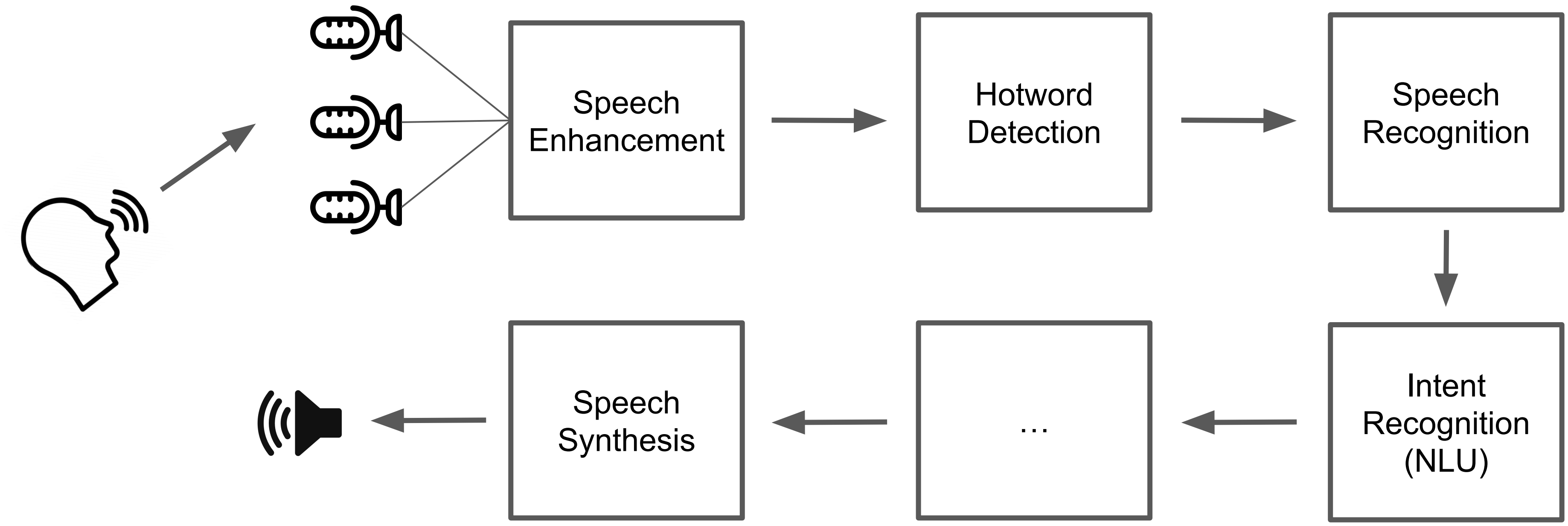
Speech Technology 2022

Lecture #4
Introduction to Signal Processing

@georgygospodinov

Voice Assistant Pipeline

- garbage in, garbage out



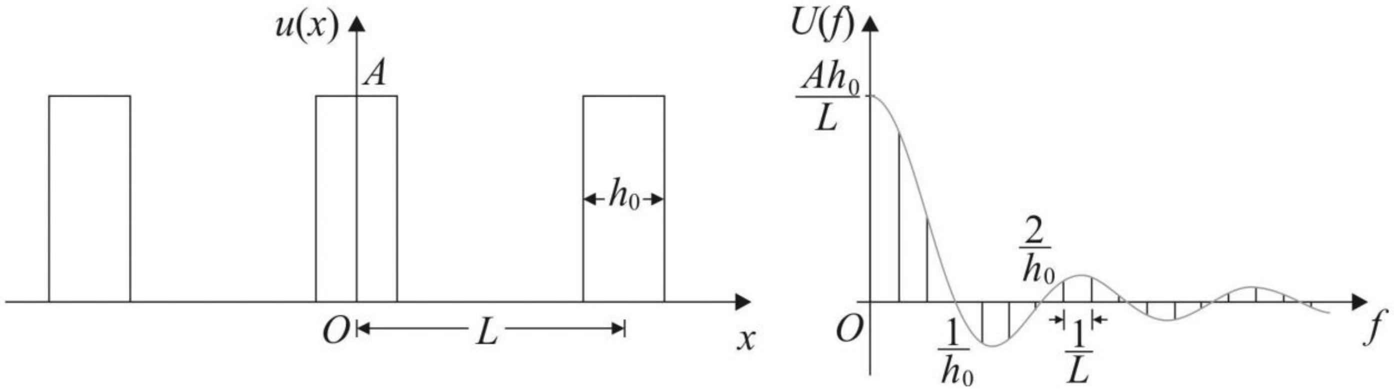
Plan

- Fourier Transform => Voice Spectrum, Digital Filters
- Convolution theorem => Filtering, Room Impulse Response
- Nyquist–Shannon sampling theorem => frequency aliasing
- sampling theorem + Biology => sample rates
- Adaptive Filters => Acoustic Echo Cancellation

Fourier Series

- periodic function: u(x) = u(x +/- L*n)
- discrete spectra

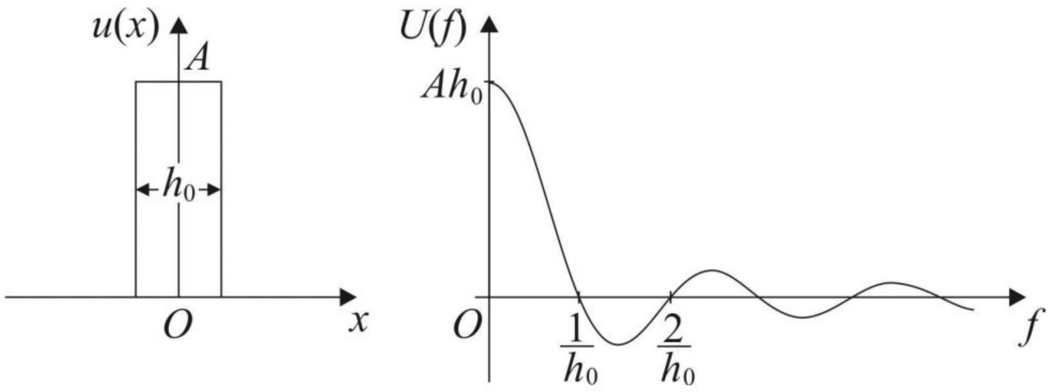
$$u(x) = \sum_{k=-\infty}^{+\infty} d_k e^{i2\pi f_k x}, \qquad d_k = \frac{1}{L}\int_{-L/2}^{L/2} u(x)e^{-2\pi i f_k x}dx$$



Fourier Transform

- non-periodic function
- continuous spectra

$$u(x) = \int_{-\infty}^{+\infty} U(f) e^{2\pi i f x}, \qquad U(f) = \int_{-\infty}^{+\infty} u(x) e^{-2\pi i f x} dx$$



