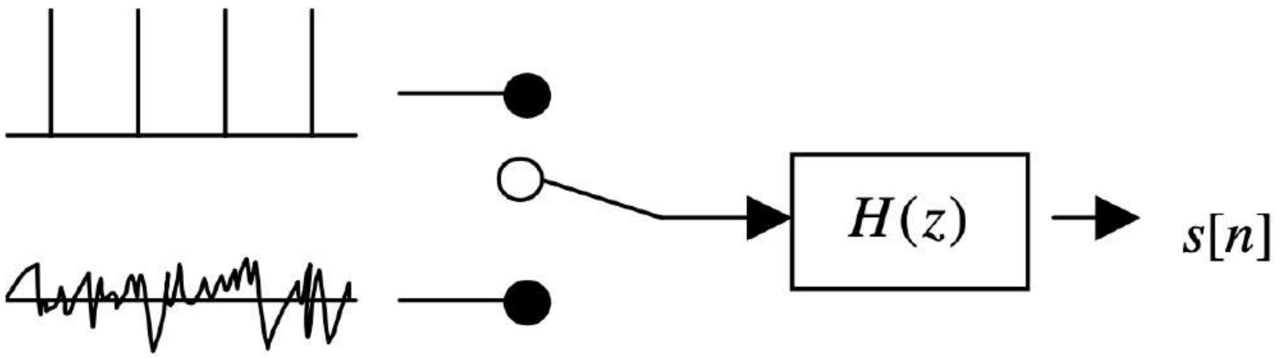
Model of Speech Production

Luis Caldas de Oliveira

Source-Filter Model

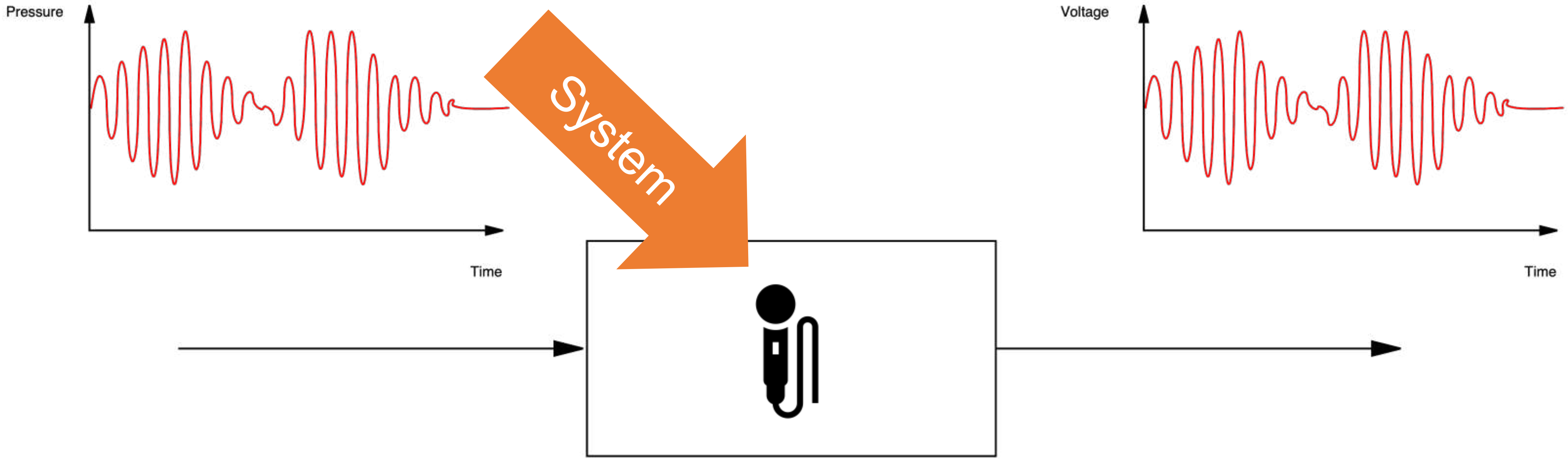
a mathematical model that represents the speech signal by a combination of a sound source with a linear acoustic filter



Systems

Microphone

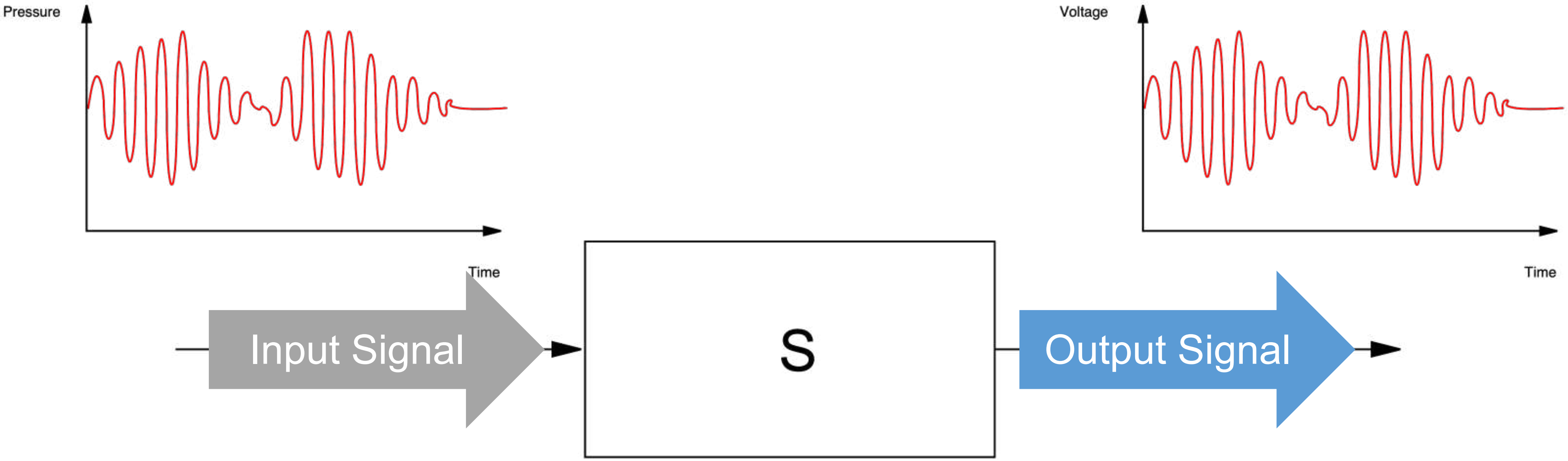
A microphone is a continuous-time system that converts an acoustic signal into an electrical signal



$$S : [Time \rightarrow Pressure] \rightarrow [Time \rightarrow Voltage]$$

System

A system can be thought of as a process that takes an input signal and produces an output signal.

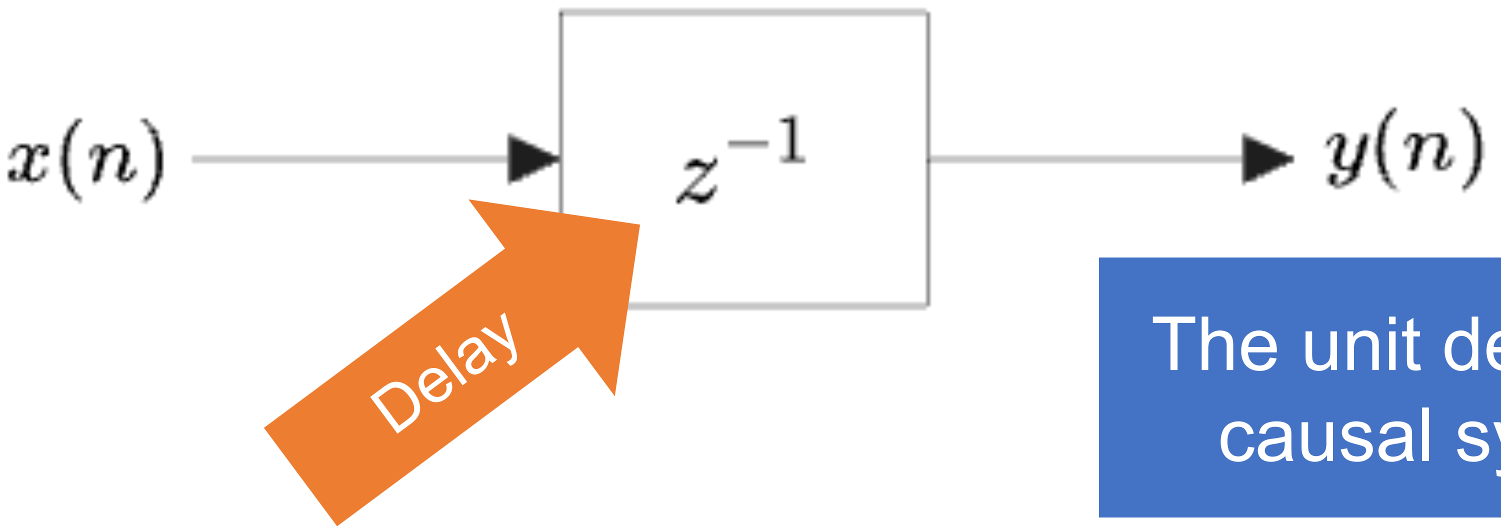


$$S : [Time \to Pressure] \to [Time \to Voltage]$$

Unit Delay

A unit delay is a system that delays the input signal by one sample period:

$$y(n) = x(n-1)$$



The unit delay is a causal system

Linear Time-Invariant Systems

Linear System

A **linear system** must simultaneously verify the properties of **additivity** and **homogeneity**.

