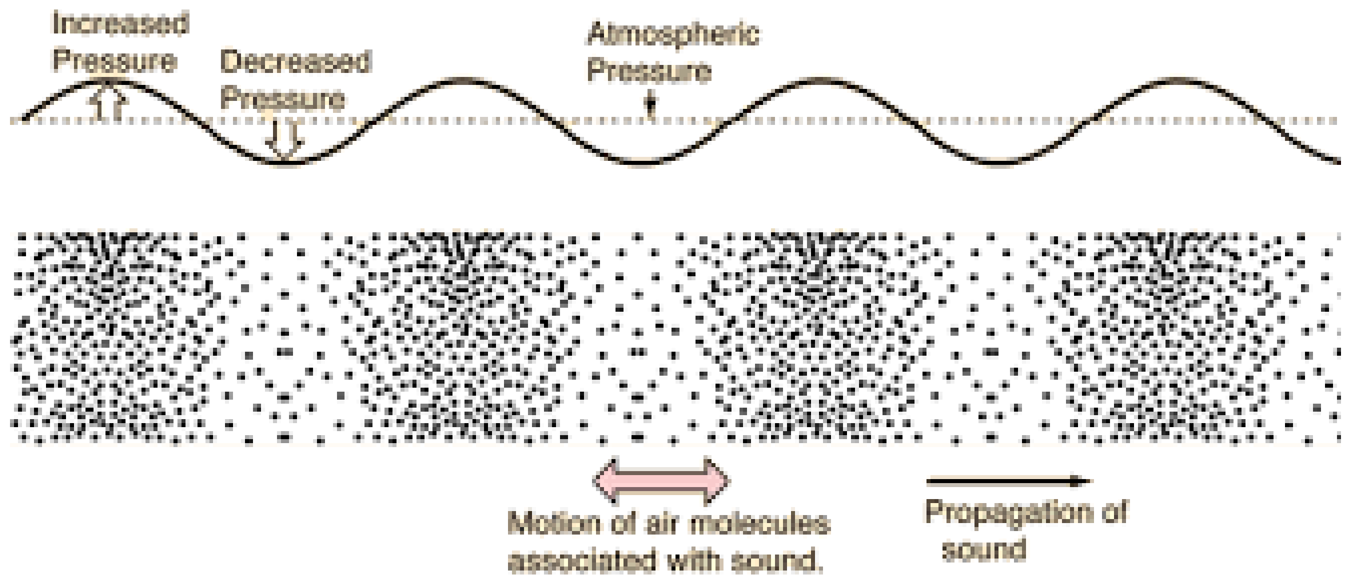
Quick Summary

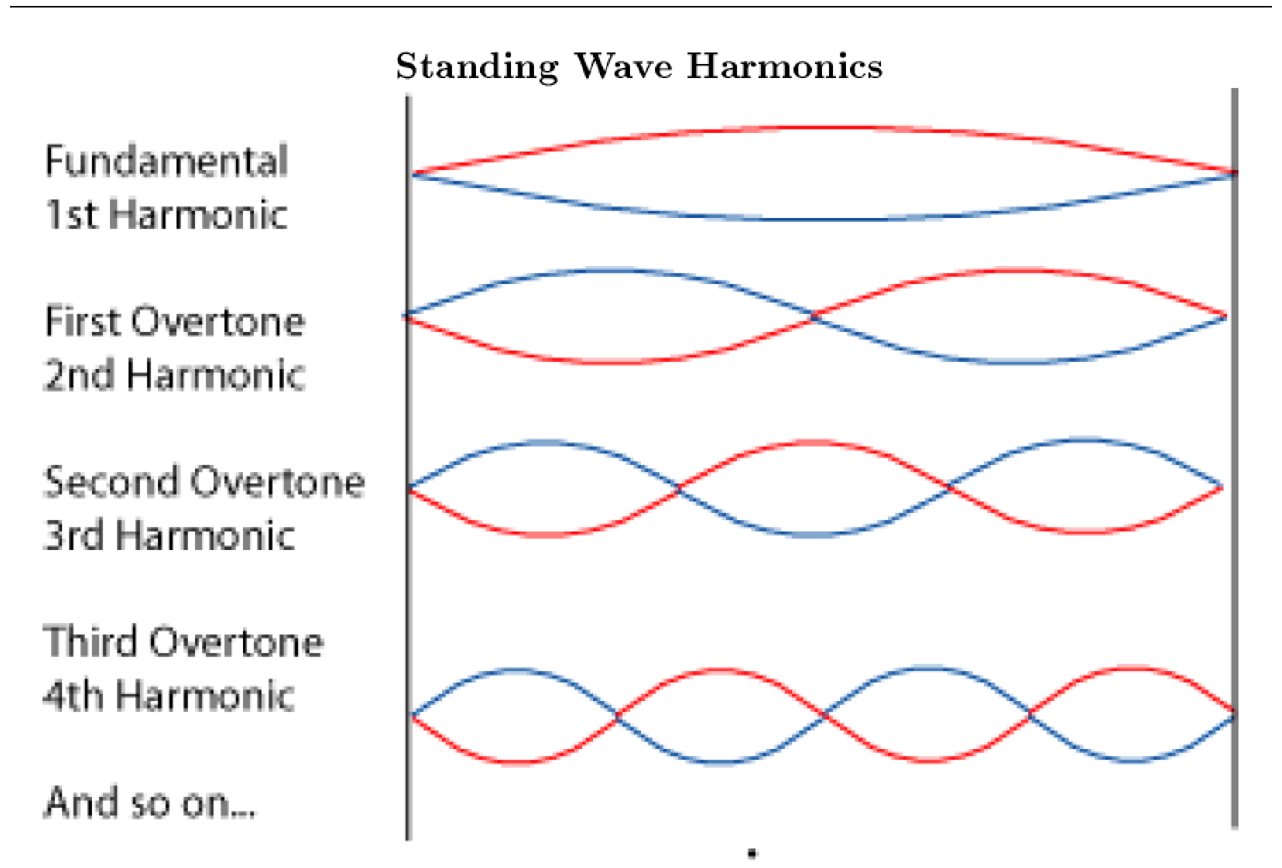
1. The physics of sound
2. Signal representation and purpose
3. Signal characteristics
4. Signal models
5. Classical ML approaches

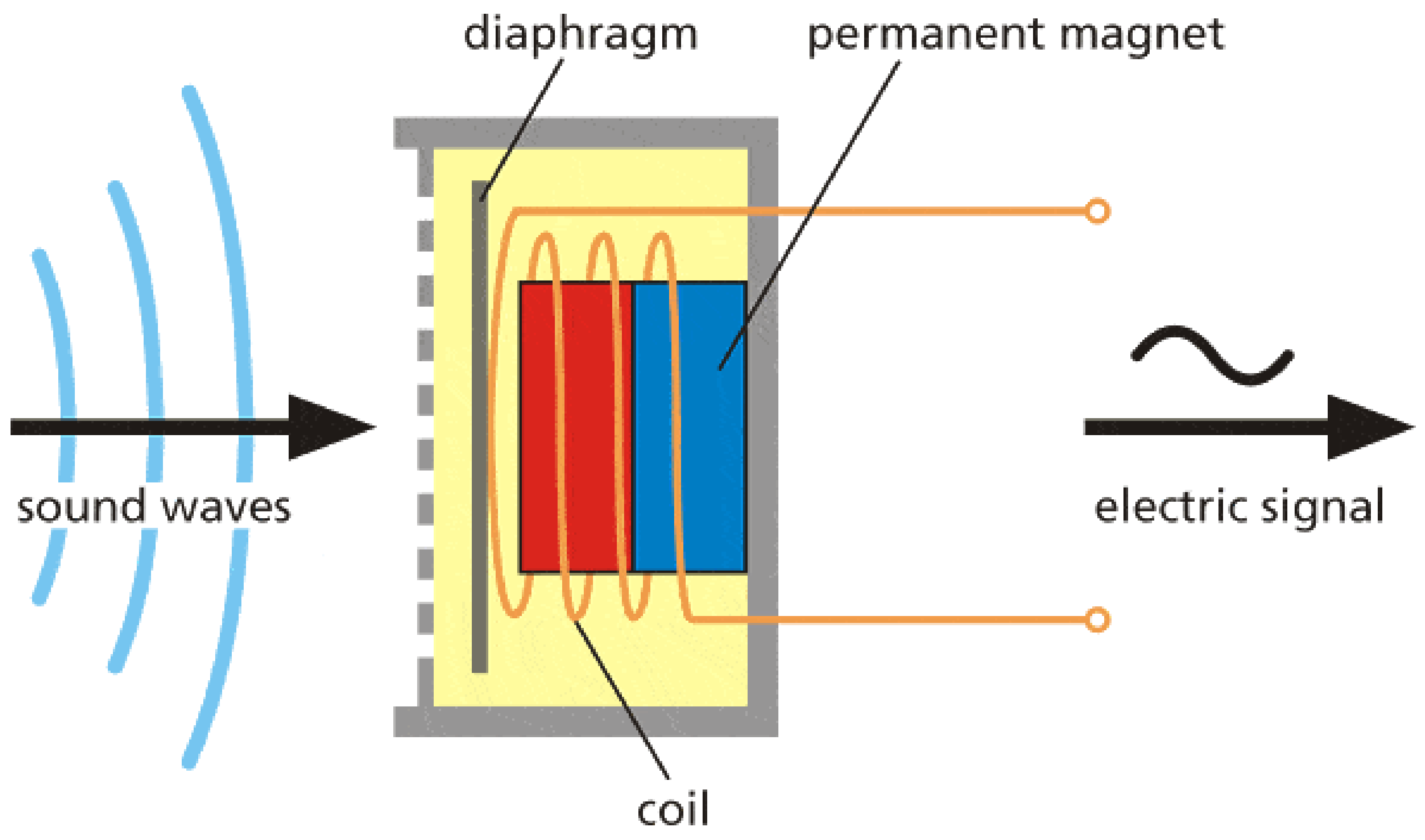
SIGNAL, AUDIO, SPEECH ENCODING.