Convolution Theorem

- Fourier transform of a convolution of two functions is the pointwise product of their Fourier transforms

$$u_1(x) * u_2(x) \longleftrightarrow U_1(f) \cdot U_2(f)$$

$$u_1(x) \cdot u_2(x) \longleftrightarrow U_1(f) * U_2(f)$$

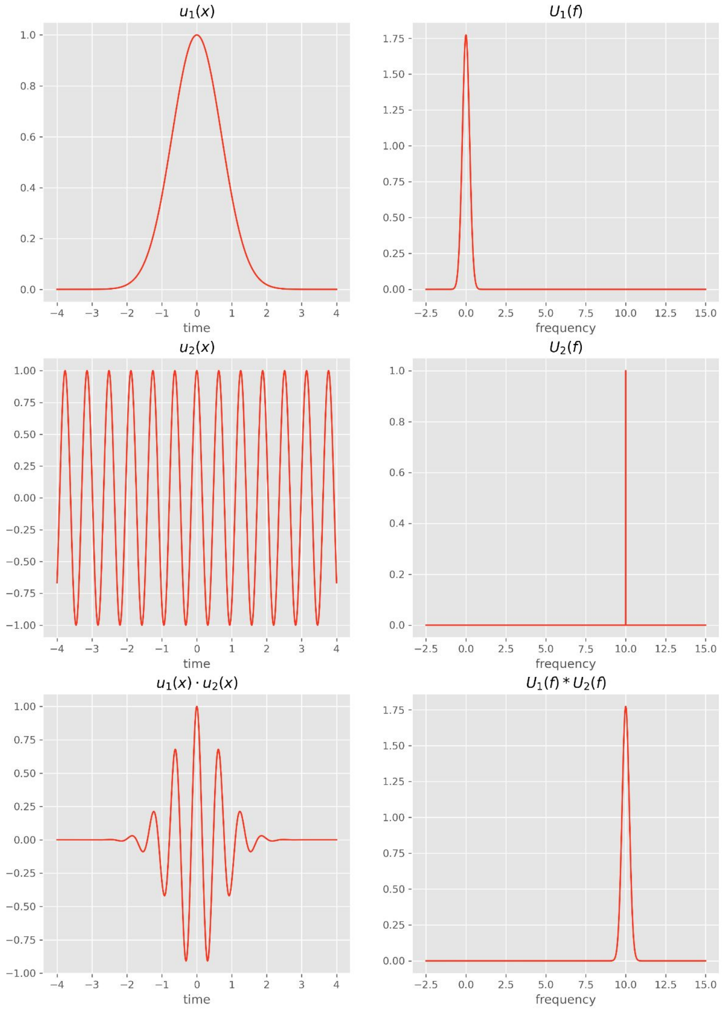
$$U_1(f) * U_2(f) = \int_{-\infty}^{\infty} U_1(f - \xi) U_2(\xi) d\xi$$

Convolution Theorem: example

$$u_1(x) = Ae^{-\frac{x^2}{a_0^2}} \longleftrightarrow U_1(f) = Aa_0\sqrt{\pi}e^{-\frac{f^2}{\left(\frac{1}{\pi a_0}\right)^2}}$$

$$u_2(x) = \cos(2\pi f_0 x) \longleftrightarrow U_2(f) = \delta(f - f_0)$$

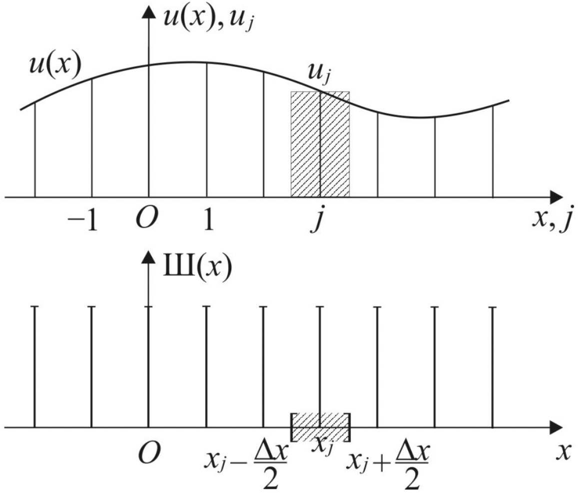
$$u_1(x)\cdot u_2(x) = Ae^{-\frac{x^2}{a_0^2}}\cos(2\pi f_0 x) \longleftrightarrow U_1(f) * U_2(f) = Aa_0\sqrt{\pi}e^{-\frac{(f-f_0)^2}{\left(\frac{1}{\pi a_0}\right)^2}}$$



Discrete Time Fourier Transform

$$u_j = u(x) \sum_{j=-\infty}^{+\infty} \delta(x - j\Delta x) \cdot \Delta x = u(x) \cdot Ш(x) \cdot \Delta x$$

$$Ш(x) = \sum_{j=-\infty}^{+\infty} \delta(x - j\Delta x) \longleftrightarrow U_{Ш}(f) = \frac{1}{\Delta x} \sum_{k=-\infty}^{+\infty} \delta(f - \frac{k}{\Delta x})$$



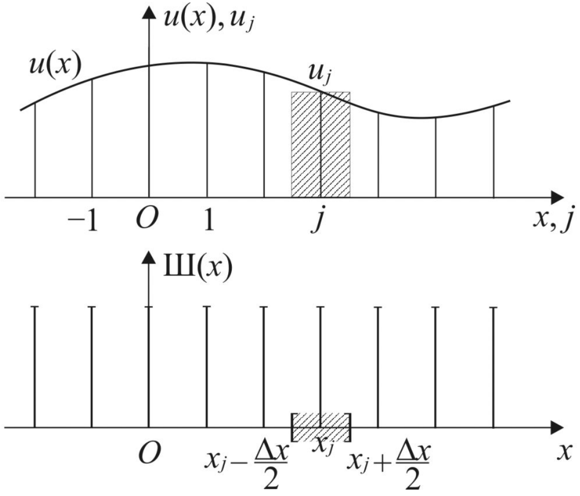
Discrete Time Fourier Transform

$$u_j = u(x)\sum_{j=-\infty}^{+\infty}\delta(x - j\Delta x)\cdot\Delta x = u(x)\cdot \text{Ш}(x)\cdot\Delta x$$

$$\text{Ш}(x) = \sum_{j=-\infty}^{+\infty}\delta(x - j\Delta x) \longleftrightarrow U_{\text{Ш}}(f) = \frac{1}{\Delta x}\sum_{k=-\infty}^{+\infty}\delta(f - \frac{k}{\Delta x})$$

$$u_j = u(x)\cdot \text{Ш}(x)\cdot\Delta x \longleftrightarrow U(f) * U_{\text{Ш}}(f)\cdot\Delta x = U_{\Delta x}(f)$$

$$U_{\Delta x}(f) = \int_{-\infty}^{+\infty} U(\xi)\frac{1}{\Delta x}\sum_{k=-\infty}^{+\infty}\delta(f - \frac{k}{\Delta x} - \xi)d\xi\cdot\Delta x = \sum_{k=-\infty}^{+\infty} U(f - \frac{k}{\Delta x})$$

