

Speech Processing Thursday

Lecture at 9am



Scan this QR code and fill
in the form to mark your
attendance

Speech Processing: The Source-Filter Model

Module 4
Catherine Lai
12 Oct 2023

Last time

Digital Speech Signals

- Sampling, quantization
- Sampling rate, Nyquist frequency, aliasing

Discrete Fourier Transform (DFT)

- Time domain to frequency domain
- Magnitude Spectrum, phase spectrum
- Time resolution vs frequency resolution tradeoff (input size)
- Role of sampling rate in determining analysis frequencies
- Leakage and windowing

Online Test in week 5!

This will cover material from modules 1-3 (not module 4, though there is some overlap).

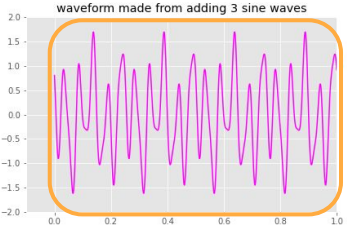
[[Show details on Learn](#)]

Today: Source-Filter Model

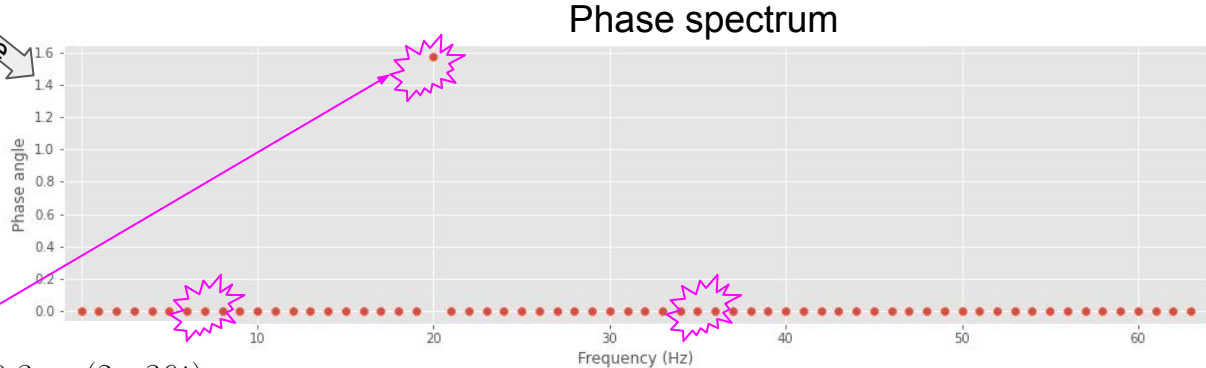
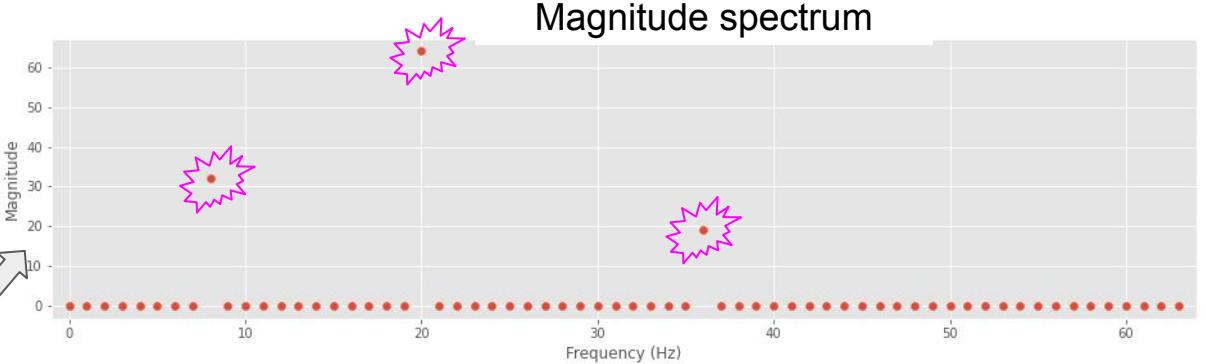
- Recap of DFT
- The physical source and filter for speech
 - Source: vocal folds (other airstreams too)
 - Filter: vocal tract + articulators
 - Tube model
- Computational models
 - Impulse, impulse train
 - Finite impulse response filter
 - Infinite impulse response filter
 - Convolution