The physics of sound

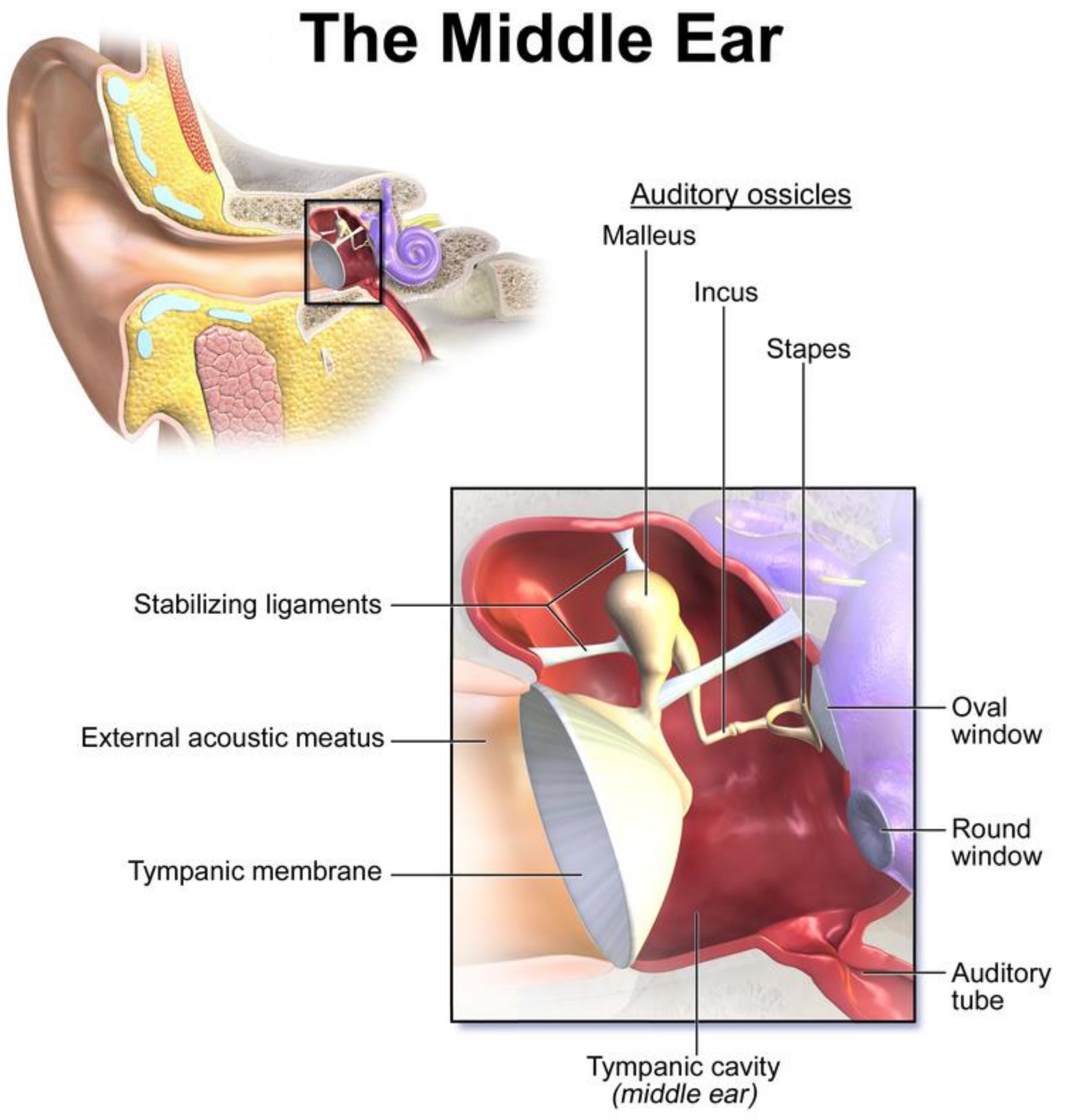
THE PHYSICS OF SOUND

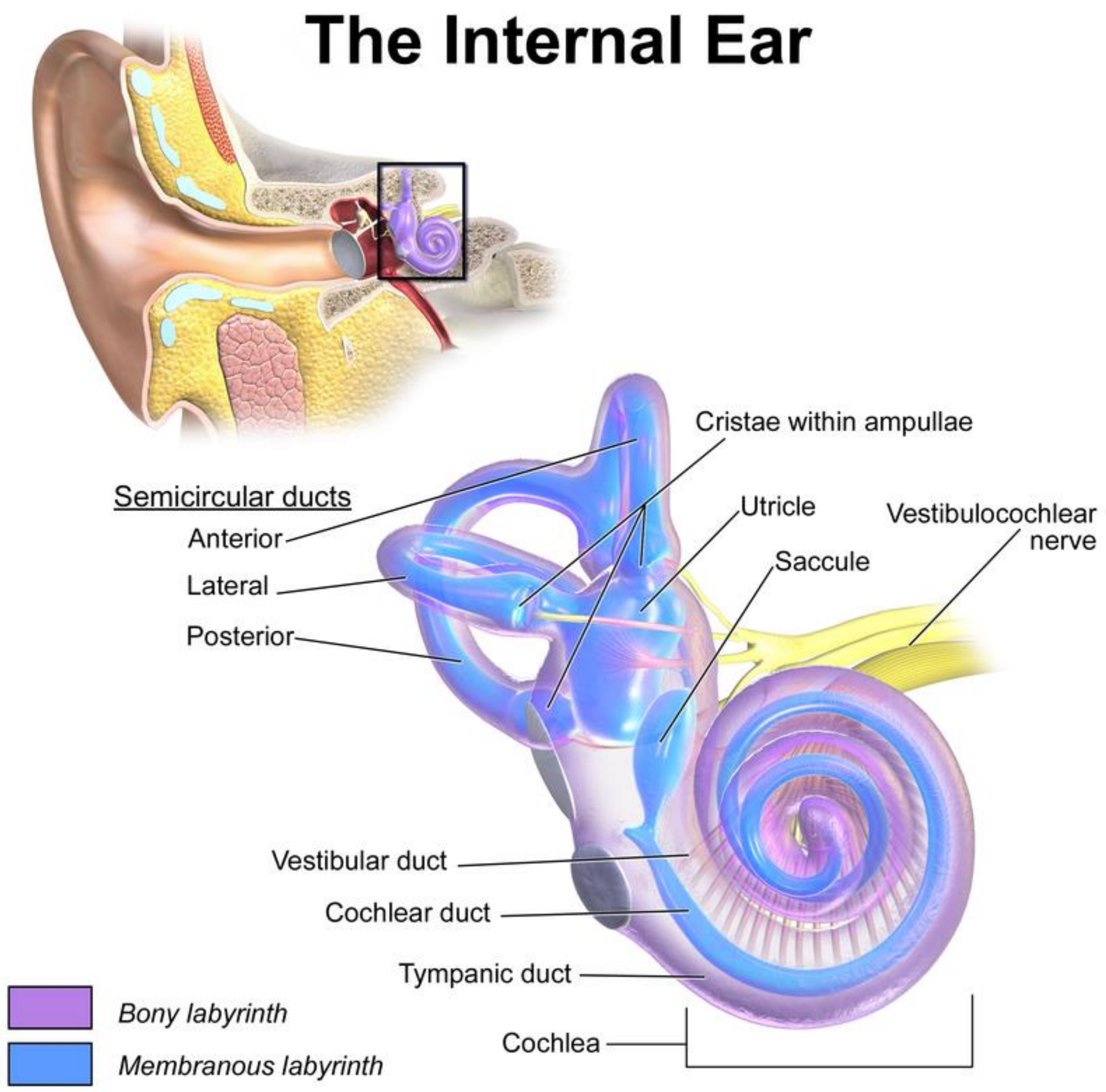






THE PHYSICS OF SOUND



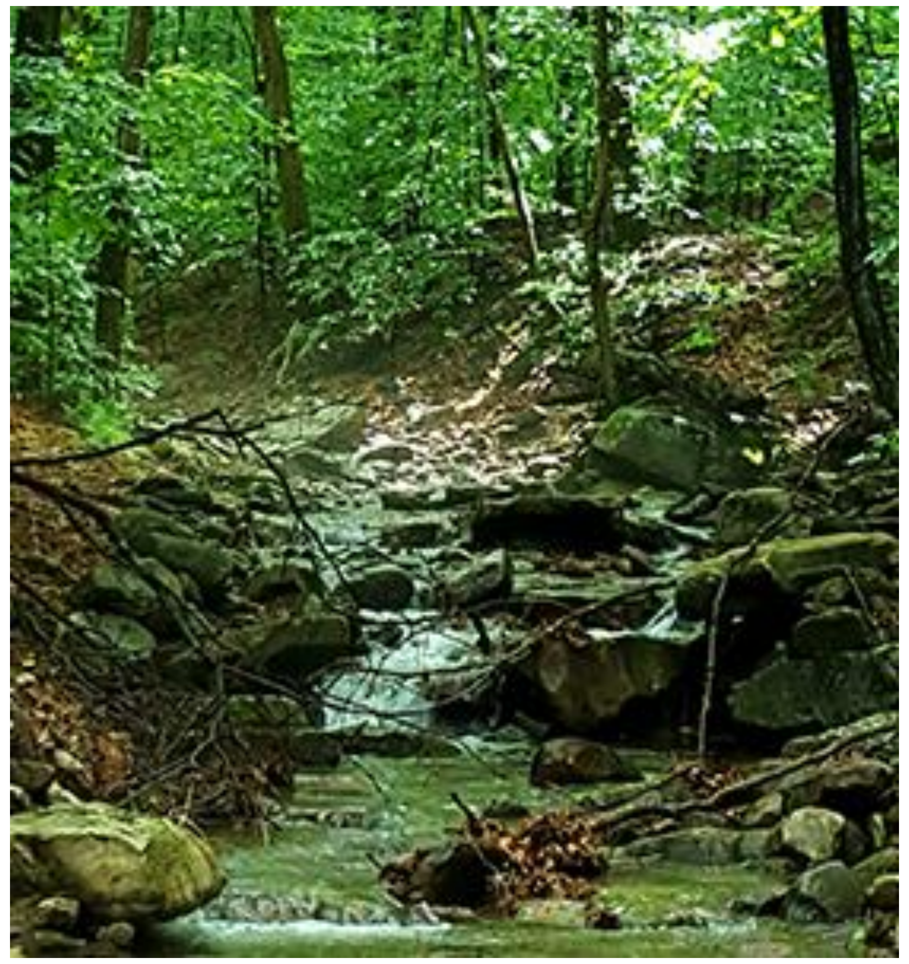


SIGNAL, AUDIO, SPEECH ENCODING.

Signal representation & purpose

REPRESENTATION & PURPOSE

Different types of audio content

Noise	Music	Speech
		
Multisource polyphonic with unstructure sound sources	Multi timbre polyphonic with structured sound sources	Monophonic with structured sound source

Extrapolation to other signals content: seismic, EEG,…

REPRESENTATION & PURPOSE

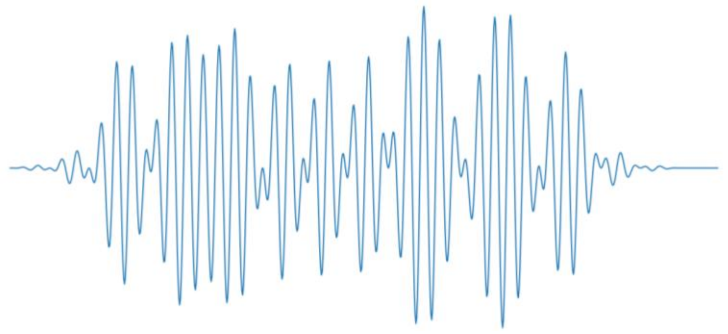
Signal & Audio symbolic representations

Codification

Production

Recording

Digitalization



(Transcription)

Vertical and horizontal organization

Frequency measurement:
- Sources
- Timbres
- Pitch polyphony
- Chord
- Tonality

Time measurment:
- Structure
- Bars
- Rythm

REPRESENTATION & PURPOSE

Applications of AI in signal processing

Musical Information Retrieval (ISMIR)
- Automatic Music Transcription
- Auto-tagging
- Music Recommendation
- Source separation
- Audio/Music generation
- Style transfer
- Fingerprinting

Spoken Content Retrieval
- Trigger word detection
- Segmentation
- Speech-to-text (ASR)
- Text-to-speech
- Emotion recognition
- Speaker recognition, diarization
- Denoising
- Active Noise Control

Other signal processing applications
- Time series prediction
  (finance, insurance, healthcare,…)

REPRESENTATION & PURPOSE

Fourier series

$$f(t) = \sum_{n=0}^{+\infty} k_n \sin(2\pi n f_0 t + \phi_n)$$

$$= \sum_{n=0}^{+\infty} A_\mathrm{n} \cos(2\pi n f_0 t) + B_\mathrm{n} \cos(2\pi n f_0 t)$$

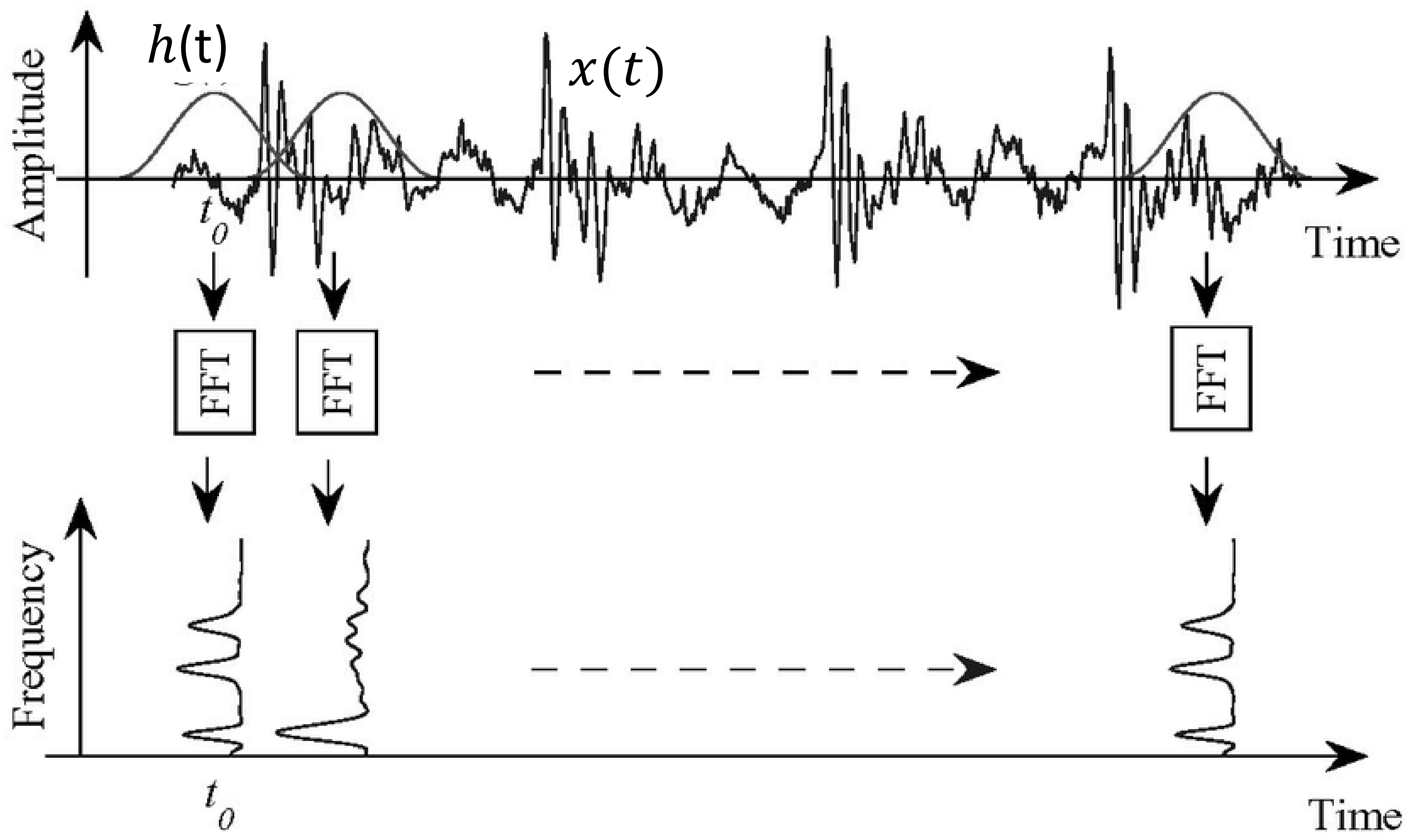
$$= \sum_{n=0}^{+\infty} c_n \, e_0^{j2\pi f t}$$

Discrete Fourier Transform

$$X(k) = \sum_{m=0}^{N-1} x(m) \, e_0^{j2\pi \frac{k}{N} m}$$

Short-Time-Fourier-Transform (STFT)