$$\begin{array}{l} S(x_1(n)) = y_1(n) \\ S(x_2(n)) = y_2(n) \end{array} \xrightarrow[additivity]{} S(x_1(n) + x_2(n)) = y_1(n) + y_2(n)$$

$$S(x(n)) = y(n) \xrightarrow[homogeneity]{} S(ax(n)) = ay(n), a \in \mathbb{R}$$

If we know the response of the system to some signals, we can compute the response to any linear combination of those signals

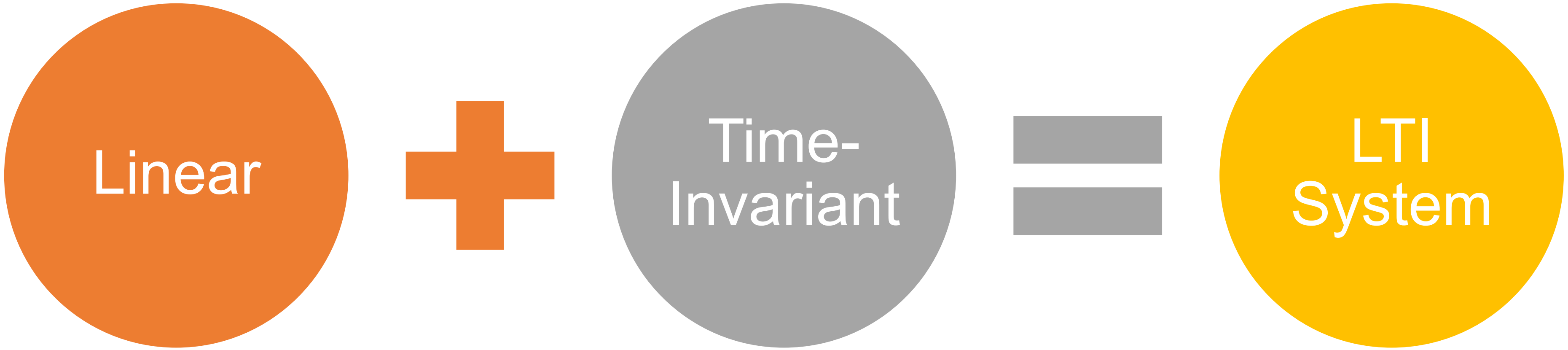
Time-Invariant System

A system is **time-invariant** if a time shift in the input signal results in an equal time shift in the output signal.

$$S(x(n)) = y(n) \xrightarrow[\text{time-invariant}]{} S(x(n - n_0)) = y(n - n_0)$$

If we know the response of the system to a signal, we can compute the response to any signal that is a time shift of that signal

Linear Time-Invariant (LTI) System

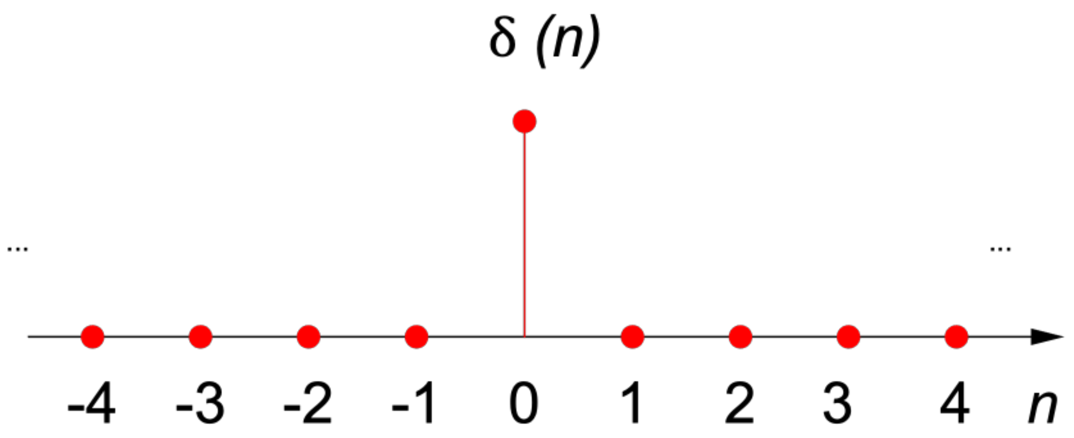


A linear time-invariant (LTI) system is both a linear system and a time-invariant system.

Discrete-Time Unit Impulse

A discrete-time signal that has a value of 1 at time n=0, and 0 everywhere else.

$$\delta(n) = \begin{cases} 0, & n \neq 0. \\ 1, & n = 0. \end{cases}$$



Impulse Response

The **discrete-time impulse response** is the output of a discrete-time system when the input signal is a discrete-time unit impulse

$$h(n) = S(\delta(n))$$

h(n) is the impulse response signal

Unit Impulse

Convolution Sum

Any signal can be represented as a sum of unit impulses

$$x(n) = \sum_{k=-\infty}^{+\infty} x(k)\delta(n-k)$$

We can compute the response of LTI system if we know its impulse response:

Impulse response

$$y(n) = x(n) * h(n) = \sum_{k=-\infty}^{+\infty} x(k)h(n-k)$$

Convolution sum

Difference Equation

A **causal** discrete-time LTI system can be defined by a difference equation:

$$\sum_{k=0}^{N} a_k y(n-k) = \sum_{k=0}^{M} b_k x(n-k)$$

current and previous samples of the output signal

current and previous sample of the input signal

The difference equation is the discrete-time equivalent to the continuous-time differential equation

Problem

Consider the discrete-time LTI system described by the following difference equation:

$$y(n) = 2x(n) - \frac{1}{2}y(n-1)$$

where the discrete-time signals x(n) and y(n) are the input and output of the system.

Find the discrete-time impulse response of the LTI system.