DFT output as magnitude and phase

View of the same window in the frequency domain



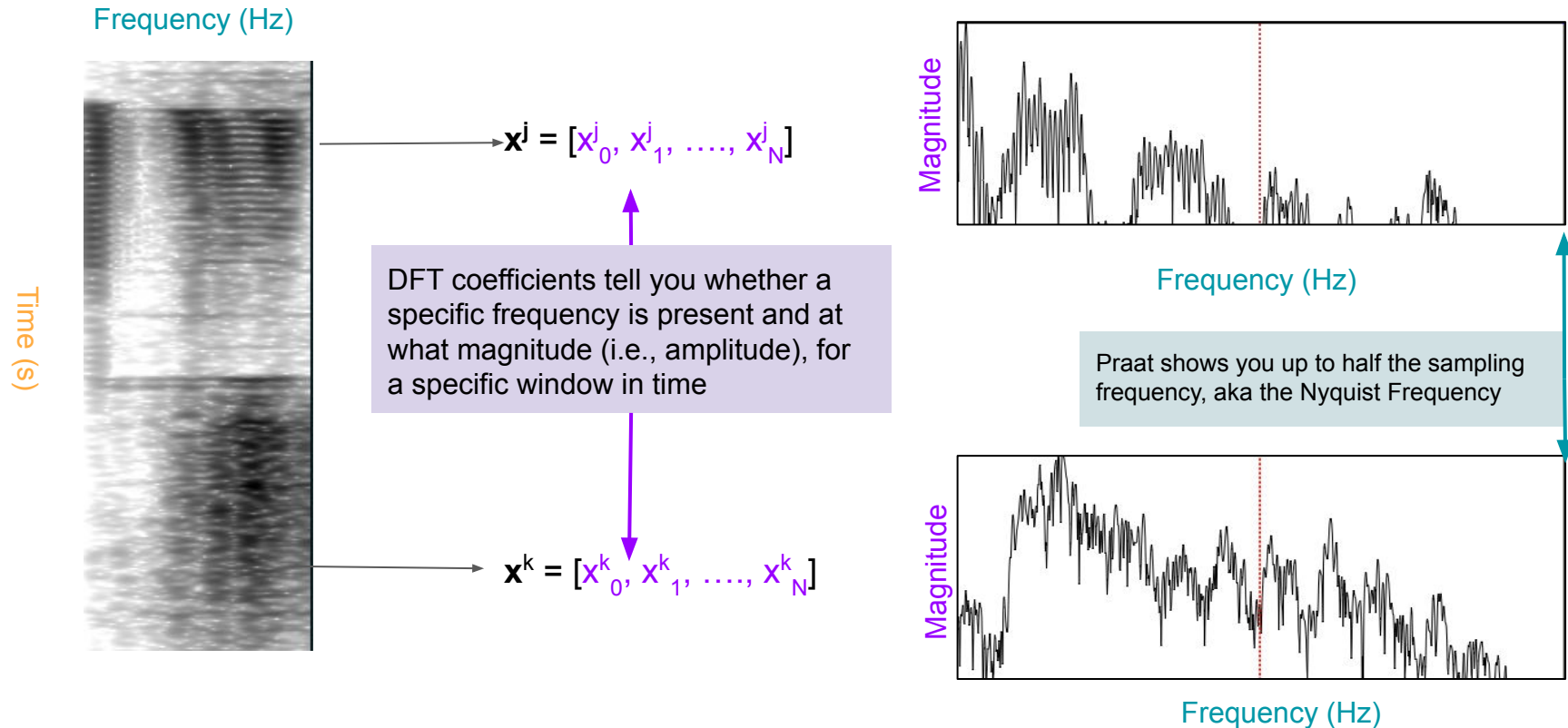
DFT



A window in the time domain

$$0.5 \cos(2\pi \cdot 8t) + \cos(2\pi \cdot 20t + \pi/2) + 0.3 \cos(2\pi \cdot 36t)$$

A Spectrogram is a sequence of feature vectors



DFT parameters

Things that affect what frequencies DFT *can detect*:

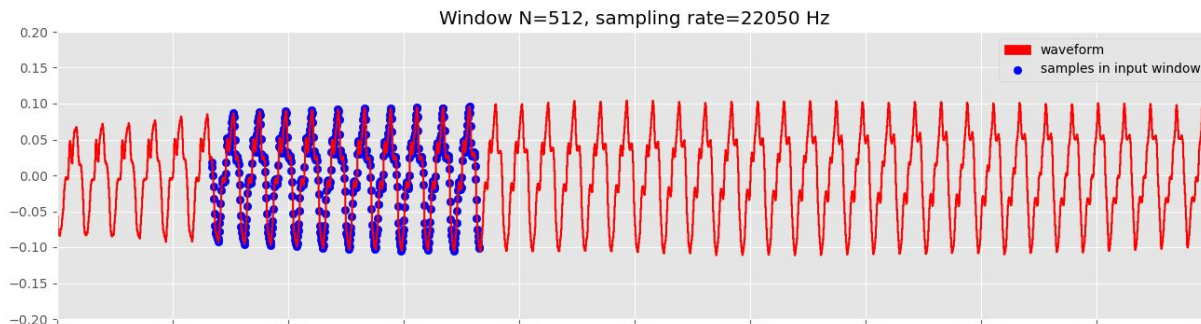
- Sampling rate: f_s
- Size of input window in terms of samples: N

For input of N samples, the DFT returns N outputs representing the magnitude (and phase) of N frequencies spaced evenly between 0 Hz and the sampling frequency f_s

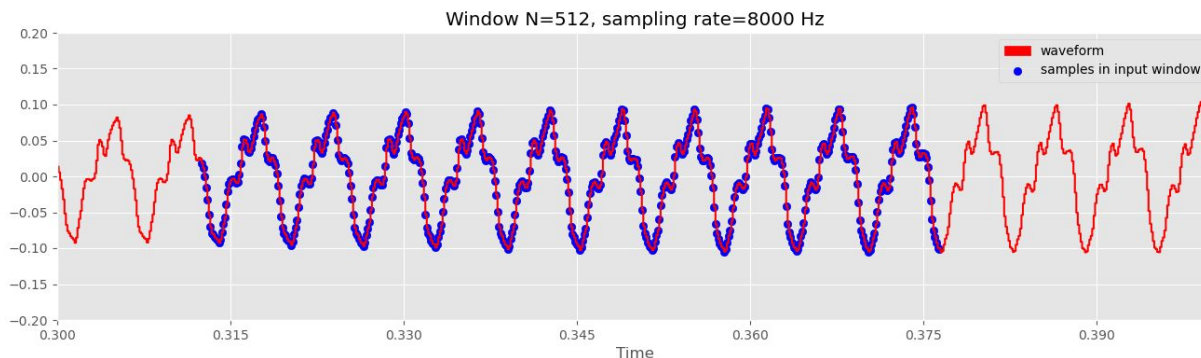
Different sampling rate

“A4” violin recording from module 3 lab

Original recording:
22050 Hz
Sampling rate



Same samples
but displayed at
8000 Hz
Sampling rate

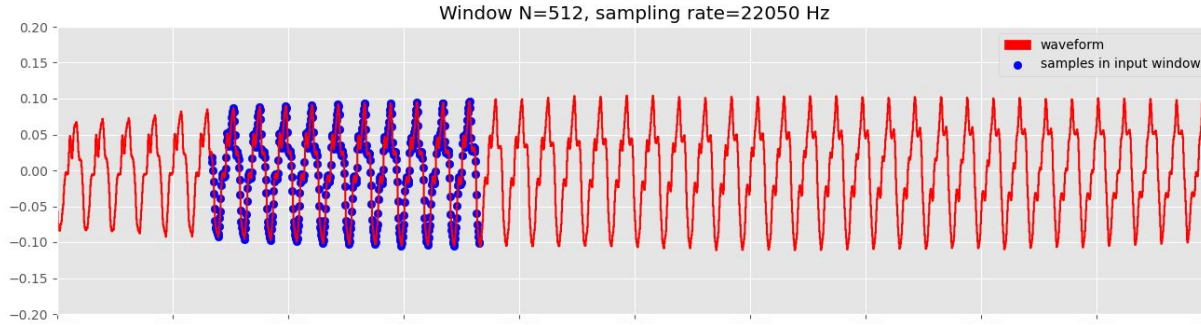


Lower sampling rate → longer sampling period

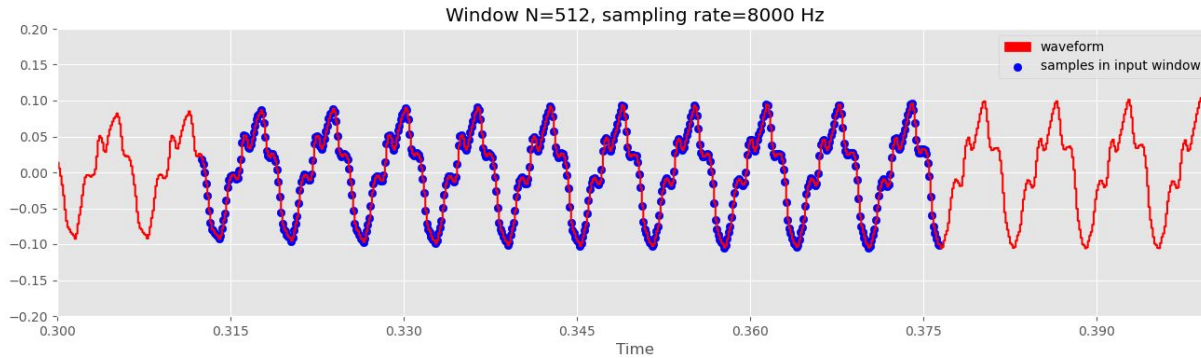
With a longer period between samples, it takes longer to complete wave cycles