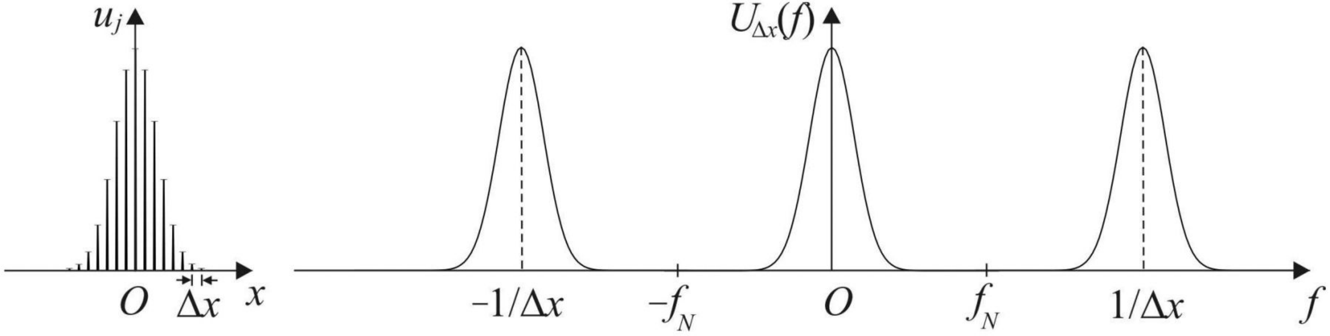


Discrete Time Fourier Transform

$$u_j = u(x) \cdot \text{Ш}(x) \cdot \Delta x \longleftrightarrow U(f) * U_{\text{Ш}}(f) \cdot \Delta x = U_{\Delta x}(f)$$

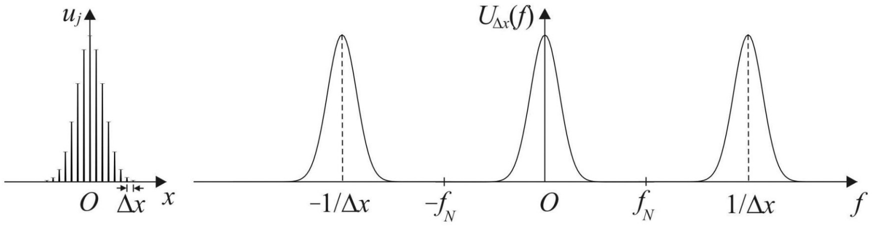
$$U_{\Delta x}(f) = \int_{-\infty}^{+\infty} U(\xi) \frac{1}{\Delta x} \sum_{k=-\infty}^{+\infty} \delta(f - \frac{k}{\Delta x} - \xi) d\xi \cdot \Delta x = \sum_{k=-\infty}^{+\infty} U(f - \frac{k}{\Delta x})$$

- Sampling in time domain replicates the spectrum in frequency domain!



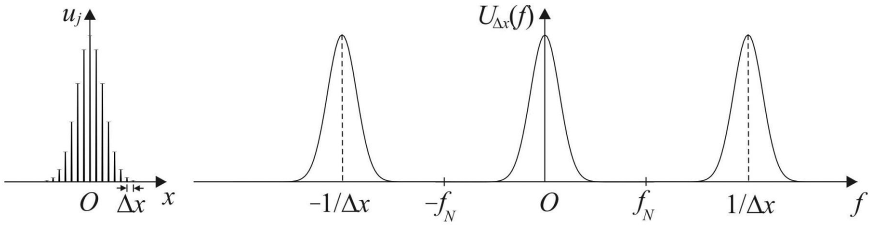
Discrete Time Fourier Transform

- Sampling in time domain replicates the spectrum in frequency domain!

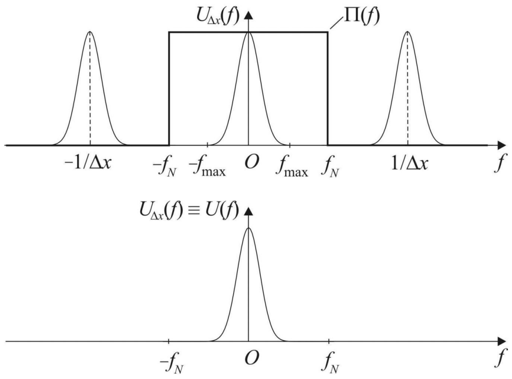


Discrete Time Fourier Transform

- Sampling in time domain replicates the spectrum in frequency domain!