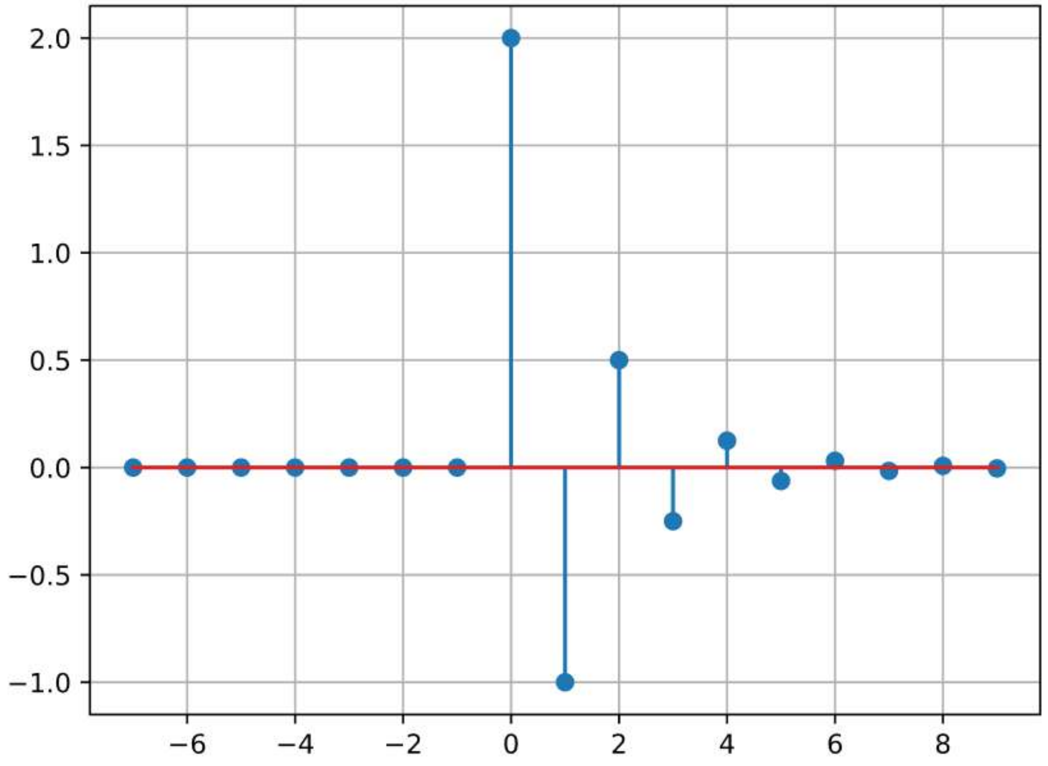
Solution

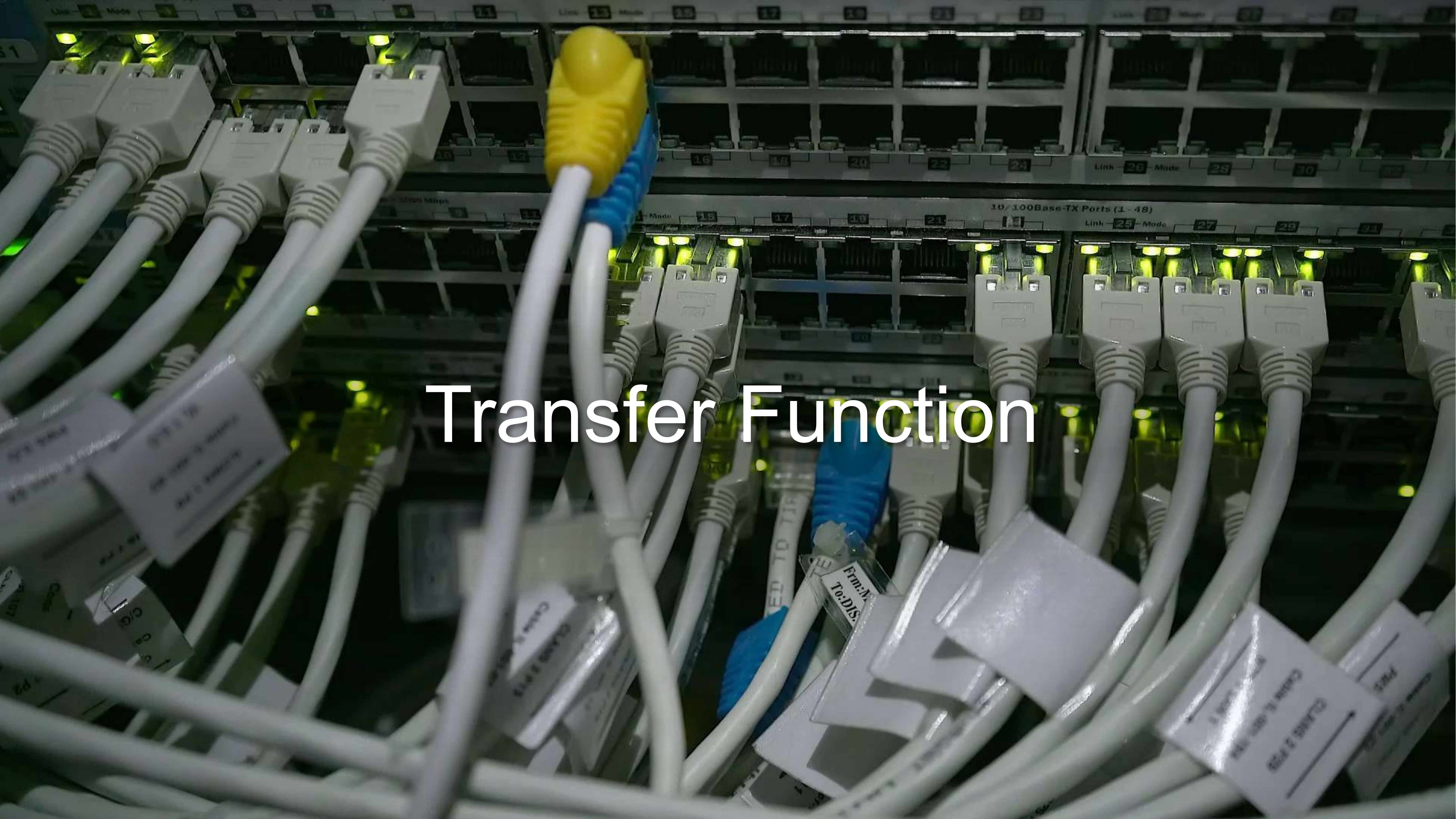
Impulse response

$$h(n) = 2\delta(n) - \frac{1}{2}h(n-1)$$

Iterative solution

$$\begin{aligned} h(-1) &= 0 \\ h(0) &= 2 \\ h(1) &= -1 \\ h(2) &= \frac{1}{2} \\ \ldots &= \ldots \\ h(n) &= (-1)^n \left(\frac{1}{2}\right)^{n-1} \end{aligned}$$





Transfer Function

Complex Exponential Response

z is a complex number

$$x(n) = z^n$$

Convolution is commutative

$$y(n) = \sum_{k=-\infty}^{+\infty} h(k) z^{(n-k)}$$

$$y(n) = z^n \underbrace{\sum_{k=-\infty}^{+\infty} h(k) z^{-k}}_{H(z)}$$

$$y(n) = H(z) z^n$$

Transfer function

Eigenfunction of an LTI system

The discrete-time complex exponential signal is an eigenfunction of a discrete-time LTI system.

$$x(n) = z^n \to y(n) = H(z)z^n$$

$$y(n) = z^n \underbrace{\sum_{k=-\infty}^{+\infty} h(k) z^{-k}}_{H(z)}$$

Transfer Function

$$x(n) = z^n \rightarrow y(n) = H(z)z^n$$

The transfer function of an LTI system, H(z), is the complex amplitude of the output signal when the input is the complex exponential signal

$$H(z) = \sum_{n=-\infty}^{+\infty} h(n) z^{-n}$$

Frequency Response

$$\tilde{x}(n) = \frac{1}{N}\sum_{k=0}^{N-1} \tilde{X}(k) e^{j\frac{2\pi}{N}kn}$$

The frequency response is a measure of how a system responds to different frequencies of input signals.

$$\tilde{y}(n) = \frac{1}{N}\sum_{k=0}^{N-1} \underbrace{\tilde{X}(k) H\left(e^{j\frac{2\pi}{N}k}\right)}_{\tilde{Y}(k)} e^{j\frac{2\pi}{N}kn}$$

Frequency Response

$$H(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} h(n) e^{-j\omega n}$$

$$z = e^{j\omega}$$

Z-Transform

The z-transform is defined as

z is any complex number

$$X(z) = \sum_{n=-\infty}^{+\infty} x(n) z^{-n}$$

The transfer function is the z-transform of the impulse response:

$$H(z) = \sum_{n=-\infty}^{+\infty} h(n) z^{-n}$$

Time-Shift Property

Time-shift of a discrete-time signal:

$$x(n-n_0) \underset{Z}{\rightarrow} z^{-n_0}X(z)$$

$$\begin{aligned}
Y(z) &= \sum_{n=-\infty}^{+\infty} y(n)z^{-n} \\
&= \sum_{n=-\infty}^{+\infty} x(n-n_0)z^{-n} \\
&= \sum_{m=-\infty}^{+\infty} x(m)z^{-(m+n_0)} \\
&= z^{-n_0} \sum_{m=-\infty}^{+\infty} x(m)z^{-m} \\
&= z^{-n_0}X(z)
\end{aligned}$$

Convolution Property

The z-transform of the convolution is the product of the z-transforms of the signals:

$$
\begin{aligned}
Y(z) &= \sum_{n=-\infty}^{+\infty} y(n) z^{-n} \\
&= \sum_{n=-\infty}^{+\infty} \sum_{k=-\infty}^{+\infty} x(n) h(n-k) z^{-n} \\
&= \sum_{n=-\infty}^{+\infty} x(n) \sum_{k=-\infty}^{+\infty} h(n-k) z^{-n} \\
&= \sum_{n=-\infty}^{+\infty} x(n) \sum_{m=-\infty}^{+\infty} h(m) z^{-m} z^{-n} \\
&= \sum_{n=-\infty}^{+\infty} x(n) z^{-n} \sum_{m=-\infty}^{+\infty} h(m) z^{-m} \\
&= X(z) H(z)
\end{aligned}
$$

$$y(n) = x(n) * h(n) \underset{Z}{\rightarrow} Y(z) = H(z)X(z)$$

Rational Transfer Function

$$\sum_{k=0}^{N} a_k y(n-k) = \sum_{k=0}^{M} b_k x(n-k)$$

Time-shift property

$$Y(z)\sum_{k=0}^{N} a_k z^{-k} = X(z)\sum_{k=0}^{M} b_k z^{-k}$$

Convolution property

$$H(z) = \frac{Y(z)}{X(z)}$$

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}} = \frac{P(z)}{Q(z)}$$

Quotient of polynomials in z

Poles and Zeros of the Transfer Function

Given a rational transfer function:

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{\sum_{k=0}^{N} a_k z^{-k}} = \frac{P(z)}{Q(z)}$$

The roots of P(z) are called **zeros** of the transfer function H(z)

The roots of Q(z) are called **poles** of the transfer function H(z)


Filtering

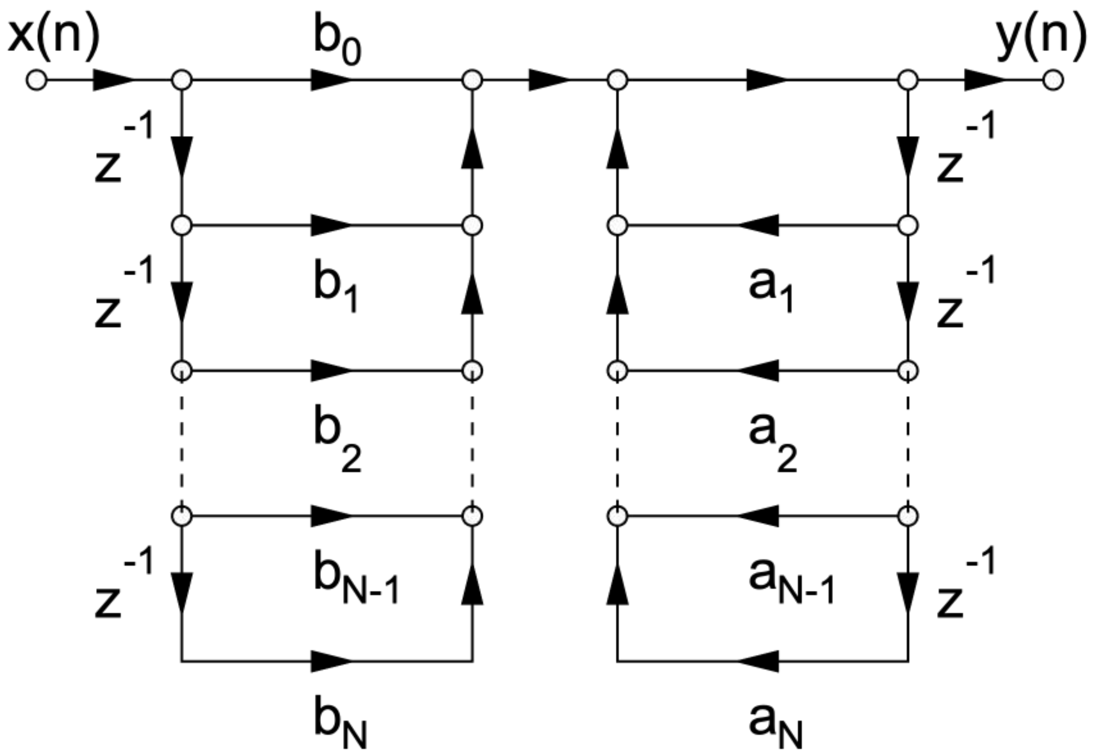
Infinite Impulse Response (IIR)