Different sampling rates

Original
recording:
22050 Hz
Sampling rate



Same samples
but displayed at
8000 Hz
Sampling rate

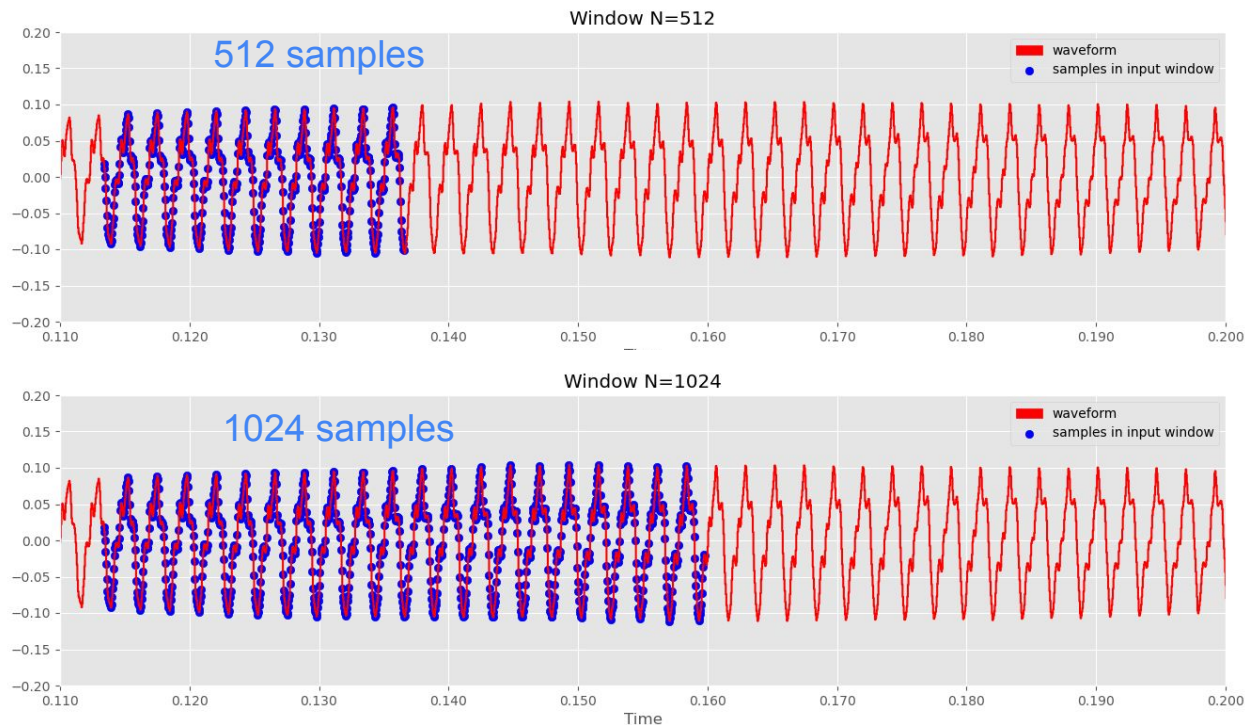


Lower sampling
rate → longer
sampling period

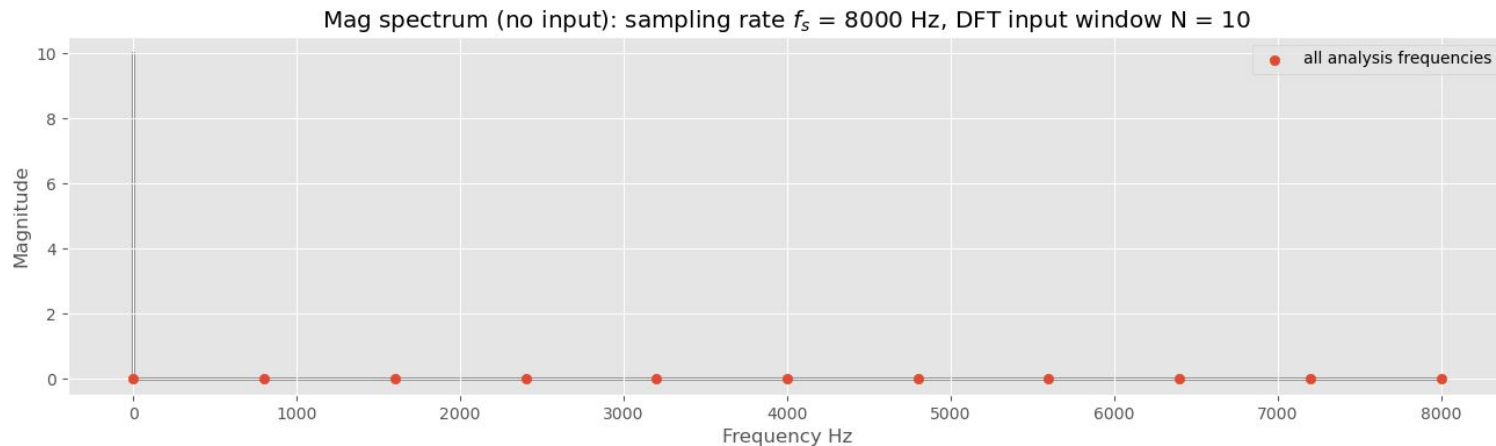
Longer period → lower frequency (less cycles per second): $f = 1/T$

Different window sizes

If we keep the sampling rate fixed, including more samples in the input window means we are applying the DFT to a longer segment of the waveform in time.