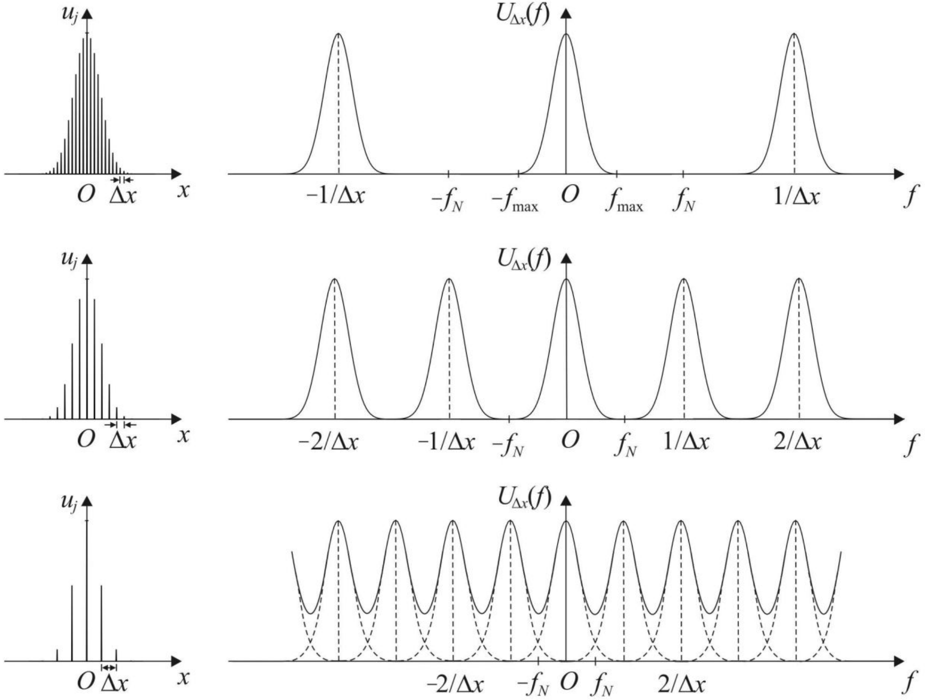


- Continuous signal reconstruction with spectral filtering

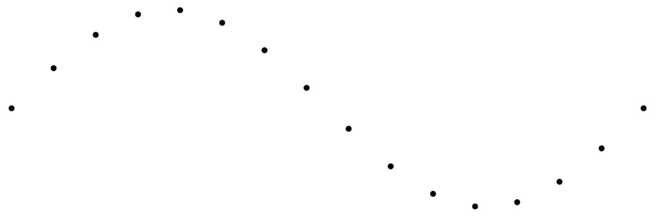


Discrete Time Fourier Transform

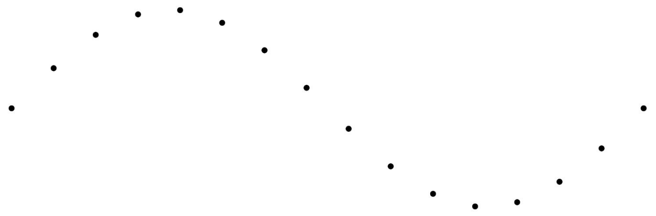
- Frequency Aliasing

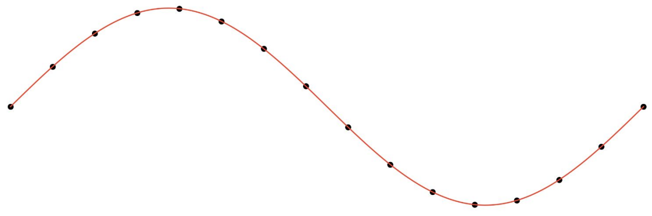


Frequency Aliasing

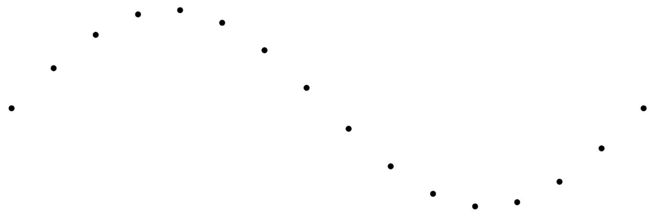


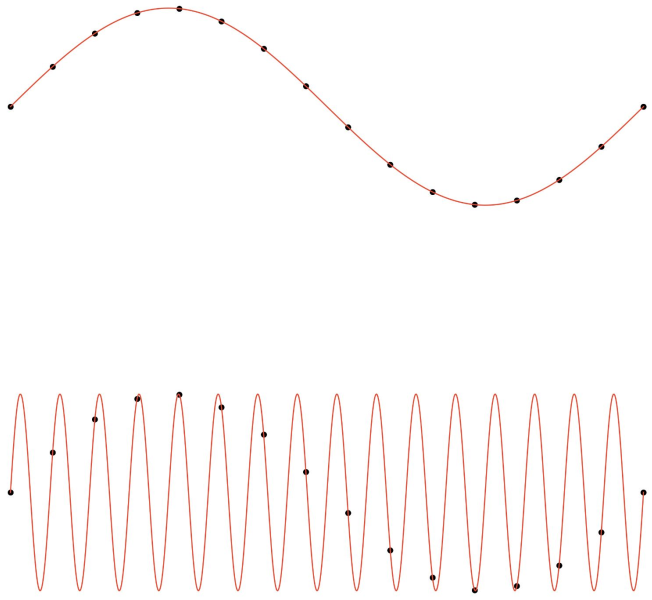
Frequency Aliasing





Frequency Aliasing





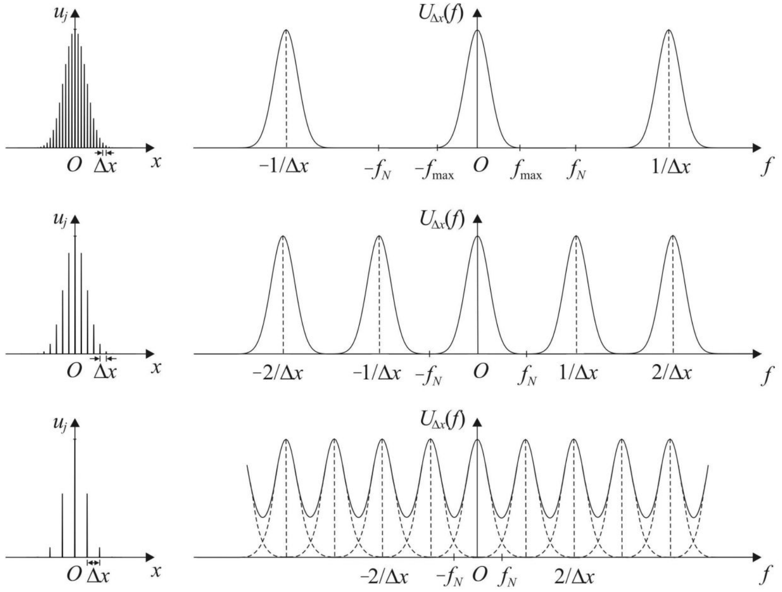
Nyquist–Shannon sampling theorem

- If a function x(t) contains no frequencies higher than B hertz, it is completely determined by giving its ordinates at a series of points spaced 1/(2B) seconds apart

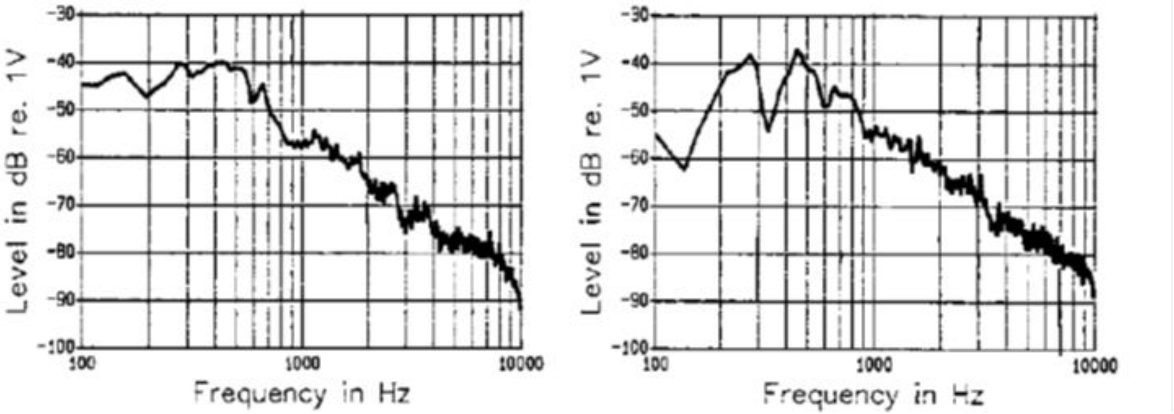
$$u(x) = \sum_{j=-\infty}^{+\infty} u(j\Delta x)\operatorname{sinc}\left(\frac{\pi\,(x - j\Delta x)}{\Delta x}\right)$$

$$f_{\text{Nyquist}} = \frac{f_{\text{sample}}}{2} = \frac{1}{2\Delta x} \geq f_{\max}$$

$$f_{\text{sample}} \geq 2 f_{\max}$$

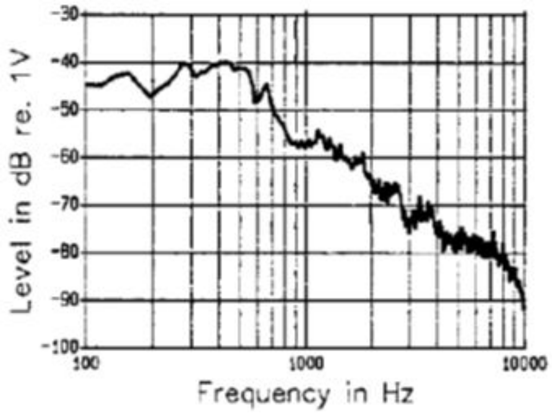


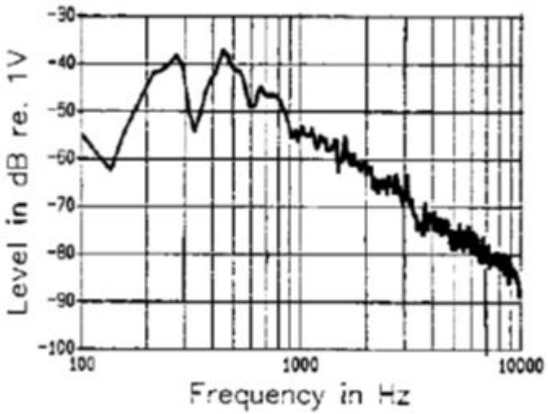
Human Voice Spectrum



- voice frequency range: 300-3400 Hz
- Nyquist–Shannon theorem => telephony sampling rate: 8kHz

