Signal Processing for Speech

DT2119 Speech and Speaker Recognition

Giampiero Salvi

KTH/CSC/TMH giampi@kth.se

VT 2019

Disclaimers

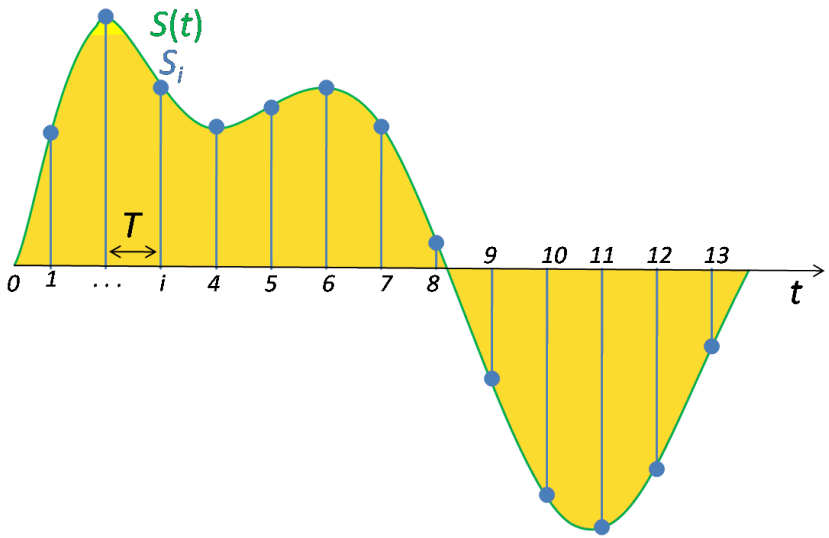
- I do not have the pretension to teach you signal processing in two hours
- The intent of this lecture is to give you an idea of what we are talking about in Lab 1
- You will not be asked to understand the details of the formulas but only the concepts

Also

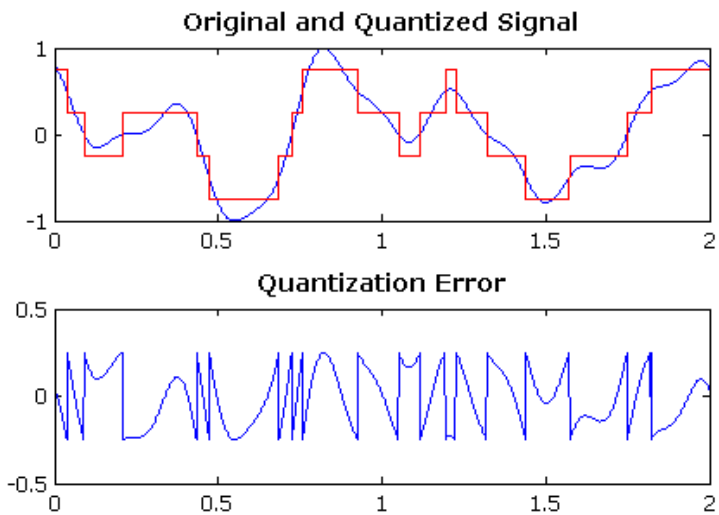
- If you know all about signal processing please forgive the necessary simplifications

Continuous vs Digital Signals

sampling: discretisation in time

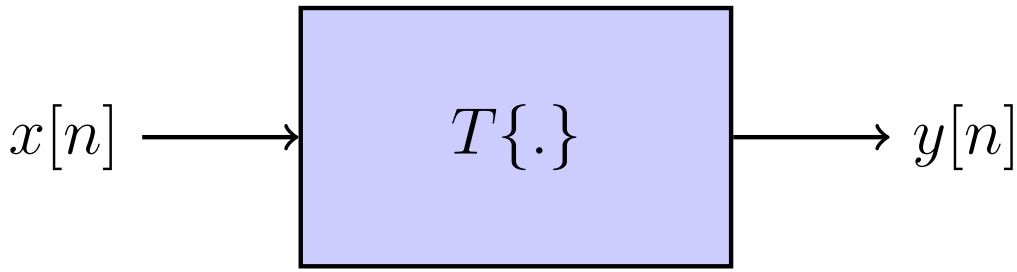


quantisation: discretisation in amplitude



(Figures from Wikipedia)

Linear Time-Invariant (LTI) Systems



In general[1]:

$$y[n] = T\{x[n]\}$$

Time invariance:

$$y[n - n_0] = T\{x[n - n_0]\}$$

Linearity:

$$T\{a_1 x_1[n] + a_2 x_2[n]\} = a_1 T\{x_1[n]\} + a_2 T\{x_2[n]\}$$

[1]Notation: $T\{\}$ may consider x at any time step to calculate $y[n]$

LTI: Impulse Response

In general we can always write:

$$x[n] = \sum_{k=-\infty}^{\infty} x[k]\delta[n-k]$$

For the linearity:

$$y[n] = T\{x[n]\} = \sum_{k=-\infty}^{\infty} x[k]T\{\delta[n-k]\}$$

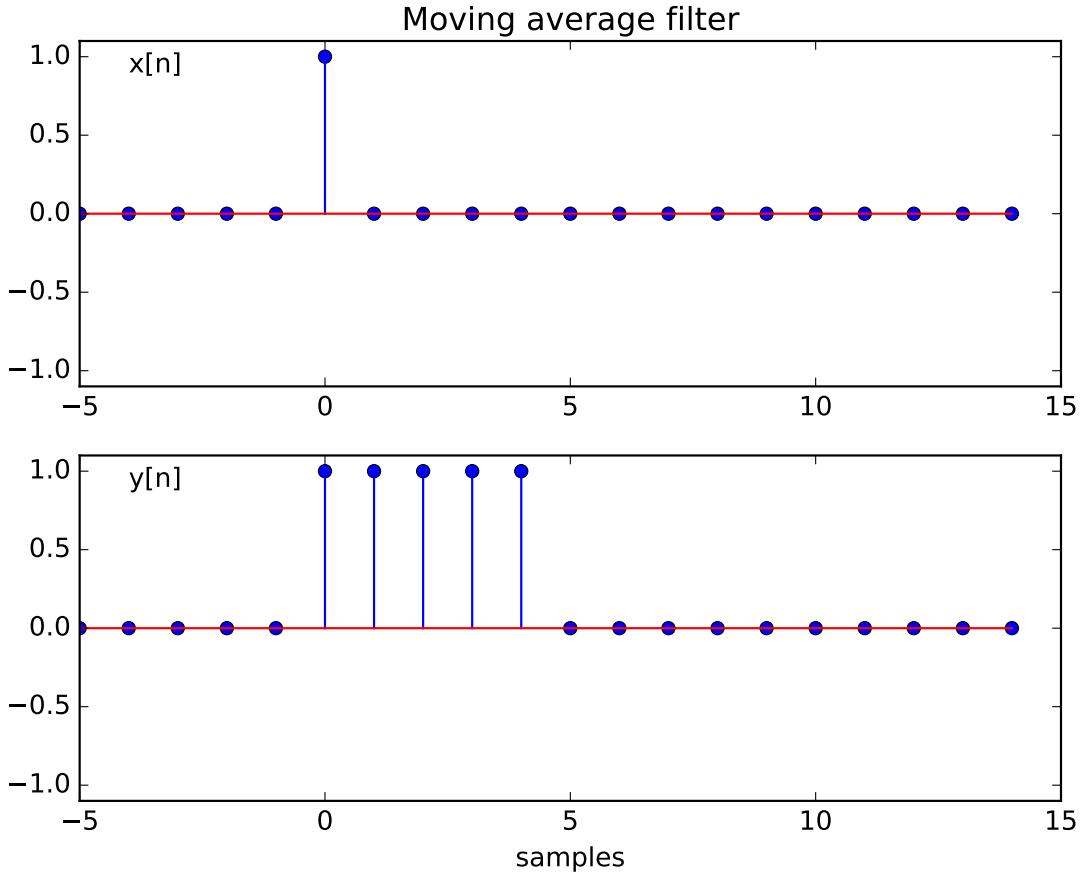
$h[n] \equiv T\{\delta[n]\}$ is the system's response to an impulse $\delta[n]$

For the time invariance:

$$T\{\delta[n-k]\} = h[n-k]$$

$h[n]$ is a complete description of the system!

Examples of LTIs: Moving Average



$$y[n] = x[n] + x[n-1] + \cdots + x[n-P]$$

Finite Impulse Response (FIR) Systems

y only depends on (delayed) samples of the input (no feedback)

$$\begin{aligned} y[n] &= b_0 x[n] + b_1 x[n-1] + \cdots + b_P x[n-P] \\ &= \sum_{i=0}^{P} b_i x[n-i] \end{aligned}$$

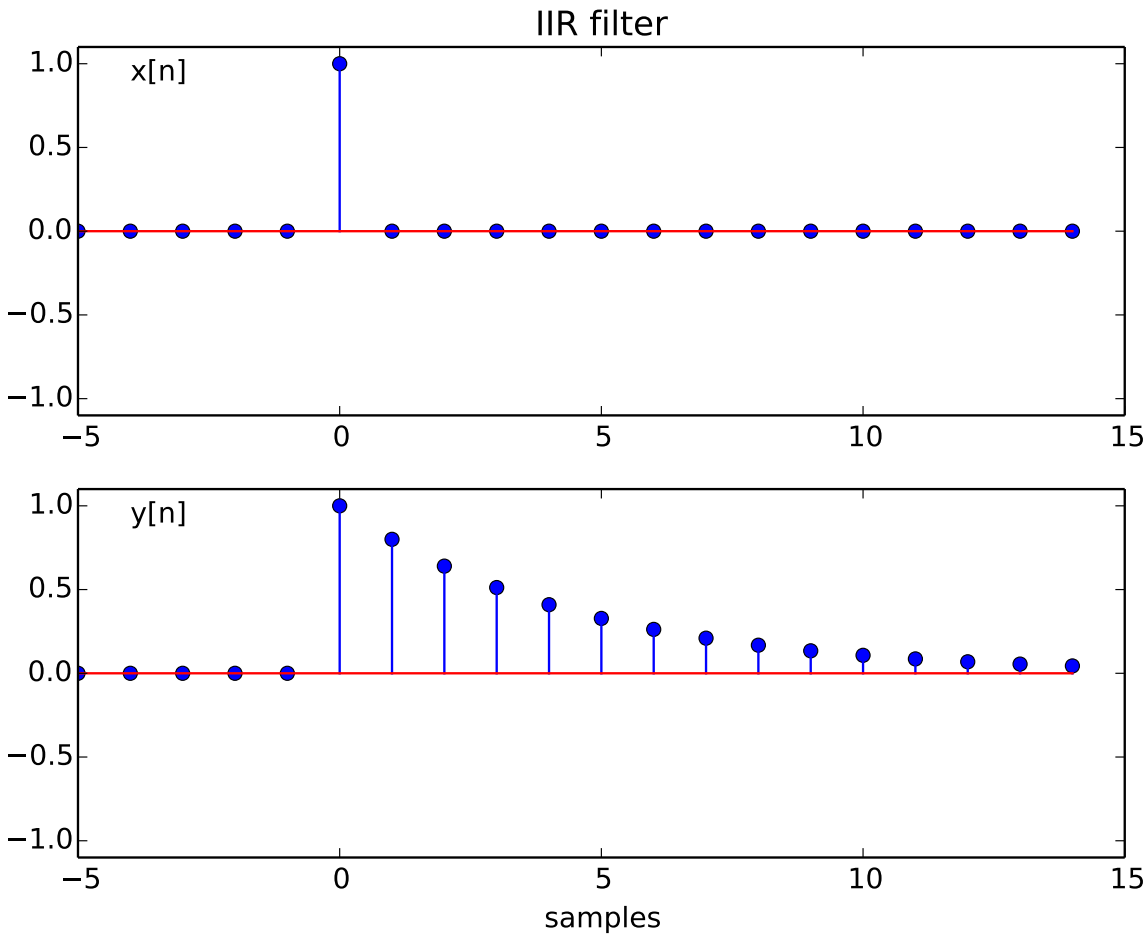
Infinite Impulse Response (IIR) Systems

Auto regressive (AR): y depends on (delayed) samples of the input, as well as the output at previous times (feedback)

$$\begin{aligned} y[n] &= \frac{1}{a_0}\left(b_0 x[n] + b_1 x[n-1] + \cdots + b_P x[n-P] + \right. \\ &\quad \left. -a_1 y[n-1] - a_2 y[n-2] - \cdots + a_Q y[n-Q]\right) \\ &= \frac{1}{a_0}\left(\sum_{i=0}^{P} b_i x[n-i] - \sum_{j=1}^{Q} a_j y[n-j]\right) \end{aligned}$$