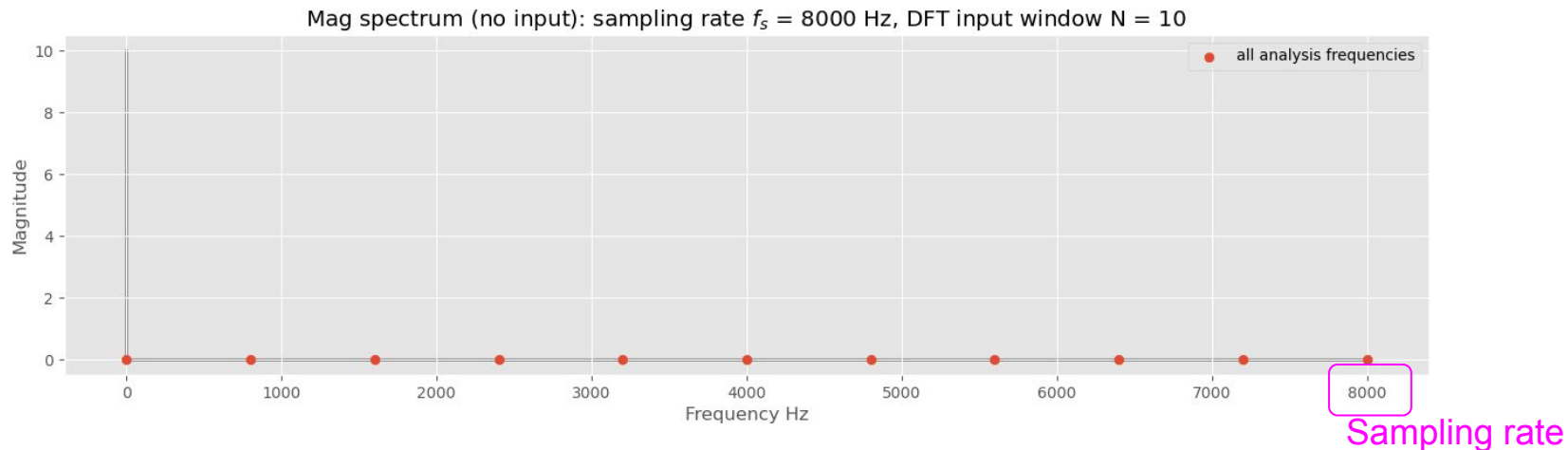


Sampling rate and DFT analysis frequencies



For an input window of N samples, the DFT returns N outputs representing the magnitude (and phase) of N frequencies spaced evenly between 0 Hz and the sampling frequency

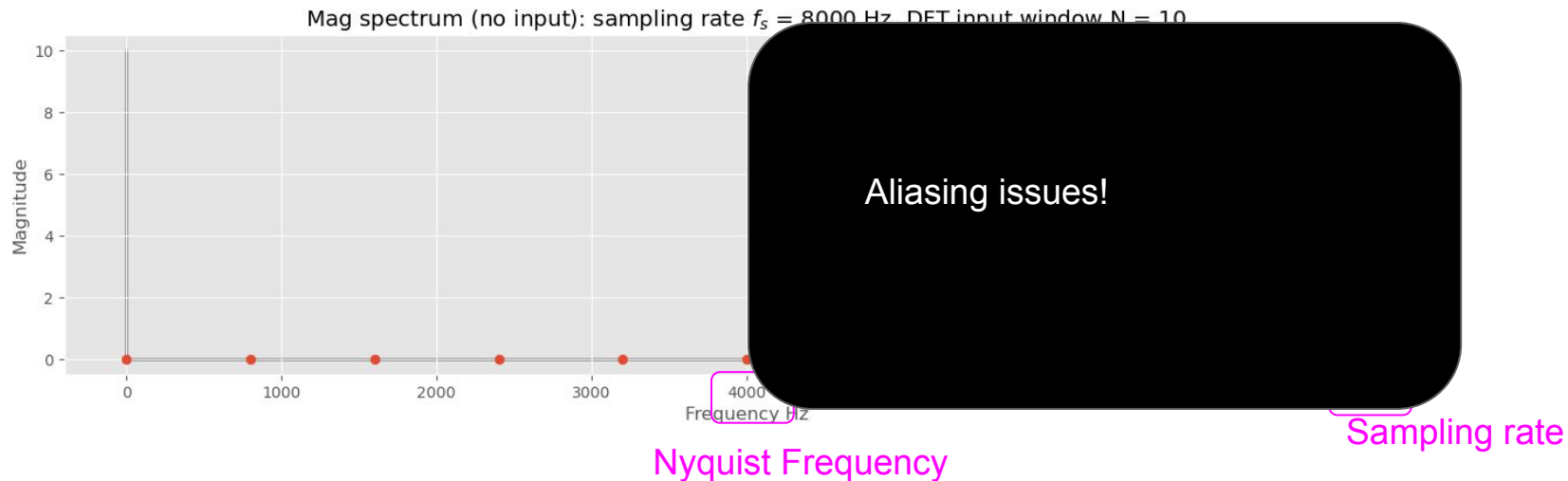
Sampling rate and DFT analysis frequencies



Example: sampling rate 8000 Hz, input window size $N=10$

→ The DFT analysis frequencies correspond to 10 points evenly spaced from 0 to 8000

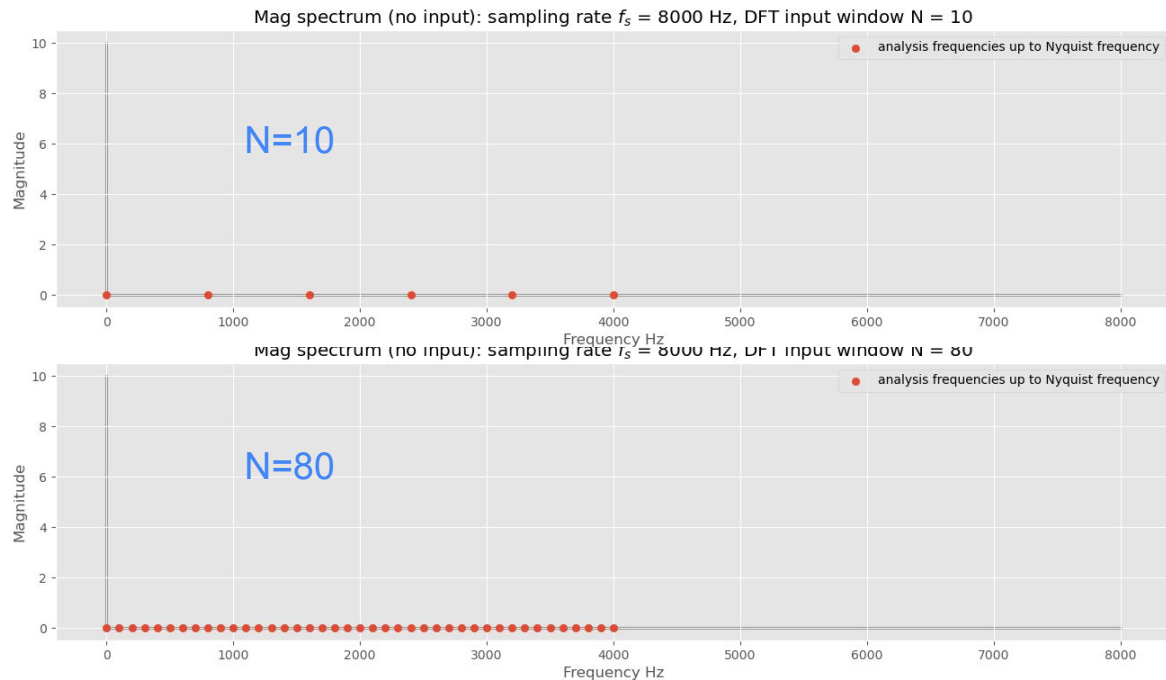
Sampling rate and DFT analysis frequencies



Example: sampling rate 8000 Hz, input window size $N=10$

→ But we can only accurately detect frequencies up to half the sampling rate (the **Nyquist Frequency**), because of aliasing