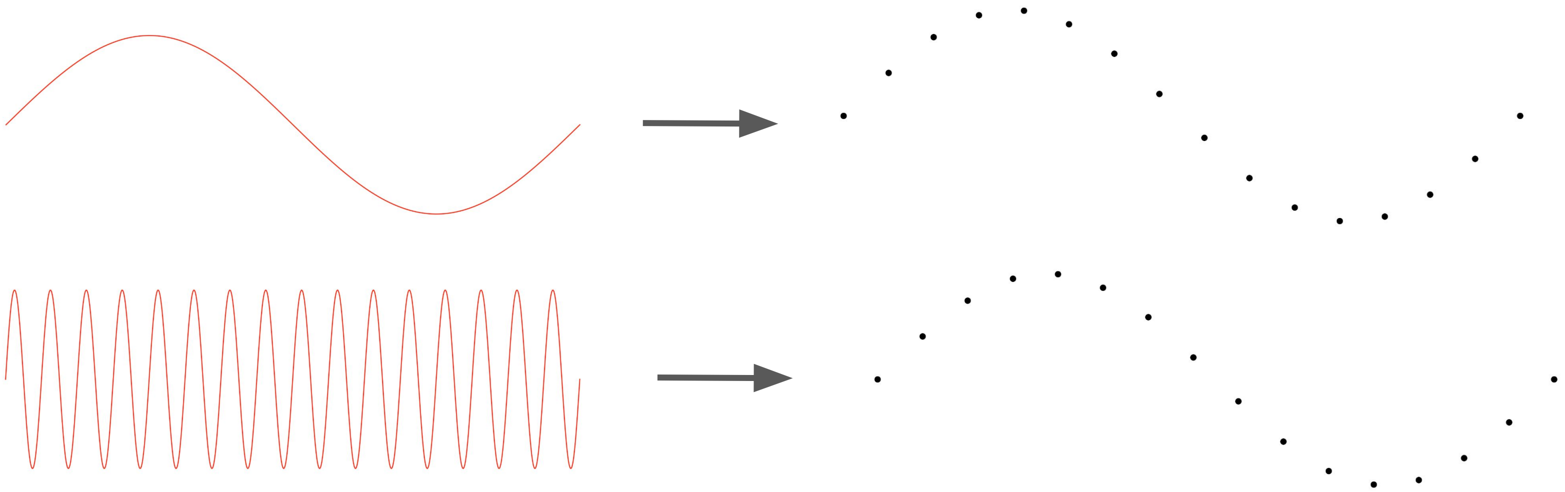
Human Voice Spectrum





- voice frequency range: 300-3400 Hz
- Nyquist–Shannon theorem => telephony sampling rate: 8kHz
- another sample rates: 16kHz, 24kHz, 48kHz. Why?

Low-Pass Filtering



- Pre-Filtering before sampling / downsampling !

Signals

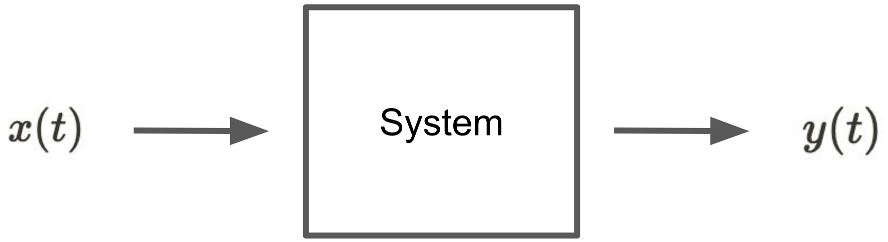

Analog Signal
Discrete Signal
Quantized Signal
Digital Signal
0.0020
0.0015
0.0010
0.0005
0.0000
−0.0005
−0.0010
−0.0015
0.000
0.002
0.004
0.006
0.008
0.010

SoX: Sound eXchange

```
(base) georgijgospodinov@MacBook-Air-Georgij data % soxi
salut_time_query.wav
Input File     : 'salut_time_query.wav'
Channels       : 1
Sample Rate    : 16000
Precision      : 25-bit
Duration       : 00:00:02.56 = 40924 samples ~ 191.831 CDDA
sectors
File Size      : 164k
Bit Rate       : 512k
Sample Encoding: 32-bit Floating Point PCM
```

Systems

- signal in, signal out
- examples:
  - electronic amplifier
  - room



Linear Time-Invariant Systems

- linear:

$$\mathcal{L}\left(\alpha x(t)\right) = \alpha \mathcal{L}(x(t))$$

$$\mathcal{L}\left(x_1(t) + x_2(t)\right) = \mathcal{L}\left((x_1(t)\right) + \mathcal{L}\left((x_2(t)\right)$$

- time invariance: effect of the system doesn’t vary over time

$$\mathcal{L}\left(x(t)\right) = y(t) \implies \mathcal{L}\left(x(t-T)\right) = y(t-T)$$

Linear Time-Invariant Systems

- linear:

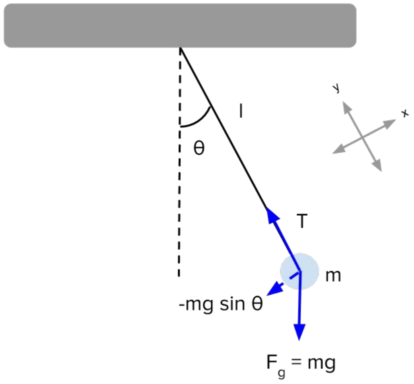
$$\mathcal{L}\left(\alpha x(t)\right) = \alpha\mathcal{L}(x(t))$$

$$\mathcal{L}\left(x_1(t) + x_2(t)\right) = \mathcal{L}\left((x_1(t)\right) + \mathcal{L}\left((x_2(t)\right)$$