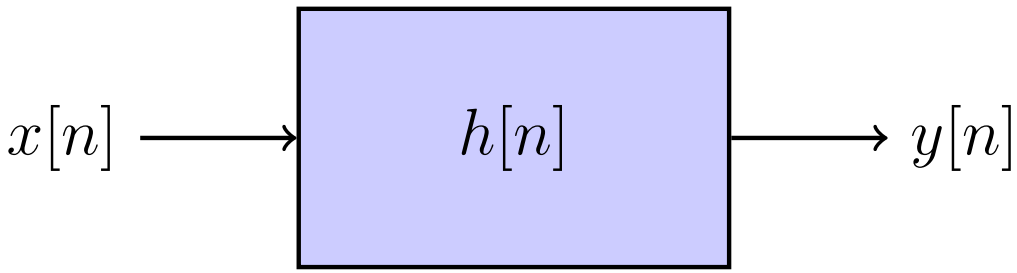
IIR Example

$$y[n] = x[n] - ay[n-1]$$



stable only if $|a| < 1$, here $a = -0.8$

Convolution



$$y[n] = T\{x[n]\} = \sum_{k=-\infty}^{\infty} x[k]h[n-k] = x[n] * h[n]$$

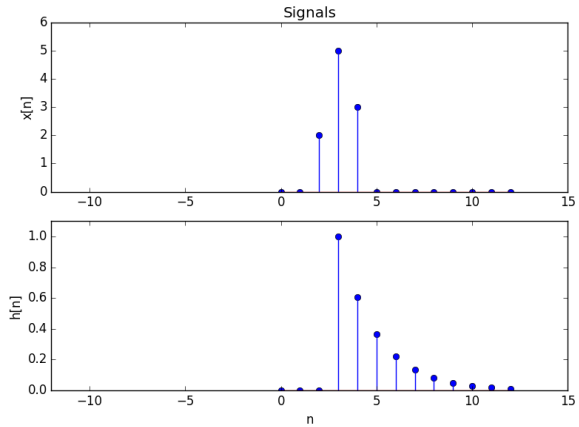
Properties:

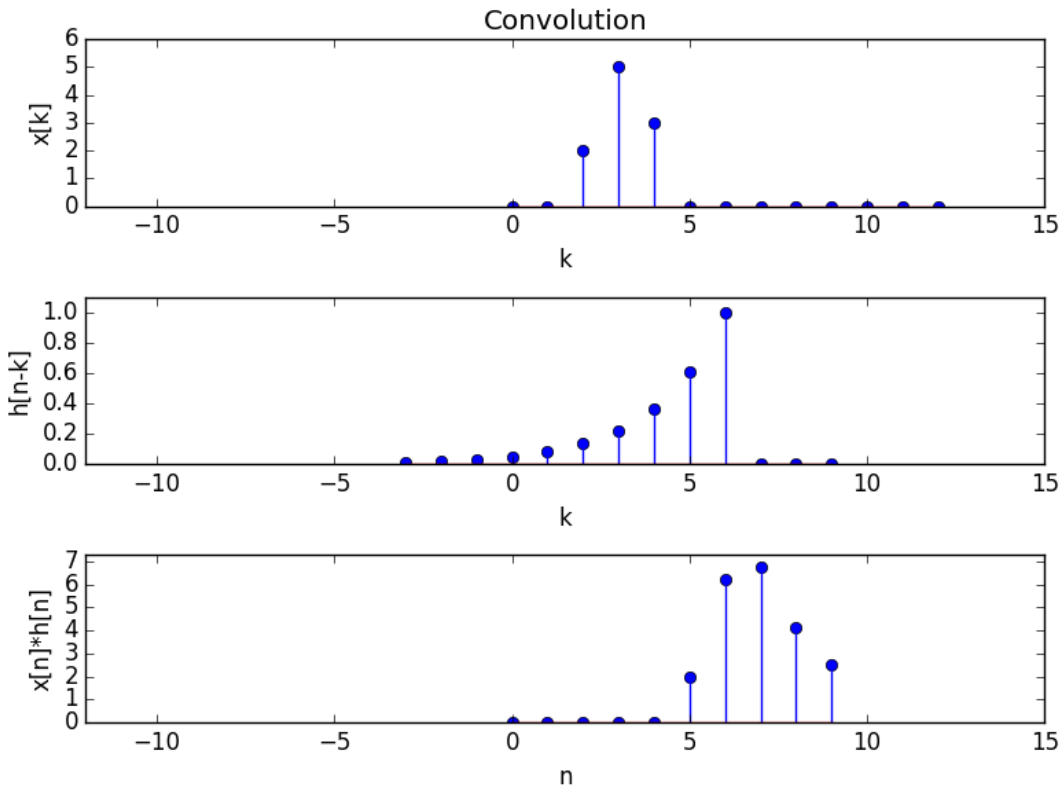
- linearity: $(a_1x_1 + a_2x_2) * h = a_1(x_1 * h) + a_2(x_2 * h)$
- simmetry: $x * h = h * x$

Kind of complicated to interpret.

Convolution: Illustration

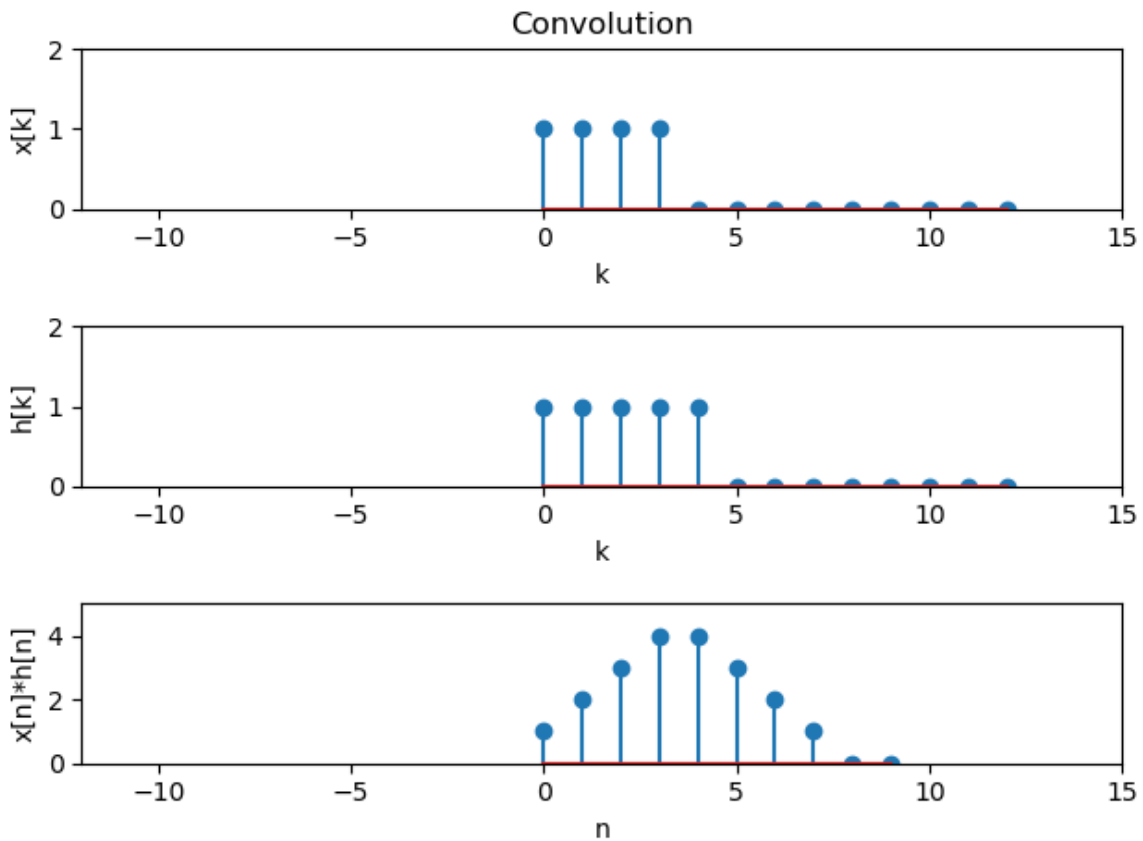
$$x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k]$$





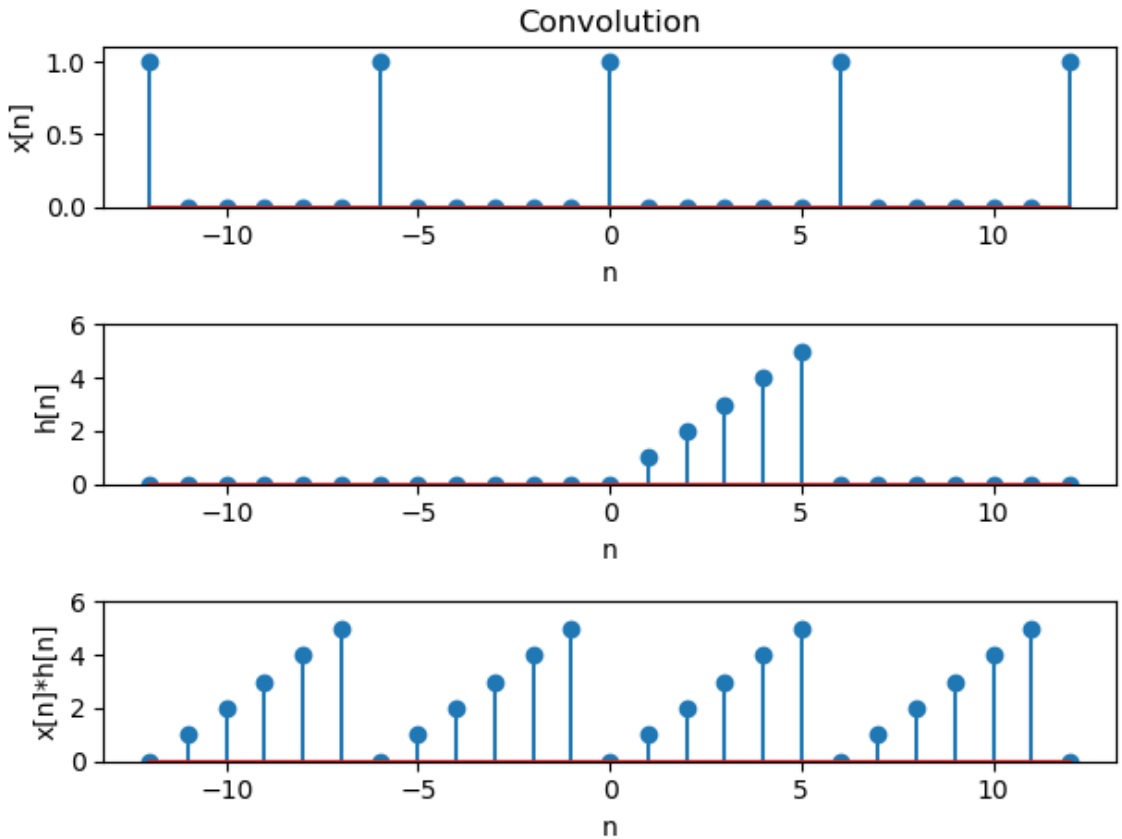
Convolution: try it yourself!

$$x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k]$$



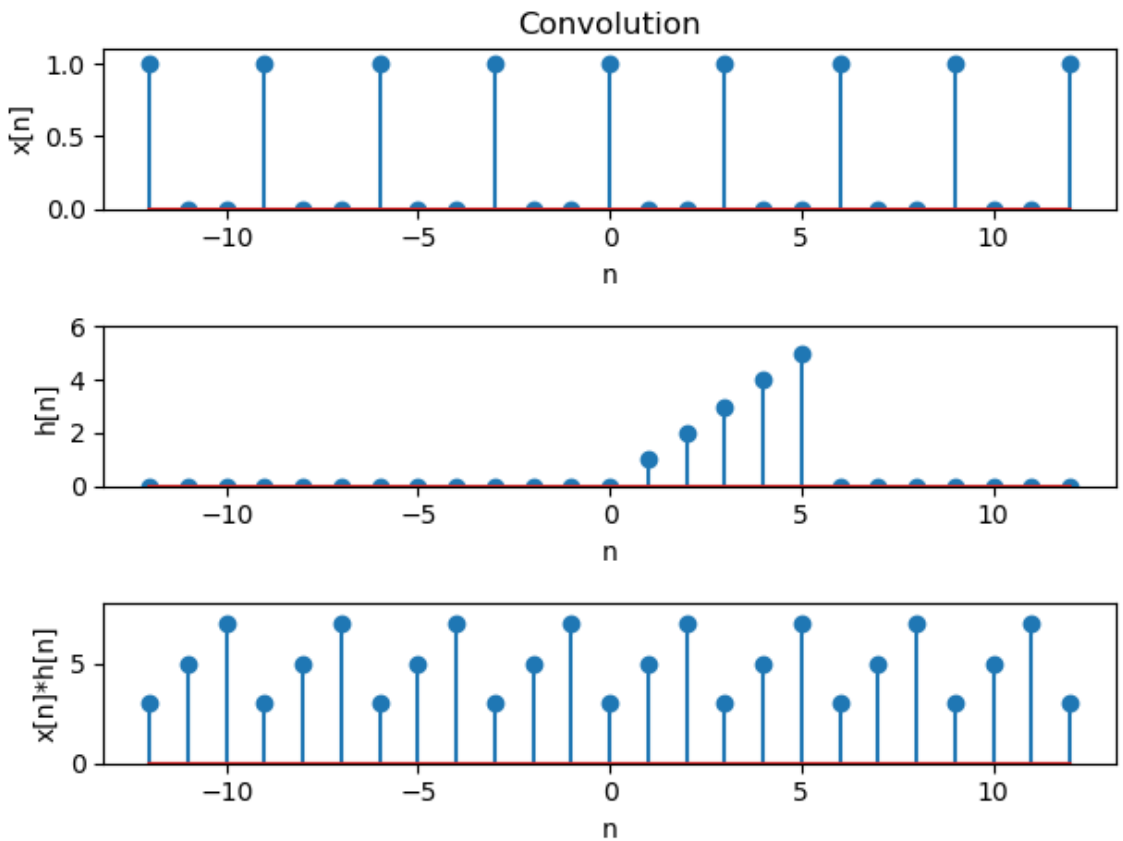
Convolution: train of pulses

$$x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k]$$



Convolution: and again

$$x[n] * h[n] = \sum_{k=-\infty}^{\infty} x[k]h[n-k]$$



Sinusoidal Signals

$$x[n] = a\cos(2\pi f \overbrace{T_s n}^{\text{time in seconds}} + \phi) = a\cos(\omega T_s n + \phi)$$