$$X(k,n) = \sum_{m=0}^{N-1} x(m) \, h(n-m) e_0^{j2\pi \frac{k}{N} m}$$

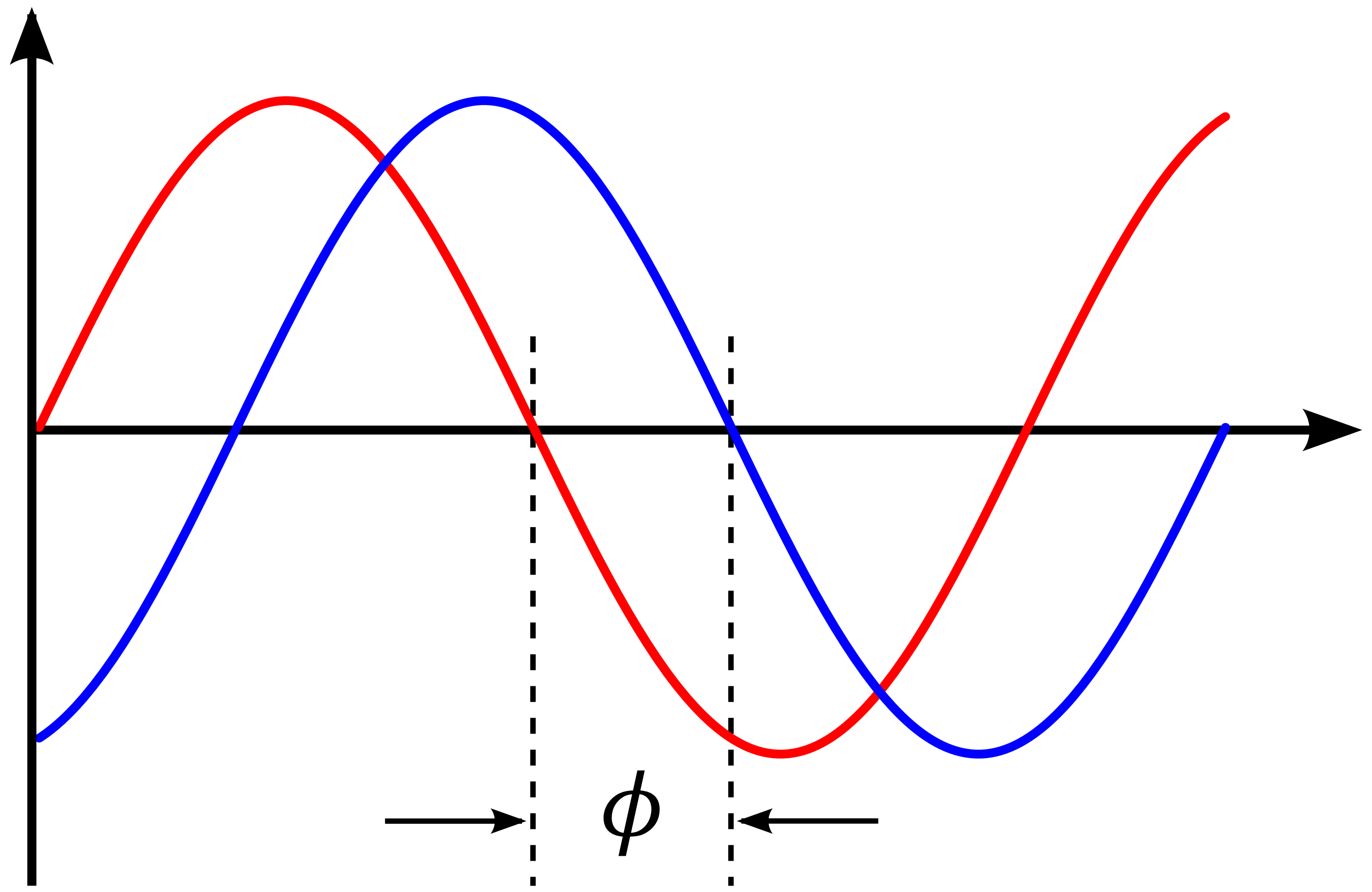


h: windowing

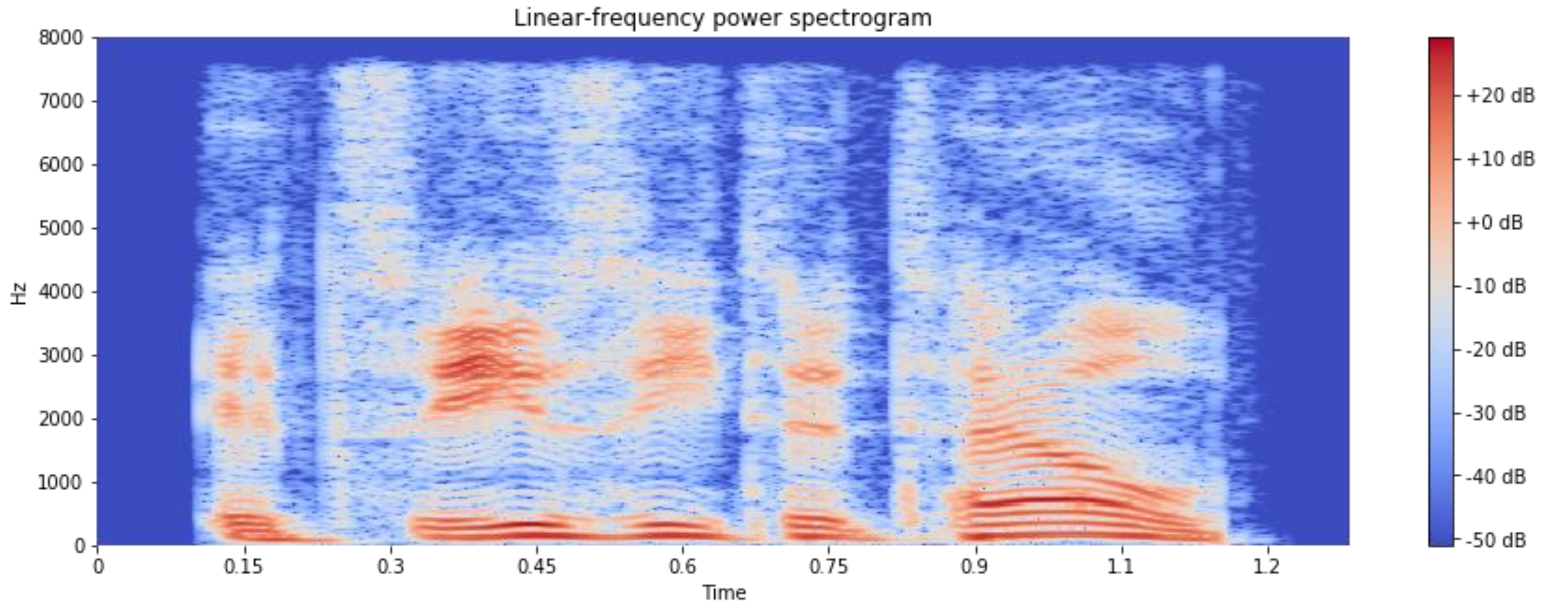
$k_n, A_n, B_n, c_n, x(m), h(m) \in \mathbb{R}$

REPRESENTATION & PURPOSE

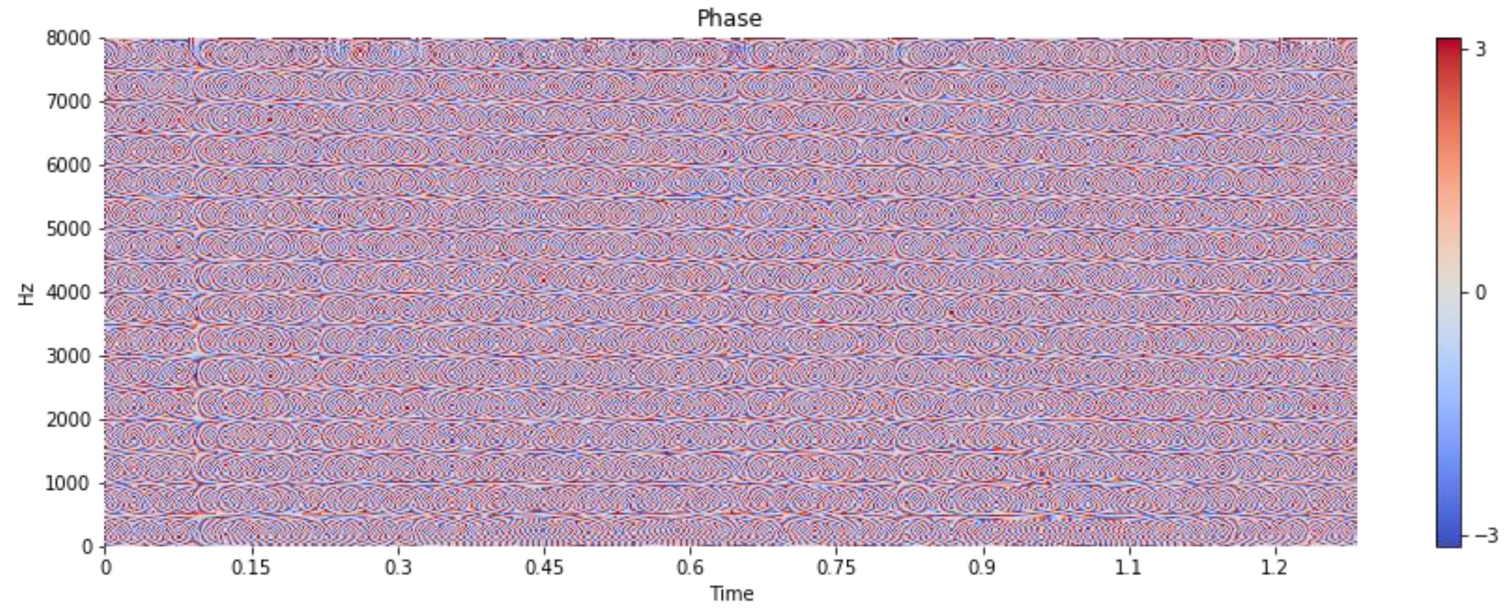
- $X(k) = |X(k)|e^{j\phi_X(k)}$
- Amplitude $|X(k)|$
- Phase $\phi_X(k)$, temporal localization of information
  (Phase is necessary to reconstruct temporal signal)
- Instantaneous frequency