- time invariance: effect of the system doesn't vary over time

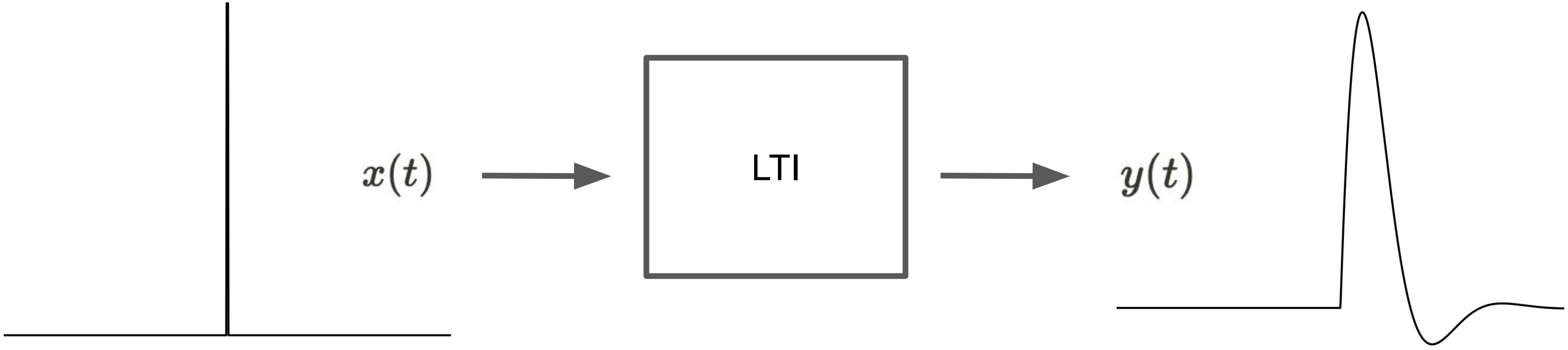
$$\mathcal{L}\left(x(t)\right) = y(t) \implies \mathcal{L}\left(x(t-T)\right) = y(t-T)$$

- examples:
  - circuits (resistors, capacitors, inductors)
  - mechanical systems
  - media that transmit the sound
- described by Linear Differential Equations



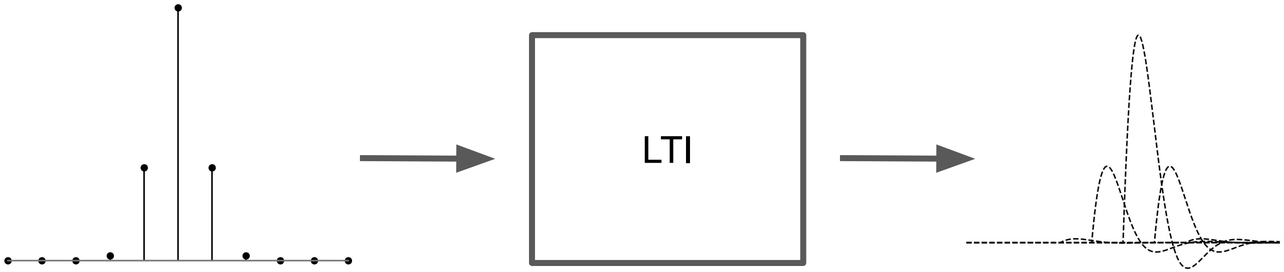
Linear Time-Invariant Systems

- Impulse Response
  - mechanical kick
  - pop a balloon, fire a gun



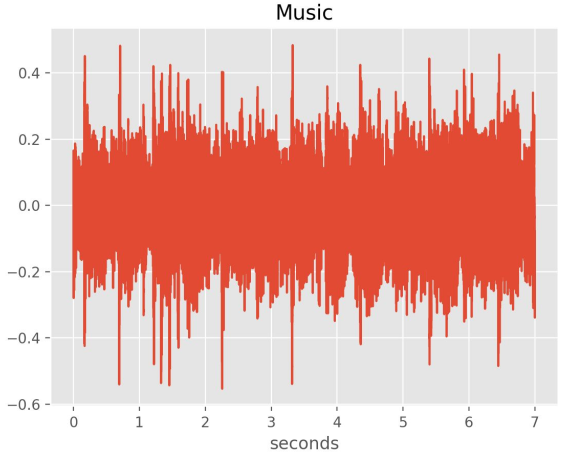
Linear Time-Invariant Systems

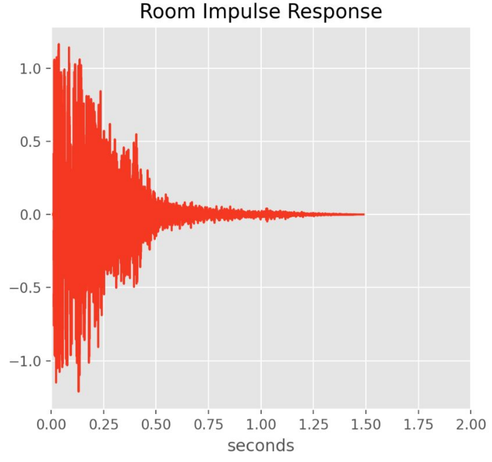
- input signal = sequence of impulses with varying amplitude
- each impulse in input yields a shifted and scaled copy of impulse response
- output signal = sum of the shifted and scaled copies of impulse response

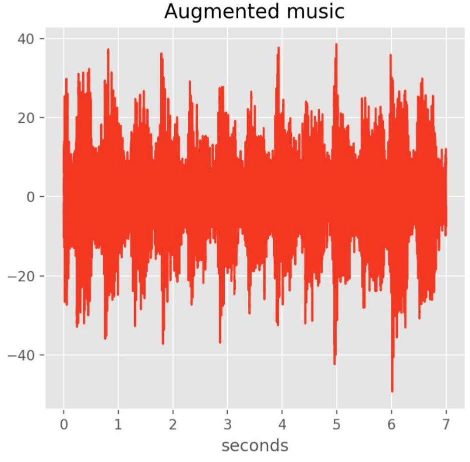


Room Impulse Response

- http://isophonics.net/content/room-impulse-response-data-set



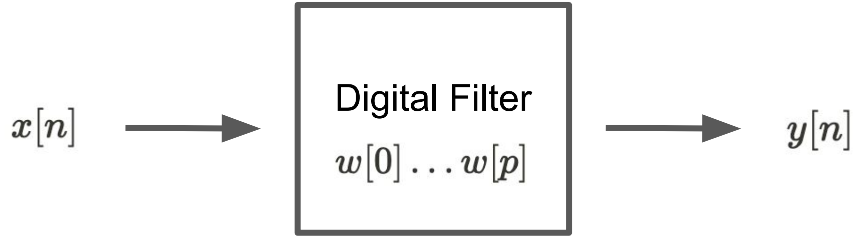




Digital Filters

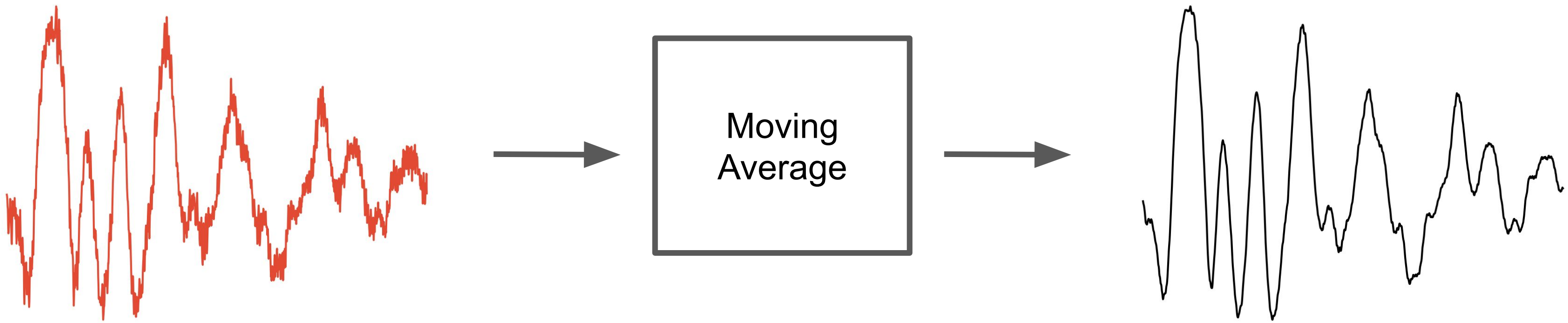
- filter: extraction information about quantity of interest
- digital: sampled, discrete-time signal
- Finite Impulse Response filters (p order):

$$y[n] = w[0]x[n] + w[1]x[n-1] + \cdots + w[p]x[n-p] = \sum_{i=0}^{p} w[i]x[n-i]$$