Where

- a is an amplitude factor
- f is the frequency ($\omega = 2\pi f$ is the angular frequency)
- ϕ is the initial phase

More convenient mathematically:
define sinusoidal signals on the complex plane:

$$c[n] = ae^{j(\omega n+\phi)} = ae^{j\phi}e^{j\omega n} \in \mathbb{C}$$

$$x[n] = \Re(c[n]) = \frac{c[n] + c^*[n]}{2} \quad \text{(real part)}$$

(Euler's formula: $e^{jt} = \cos(t) + j\sin(t)$)

Sinusoidal Signals and Linear Systems

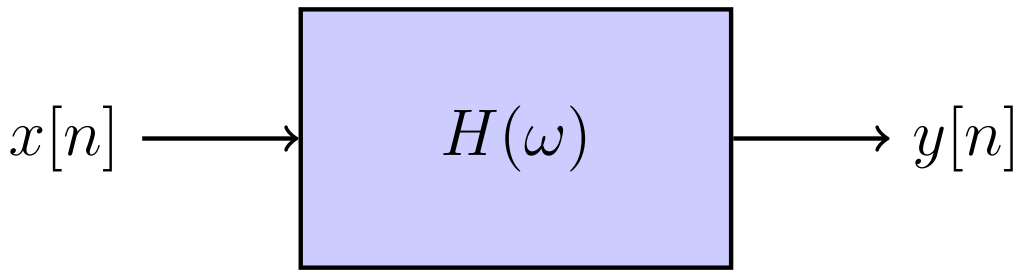
If $x[n] = e^{j\omega_0 n}$ then

$$\begin{aligned} y[n] &= x[n] * h[n] = \sum_{m=-\infty}^{\infty} x[m]h[n-m] = (k \equiv n-m) \\ &= \sum_{k=-\infty}^{\infty} h[k]x[n-k] = \sum_{k=-\infty}^{\infty} h[k]e^{j\omega_0(n-k)} \\ &= \sum_{k=-\infty}^{\infty} h[k]e^{-j\omega_0 k}e^{j\omega_0 n} = e^{j\omega_0 n}\sum_{k=-\infty}^{\infty} h[k]e^{-j\omega_0 k} \\ &= H(\omega_0)e^{j\omega_0 n} \end{aligned}$$

- Sinusoidal signals are eigensignals for LTI systems
- $H(\omega_0)$ is called Transfer Function of the system

Transfer Function

Expresses how the system modifies sinusoidal signals (amplitude and phase)



$$H(\omega) = \sum_{k=-\infty}^{\infty} h[k]e^{-j\omega k}$$

It corresponds to the Fourier Transform of the impulse response $h[n]$

Fourier Transforms

Fourier transform of continuous signals

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt$$

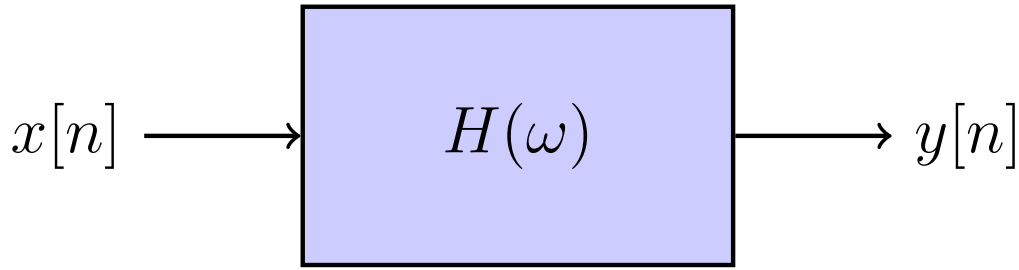
Fourier transform of discrete signals

$$X(\omega) = \sum_{k=-\infty}^{\infty} x[k] e^{-j\omega k}$$

- Any signal $x[n]$ can be decomposed into sums of pure tones
- $X(\omega)$ indicates how large is the contribution of each tone and with which phase.

Great illustration from 3Blue1Brown: `https://youtu.be/spUNpyF58BY`

Transfer Function for Generic Signals



Sinusoidal signals:

$$x[n] = e^{j\omega_0 n} \rightarrow y[n] = H(\omega_0)e^{j\omega_0 n}$$

Generic signals (can be decomposed in sinusoids):

$$x[n] = X(\omega_1)e^{j\omega_1 n} + X(\omega_2)e^{j\omega_2 n} + \ldots$$

Then

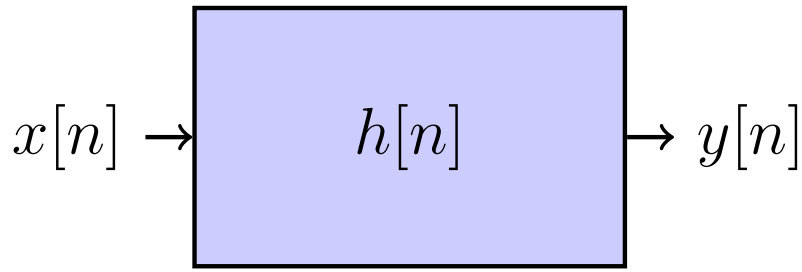
$$y[n] = H(\omega_1)X(\omega_1)e^{j\omega_1 n} + H(\omega_2)X(\omega_2)e^{j\omega_2 n} + \ldots$$

Which means that:

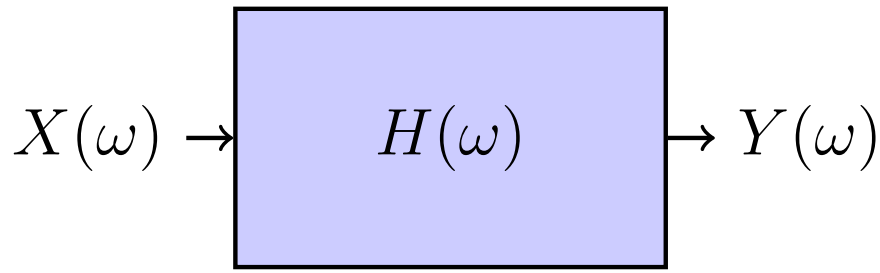
$$Y(\omega) = H(\omega)X(\omega)$$

Linear Time-Invariant Systems

Time Domain



Frequency Domain



Convolution

$$\begin{aligned} y[n] &= x[n] * h[n] \\ &= \sum_{k=-\infty}^{\infty} x[k]h[n-k] \end{aligned}$$