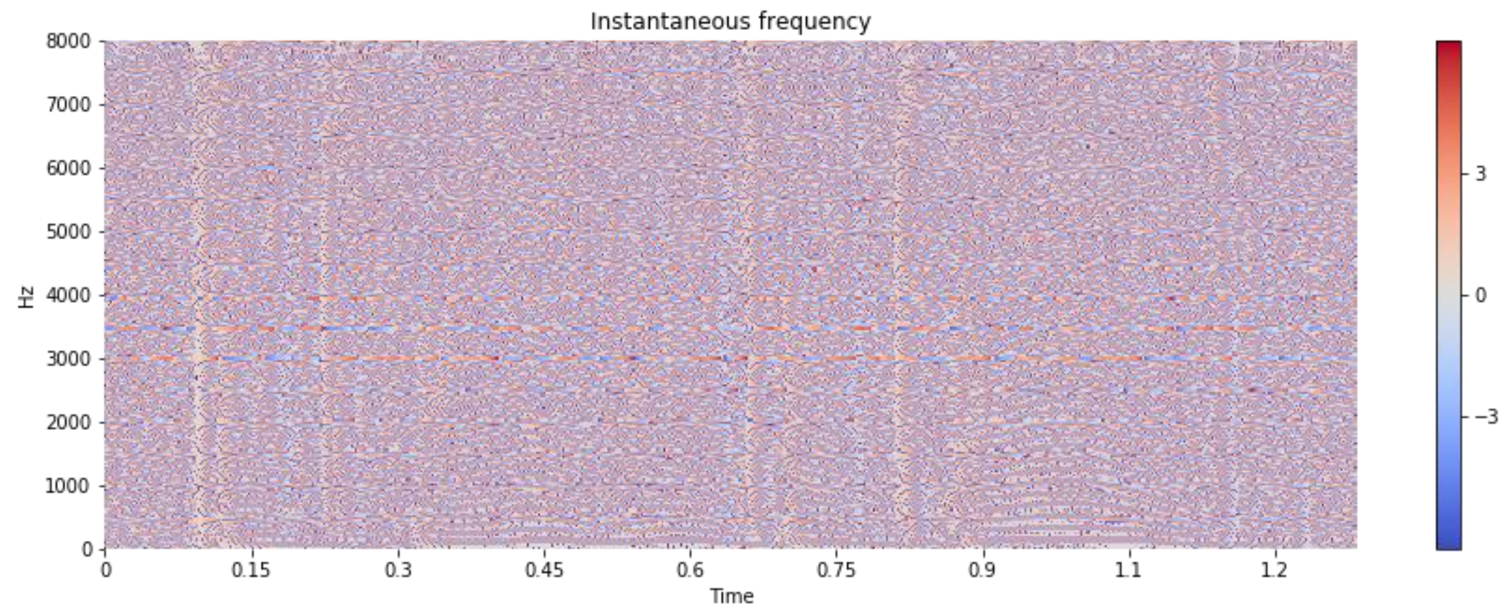
$10\log_{10}|X(k,n)|^2$



$\phi_X(k,n)$



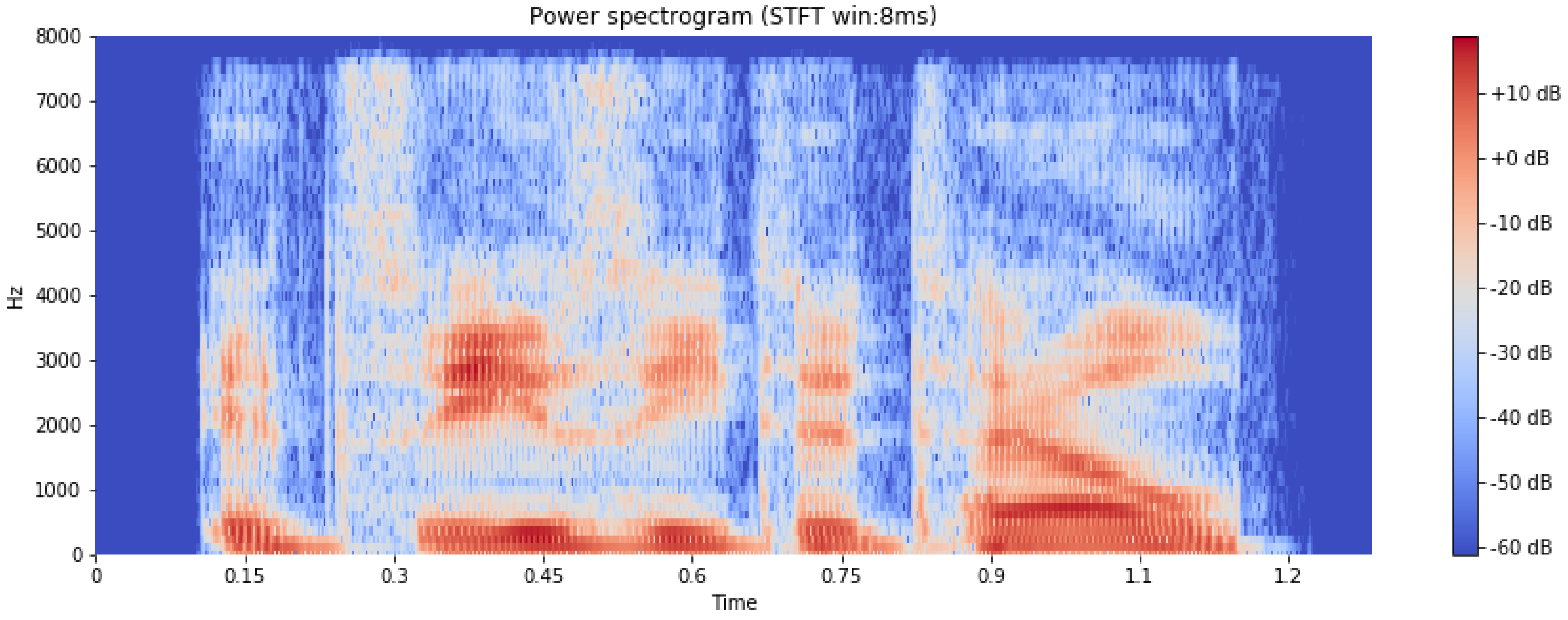
$\phi_X(k,n) - \phi_X(k,n-1)$

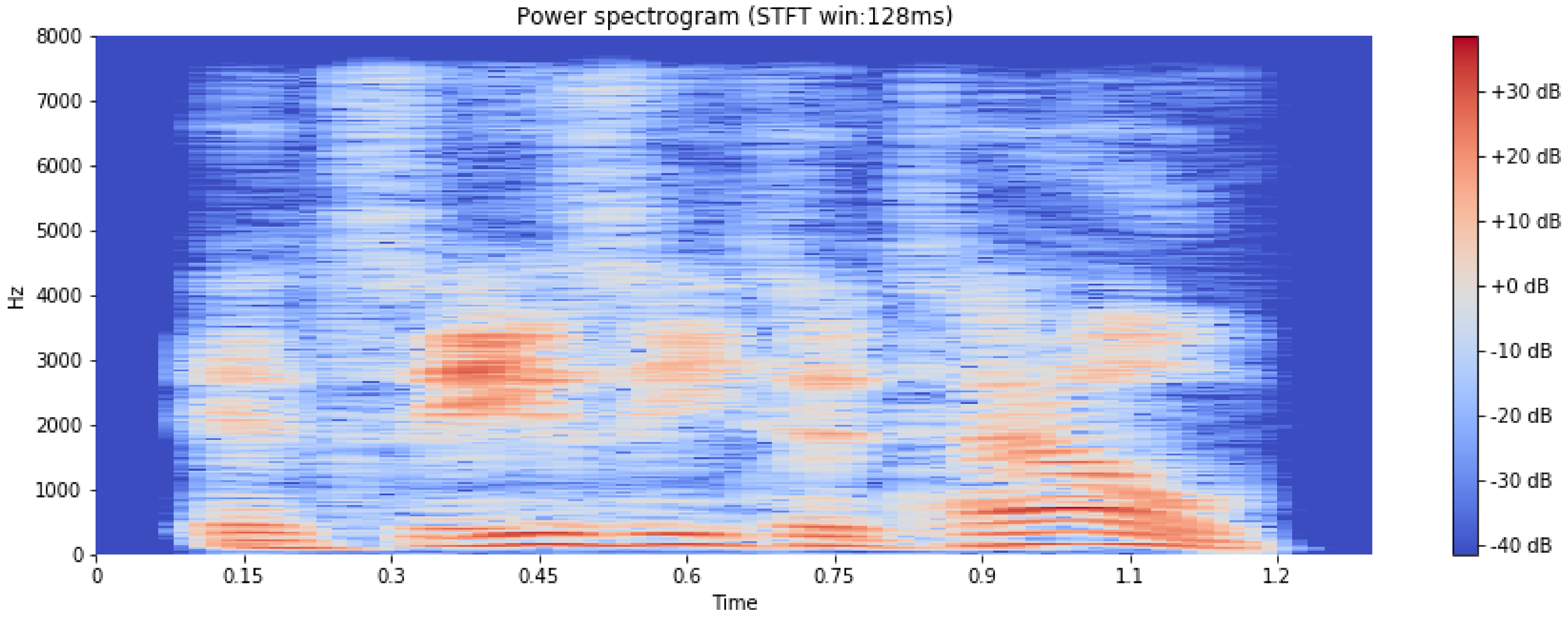


k: frequency bins
n: time windows index

REPRESENTATION & PURPOSE

Remark on trade-off temporal/frequential resolution





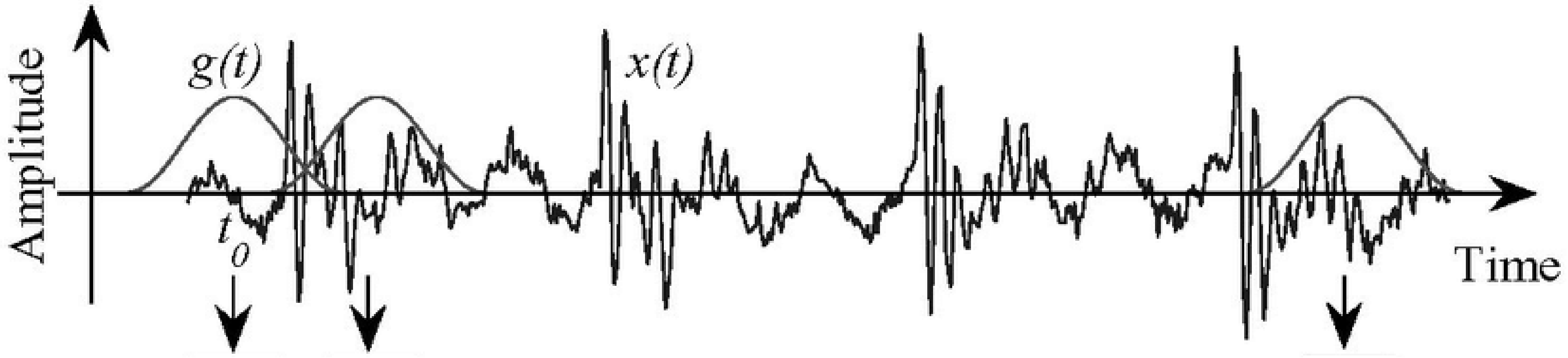
Remark on sound perception

- log perception of sound amplitude
  → magnitude spectrogram usually in dB

$$10 \log_{10} |X(k,n)|^2$$

- log perception of sound frequency
  → magnitude spectrogram often in mel scale

$$10 \log_{10} |X(k|_{mel}, n)|^2$$

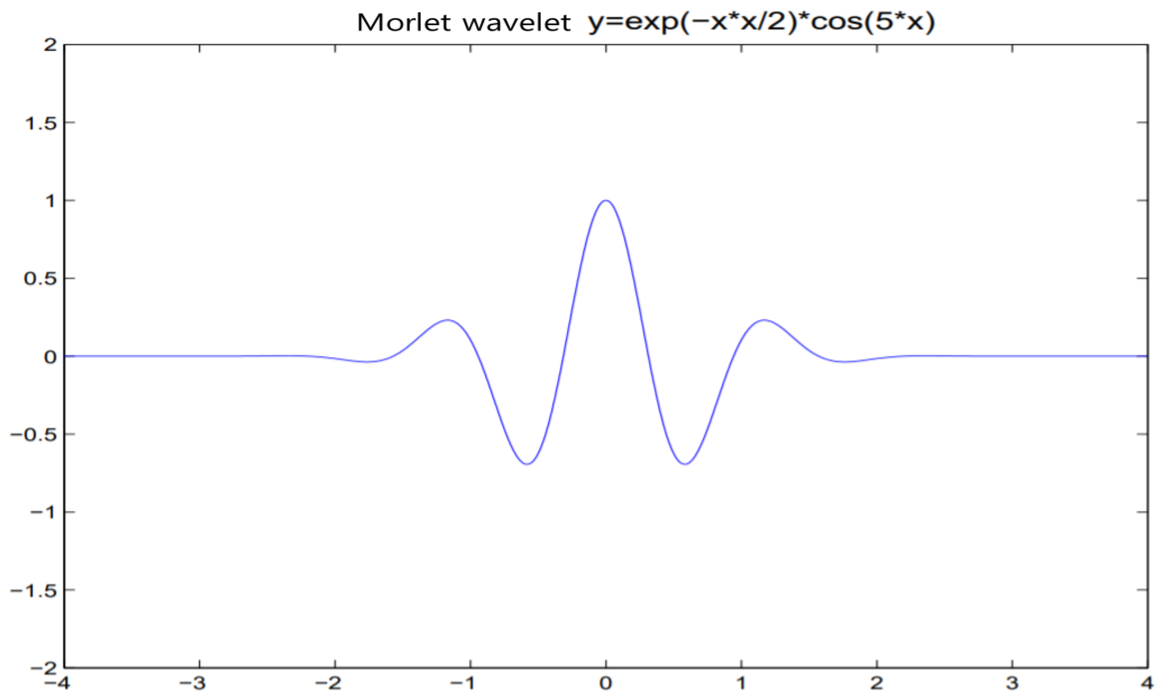


with mel scale: $m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$

REPRESENTATION & PURPOSE

Wavelet transform (WT)

(not used in practice, but concept is used to define CQT, see next slide)



- Continuous wavelet basis definition

$$\forall t \in \mathbb{R}, \psi_{s,\tau}(t) = \frac{1}{\sqrt{s}}\psi\left(\frac{t-\tau}{s}\right)$$

ψ is the "mother" wavelet
$\psi_{s,\tau}$ are the "child" wavelets that form a basis
s and τ are the dilation and translation variables

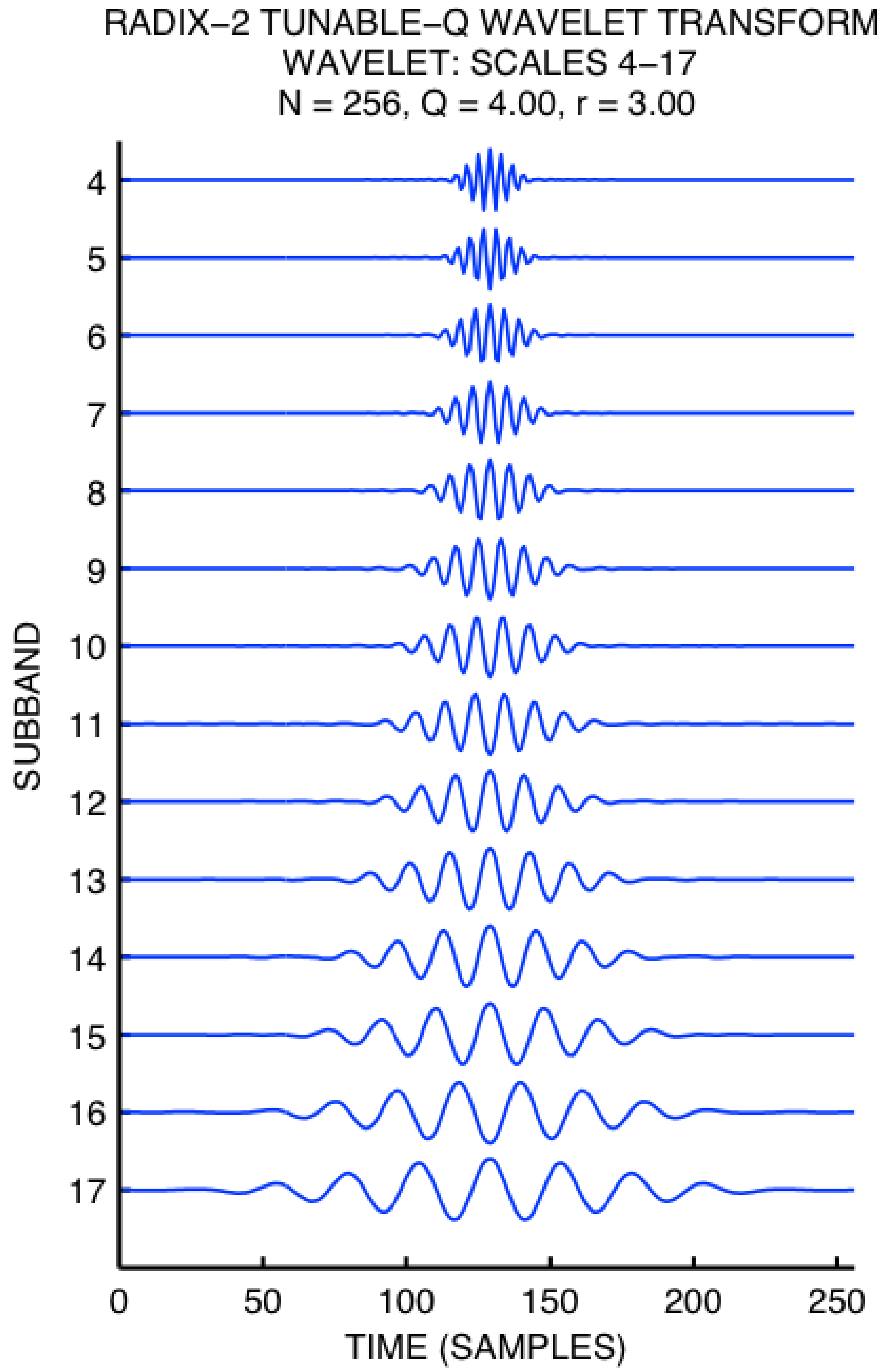
- Discrete wavelet basis definition

- discretization in wavelet (time-frequency) domain: $\tau \rightarrow n$ and $s \rightarrow k$
- discretization in time domain: $t \rightarrow m$

$$\psi_{k,n}[m] = s_0^{-\frac{k}{2}}\psi\left(s_0^{-k}m - n\tau_0\right)$$ with s_0 and τ_0 constants

- Discrete wavelet transform definition

$$W_\psi[k,n] = \frac{1}{\sqrt{N}}\sum_{m=0}^{N-1} x[m]\psi_{k,n}[m]$$ with n time index and k frequency bin



REPRESENTATION & PURPOSE

Constant-Q transform (CQT)

(behavior of WT but no orthogonality CQ functions)

- Frequency resolution changing with frequency considered

$$N_{[k]} = Q \frac{f_s}{f_k}$$

- A pitch translation corresponds to a frequency translation on CQT

$$X_{[k,n]}^{CQ} = \sum_{m=n-\frac{N_k}{2}}^{n+\frac{N_k}{2}} \frac{h\left(\frac{n-j}{N_k} - {}^{1}/{}_{2}\right)}{N_k} e^{-i2\pi Q\left(\frac{n-j}{N_k} - {}^{1}/{}_{2}\right)}$$