Window size and DFT analysis frequencies

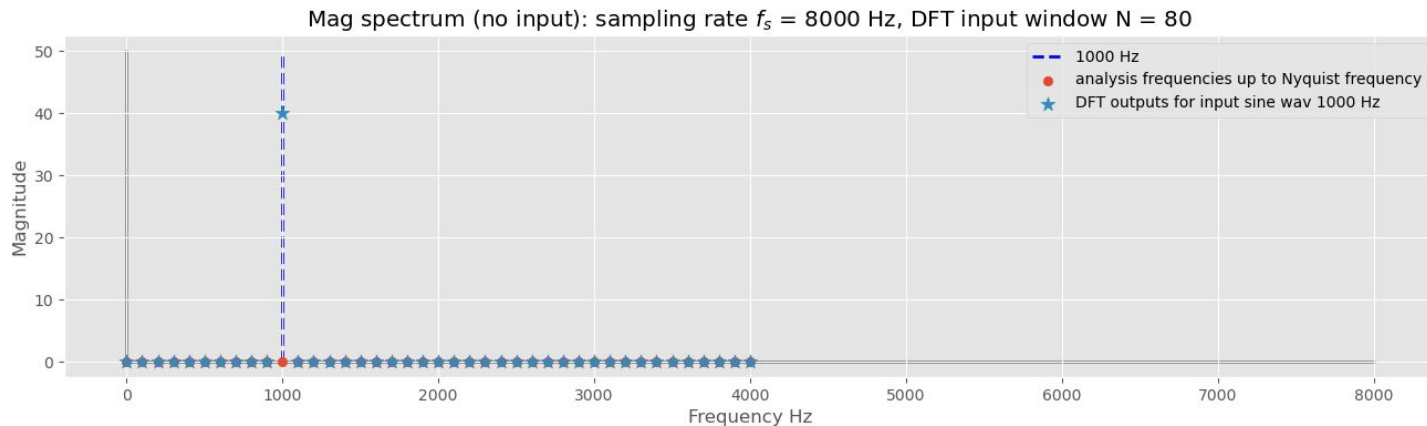


Compare: sampling rate 8000 Hz; input window size $N=10$ versus $N=80$

→ We can detect many more frequencies faithfully in the $N=80$ version

→ higher frequency resolution with more input samples (assuming fixed sampling rate)

Frequency response with different DFT setups

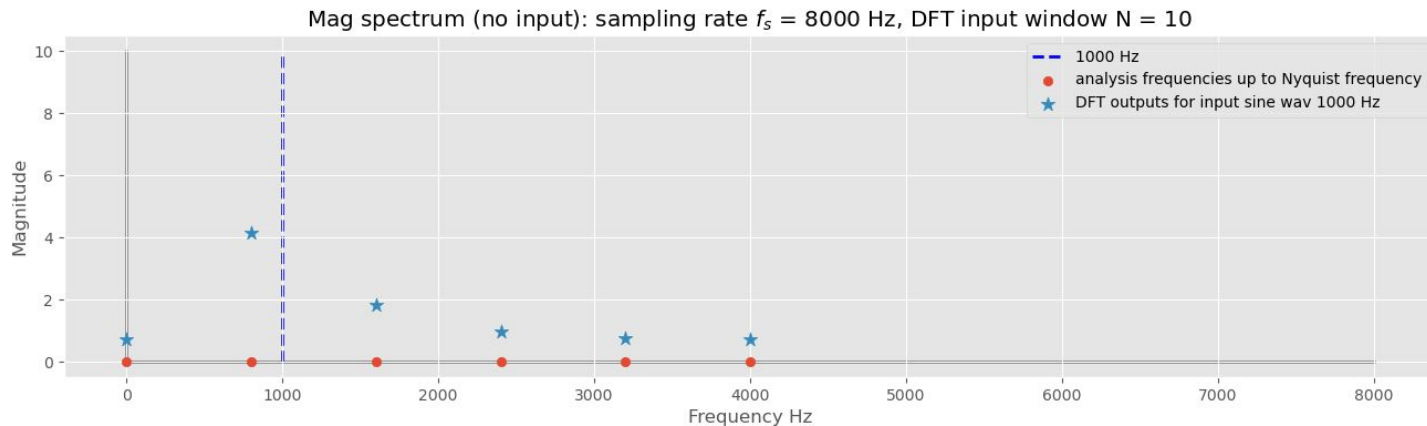


If we now apply the DFT to 1000 Hz sine wav, with a window size **N=80**.

1000 Hz is one of the analysis frequencies, so we see a positive magnitude for this and only this frequency in the magnitude spectrum

→ We can accurately capture frequency components that match the analysis frequencies

Frequency response with different DFT setups

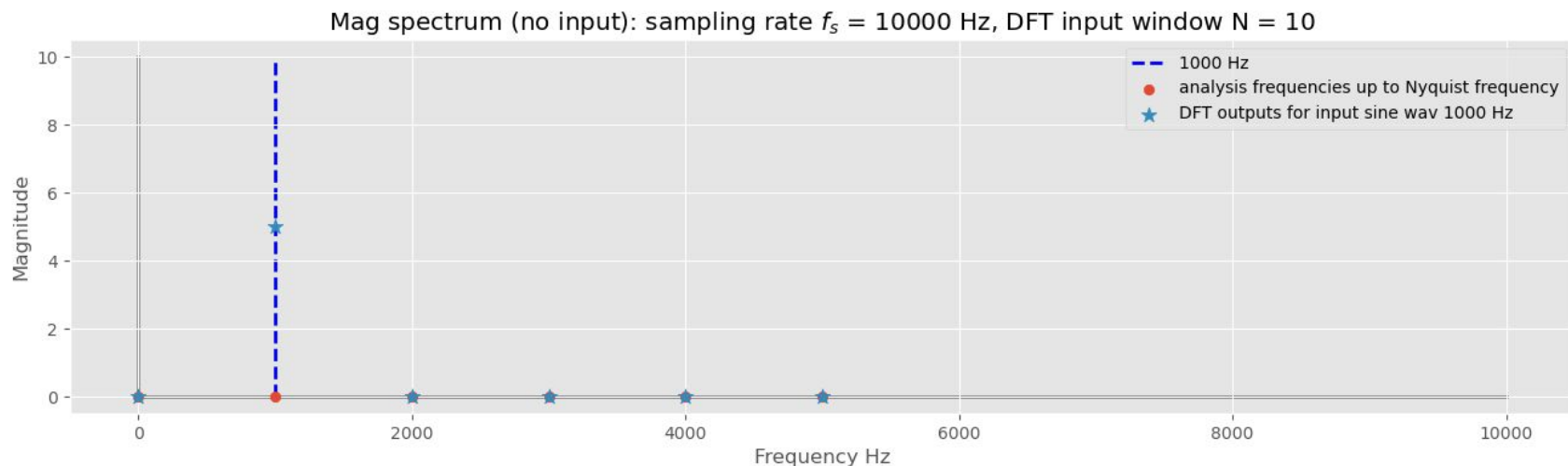


In contrast, if we apply the DFT to a 1000 Hz sine wav, with window size **N=10**, 1000 Hz is **NOT** one of the analysis frequencies