Multiplication

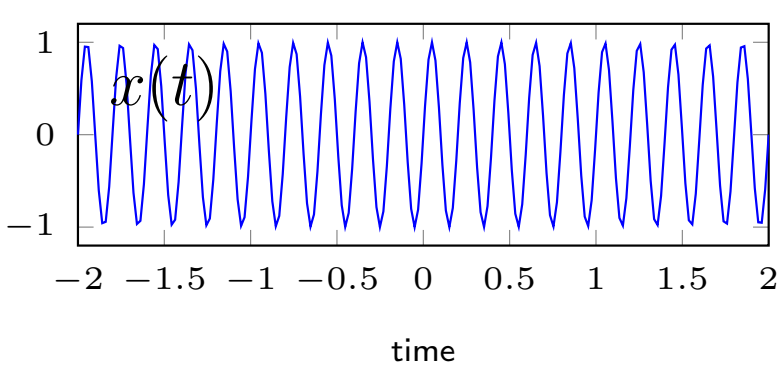
$$Y(\omega) = X(\omega)H(\omega)$$

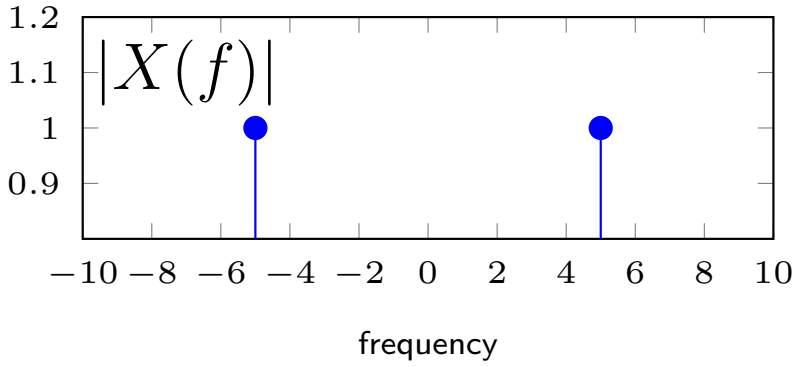
$h[n]$ Impulse Response

$H(\omega)$ Transfer Function

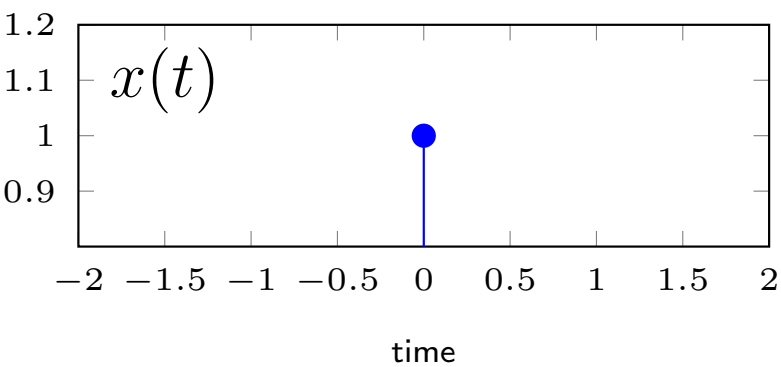
Fourier Transform of Useful Signals

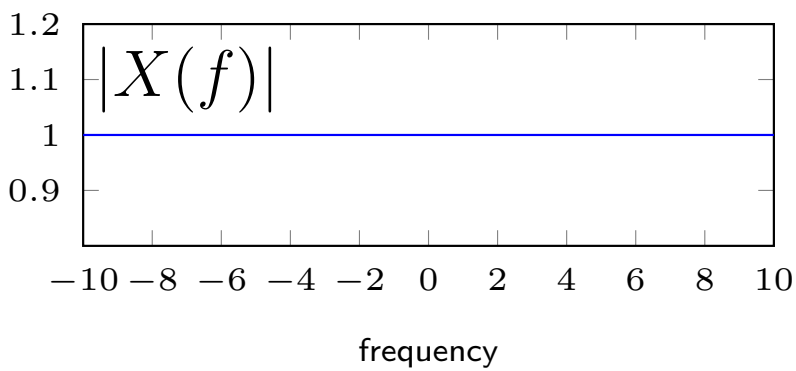
Sinusoidal at frequency $f = 5Hz$





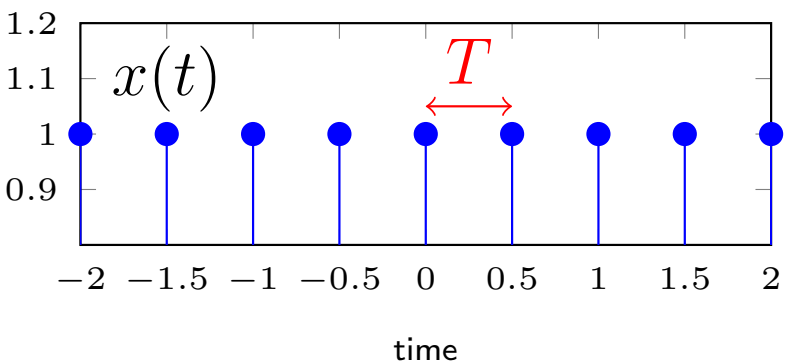
Single (Ideal) Impulse

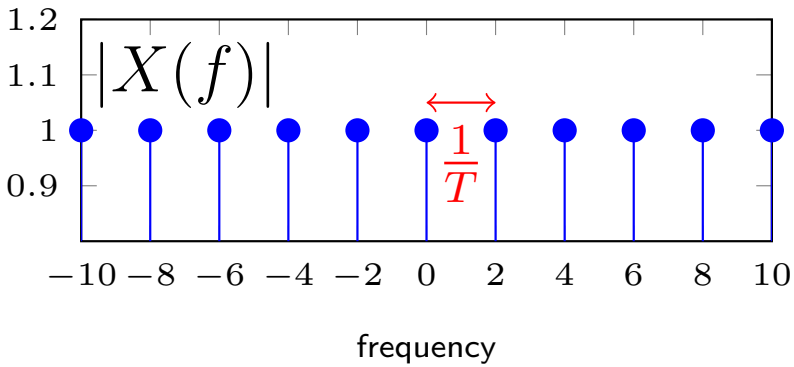




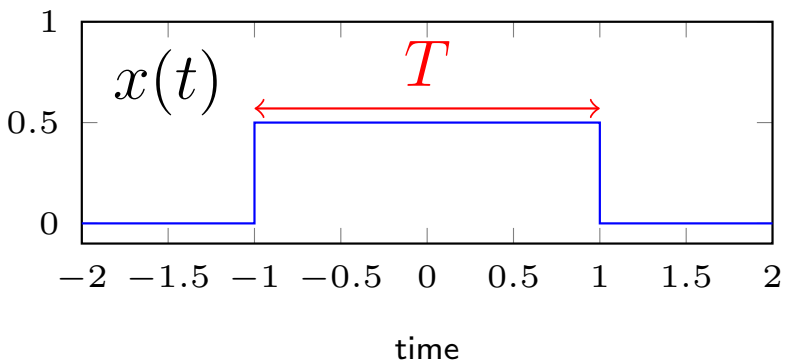
Fourier Transform of Useful Signals

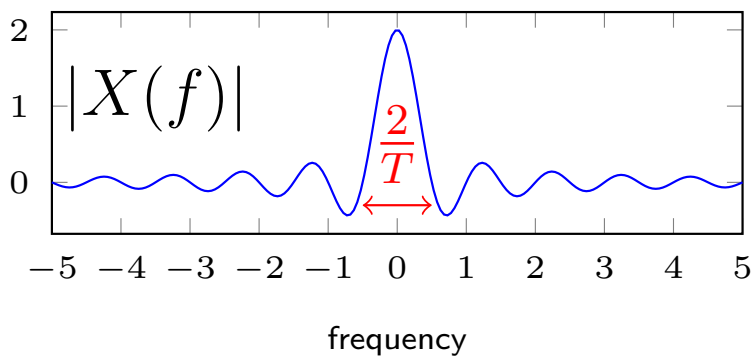
Train of (Ideal) Impulses





Square function

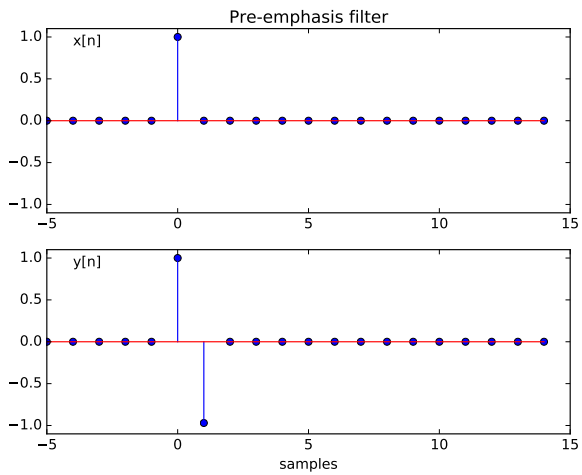


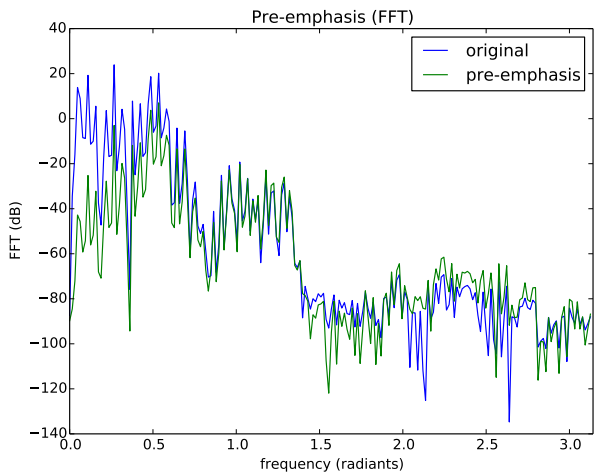


Properties of Fourier Transform

Time domain		Frequency domain
$x[n] * h[n]$	$\Longleftrightarrow$	$X(\omega)H(\omega)$

Example: any filter (pre-emphasis, moving average...)

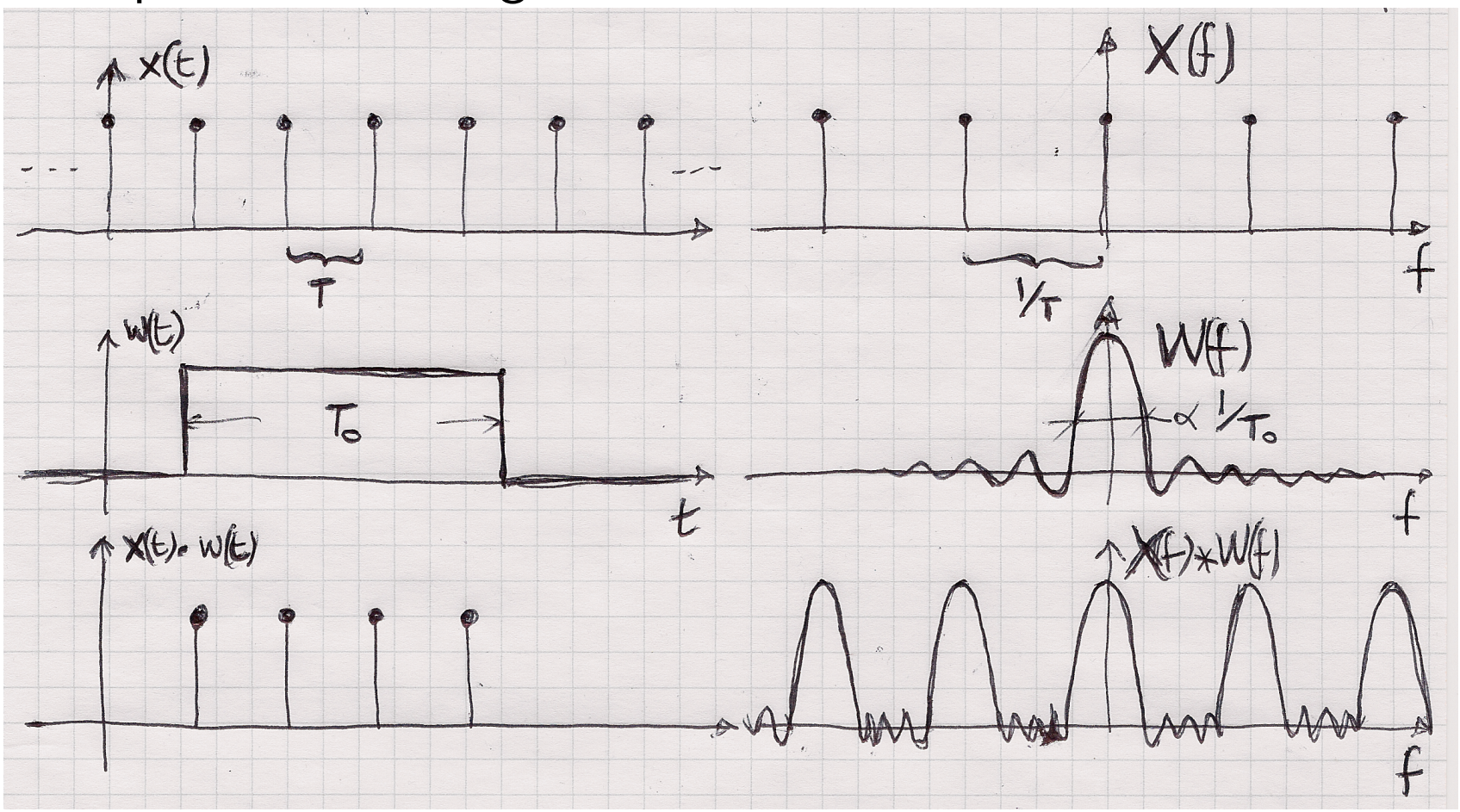




Properties of Fourier Transform

Time domain		Frequency domain
$x[n]w[n]$	$\Longleftrightarrow$	$X(\omega) * W(\omega)$

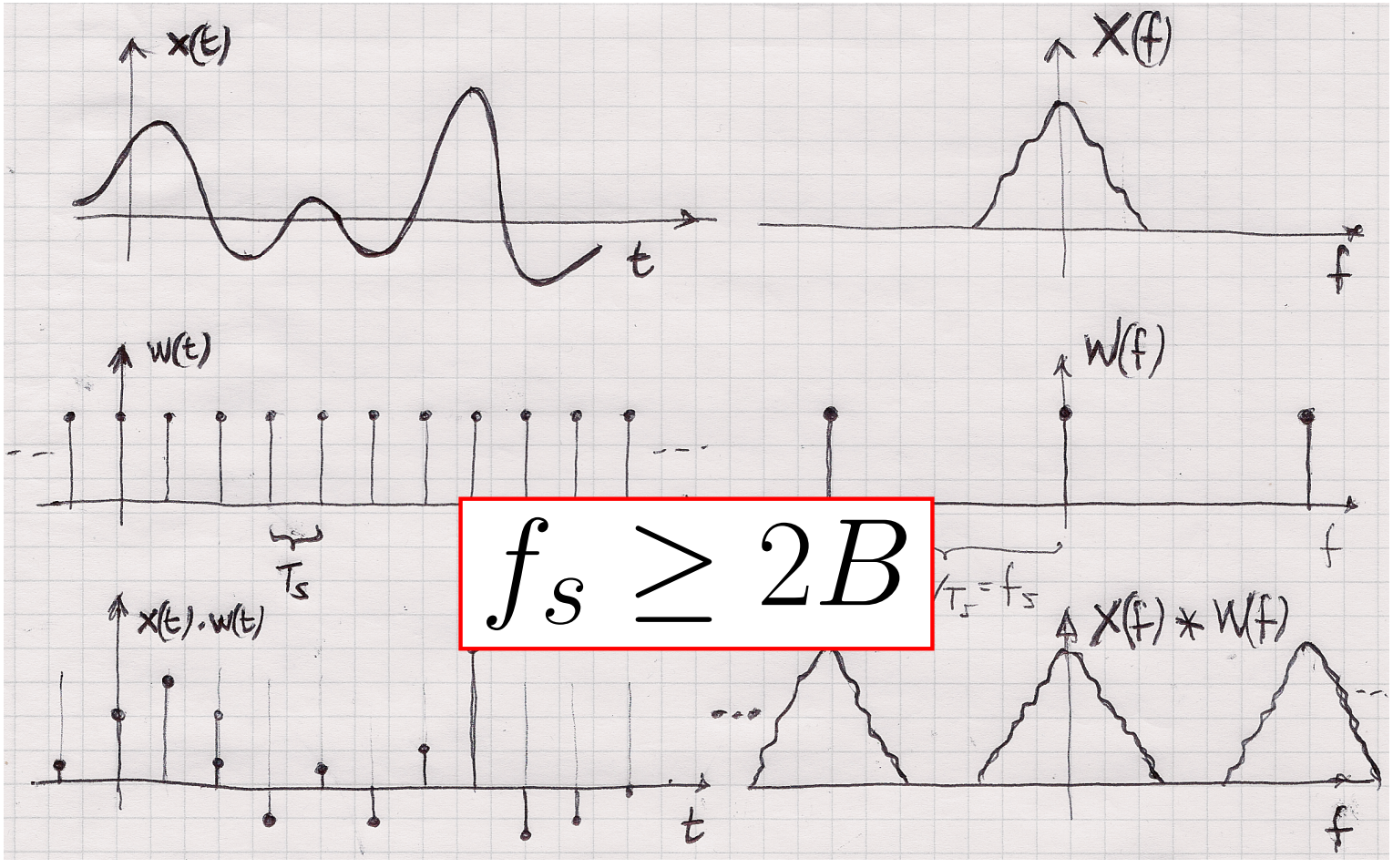
Example 1: windowing



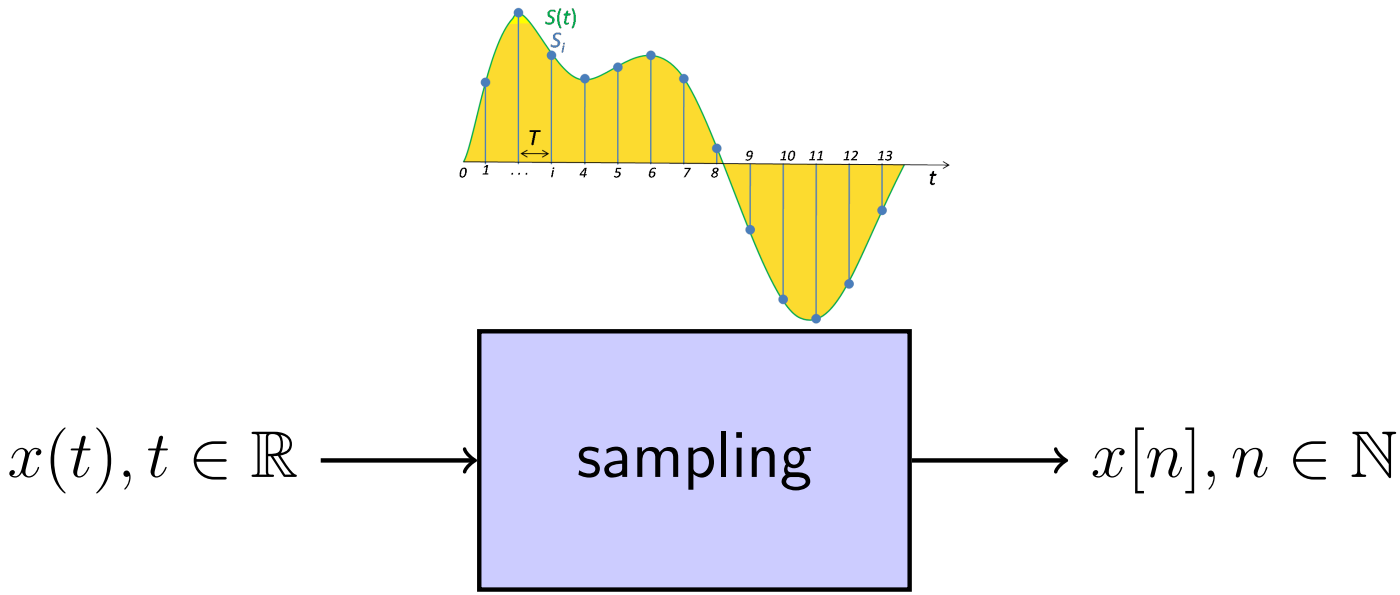
Properties of Fourier Transform

Time domain		Frequency domain
$x[n]w[n]$	$\Longleftrightarrow$	$X(\omega) * W(\omega)$