A type of LTI system where the output depends both on a finite number of input and output samples in the form of a difference equation.

$$y(n) = \frac{1}{a_0}\left(\sum_{k=0}^{M} b_k x(n-k) - \sum_{k=1}^{N} a_k y(n-k)\right)$$

often 1

scipy.signal.lfilter() implements this equation



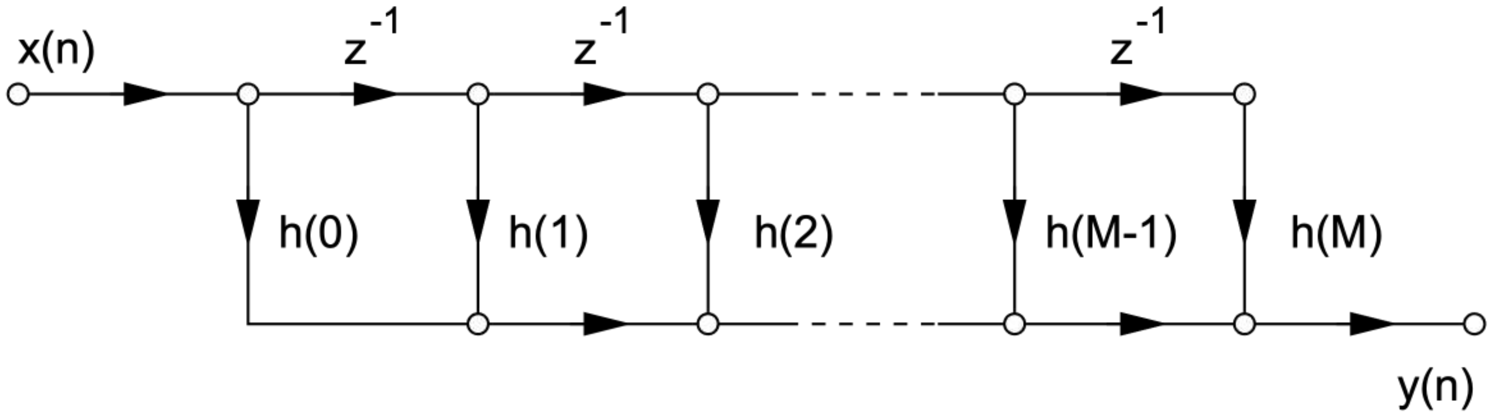
Finite Impulse Response (FIR)

A type of LTI system where the output depends only on a finite number of input samples

$$y(n) = \sum_{k=0}^{M} b_k x(n-k)$$

$$h(n) = b_n$$

Impulse response



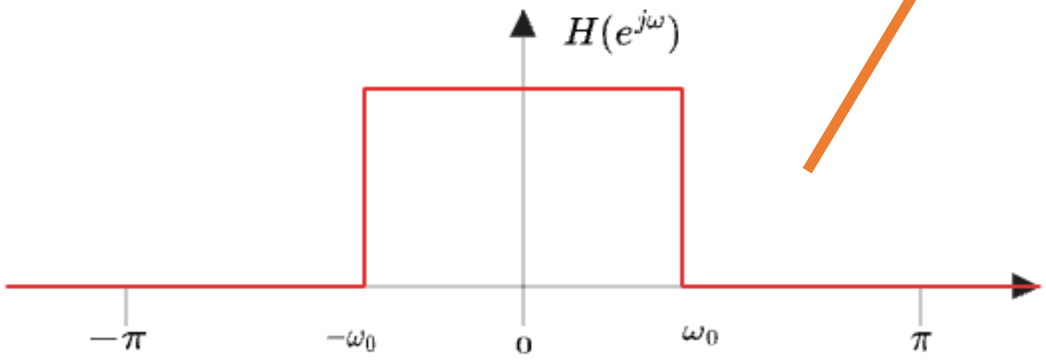
Ideal Low-Pass Filter

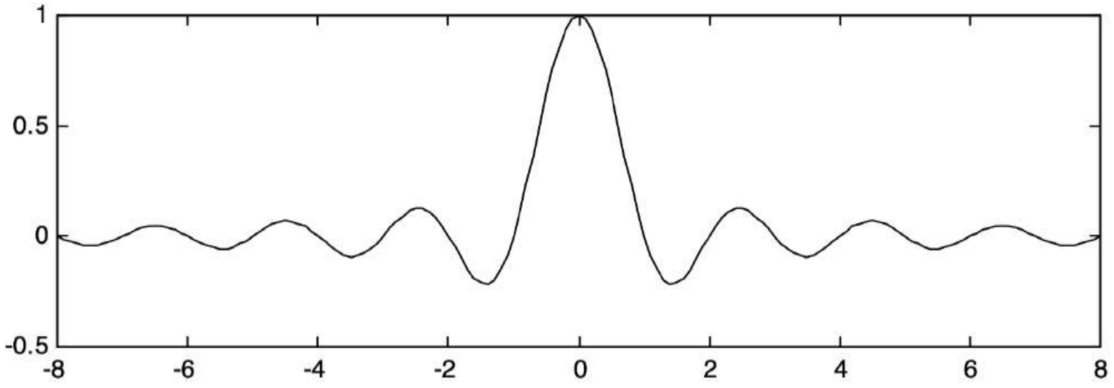
A type of filter that passes all frequency components of a signal below the cutoff frequency and blocks all frequency components above that.

$$H(e^{j\omega}) = \begin{cases} 1, & |\omega| < \omega_0 \\ 0, & \omega_0 < |\omega| < \pi \end{cases}$$

$$\begin{aligned} h(n) &= \frac{1}{2\pi}\int_{-\omega_0}^{\omega_0} e^{j\omega n}\, d\omega \\ &= \frac{e^{j\omega_0 n} - e^{-j\omega n}}{2\pi j n} \\ &= \frac{\sin(\omega_0 n)}{\pi n} \end{aligned}$$

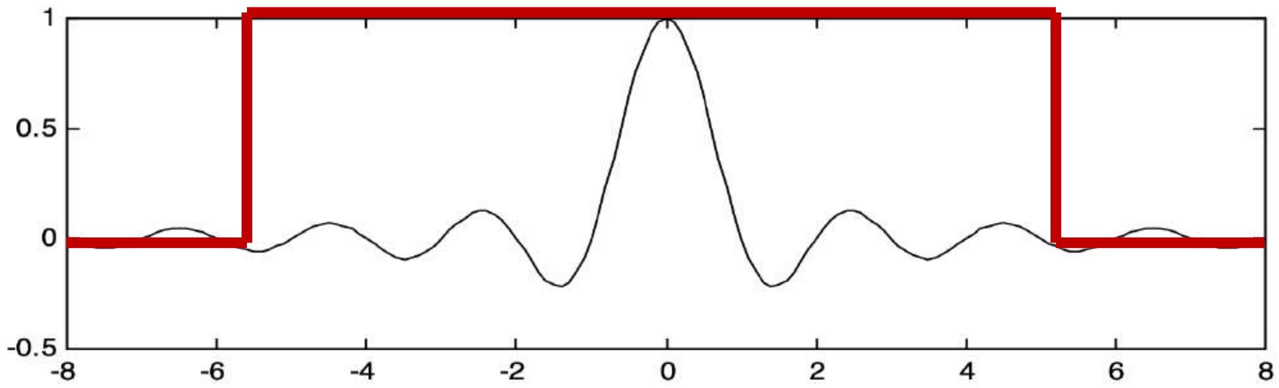
Infinite impulse response

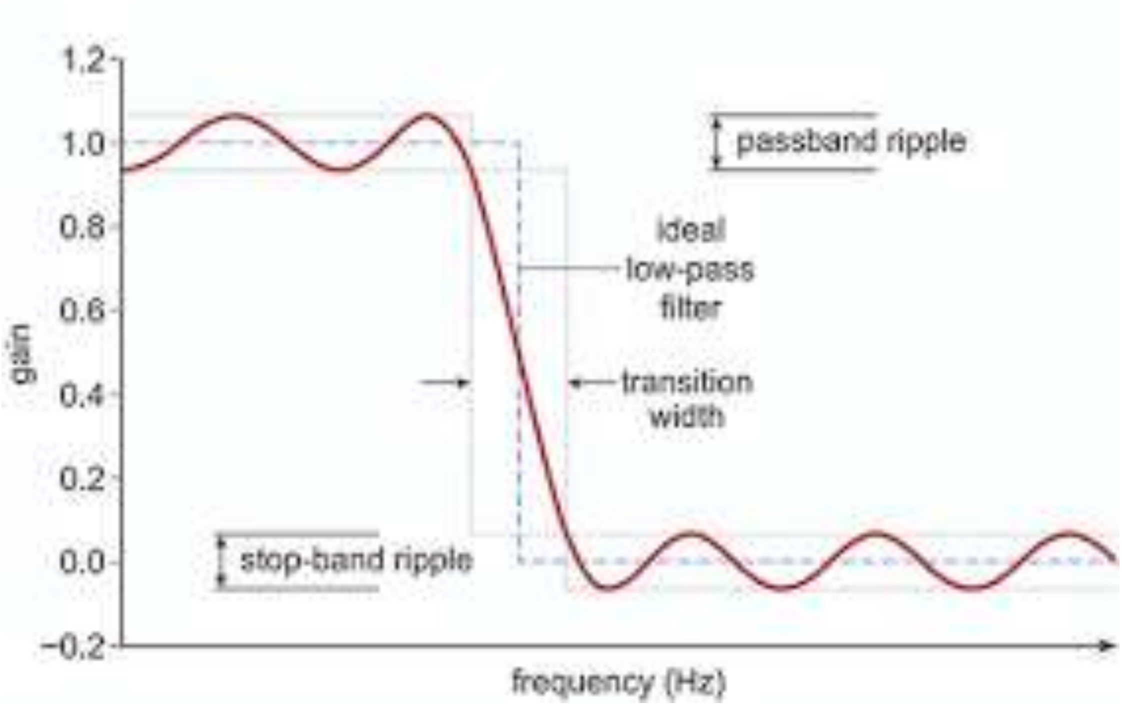




Impulse Response Windowing

Provides an approximation of the ideal low-pass filter that depends on the size and type of window