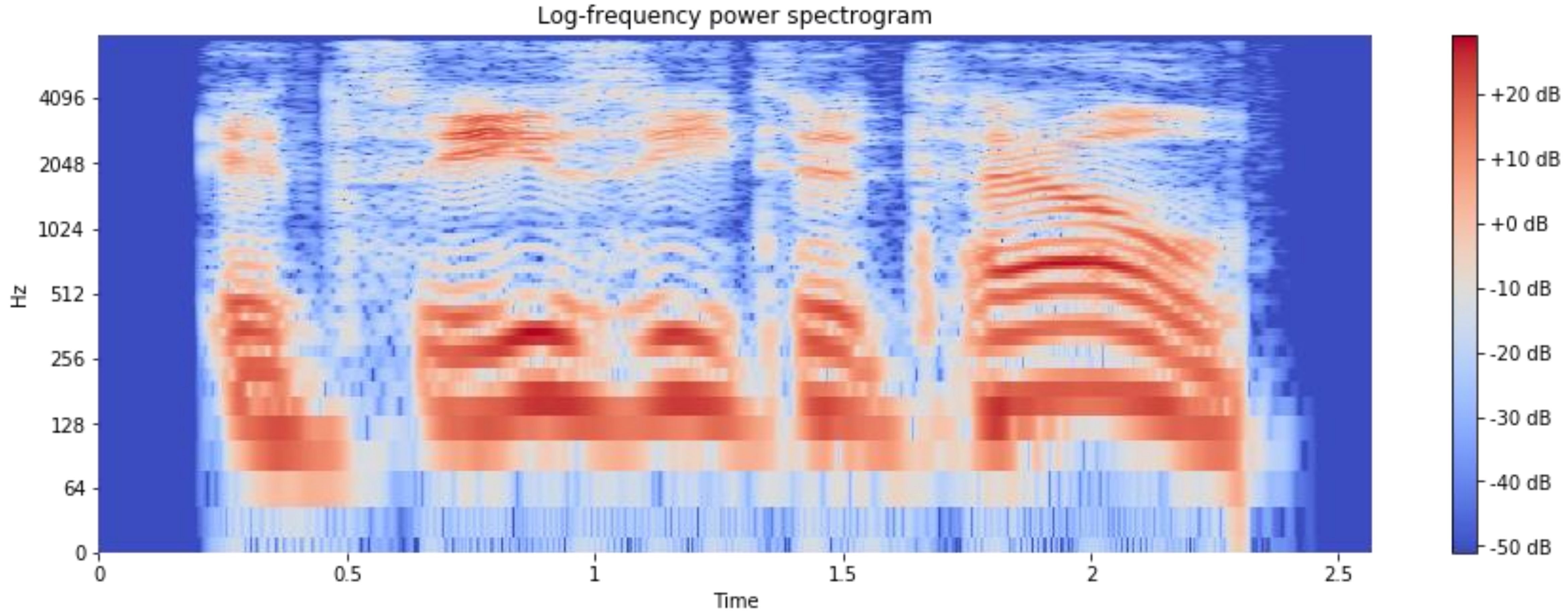
h is a window (Hanning, Blackmann Harris, …)

C. Schorkhube, A. Klapuri, Constant-Q transform toolbox for music processing. 2010.

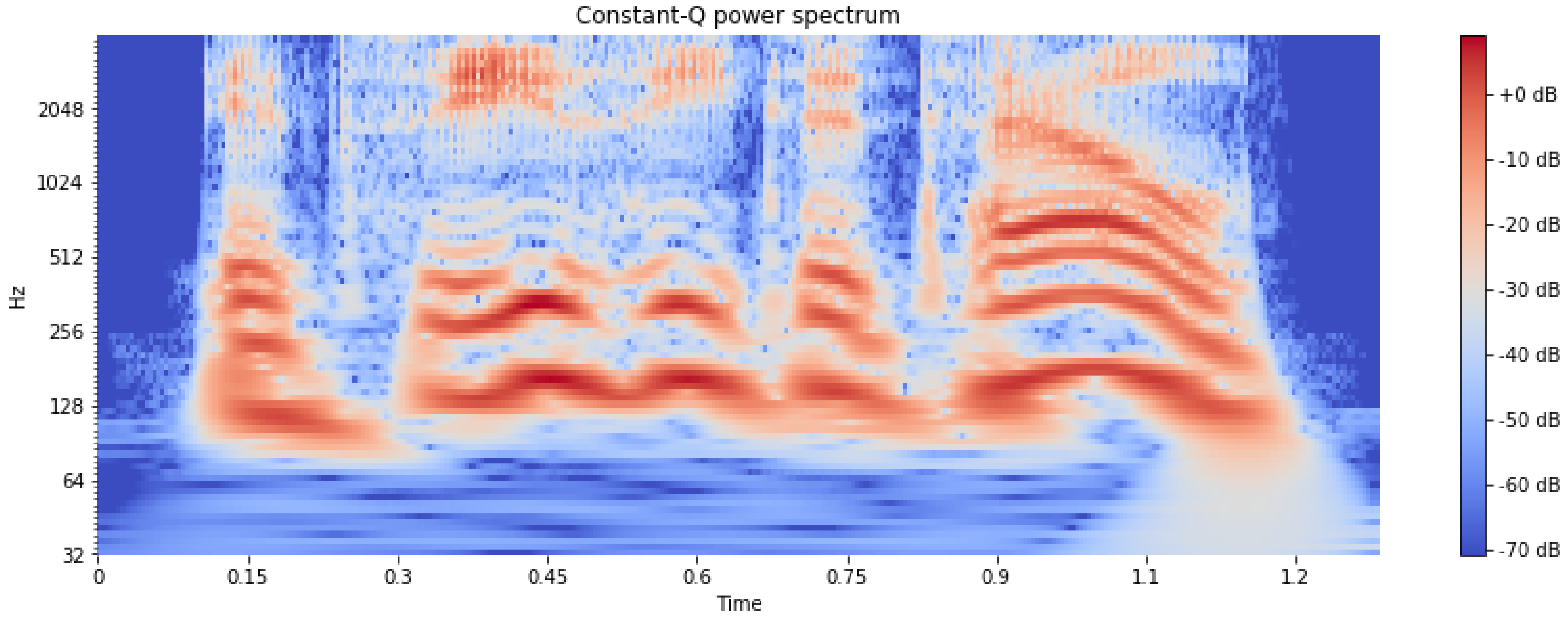
REPRESENTATION & PURPOSE

FT vs CQT

Fourier transform



Constant-Q transform



SIGNAL, AUDIO, SPEECH ENCODING.

Signal characteristics

CHARACTERISTICS

Digital waveform characteristics

- Sample rate
- Bits depth
- level dBfs
- mean RMS (dBfs)
- Compression(dBA, dBC, SPL)

Artifacts

- noise (SNR)
- amplitude distortion (signal balance) target curve
- phase distortion (unprecise onset, spatialization)
- non-linearities (clipping for instance) THD
- delay (real-time, digitalization or processing)

Sources of artifacts

- Sound or signal capture,
- sound or signal production
- sound propagation in air
- signal propagation in wires,
- processing

SIGNAL, AUDIO, SPEECH ENCODING.

Signal models

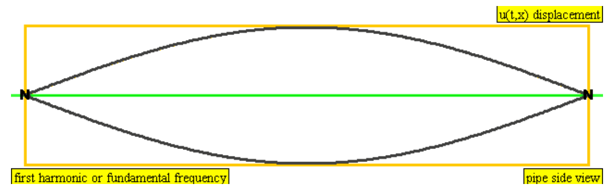
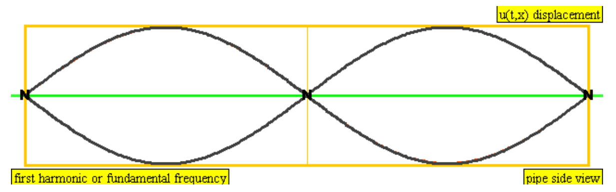
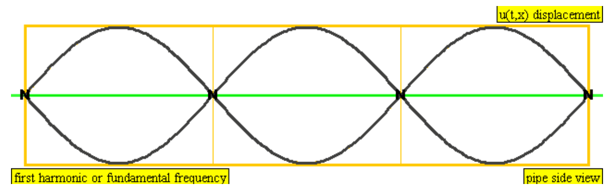
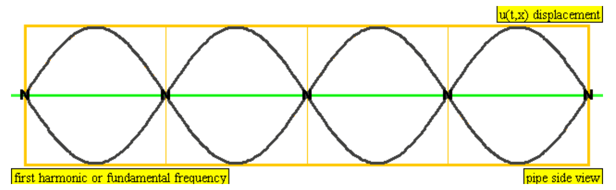
SIGNAL MODELS

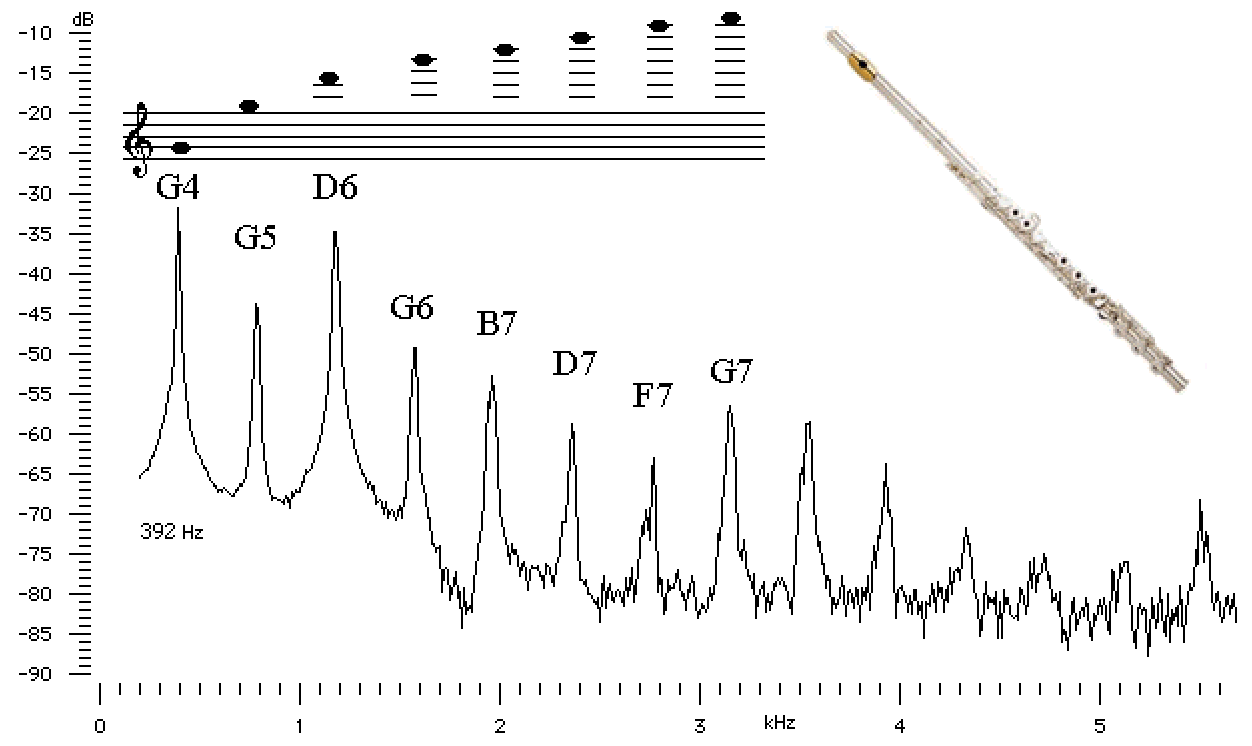
Sine wave harmonics and noise model:

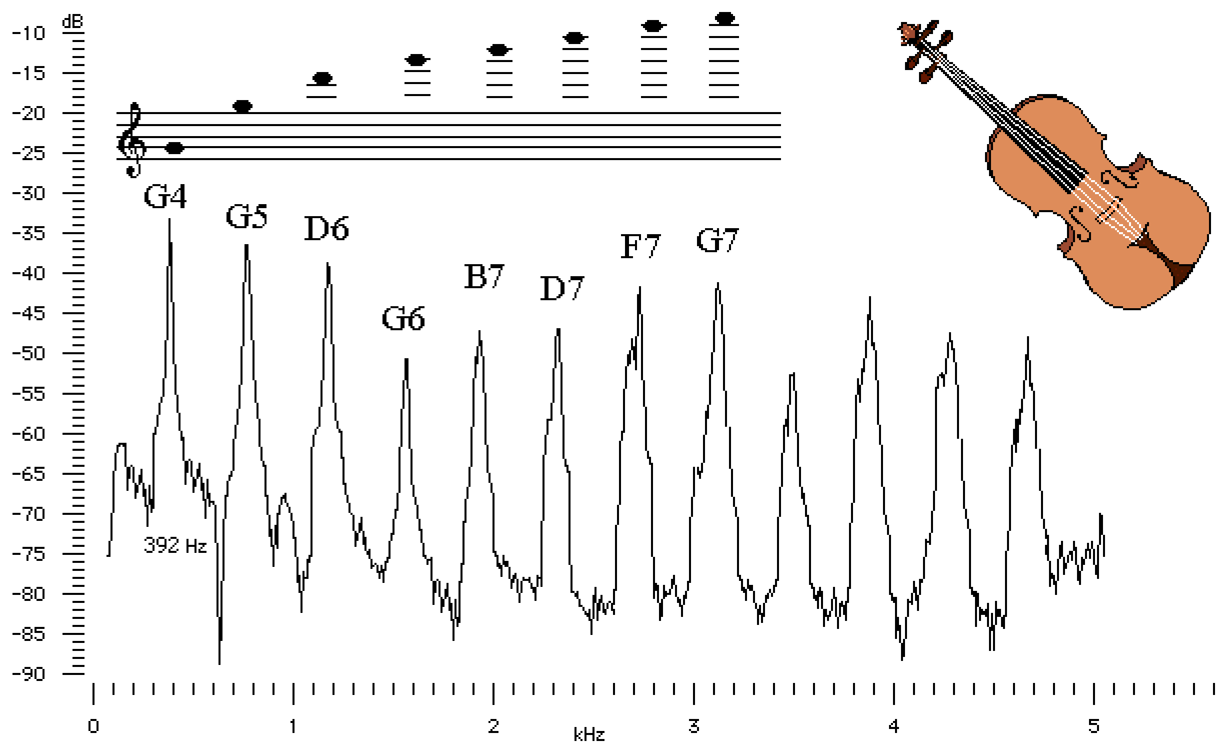
$$x(t) = \sum_{h=1}^{H(t)} A_{\mathrm{h}}(\mathrm{t}) \cos\big(\phi_h(t)\big) + b(t)$$

Assumption that the audio signals follows this model.