Example 2: sampling



Sampling Theorem (Nyquist-Shannon)



If $x(t)$ contains energy up to B_x, in order to reconstruct the signal we need to sample with

$$f_s > 2B_x$$

Aliasing

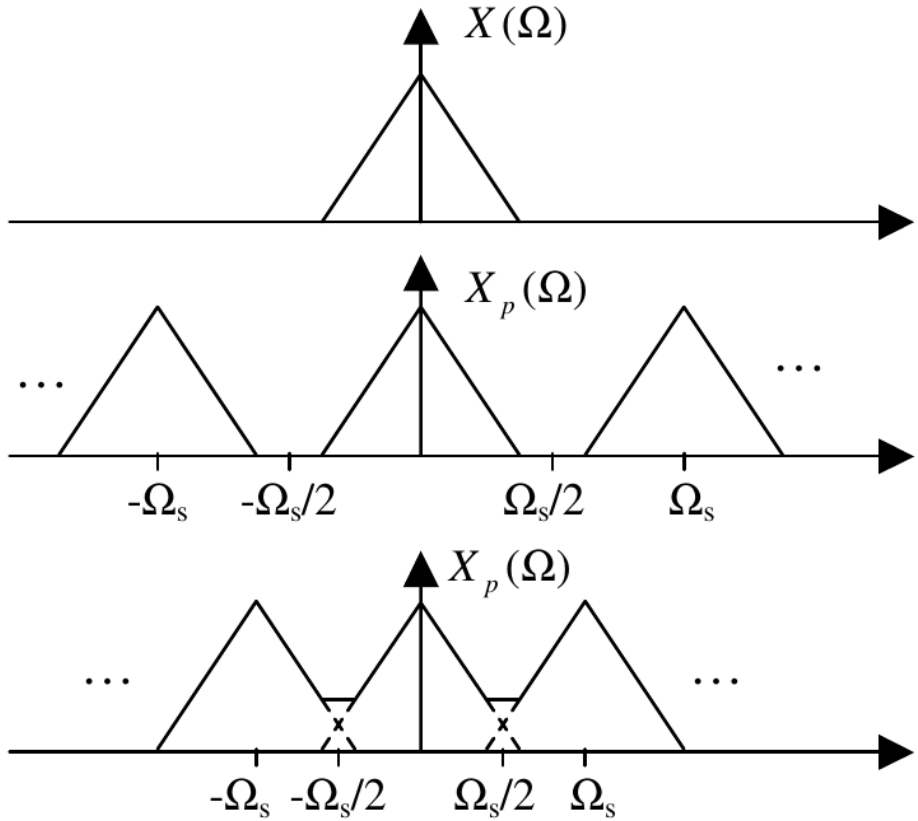


Figure from Huang, Acero and Hon (2001)

Aliasing: Illustration

Video from `https://youtu.be/usN47Jvy9PY`

First step: represent speech signal

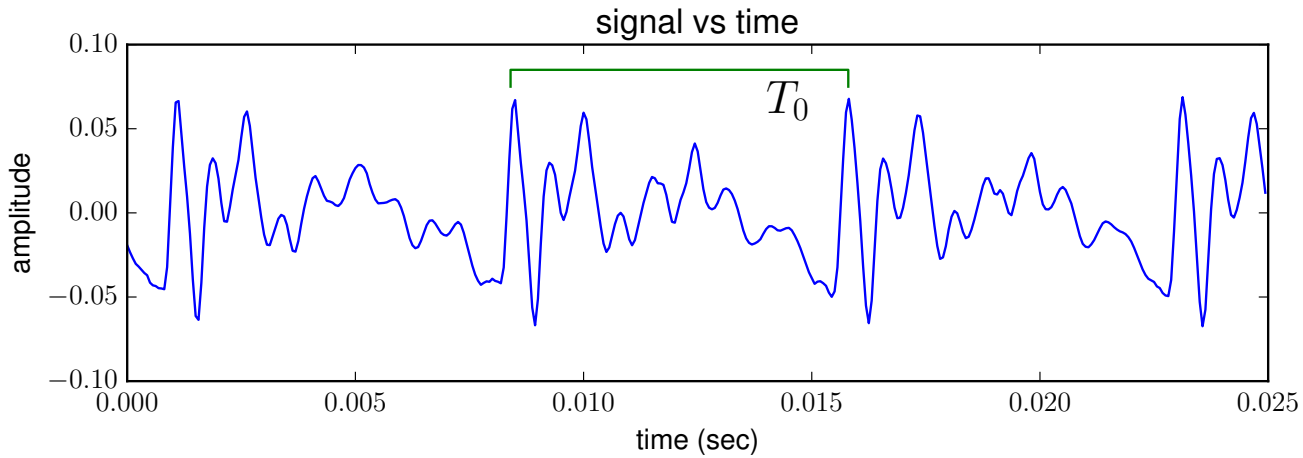
Sampling

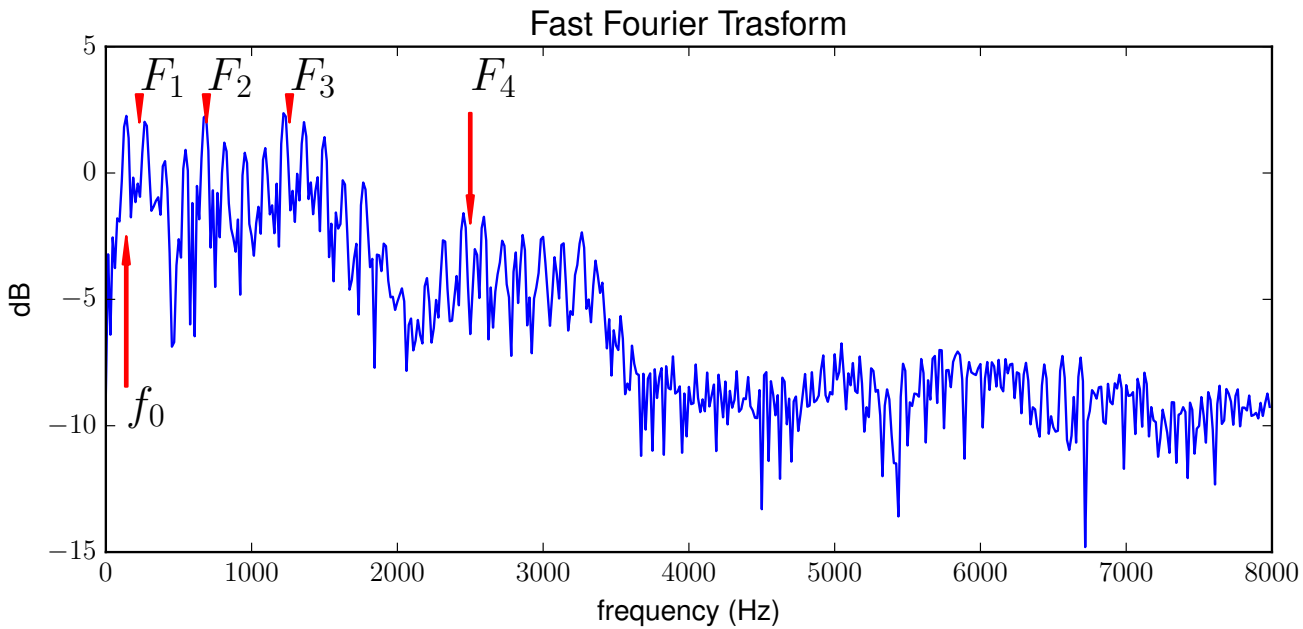
- Nyquist-Shannon Theorem: sample at twice the band
- 8kHz (4kHz band, telephone), 16kHz (8 kHz band, high quality)
- TIDIGITS sampled at 20kHz
- TIMIT sampled at 16kHz

Quantisation

- Type of quantisation: linear, a-law, μ-law
- 8, 16 bits (more rare 32, floating point)
- TIDIGITS and TIMIT are quantised with 16 bits linear

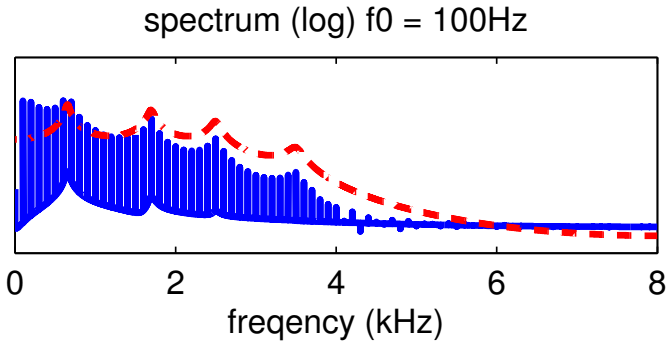
Example: vowel

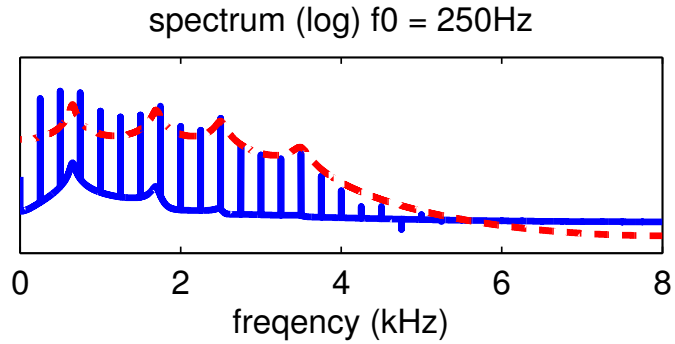




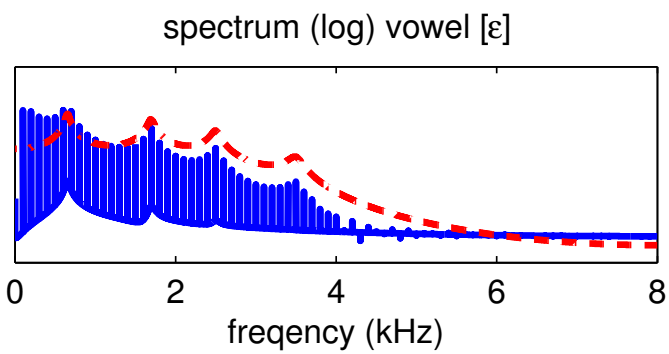
F_0 and Formants

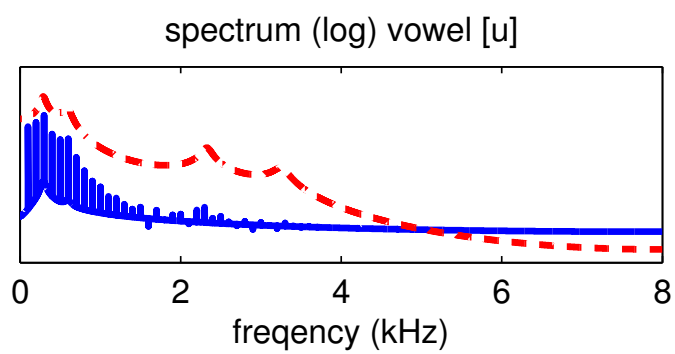
- Varying F_0 (vocal fold oscillation rate)





- Varying Formants (vocal tract shape)