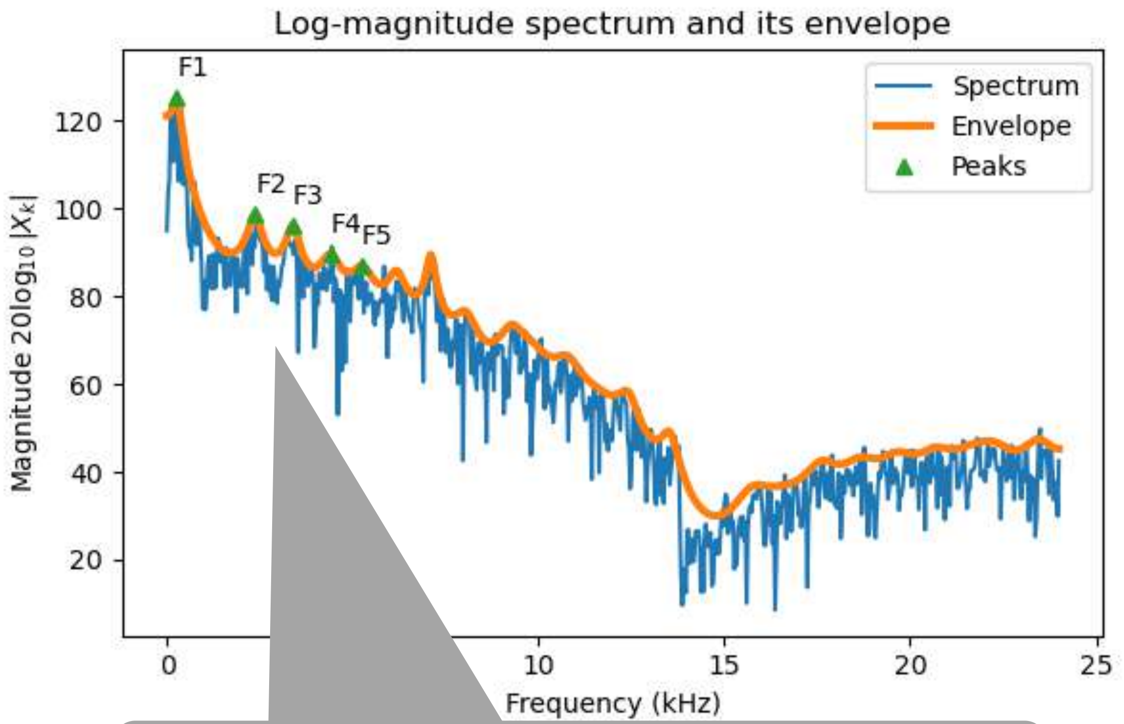
Window Spectral Features

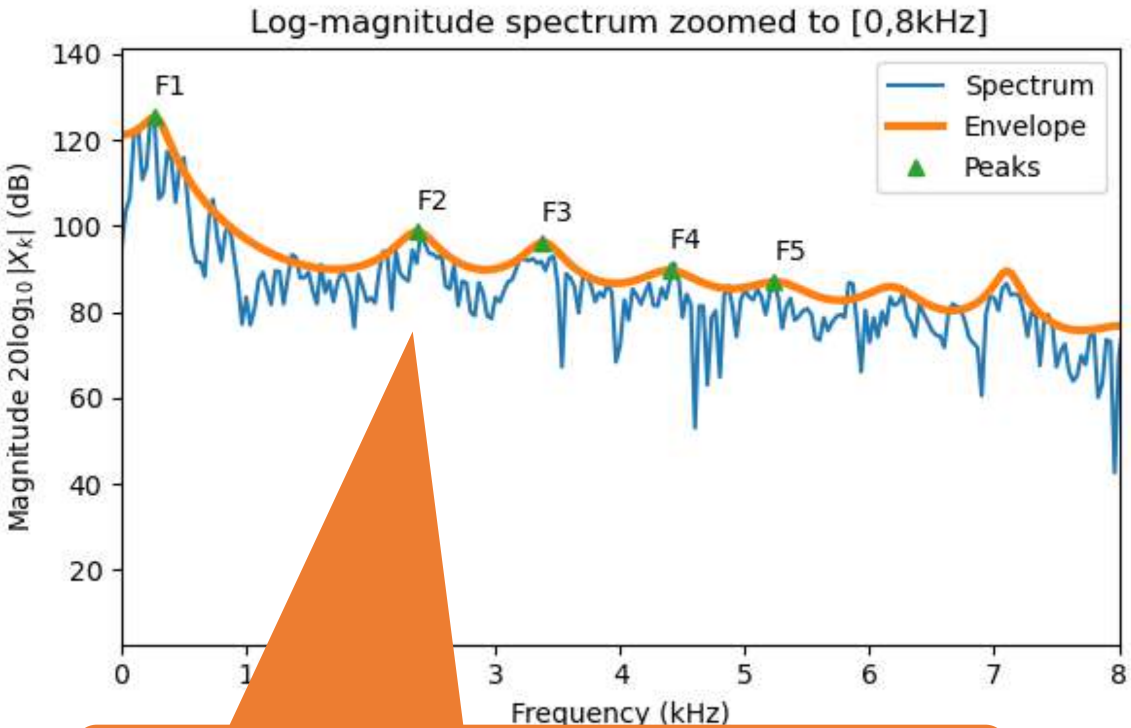
Type of Window	Side Lobe Amplitude	Width of Main Lobe	Transition Width	Passband Ripple	Stopband Attenuation
Rectangular	$-13dB$	$4\pi/M$	$0.9/(MT)$	$0.7416dB$	$> 21dB$
Hanning	$-31dB$	$8\pi/M$	$3.1/(MT)$	$0.0546dB$	$> 44dB$
Hamming	$-41dB$	$8\pi/M$	$3.3/(MT)$	$0.0194dB$	$> 53dB$
Blackman	$-74dB$	$12\pi/M$	$5.5/(MT)$	$0.0274dB$	$> 74dB$

Acoustical Model

Formant

A formant is a resonance in the vocal tract that results in a peak of energy in the speech signal at a particular frequency.