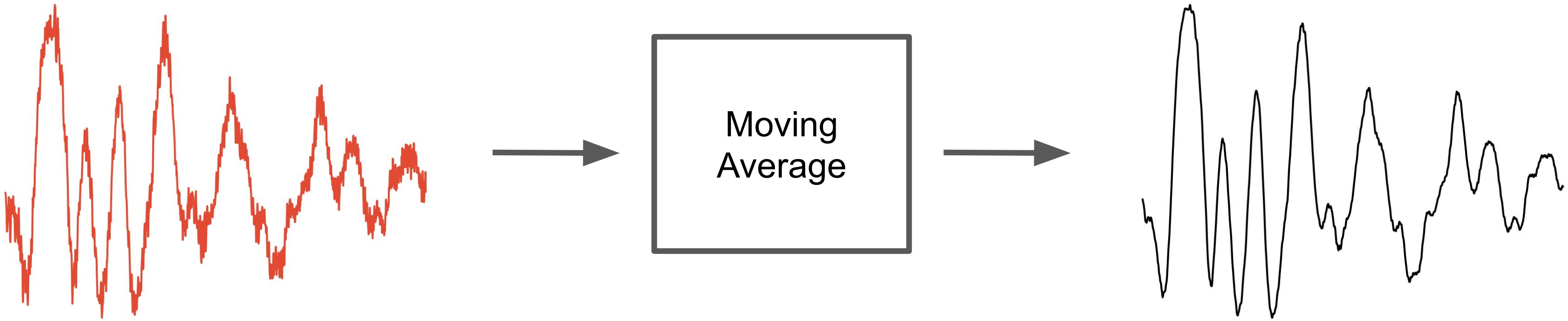
Filter Example: Moving Average

$$y[n] = \frac{x[n] + x[n-1] + \cdots + x[n-N+1]}{N}$$



Filter Example: Moving Average

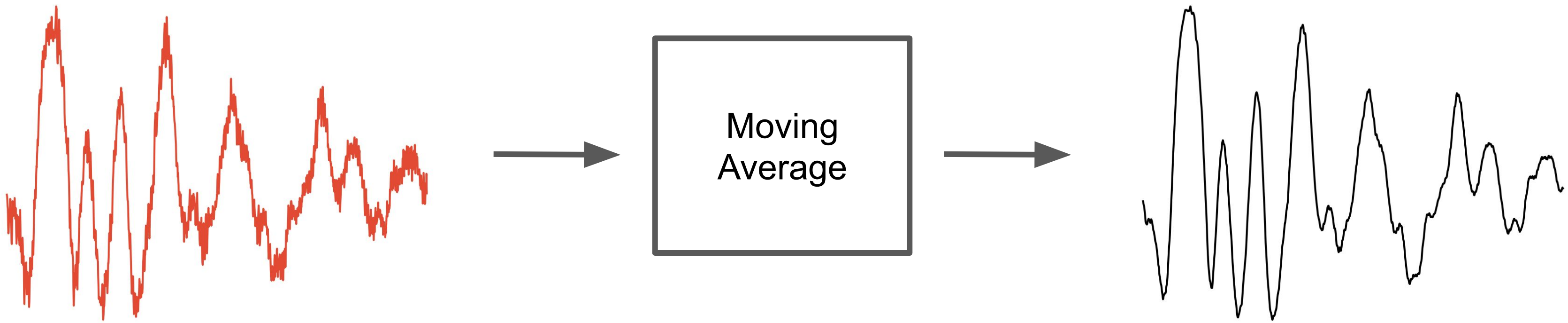
$$y[n] = \frac{x[n] + x[n-1] + \cdots + x[n-N+1]}{N}$$

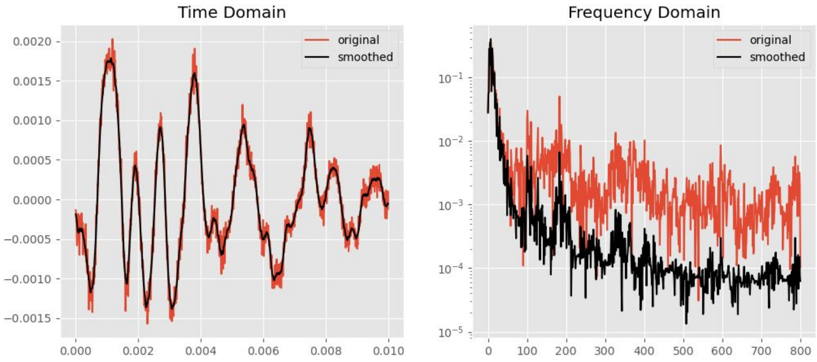


Low Pass Filter

Filter Example: Moving Average

$$y[n] = \frac{x[n] + x[n-1] + \cdots + x[n-N+1]}{N}$$

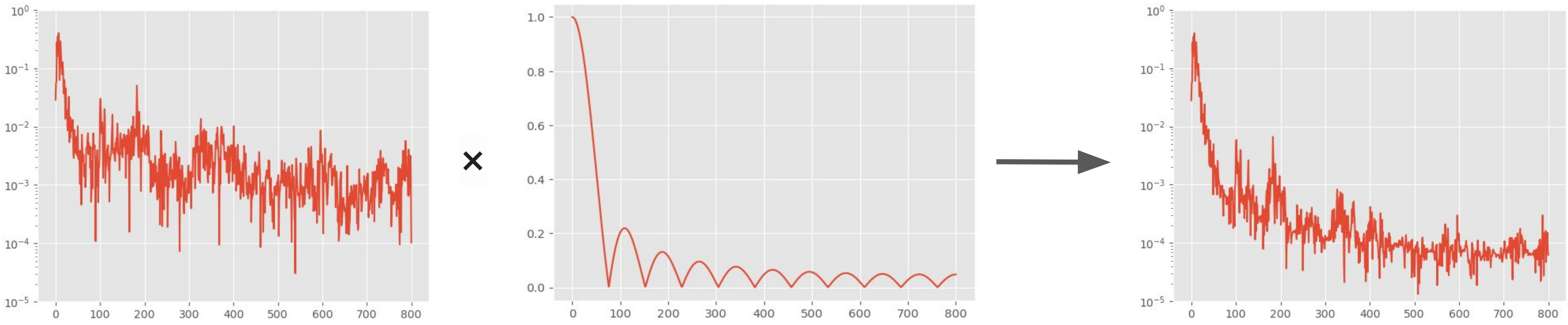




Filter Example: Moving Average

- Convolution theorem: time domain convolution ⇔ frequency domain multiplication
- simple average => sidelobes; Gaussian window is better

$$y[n] = \frac{x[n] + x[n-1] + \cdots + x[n-N+1]}{N} \qquad w = \left(\frac{1}{N}, \dots, \frac{1}{N}\right)$$