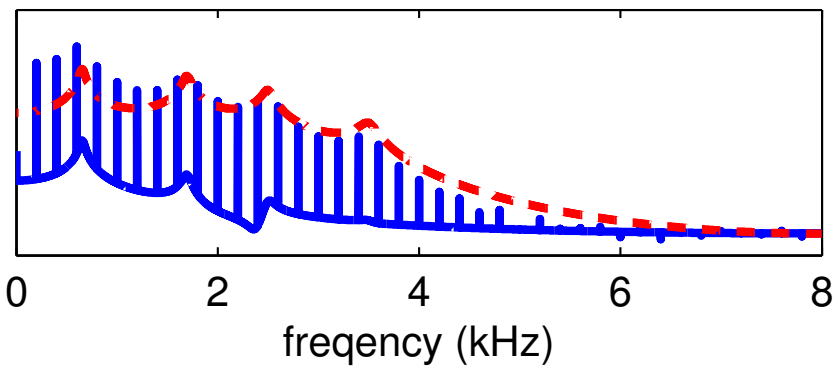


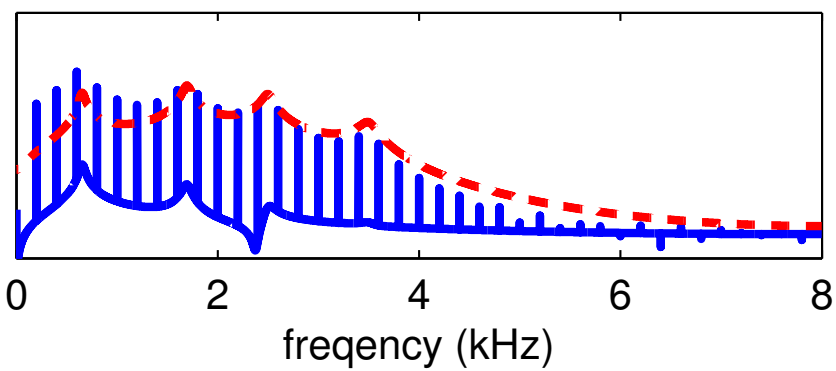
Pre-emphasis

Compensate for the 6db/octave drop (radiation at the lips)

$$y[n] = x[n] - \alpha x[n-1]$$

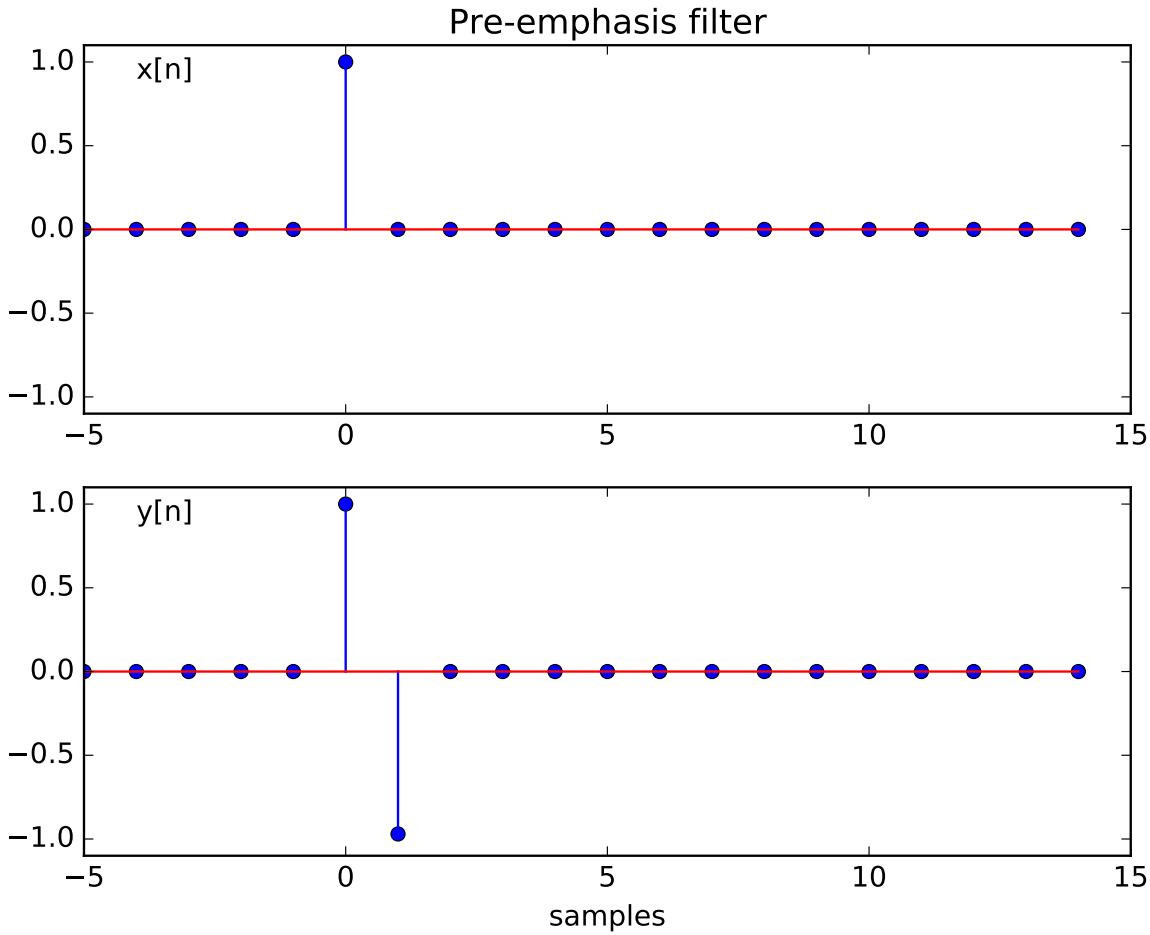
Corresponds to a linear filter with $A = 1$ and $B = [1 \quad -\alpha]$