The spectral enveloppe $\{A_h\}$ defines the timbre of the instrument.
Example of voice or instruments harmonic spectrum (spectral envelop)

Frequency	Order	Name	Standing wave representation
$1 \times f$ = 440 Hz	n = 1	1st harmonic (fundamental tone)	
$2 \times f$ = 880 Hz	n = 2	2nd harmonic (1^{st} overtone)	
$3 \times f$ = 1320 Hz	n = 3	3rd harmonic (2^{nd} overtone)	
$4 \times f$ = 1760 Hz	n = 4	4th harmonic (3^{rd} overtone)	





SIGNAL MODELS

Source/filter model

- Periodical pulse convolved with a resonant filter

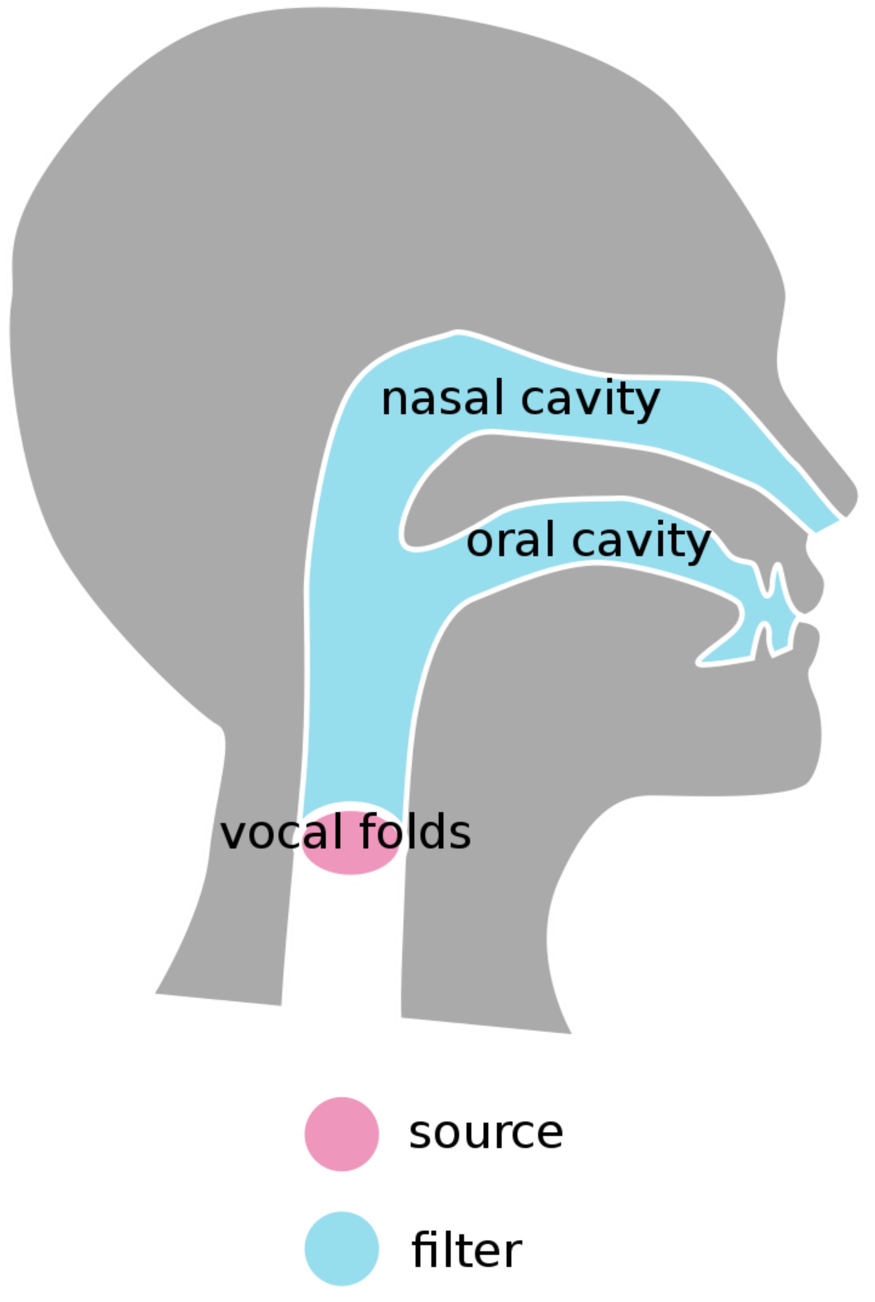
$$x(t) = e(t) * h(t)$$
$$X(f) = E(f) \cdot H(f)$$

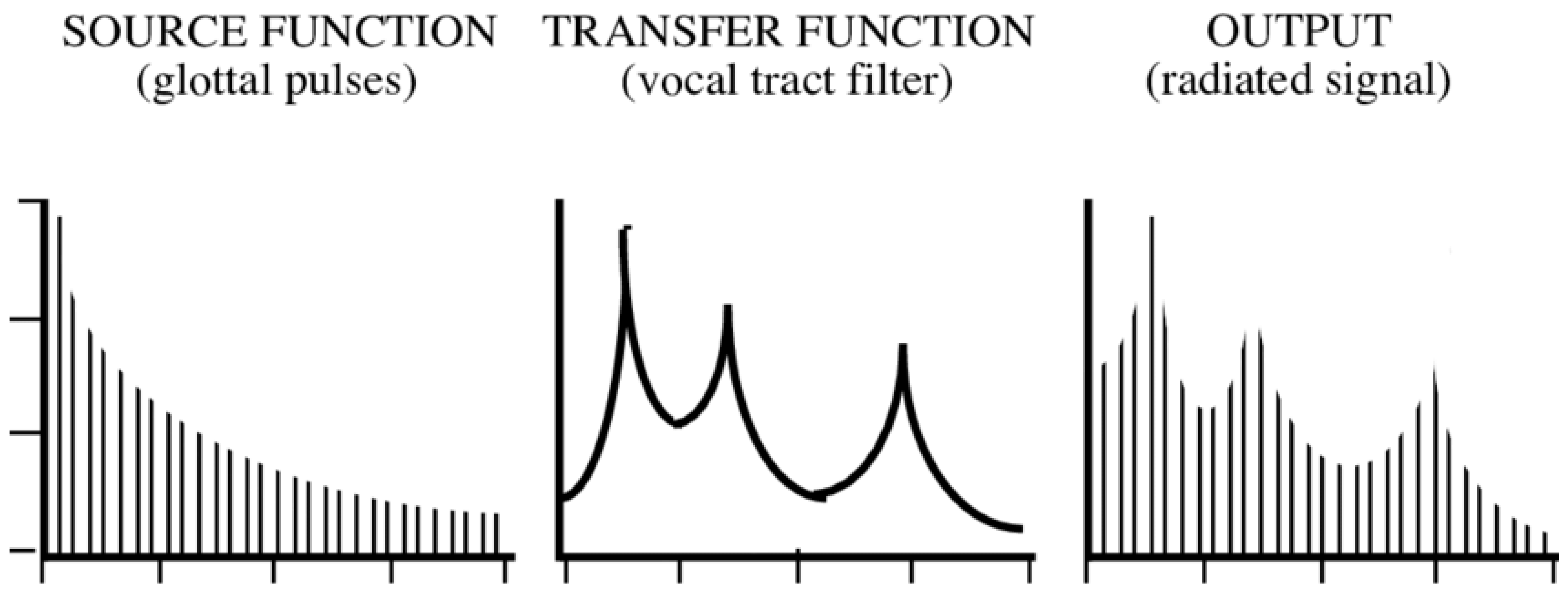
- Resonant filter is an AR filter:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{K} a_k z^{-k}}$$

x(n) can be predicted with a linear combinaison of previous values

$$x(n) \approx \sum_{k=1}^{K} a_k \, x(n-k)$$



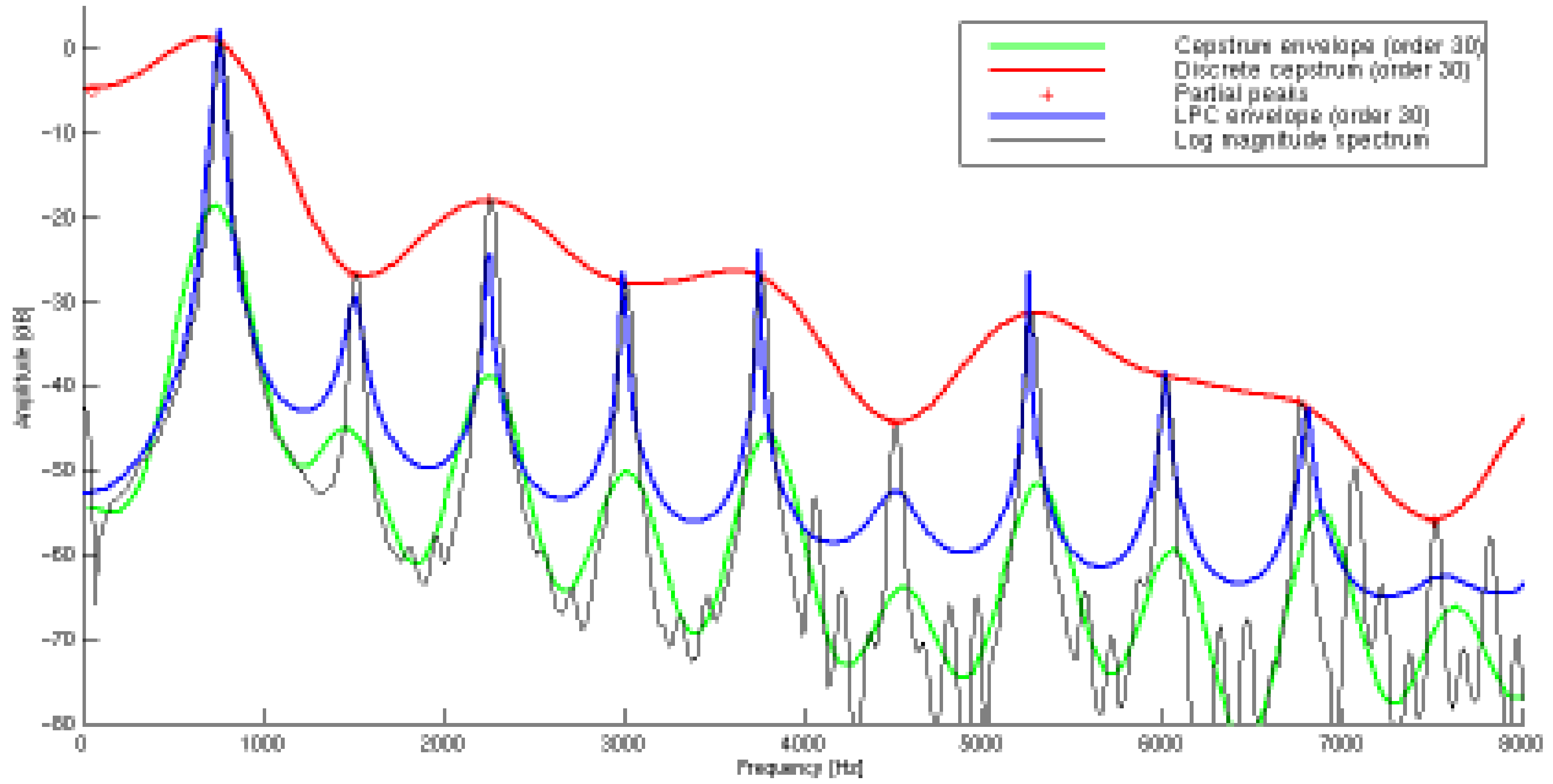


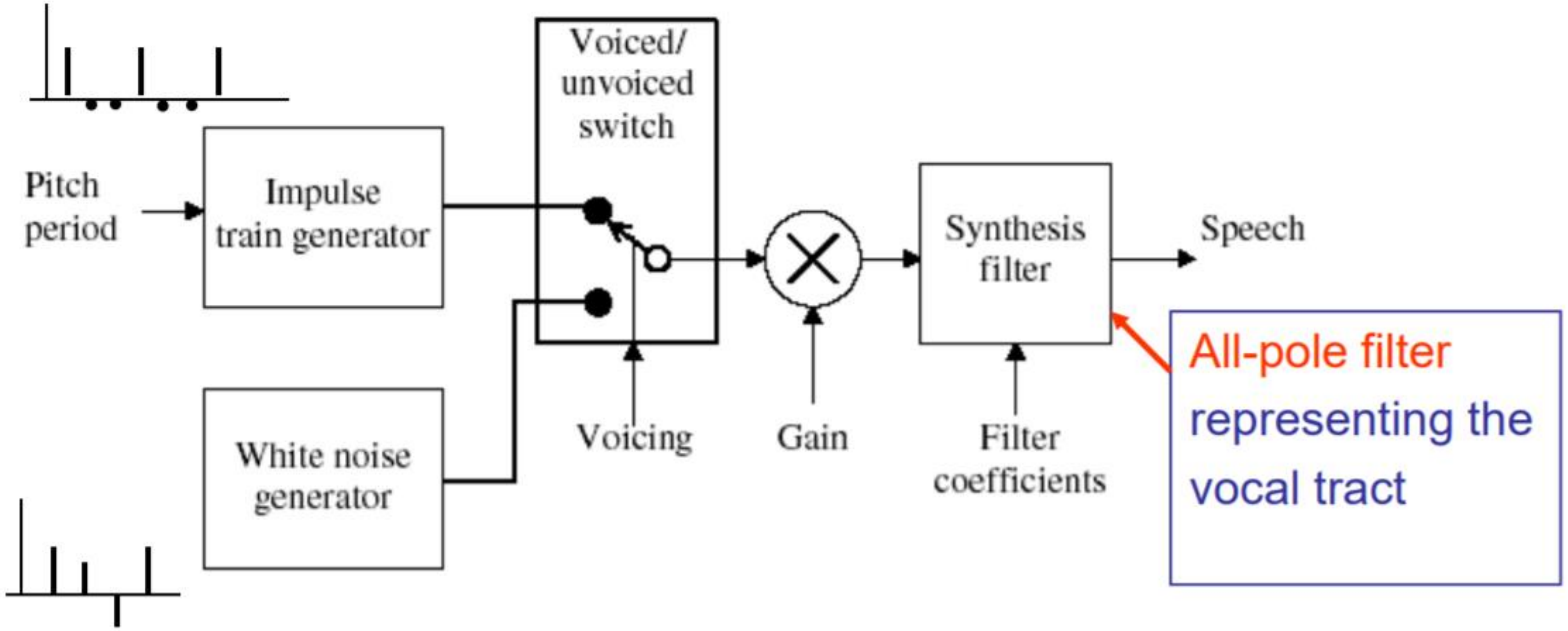
SIGNAL MODELS

Classical speech production model: LPC

- Impulse train or white noise (voiced/unvoiced speech frame)
- Gain: energy level of frame
- AR synthesis filter
- Pitch period (pitch height)

$$y(n) = \sum_{i=1}^{p} a_i y(n-i) \pm Gx(n)$$





Variants: LPC-10, CELP, MELP, RELP, VSELP, ASELP, LD-CELP

Ref: Nimrod Peleg/IRCAM

SIGNAL, AUDIO, SPEECH ENCODING.

Classical ML approaches

CLASSICAL ML APPROACHES

Input → Hand designed features → Mapping from features → Output

<u>Hand design features</u>
Knowledge a priori for the design of audio features

MFCC

- Source-filter model $X(f) = E(f) \cdot H(f)$
- Cepstrum

$$cepstr(x)(\tau) = DCT^{-1}\left(\log(|X(f)|^2)\right)$$

$$f(\tau) = DCT^{-1}\left(\log(|H(f)|^2)\right) + DCT^{-1}\left(\log(|E(f)|^2)\right)$$

- Filter/source contribution separation
- Perception frequency scale: Mel

Chromas

Contribution of frequency height class

$$f(C) = X^2\left(f_{C_1}\right) + X^2\left(f_{C_2}\right) + \cdots$$

What are MFCC ?
What is the difference between MFCC and filter bank ?