→ So, we see “leakage” onto the surrounding analysis frequencies

DFT: Sampling rate

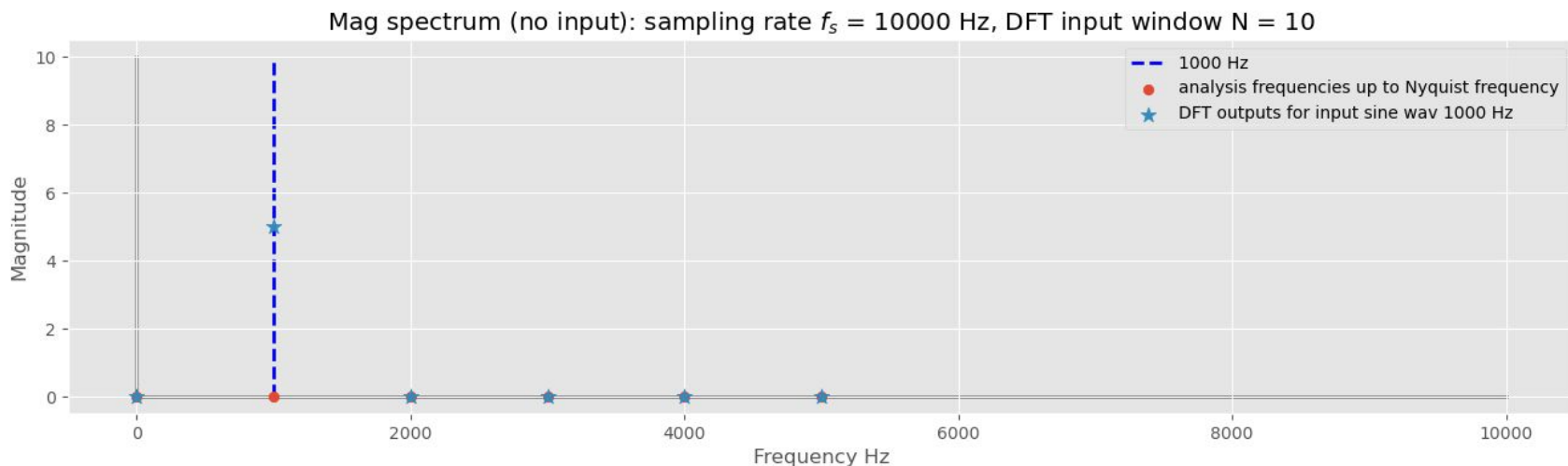
The sampling rate determines what **range** of frequencies the DFT can pick up.



With a 10000 Hz sampling rate we can pick up frequencies between 0-5000Hz (accounting for aliasing)

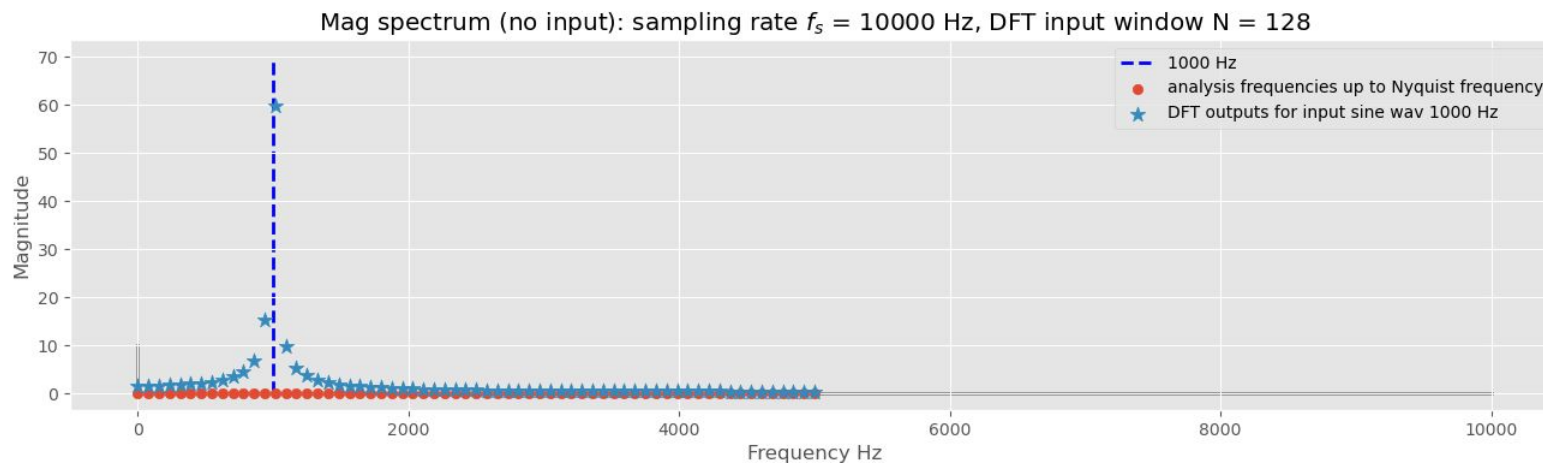
DFT: Sampling rate

The sampling rate determines what **range** of frequencies the DFT can pick up.



Now with input window size $N=10$, and sampling rate 10000 Hz, 1000 Hz is now an analysis frequency. So, we see a spike at 1000 Hz and zero magnitude for the rest of the DFT outputs

Another leakage example



But with input window size $N=128$ and sampling rate 10000 Hz, 1000 Hz is NOT an analysis frequency and we see leakage again.

Discrete Fourier Transform

for input $x[n]$ with $n=0, \dots, N-1$ (N inputs), for $k=0, \dots, N-1$ (N analysis frequencies)

$$\text{DFT}[k] = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi n k}{N}}$$

The n th element of the input sequence

A complex sinusoid rotating at a specific frequency: magnitude 1, angle $2\pi n k / N$

The k th output of the DFT corresponds to a sinusoid of frequency: $k \cdot \text{sampling_rate} / N$