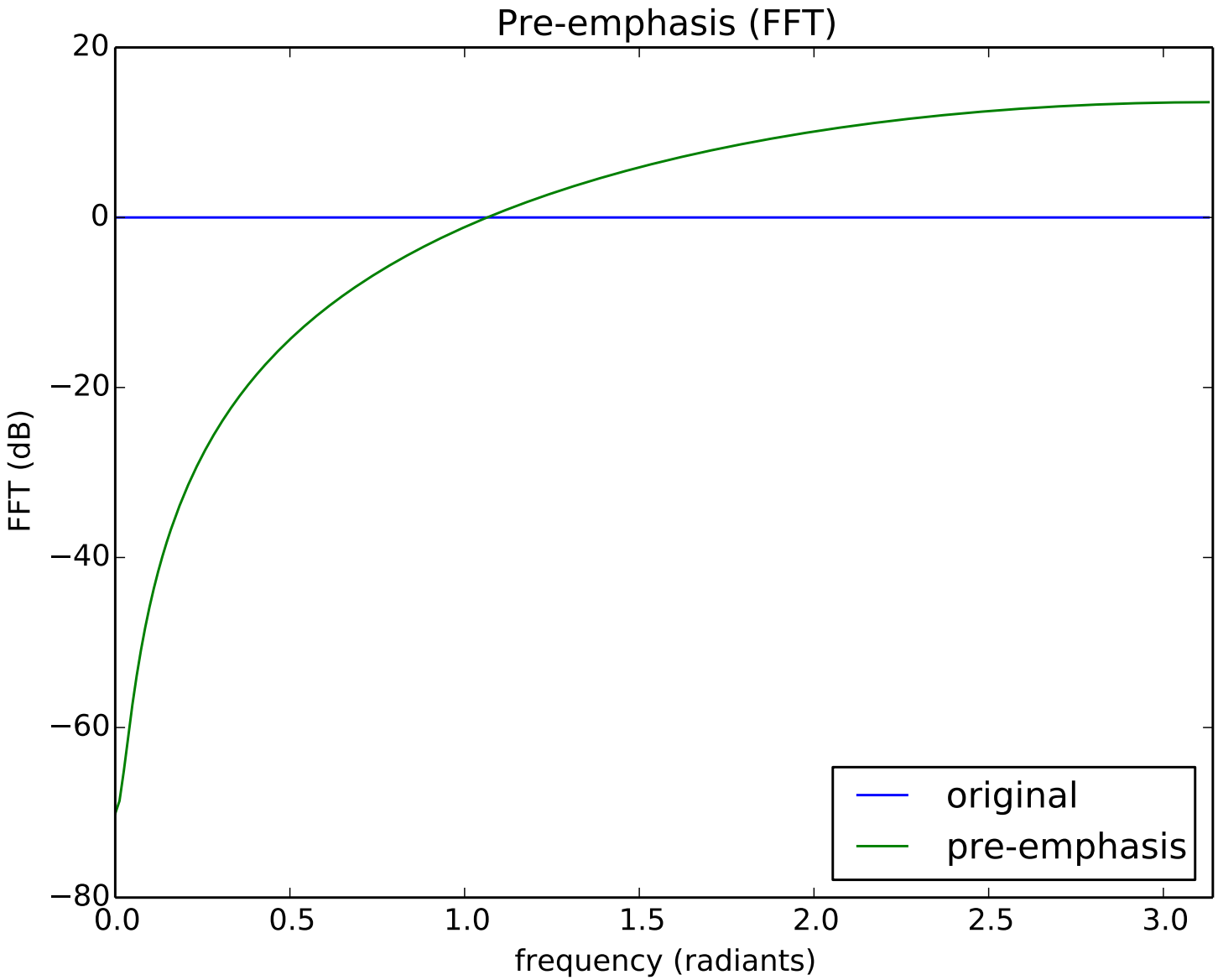
α is usually 0.95–0.97

Examples of LTIs: Pre-emphasis

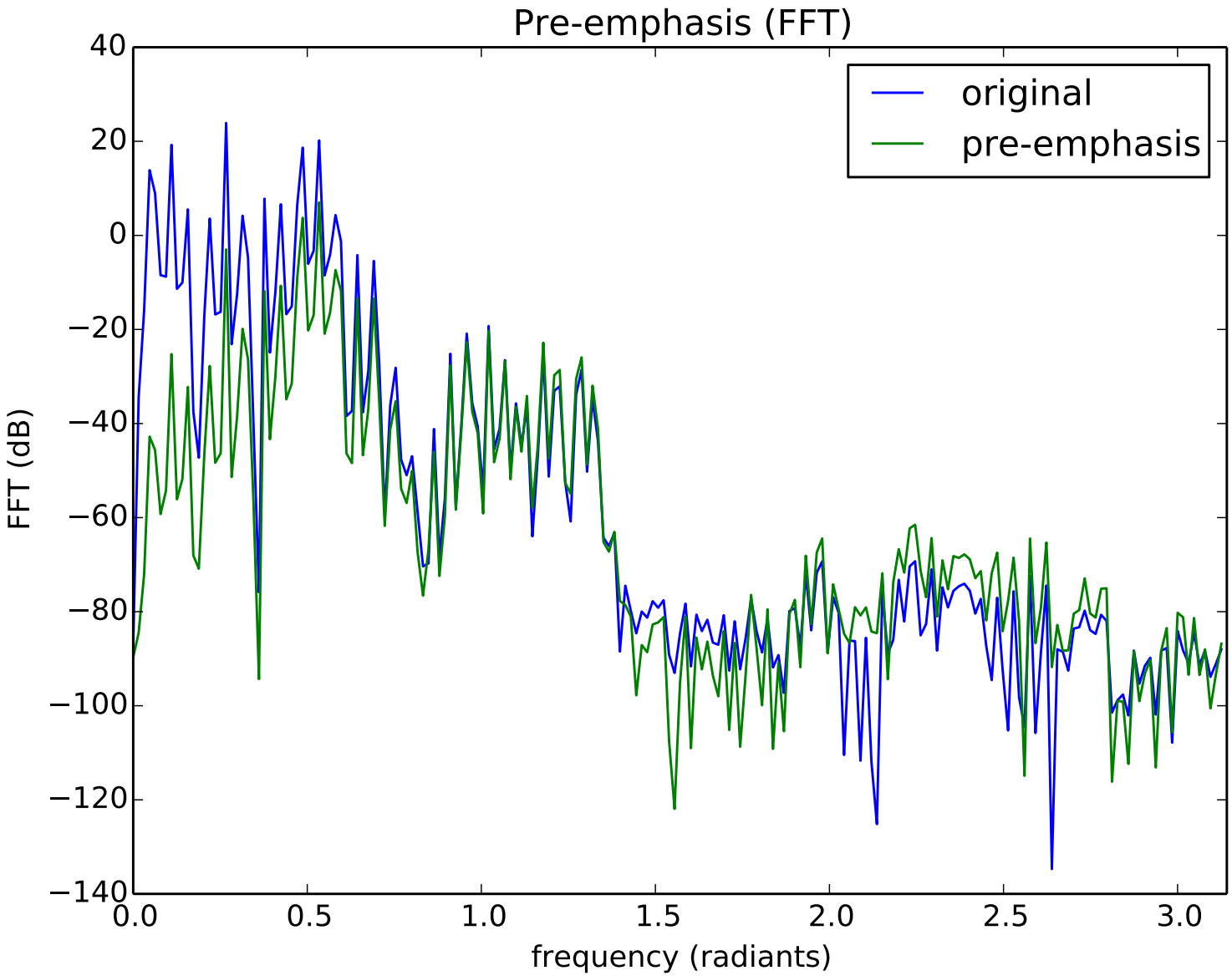


$$y[n] = x[n] - \alpha x[n-1], \quad \text{with } \alpha = 0.97$$

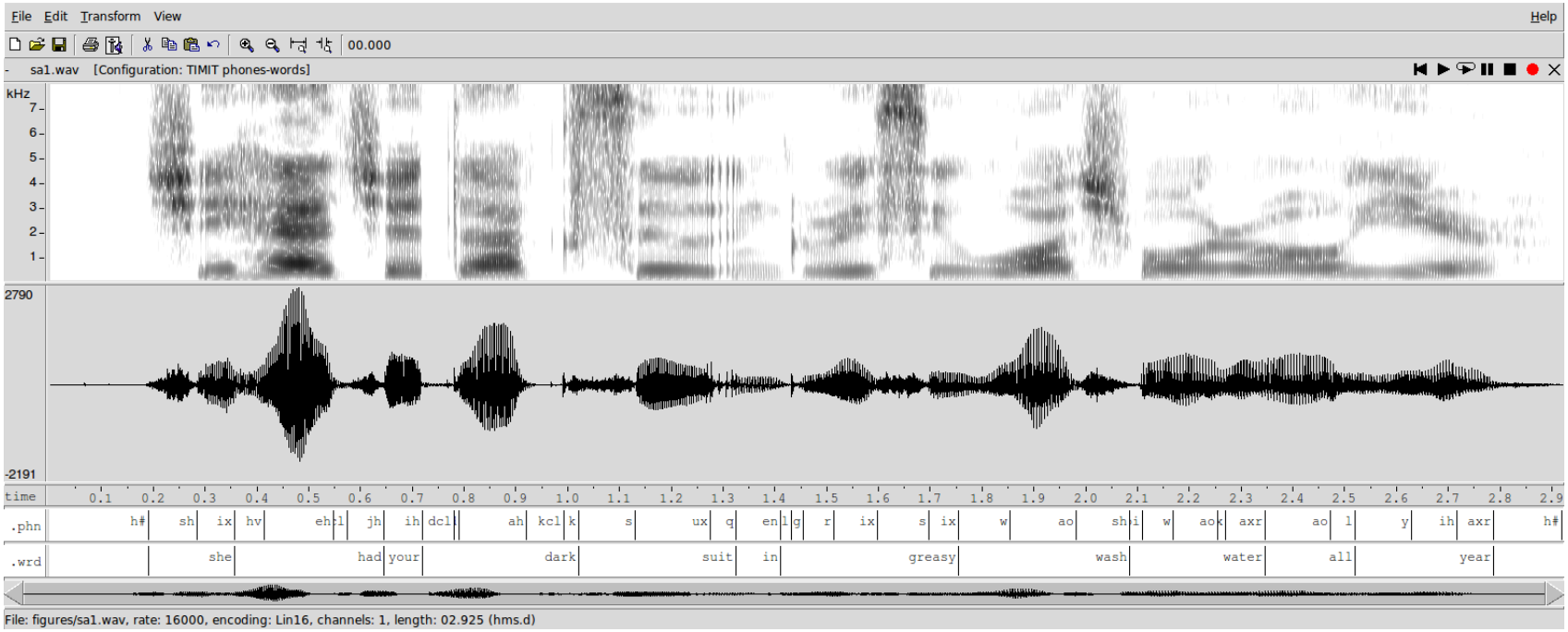
Pre-emphasis in frequency domain



Pre-emphasis applied to vowel

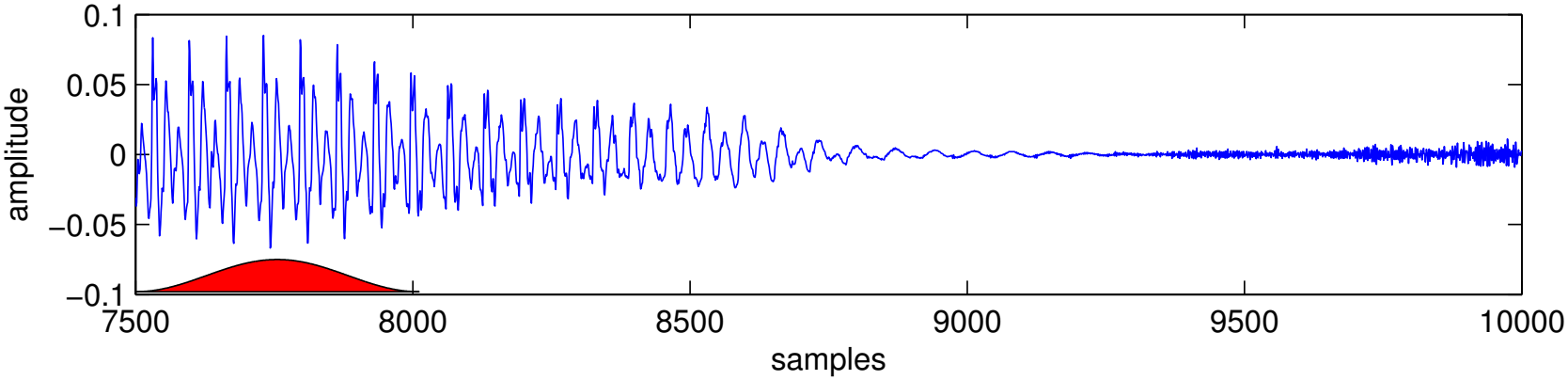


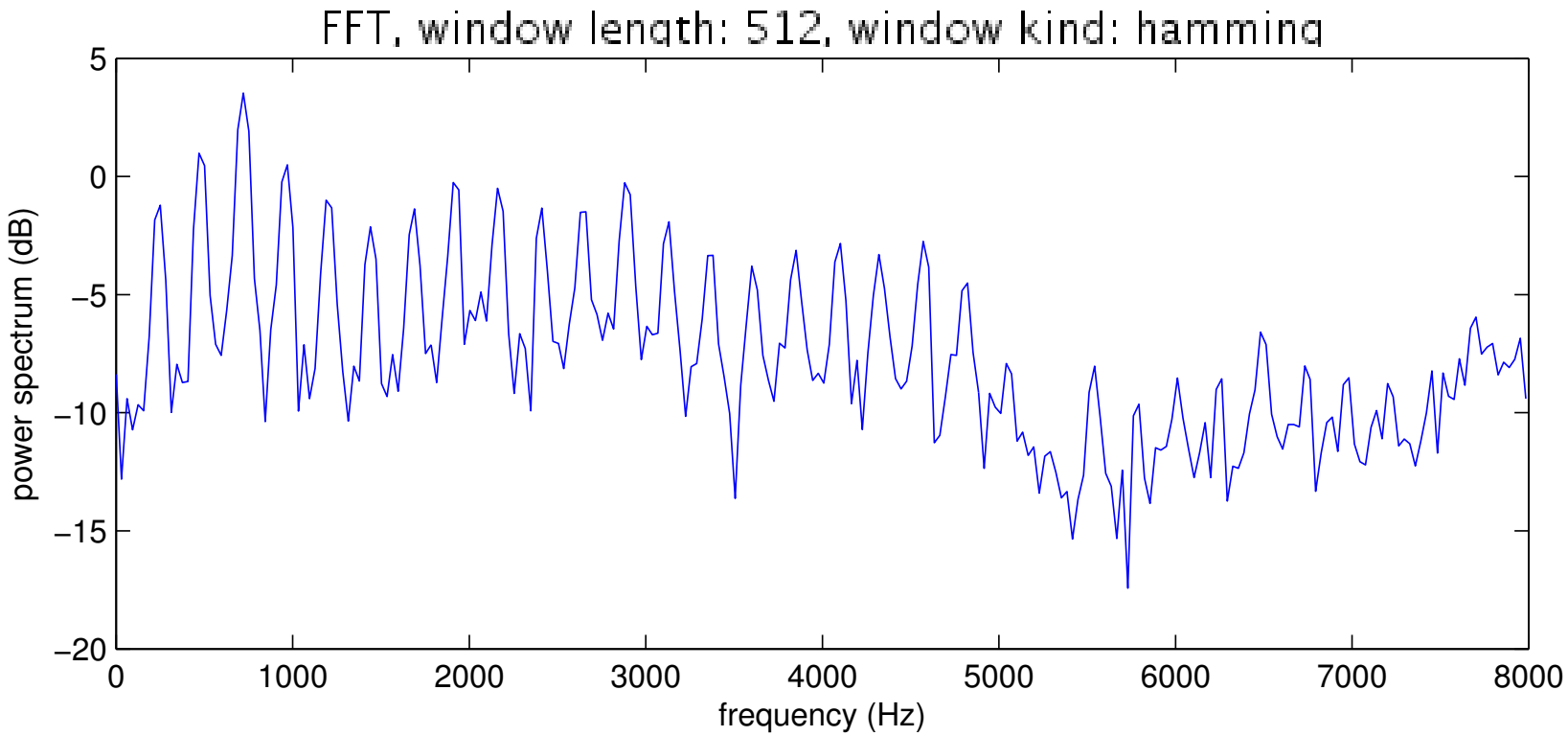
A time varying signal



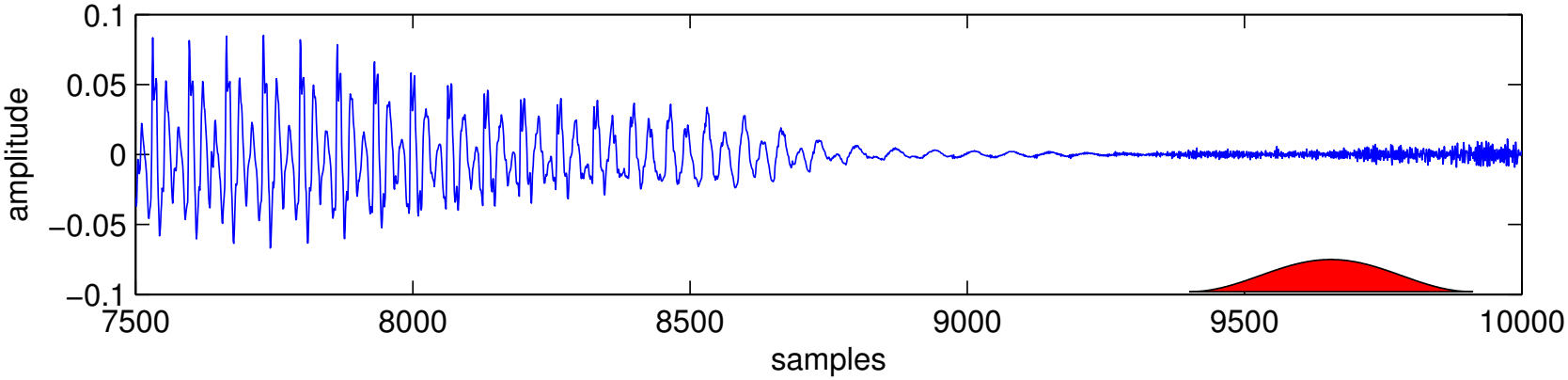
- speech is time varying
- short segment are quasi-stationary
- use short time analysis