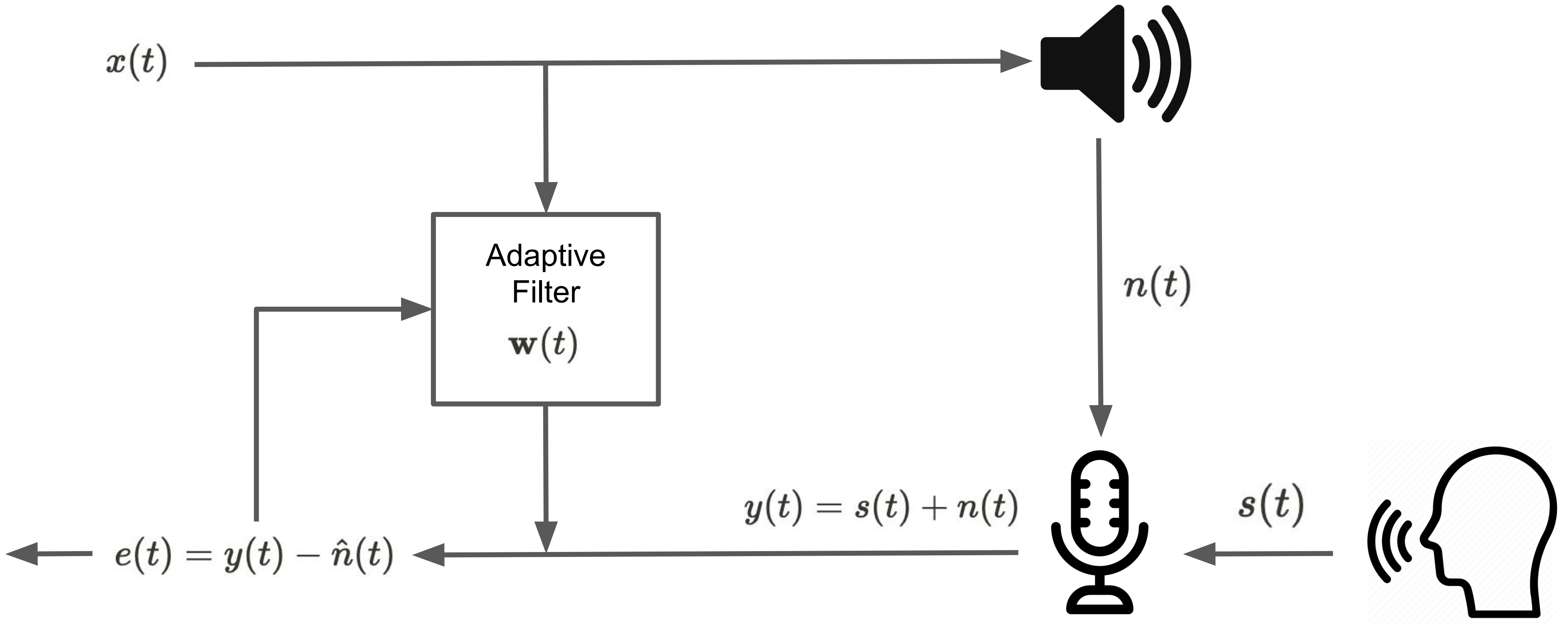
Limitations of fixed-coefficient digital filters

- Time-varying noise
- Overlapping bands of signal and noise
- Unknown parameters (eg room)

Adaptive Filters

- Digital Filter (with adjustable weights)
- Adaptive Algorithm

Acoustic Echo Cancellation: Scheme



Acoustic Echo Cancellation

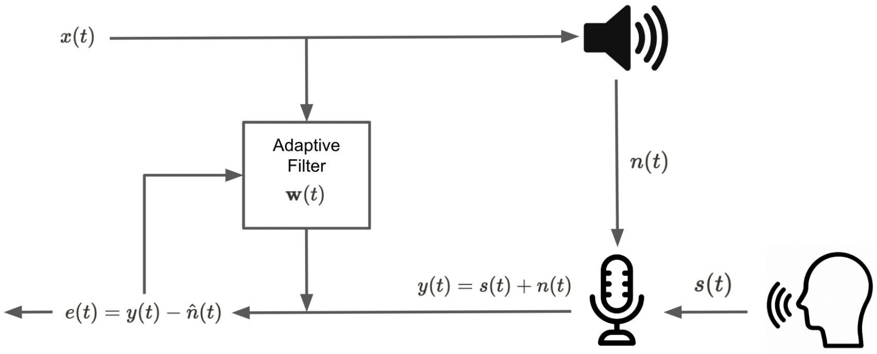
$$e_k = y_k - \hat{n}_k = s_k + n_k - \hat{n}_k$$

$$e_k^2 = s_k^2 + (n_k - \hat{n}_k)^2 + 2s_k(n_k - \hat{n}_k)$$

$$\mathbb{E}e_k^2 = \mathbb{E}s_k^2 + \mathbb{E}(n_k - \hat{n}_k)^2 + 2\mathbb{E}s_k(n_k - \hat{n}_k)$$

$$\mathbb{E}e_k^2 = \mathbb{E}s_k^2 + \mathbb{E}(n_k - \hat{n}_k)^2$$

$$\min \mathbb{E}e_k^2 = \mathbb{E}s_k^2 + \min \mathbb{E}(n_k - \hat{n}_k)^2$$



- minimize total power at the output maximize the output signal-to-noise ratio

Acoustic Echo Cancellation: LMS

$$e_k = y_k - \mathbf{w}^T \mathbf{x}_k$$

$$J(\mathbf{w}) = (y_k - \mathbf{w}^T \mathbf{x}_k)^2 \longrightarrow \min_{\mathbf{w}}$$

Acoustic Echo Cancellation: LMS

$$e_k = y_k - \mathbf{w}^T \mathbf{x}_k$$

$$J(\mathbf{w}) = (y_k - \mathbf{w}^T \mathbf{x}_k)^2 \longrightarrow \min_{\mathbf{w}}$$

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \mu \nabla_{\mathbf{w}} J$$

$$\nabla_{\mathbf{w}} J = -2(y_k - \mathbf{w}^T \mathbf{x}_k)\mathbf{x}_k = -2e_k \mathbf{x}_k$$

$$\mathbf{w}_{k+1} = \mathbf{w}_k + 2\mu e_k \mathbf{x}_k$$

Acoustic Echo Cancellation: NLMS

- Normalized Least Mean Squares
- LMS algorithm is sensitive to the scaling of input signal

$$\mathbf{w}_{k+1} = \mathbf{w}_k + 2\mu \frac{e_k \mathbf{x}_k}{\mathbf{x}_k^T \mathbf{x}_k}$$

$$\mathbf{w}_{k+1} = \mathbf{w}_k + 2\mu \frac{e_k \mathbf{x}_k}{\delta + \mathbf{x}_k^T \mathbf{x}_k}$$

Thank you for your attention!

@georgygospodinov