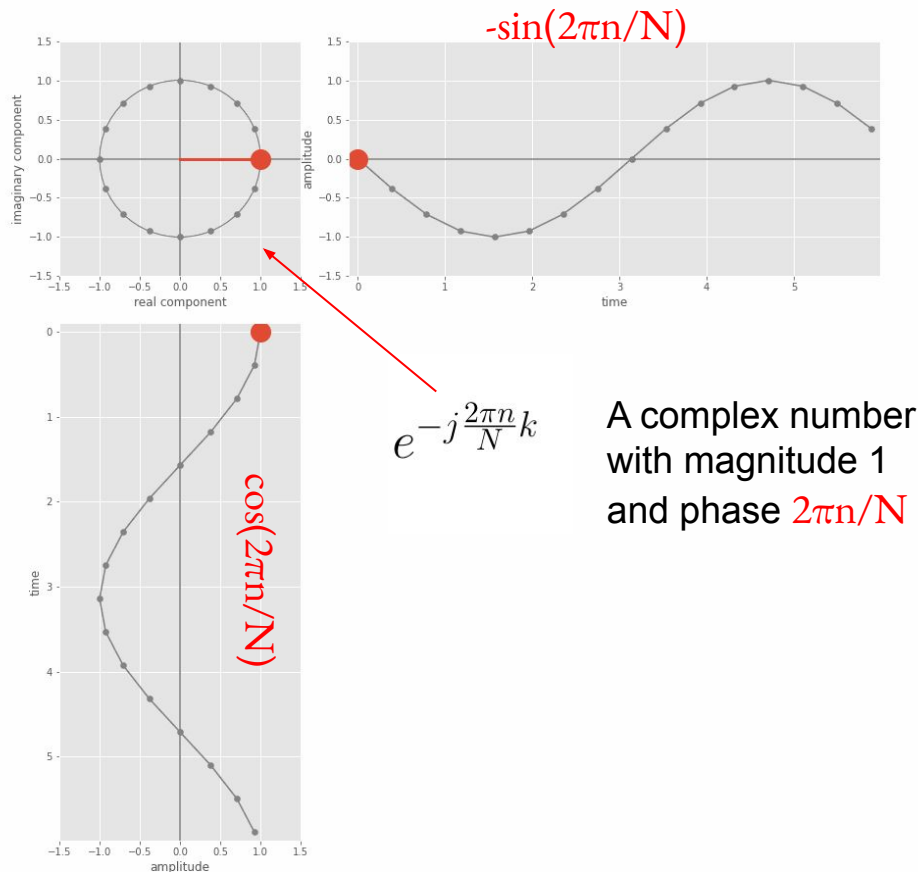
Dot-product: a measure of similarity between two sequences

Sine and cosine

We now define sine and cosine in terms of the vector rotation in the complex plane: A complex sinusoid

- **Sine** is the vertical projection of the rotating vector
- **Cosine** is the horizontal projection of the rotating vector

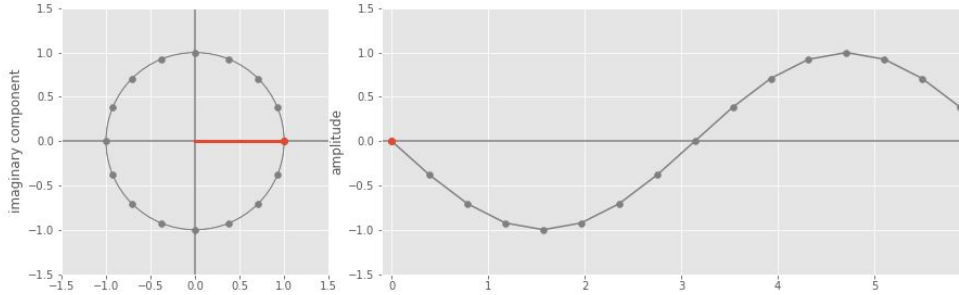
Infinite repetition in a finite space!



DFT Analysis Frequencies as sinusoids

Example Input size $N=16$, sampling rate of 800 Hz. So, $N=16$ DFT outputs...

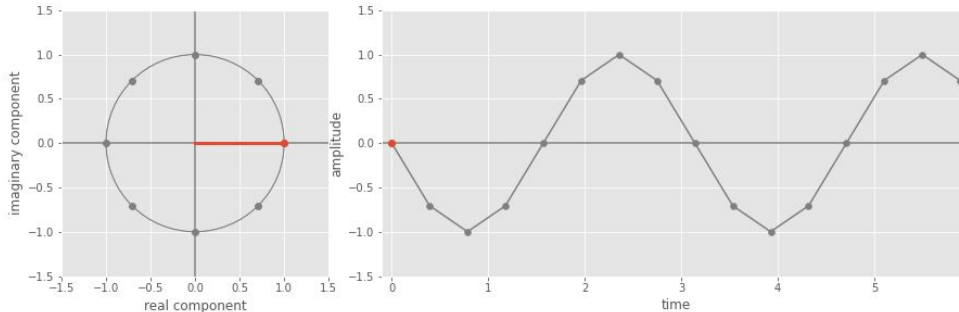
DFT[1]



16 steps for 1 cycle, 50 Hz

$$\text{DFT}[1] = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi n}{N}} \times 1$$

DFT[2]



16 steps for 2 cycles, 100 Hz

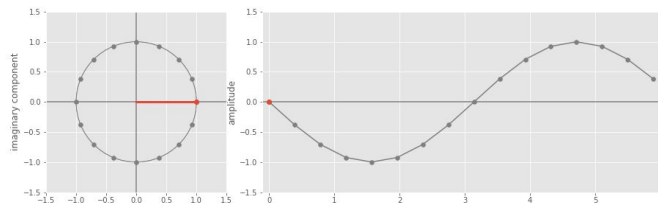
$$\text{DFT}[2] = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi n}{N}} \times 2$$

Think of this as landing on every 2nd point of the DFT[1] phasor

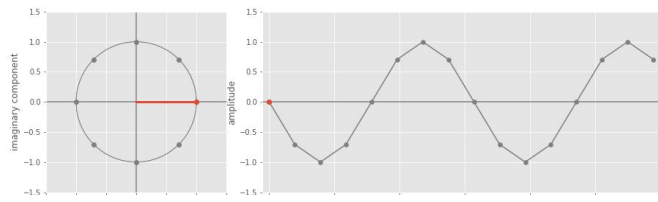
Aliasing again

Input size $N=16$
So, $N=16$ DFT outputs

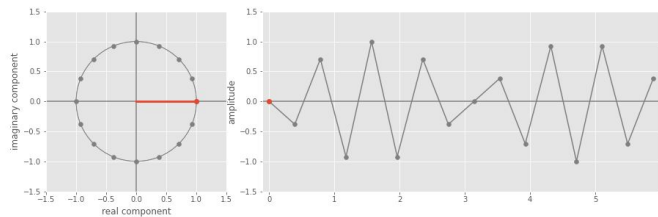
DFT[1]
50 Hz



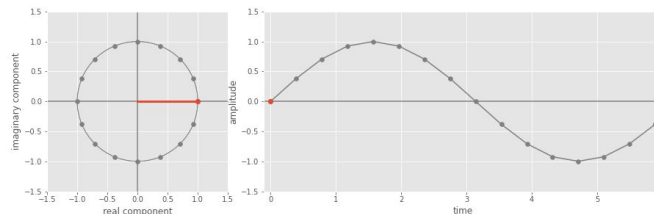
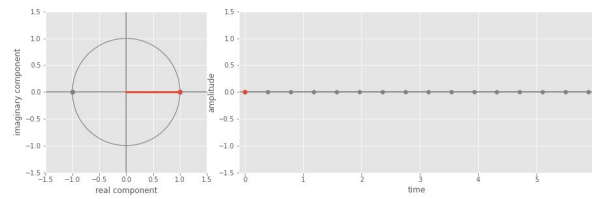
DFT[2]
100 Hz