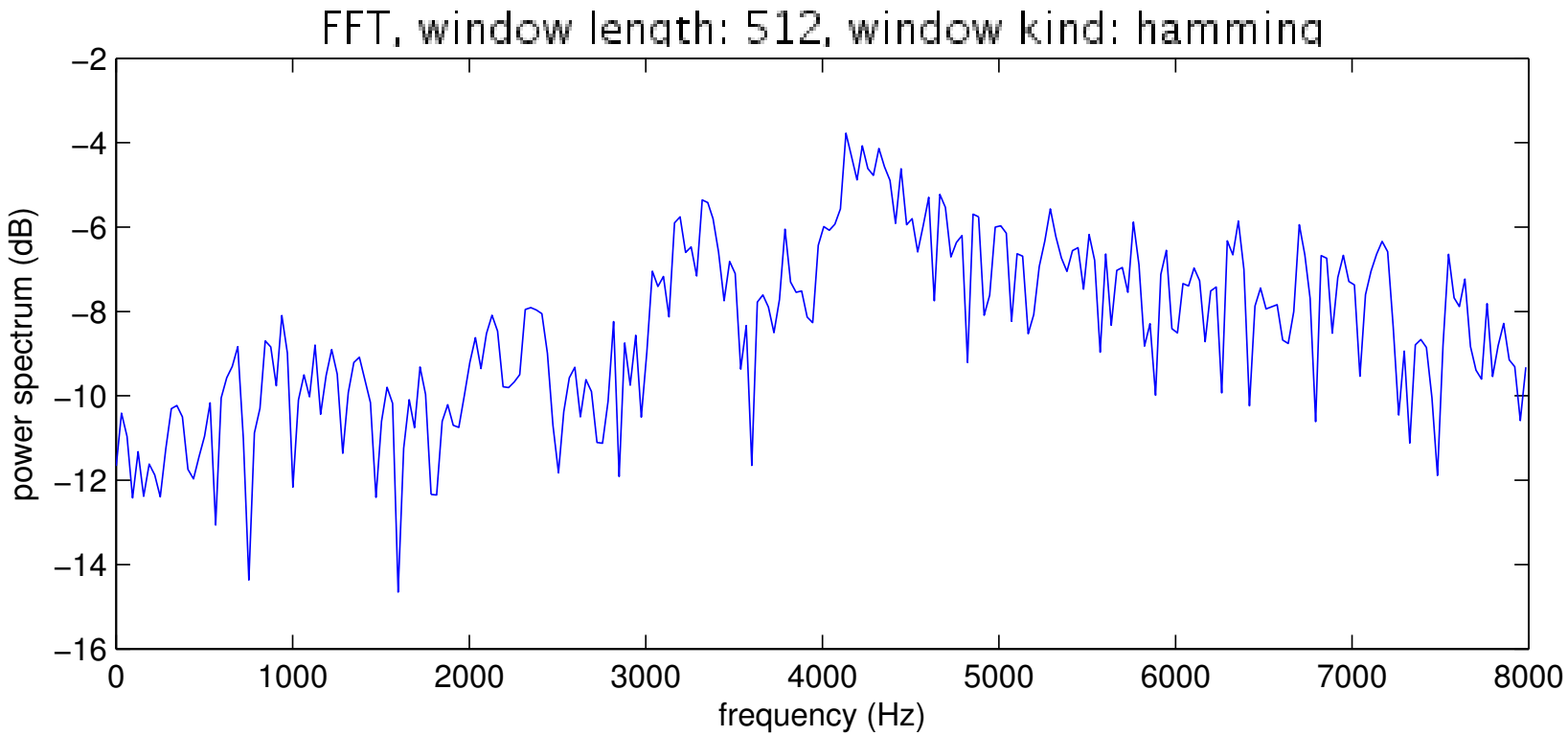
Short-Time Fourier Analysis





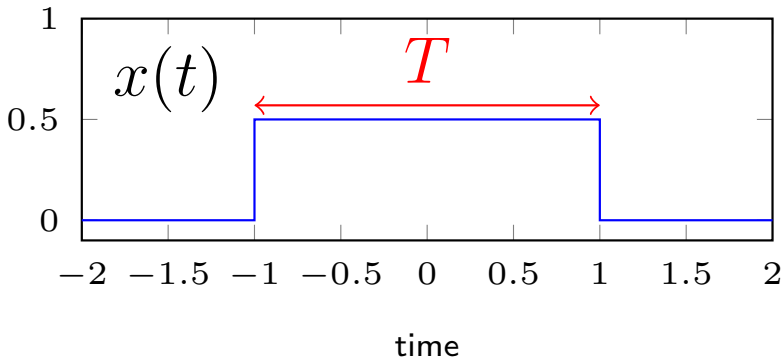
Short-Time Fourier Analysis

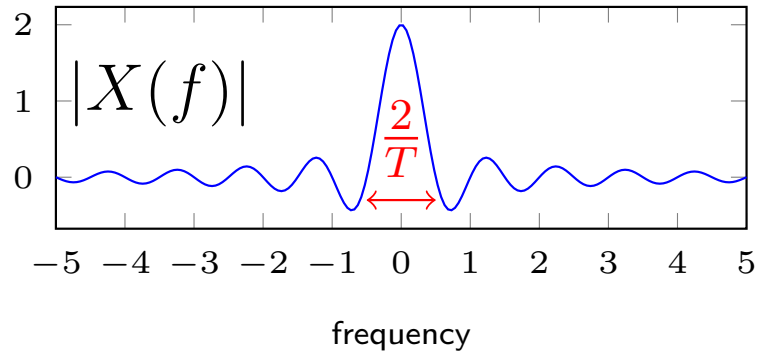




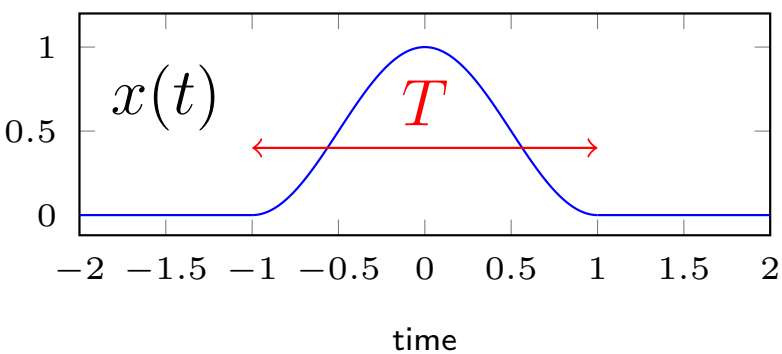
Effect of Windowing

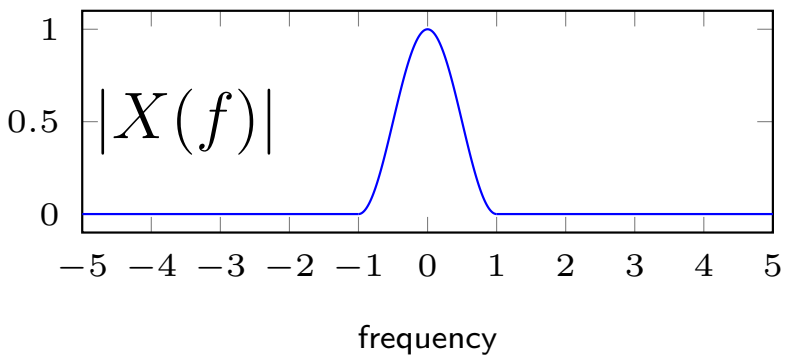
Square window



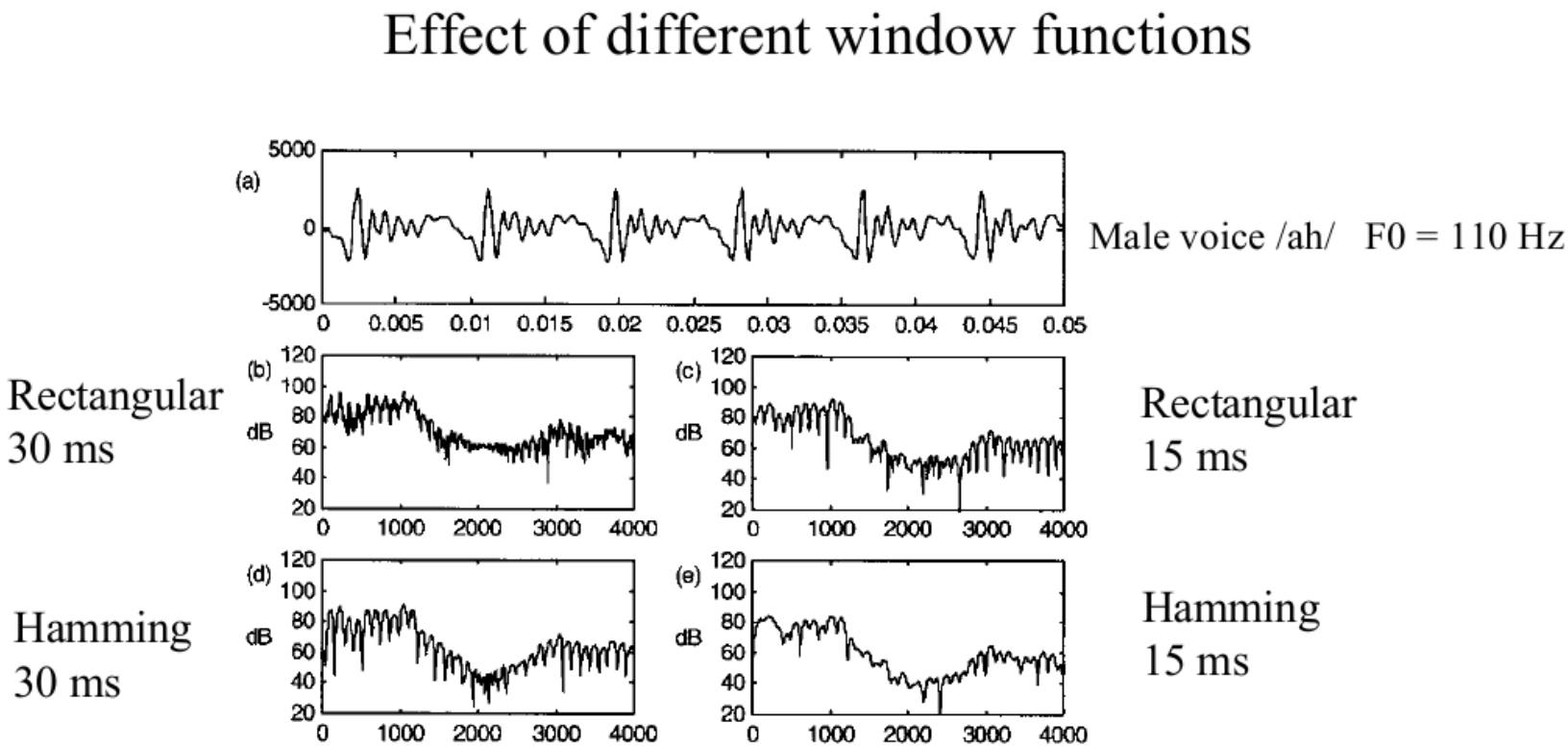


Hamming window





Effect of Windowing on Speech



Window should be long enough to cover 2 pitch pulses
Short enough to capture short events and transitions

Windowing, typical values

- signal sampling frequency: 8–20kHz
- analysis window: 10–50ms
- frame step: 10–25ms (100–40Hz)

Discrete Fourier Transform

Fourier transform of continuous signals

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt$$

Fourier transform of discrete signals

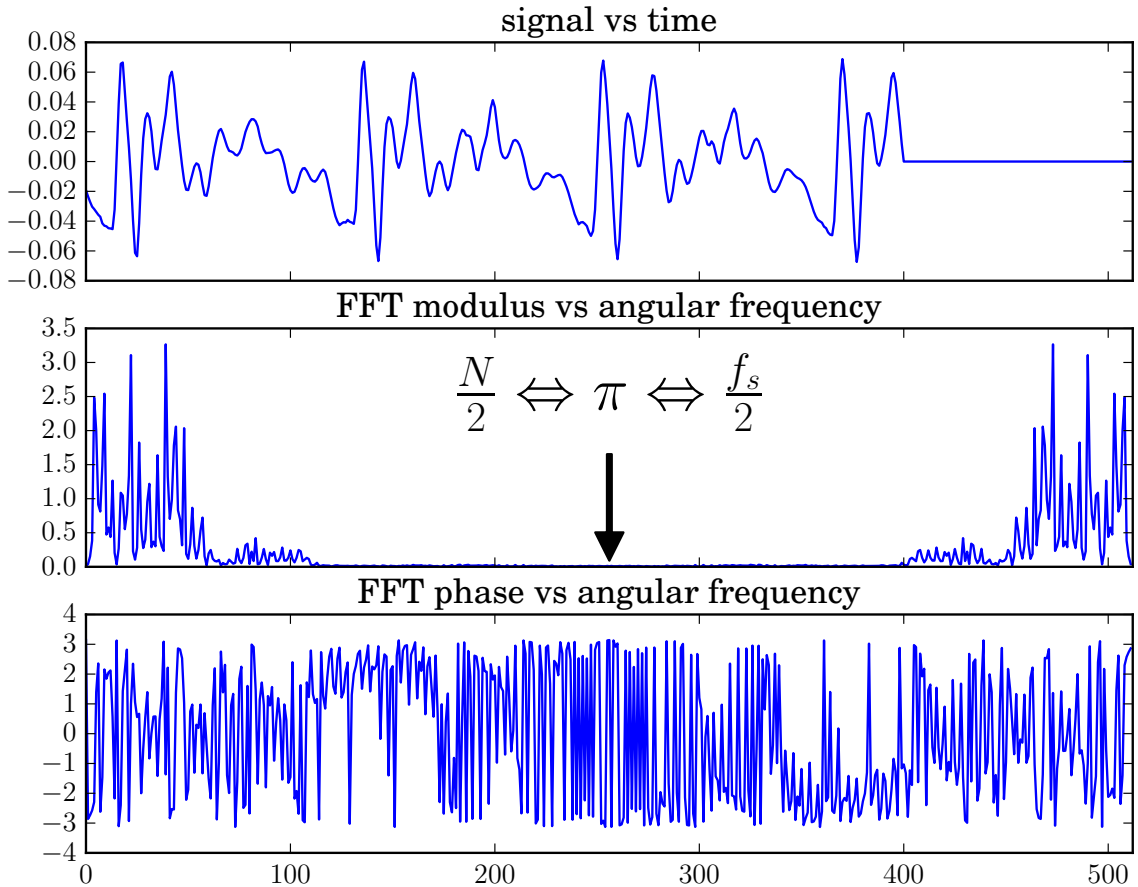
$$X(\omega) = \sum_{k=-\infty}^{\infty} x[k] e^{-j\omega k}$$

Discrete Fourier Transform (and Fast Fourier Transform)

$$X[n] = \sum_{k=0}^{N-1} x[k] e^{-j2\pi \frac{n}{N} k}$$

Discrete Fourier Transform (DFT) ($N = 512$)

$$X[n] = \sum_{k=0}^{N-1} x[k] e^{-j2\pi \frac{n}{N} k}$$



Feature Extraction Steps (Lab 1)