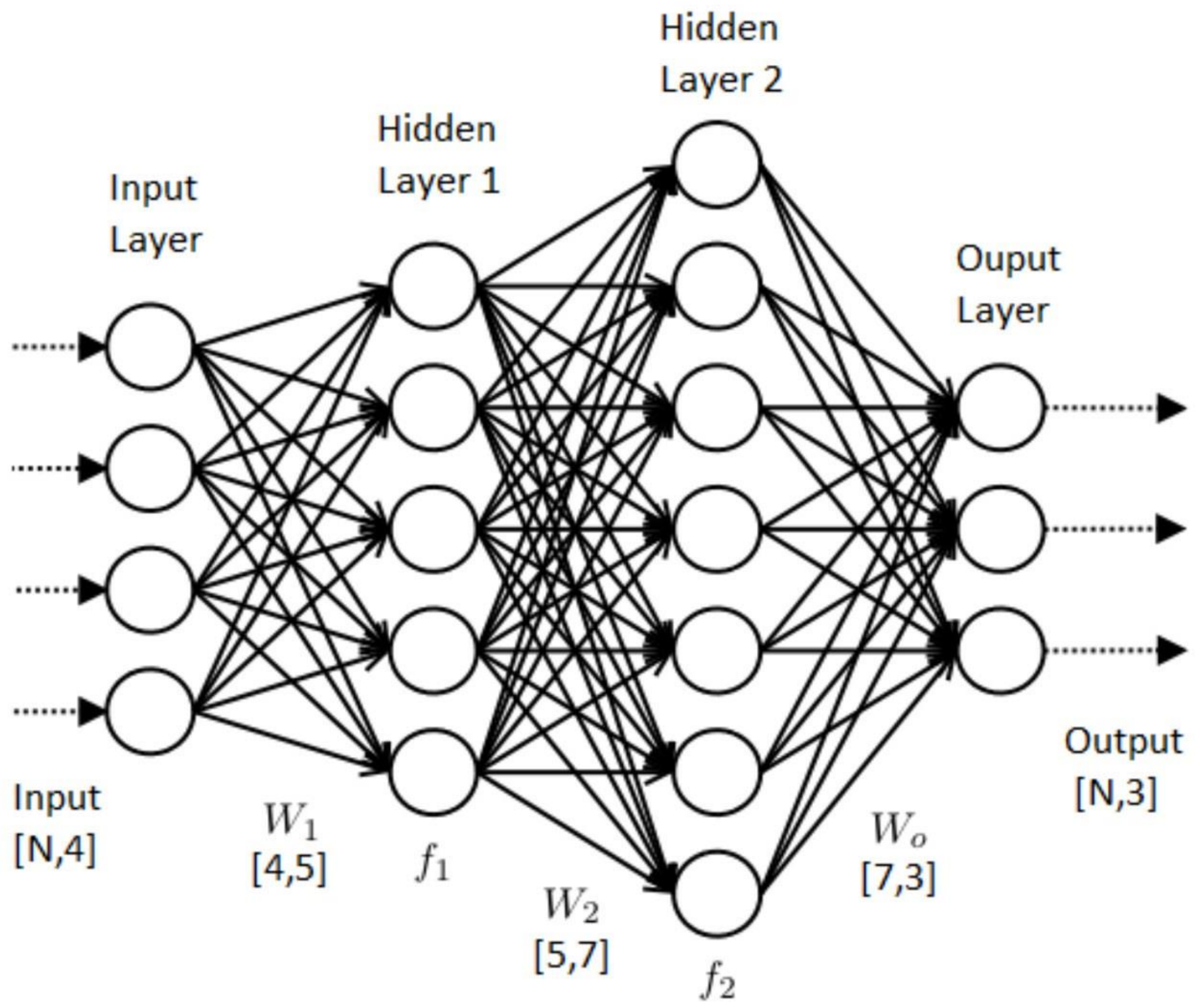
CLASSICAL ML APPROACHES

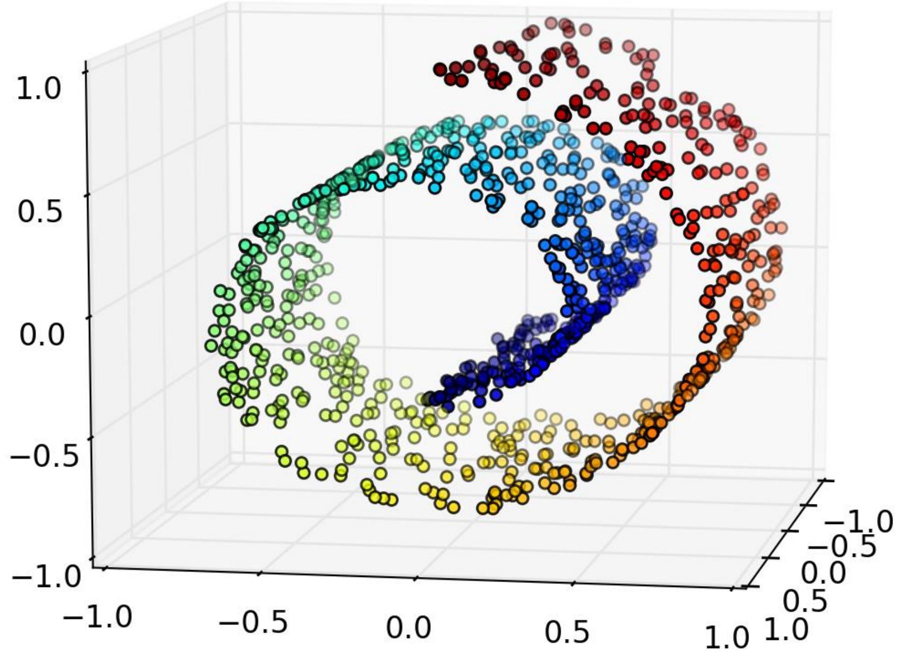
<u>Mapping from features</u>

- Preactivation (linear)
- Activation (non-linear)

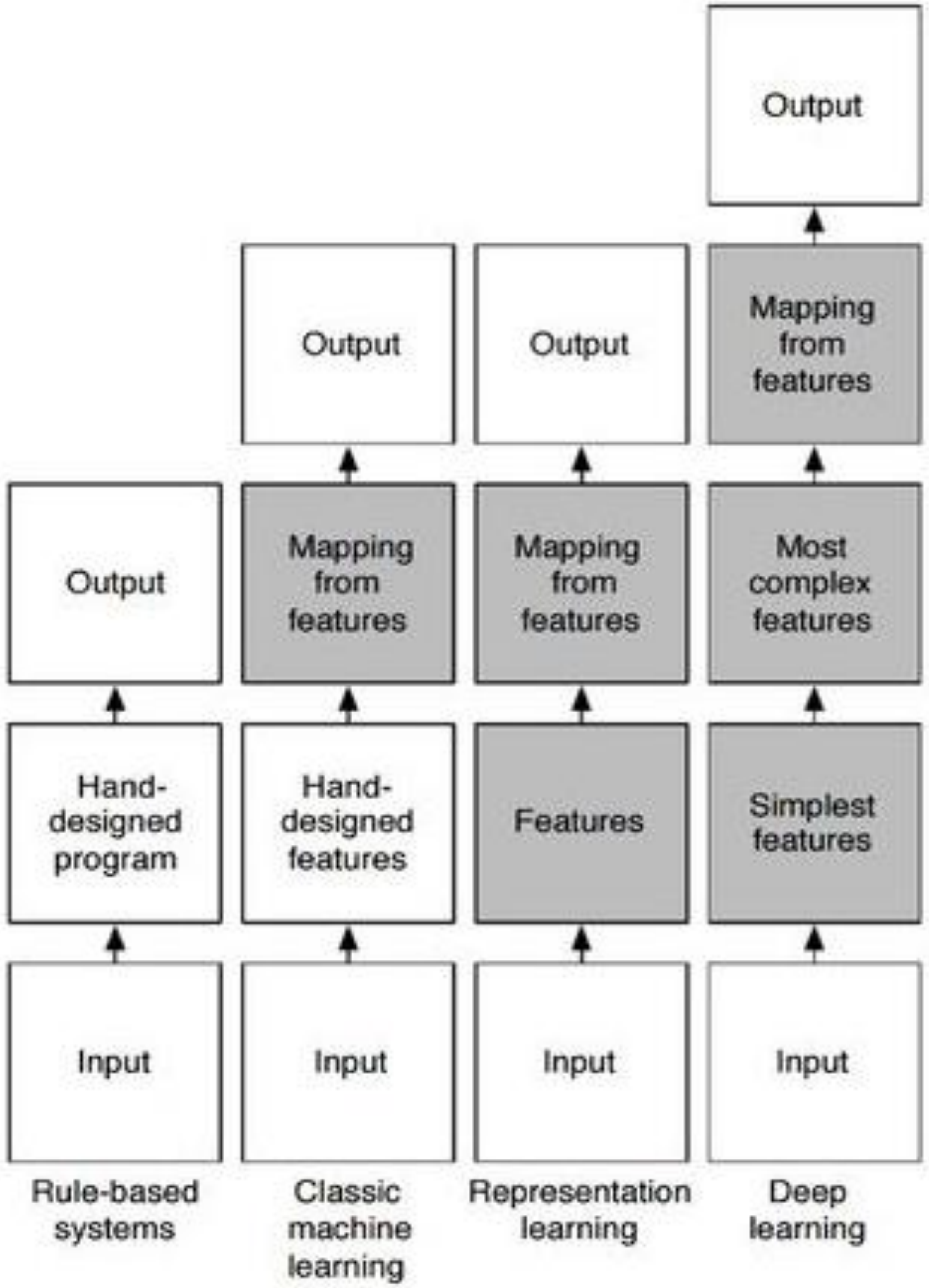
- Non-linear operator enabling a linear separation

- The Manifold Hypothesis states that real-world high-dimensional data lie on low-dimensional manifolds embedded within the high-dimensional space.





CLASSICAL ML APPROACHES



J. Humphrey, J. P. Bello and Y. Lecun. Moving beyond feature design: Deep architecures and automatic feature learning in music informatics. ISMIR 2012.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, The MIT Press, 2016.

SIGNAL, AUDIO, SPEECH ENCODING.

Thank you for your attention.

Refenencences:

- Geoffroy Peeters, Telecom Paris Tech
- Ian Goodfellow, Deep learning, 2018

ANNEXE 1: FOURIER TRANSFORM

The Fourier transform (FT) is an integrable function $\hat{f}: \mathbb{R} \rightarrow \mathbb{C}$, defined as following (one among several convention):

$$F[\omega] = [\int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt]$$

with pulsation: $\omega = 2\pi \cdot f_{req}$

The Fourier inverse function is the following:

$$f(t) = \int_{-\infty}^{\infty} F[\omega]e^{j\omega t}d\omega$$

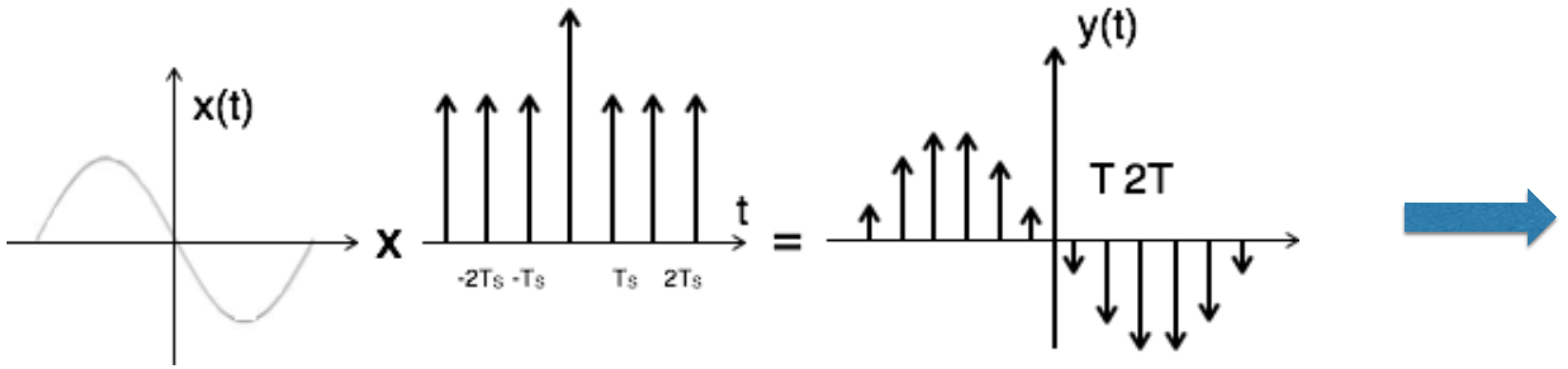
<u>Usefull properties</u> (let be x, y real functions):

- Linearity $ax(t) + by(t) \overset{\text{F.T}}{\longleftrightarrow} aX(\omega) + bY(\omega)$
- FT of an impulse function $\delta(\omega) = 1$
- Time scaling $x(at)\frac{1}{|a|}X\frac{\omega}{a}$
- Time shifting / frequency shifting $x(t - t_0) \overset{\text{F.T}}{\longleftrightarrow} e^{-j\omega t_0}X(\omega)$ and $e^{j\omega_0 t}.x(t) \overset{\text{F.T}}{\longleftrightarrow} X(\omega - \omega_0)$
- Multiplication and convolution $x(t).y(t) \overset{\text{F.T}}{\longleftrightarrow} X(\omega) * Y(\omega)$ and $x(t) * y(t) \overset{\text{F.T}}{\longleftrightarrow} \frac{1}{2\pi}X(\omega).Y(\omega)$
- Vectorized representation of FT values $X(\omega) = |X(\omega)|e^{j\theta(\omega)}$ where $\theta(\omega) = argX(\omega)$
  and $|X(\omega)|, \theta(\omega)$ are called magnitude and phase spectrum of X(ω).