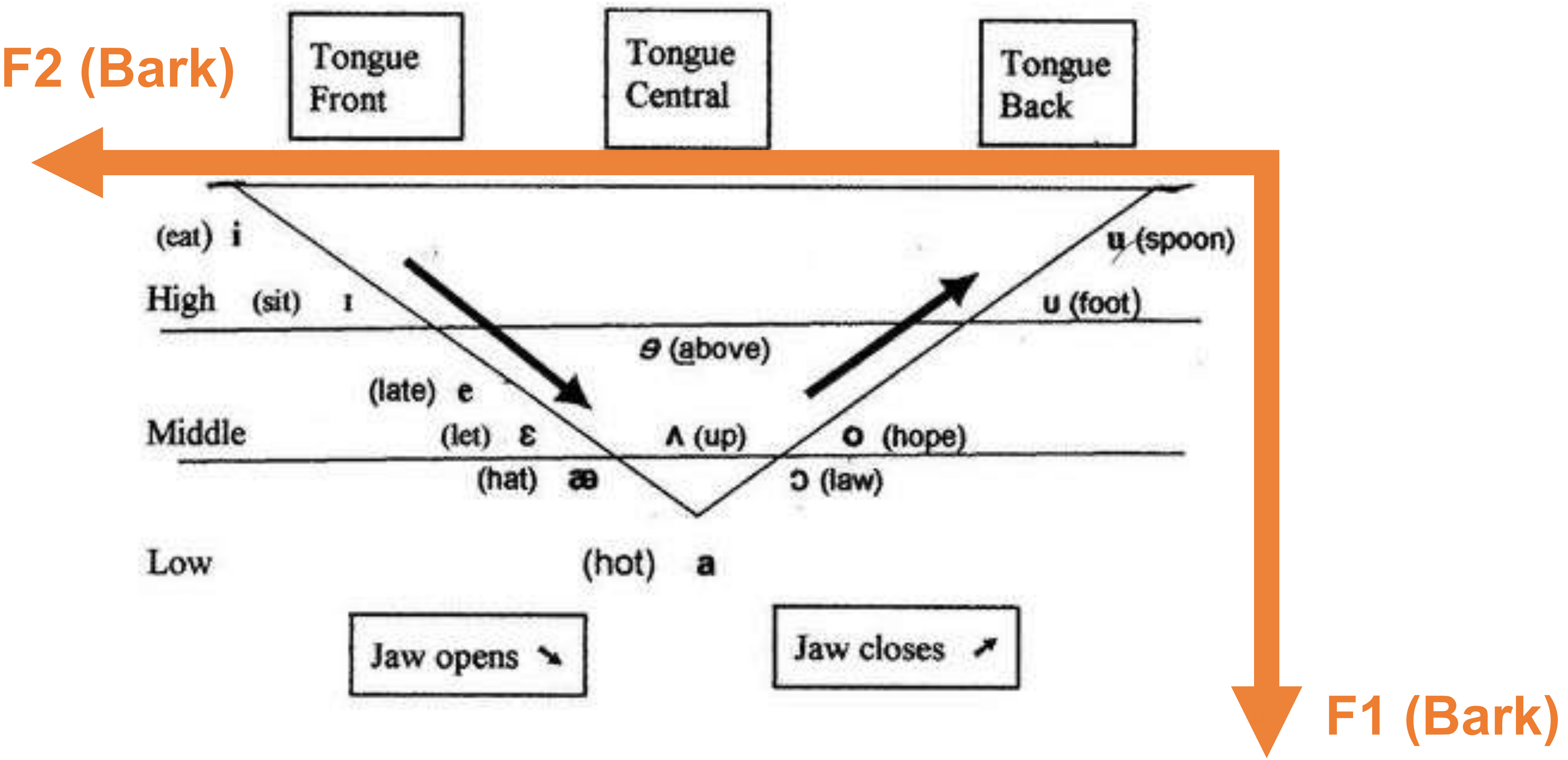




Formants are numbered from lower to higher frequencies

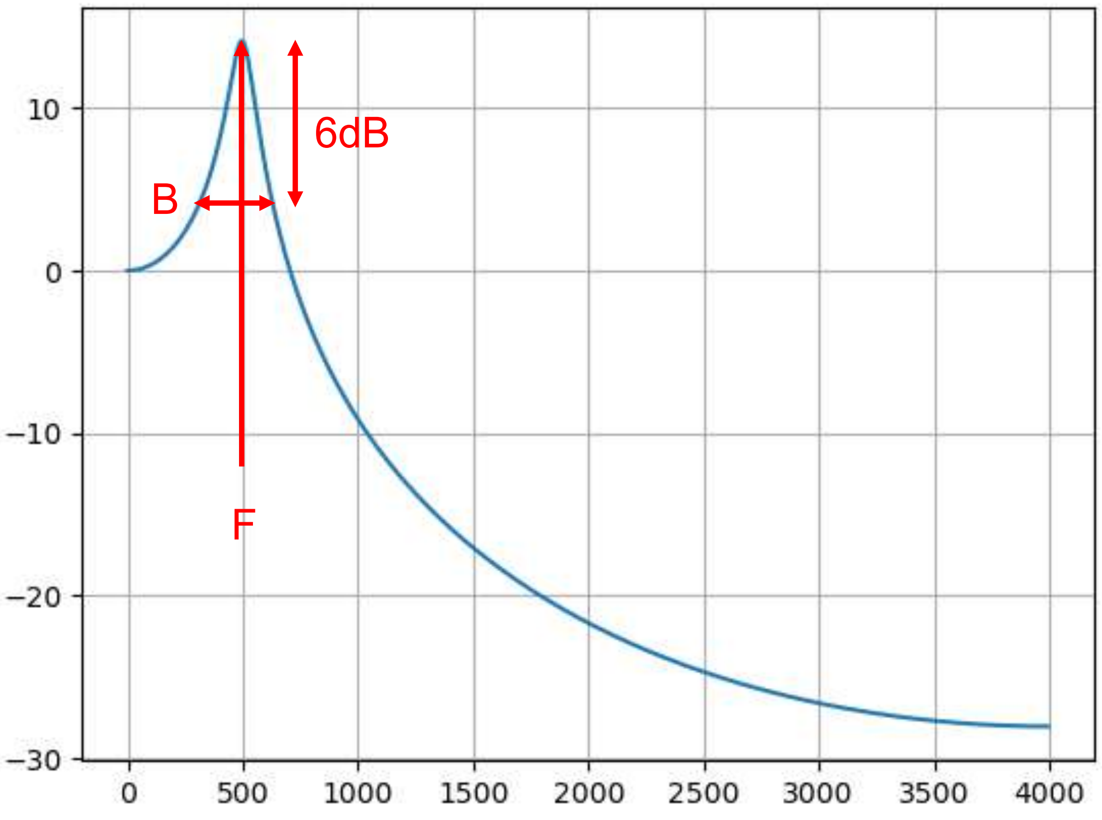
The frequency of the formants uniquely identifies a vowel

Vowel Triangle (Articulation)



Continuous-Time Resonator

A system that models a resonance with two poles and no zeros in the transfer function



$$\frac{d^2y(t)}{dt^2} + 2\zeta\omega_n \frac{dy(t)}{dt} + \omega_n^2 y(t) = \omega_n^2 x(t)$$

$$H(s) = \frac{\omega_n^2}{s^2 + 2\zeta\omega_n s + \omega_n^2} = \frac{\omega_n^2}{(s-c_1)(s-c_2)}$$

Laplace transform

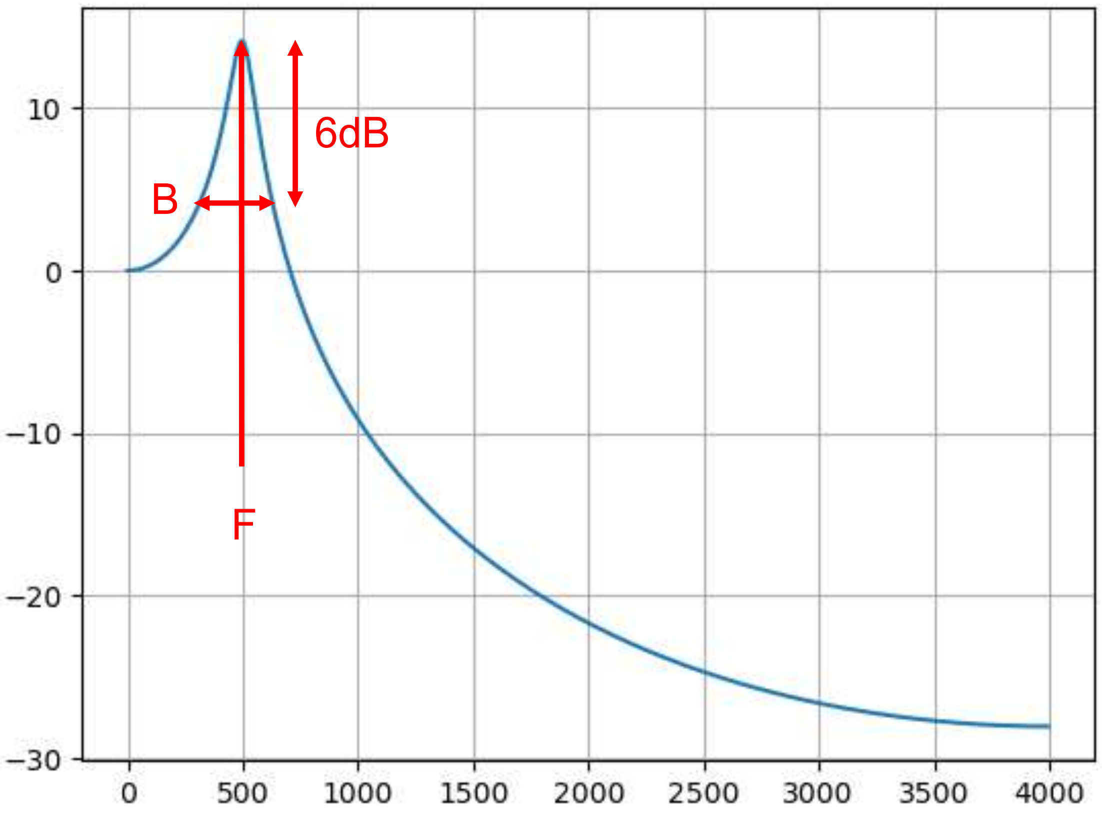
$$\omega_n = 2\pi F \qquad \zeta = \frac{\pi B}{\omega_n}$$

$$c_1 = -\zeta\omega_n + \omega_n\sqrt{\zeta^2 - 1}$$
$$c_2 = -\zeta\omega_n - \omega_n\sqrt{\zeta^2 - 1}$$

poles of the transfer function

Second Order All-Pole IIR System

A discrete-time system that has two poles and no zeros in the transfer function



$$y(n) = (a_1 + a_2)x(n) + a_1 y(n-1) + a_2 y(n-2)$$

$$H(z) = \frac{a_1 + a_2}{1 - a_1 z^{-1} - a_2 z^{-2}} = \frac{a_1 + a_2}{(1 - p_1 z^{-1})(1 - p_2 z^{-1})}$$

z-transform

$$\omega_n = 2\pi F \qquad \zeta = \frac{\pi B}{\omega_n}$$

$$p_1 = e^{c_1/f_s} = e^{(-\zeta\omega_n + \omega_n\sqrt{\zeta^2 - 1})/f_s}$$

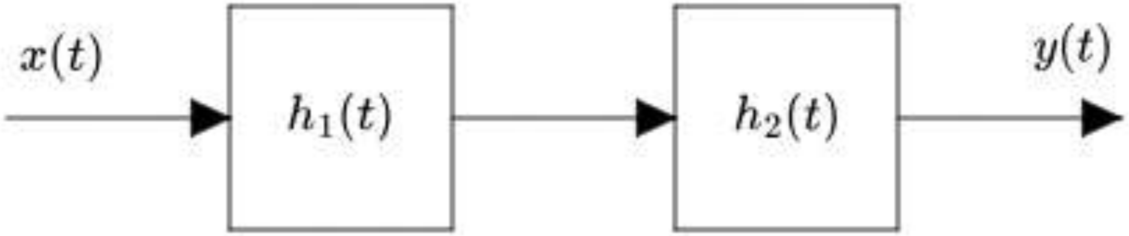
$$p_2 = e^{c_2/f_s} = e^{(-\zeta\omega_n - \omega_n\sqrt{\zeta^2 - 1})/f_s}$$

poles of the transfer function

sampling frequency

Cascade Combination

The cascade combination of two systems means that the output of the first system is fed as input to the second system.