ANNEXE 2: SIGNAL SAMPLING

Sampling of a continuous signal into a discrete signal:

$y(t) = x(t) \times$ impulse train

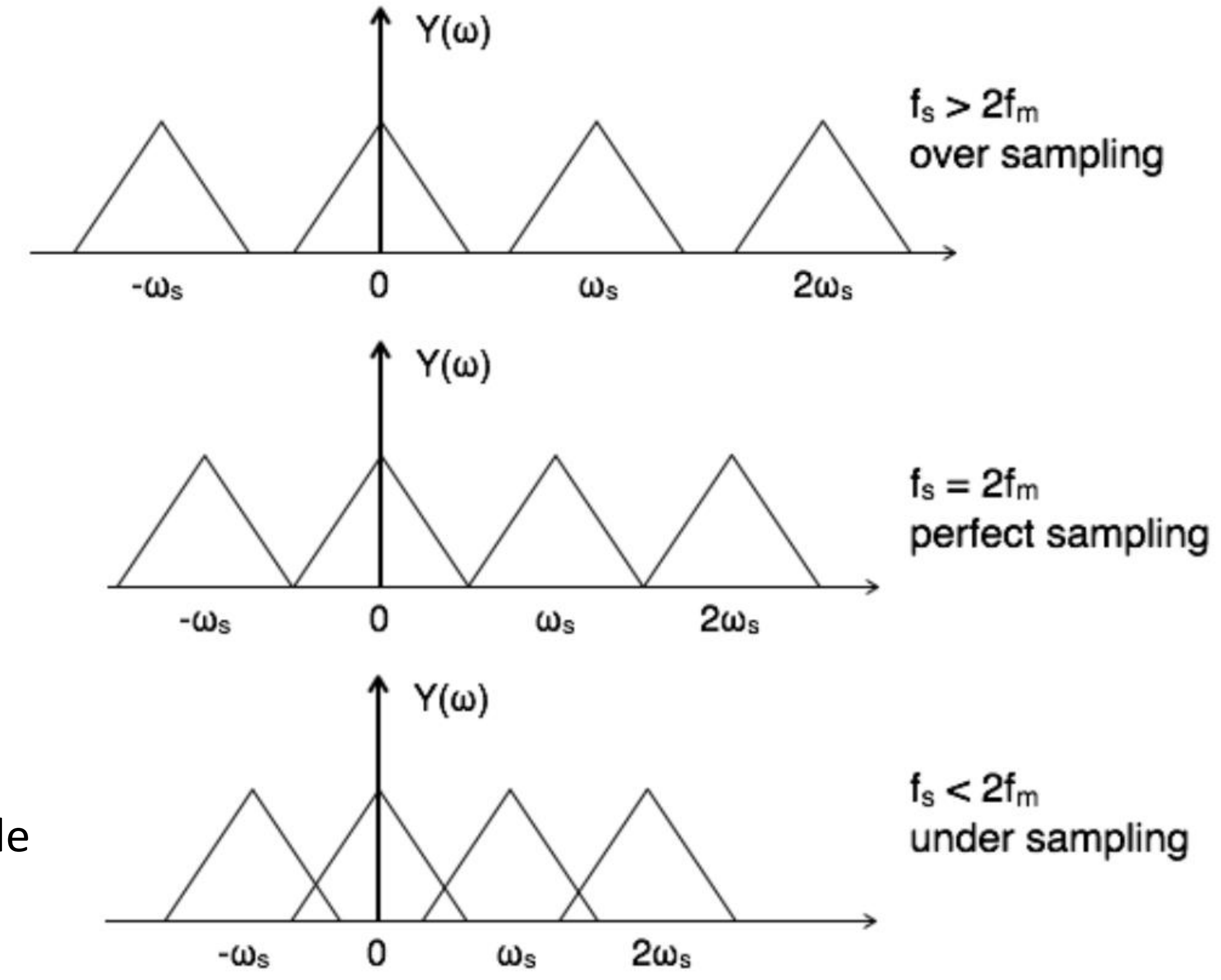


$x(n)$

Nyquist theorem: A continuous time signal can be represented in its samples and can be recovered back when sampling frequency f_s is greater than or equal to the twice the highest frequency com ponent of message signal. i. e.

$$f_s \geq 2f_m$$

frequency overlap here is calle d aliasing

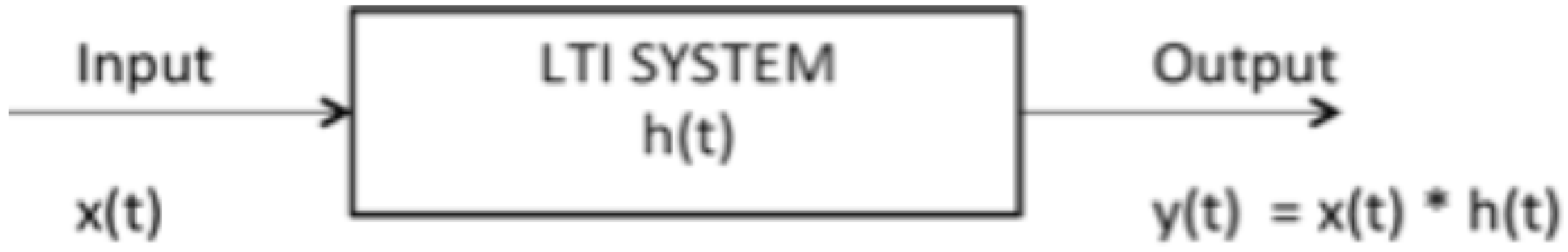


ANNEXE 3: CONVOLUTION & CORRELATION

Convolution of x by H (or H by x):

$$y(t) = x(t) * h(t)$$

$$= \int_{-\infty}^{\infty} x(\tau)h(t-\tau)d\tau$$



Property of Fourier Transform (see annexe 1)

$$x(t).y(t) \overset{\text{F.T}}{\longleftrightarrow} X(\omega) * Y(\omega)$$

$$x(t) * y(t) \overset{\text{F.T}}{\longleftrightarrow} \frac{1}{2\pi} X(\omega).Y(\omega)$$

Correlation of 2 signals x_1 and x_2:

$$\int_{-\infty}^{\infty} x_1(t)x_2(t-\tau)dt$$

ANNEXE 4: Z-TRANSFORM AND FILTERING

Definition of Z-Transform:

$$X(z) \quad = \sum_{n=-\infty}^{\infty} x[n]z^{-n}$$ with $z = re^{j\omega}$

Possibility to create y signal that depends on x(n), x(n-1), x(n-2), … and y(n-1), y(n-2), y(n-3), ….
This operation of y creation is called x filtering. And the Z-transform helps to characterize this filter (linear and invariant in time):

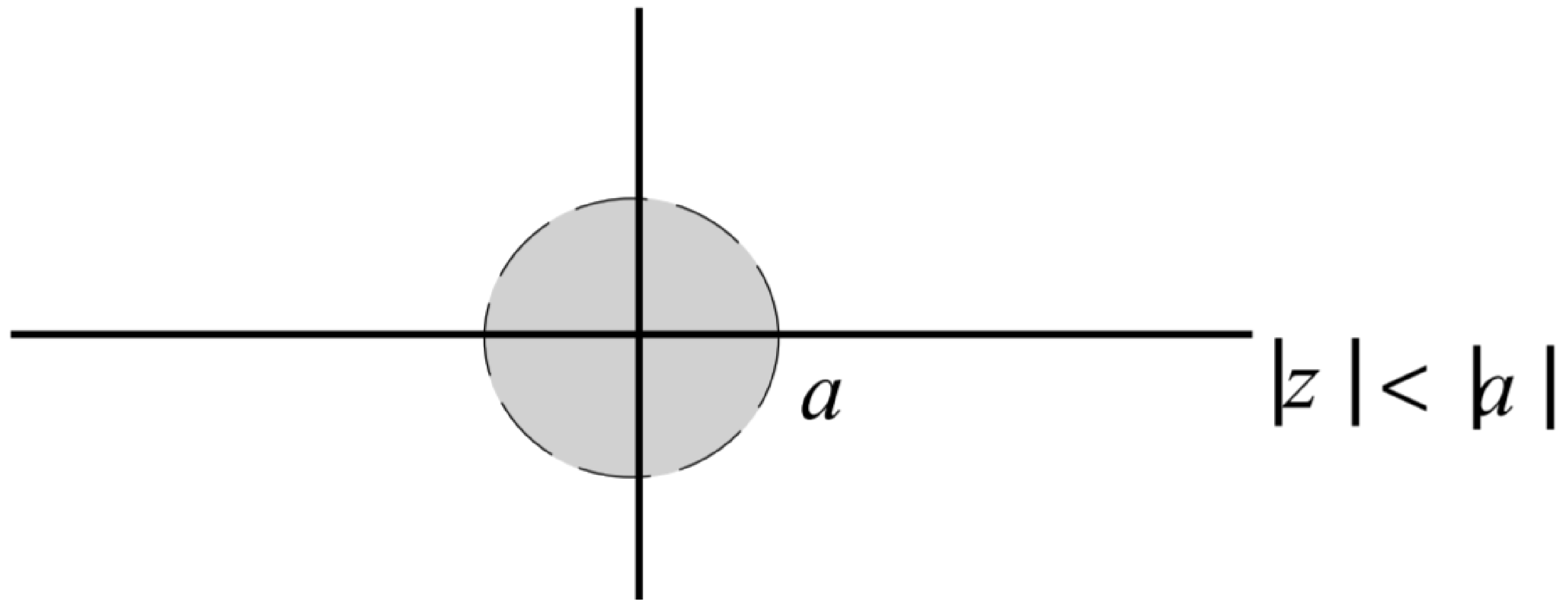
$$\sum_{l=-N}^{N} a_l y[n-l] \quad = \quad \sum_{k=-M}^{M} b_k x[n-k]$$

$$\sum_{l=-N}^{N} a_l z^{-l} Y(z) \quad = \quad \sum_{k=-M}^{M} b_k z^{-k} X(z)$$

$$H(z) \; = \; \frac{Y(z)}{X(z)} \; = \; \frac{\displaystyle\sum_{k=-M}^{M} b_k z^{-k}}{\displaystyle\sum_{l=-N}^{N} a_l z^{-l}}$$

Specific properties of H:

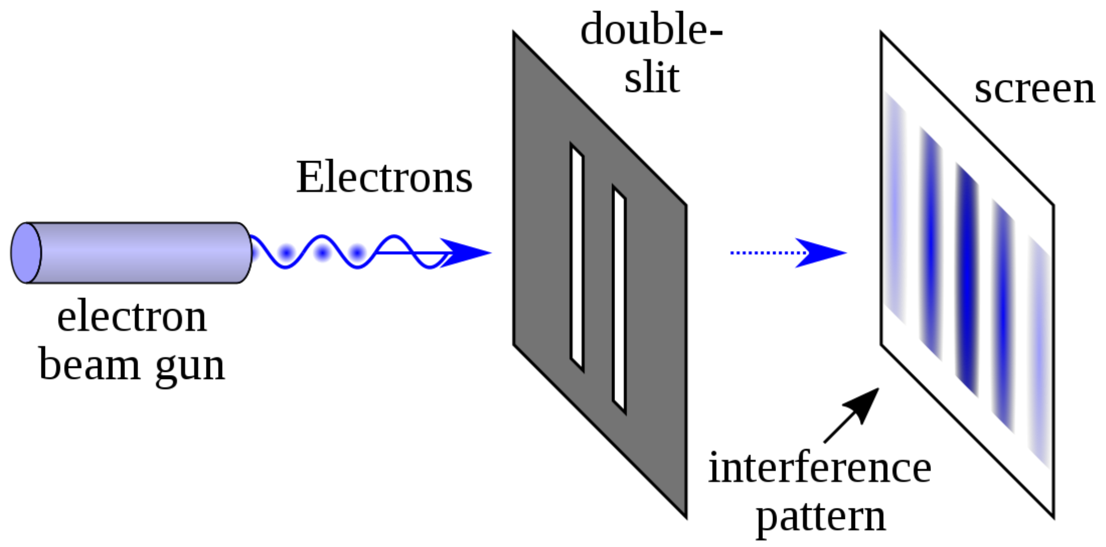
- if H denominators (complex) zeros have a modu lus greater than 1, H is not stable.
- Resonances at pulsation close to argument of d enominators zeros



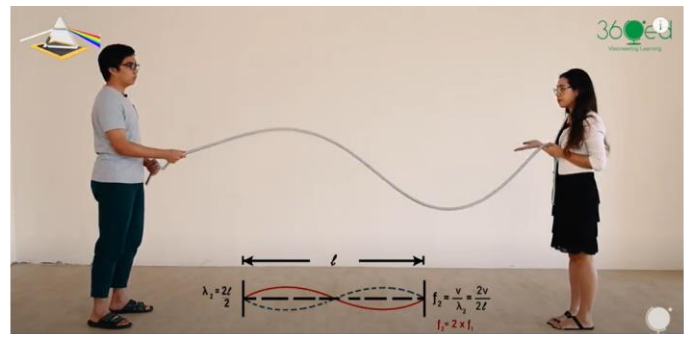
ANNEXE 5: WAVES INTERFERENCES & WAVEGUIDES

Sound wave
Propagation of a local oscillation of air pressure in a medium (air for instance).
It takes the form of longitudinal or transverse (for solids only) wave.

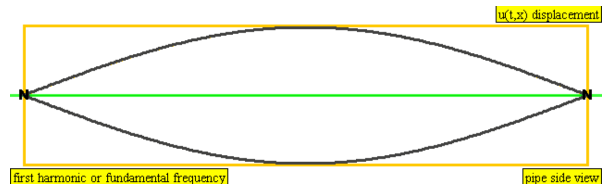
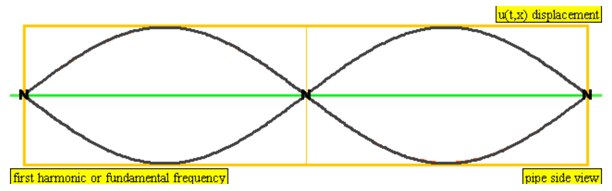
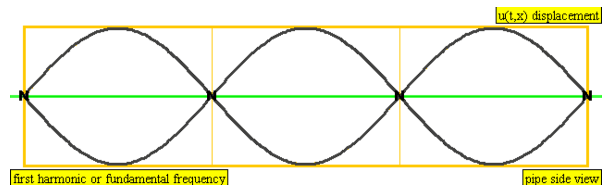
Wave interferences
Example of light wave interference: Young experiment.



Harmonics and wave guides



If λ is the wavelength (depending on frequency f) and v is sound speed (340 m/s): $\lambda = \frac{v}{f}$

Frequency	Order	Name	Standing wave representation
$1 \times f = 440$ Hz	$n = 1$	1st harmonic (fundamental tone)	
$2 \times f = 880$ Hz	$n = 2$	2nd harmonic (1^{st} overtone)	
$3 \times f = 1320$ Hz	$n = 3$	3rd harmonic (2^{nd} overtone)	
$4 \times f = 1760$ Hz	$n = 4$	4th harmonic (3^{rd} overtone)