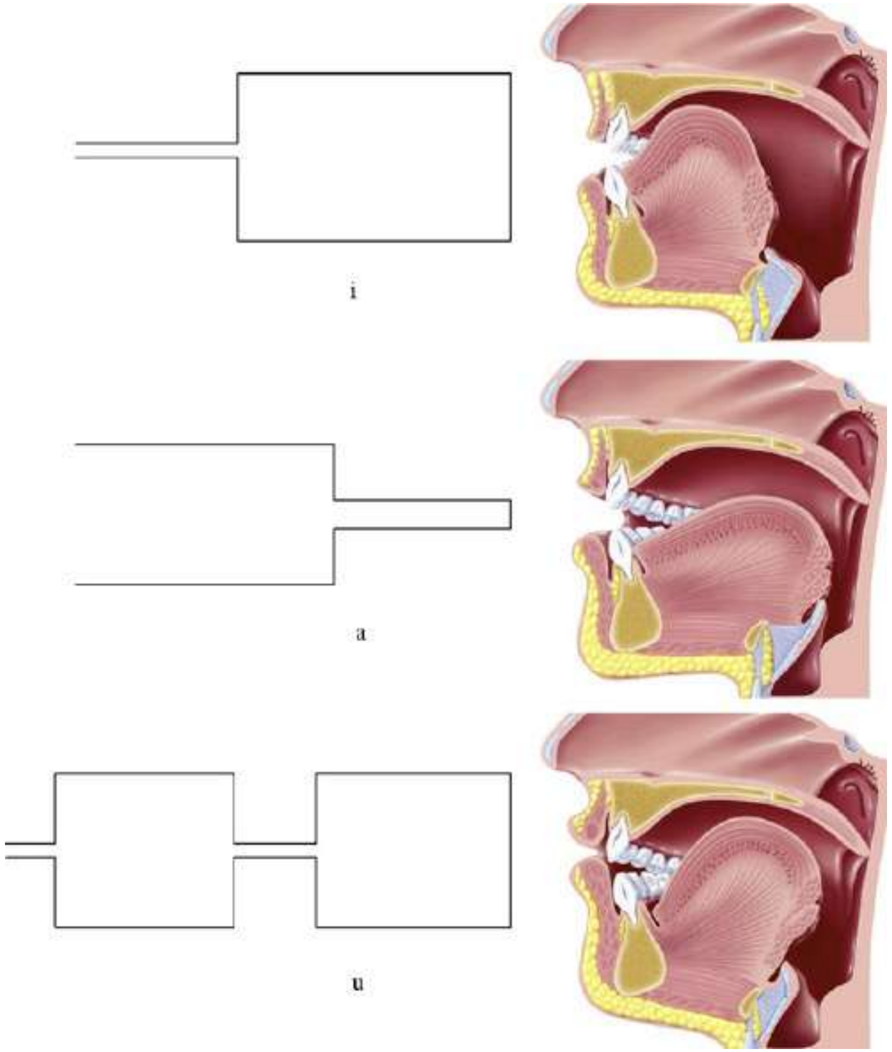
$$h(t) = h_1(t) * h_2(t)$$

Convolution of the impulse responses

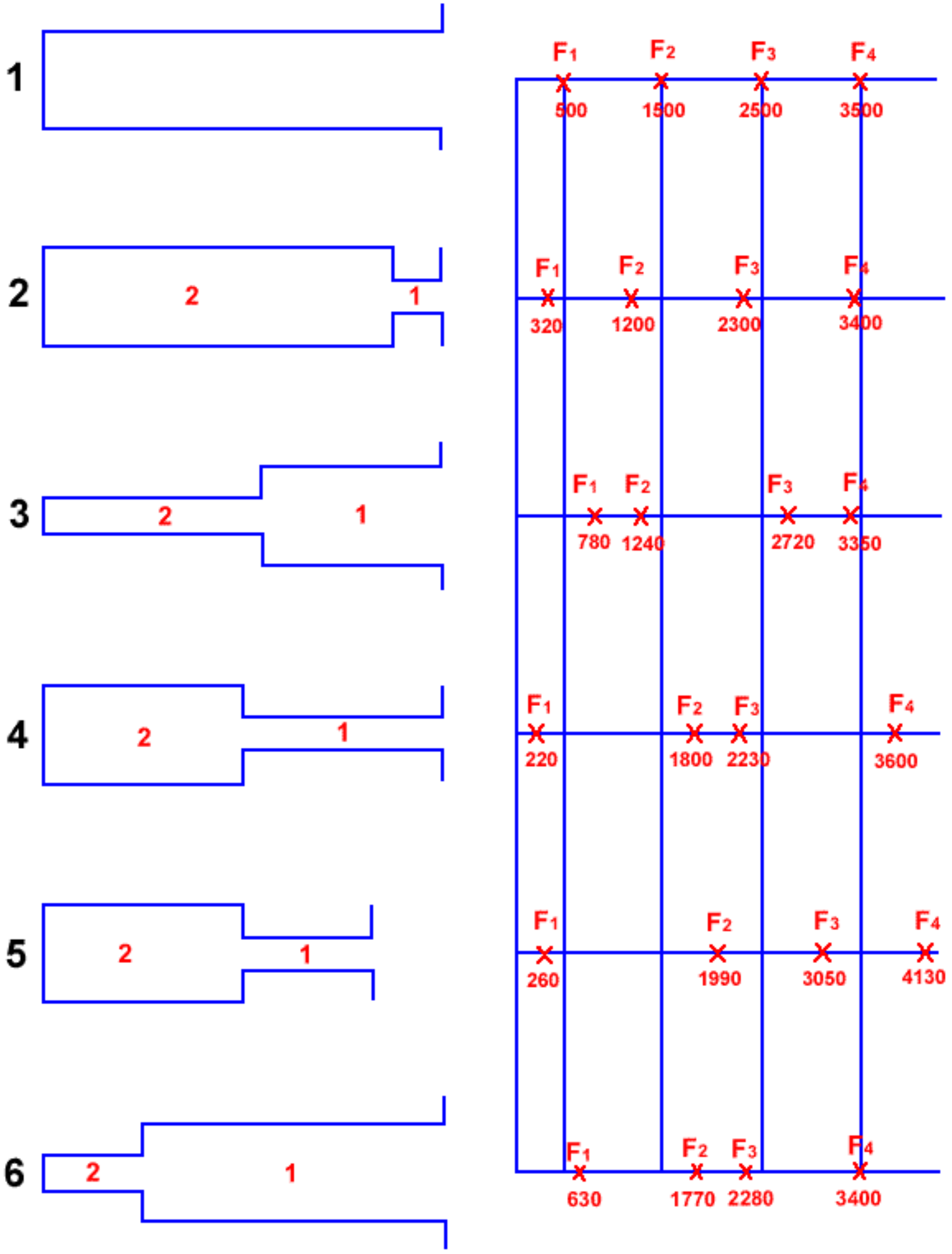
$$H(z) = H_1(z)H_2(z)$$

Product of the transfer functions

Multiple tubes



Louis-Jean Boe and Sophie Jacopin



Dep Liguistics, Macquarie Univ

Linear Prediction

Linear Prediction

Tries to predict the value of signal sample s(n) using a linear combination of the signal's past samples

linear prediction order

$$\hat{s}(n) = \sum_{k=1}^{P} a_k s(n-k)$$

predicted sample

linear prediction coefficients

Prediction Error or Residue

the difference between the real and the predicted value for the sample

actual sample

predicted sample

prediction error

$$e(n) = s(n) - \sum_{k=1}^{P} a_k s(n-k)$$

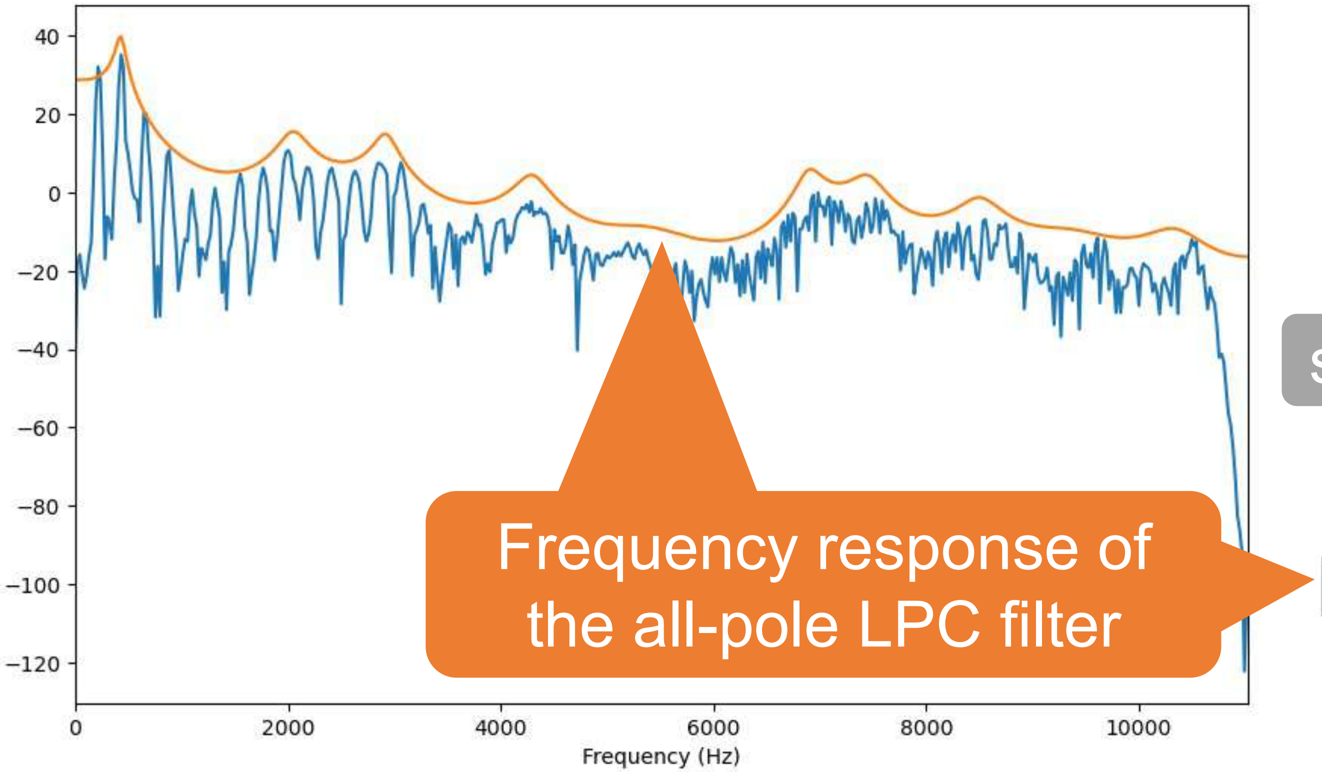
applying the z-transform results in an all-zero transfer function:

$$A(z) = \frac{E(z)}{S(z)} = 1 - \sum_{k=1}^{P} a_k z^{-k}$$

polynomial in z

Linear Predictive Synthesis

The inverse filter has an all-pole transfer function that can be seen as similar to the vocal tract transfer function.



source signal

$$S(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} E(z)$$

speech signal

vocal tract filter

$$|H(e^{j\omega})|_{dB} = 20 \log \left| \frac{1}{1 - \sum_{k=1}^{P} a_k e^{-j\omega k}} \right|$$

Linear Prediction Optimization

$$a_k = \operatorname{argmin} \sum_{n=-\infty}^{+\infty} (e(n))^2$$

prediction error energy

Minimize the energy of the prediction error

$$\mathcal{E} = \sum_{n=-\infty}^{+\infty} \left(s(n) - \sum_{k=1}^{P} a_k s(n-k) \right)^2$$

$$\frac{d\mathcal{E}}{da_k} = 2 \sum_{n=-\infty}^{+\infty} \left(s(n) - \sum_{k=1}^{P} a_k s(n-k) \right) s(n-k) = 0$$

system of P equations

System of Equations

The optimization process results in a set of P equations to determine the P linear prediction coefficients.

$$\sum_{n=-\infty}^{+\infty} s(n-i)s(n) = \sum_{k=1}^{P} a_k \sum_{n=-\infty}^{+\infty} s(n-i)s(n-k), \; i = 1, \dots, P$$

P coefficients

$$\sum_{k=1}^{P} a_k \phi(i,k) = \phi(i,0) \qquad \mathbf{Ra} = \gamma$$

$$\phi(i,k) = \sum_{n=-\infty}^{+\infty} s(n-i)s(n-k)$$

Yule-Walker equations

PxP

Px1

Autocorrelation Method

Assumes that the input signal is finite-duration signal resulting from a windowing process

window

$$s_m(n) = s(m+n)w(n)$$

sum starts at n=0

$$\mathcal{E}_m = \sum_{n=0}^{N-P-1} \left(s_m(n) - \sum_{k=1}^{P} a_k s_m(n-k) \right)^2$$

$$\phi(i,k) = \phi(k,i) = R_m(|k-i|)$$

autocorrelation function

$$R_m(l) = \sum_{n=-\infty}^{+\infty} s_m(n) s_m(n-l)$$

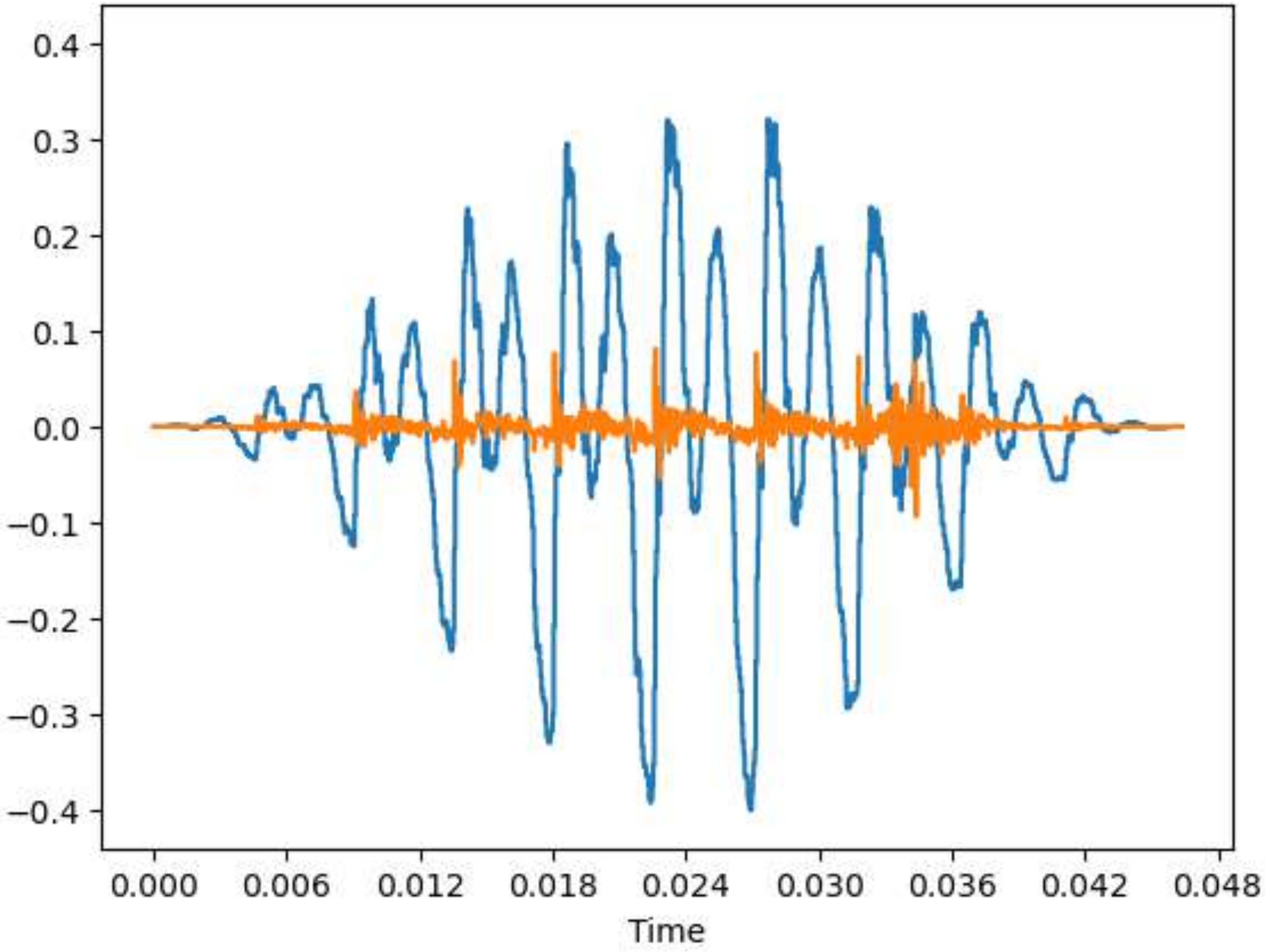
can be solved with the Levinson-Durbin recursion

$$\begin{pmatrix} R_m[0] & R_m[1] & R_m[2] & \cdots & R_m[p-1] \\ R_m[1] & R_m[0] & R_m[1] & \cdots & R_m[p-2] \\ R_m[2] & R_m[1] & R_m[0] & \cdots & R_m[p-3] \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ R_m[p-1] & R_m[p-2] & R_m[p-3] & \cdots & R_m[0] \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_p \end{pmatrix} = \begin{pmatrix} R_m[1] \\ R_m[2] \\ R_m[3] \\ \cdots \\ R_m[p] \end{pmatrix}$$

R is a Toeplitz matrix

Residue for Voiced Speech

The linear prediction error approximates an impulse train



Summary

System Modelling

- Source-filter model, system, causal system

LTI Systems

- impulse response, difference equation

Transfer Function

- z-transform, rational transfer function

Summary (cont.)

Filtering

- IIR and FIR systems

Acoustic Model

- Formant, resonator, cascade combination

Linear Prediction

- Residue, LPC coefficients, autocorrelation method

Obrigado