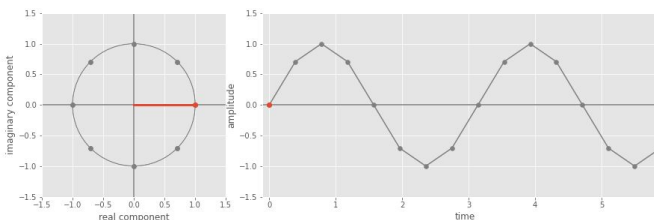
DFT[7]
350 Hz



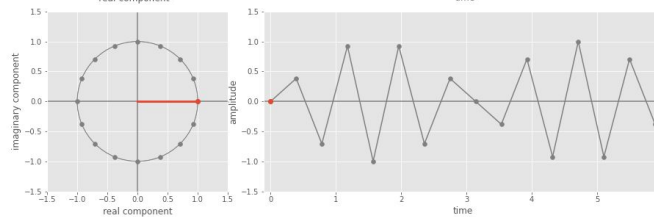
DFT[8]
400 Hz



DFT[15]
750 Hz?



DFT[14]
700 Hz?

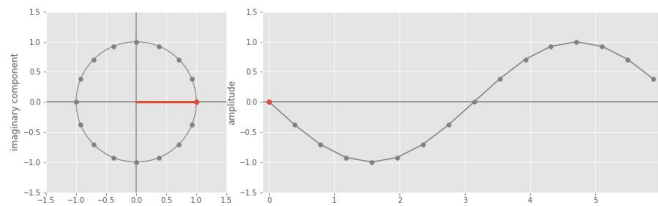


DFT[9]
450 Hz?

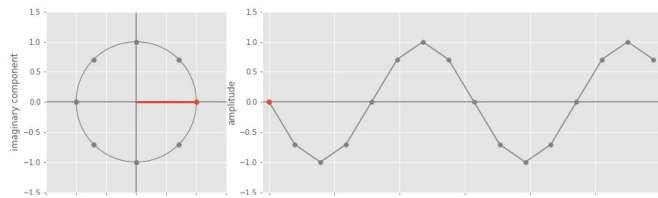
Aliasing again

Input size $N=16$
So, $N=16$ DFT outputs

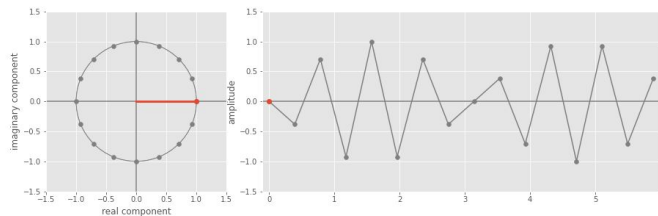
DFT[1]
50 Hz



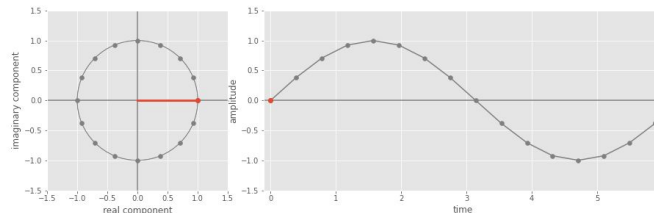
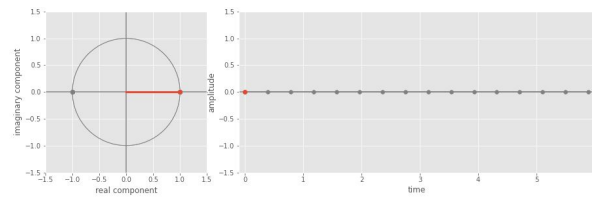
DFT[2]
100 Hz



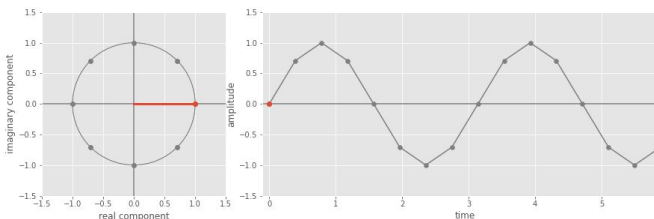
DFT[7]
350 Hz



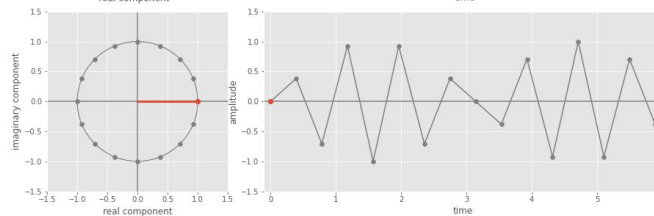
DFT[8]
400 Hz



DFT[15]
~~750 Hz?~~
50 Hz



DFT[14]
~~700 Hz?~~
100 Hz



DFT[9]
~~450 Hz?~~
350 Hz

Discrete Fourier Transform

for input $x[n]$ with $n=0, \dots, N-1$ (N inputs), for $k=0, \dots, N-1$ (N analysis frequencies)

$$\text{DFT}[k] = \sum_{n=0}^{N-1} x[n] e^{-j \frac{2\pi n k}{N}} = \boxed{M e^{j\varphi}}$$

A magnitude (scale factor) A phase angle (shift factor)

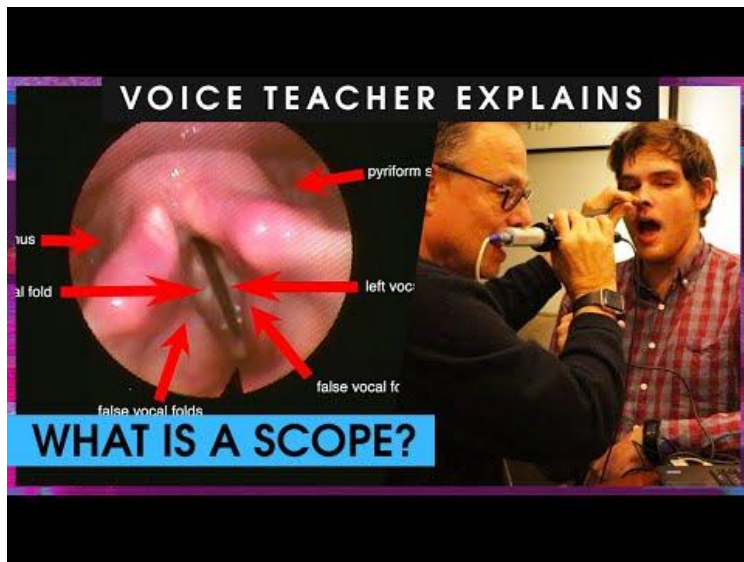
A complex number

The DFT formula calculates the **similarity** between the input and the complex sinusoid of a specific frequency. It's output is a **complex number** that tells you how you would **scale** and **shift** that sinusoid in order to reconstruct the original input (summing the complex sinusoids corresponding to the analysis frequencies)

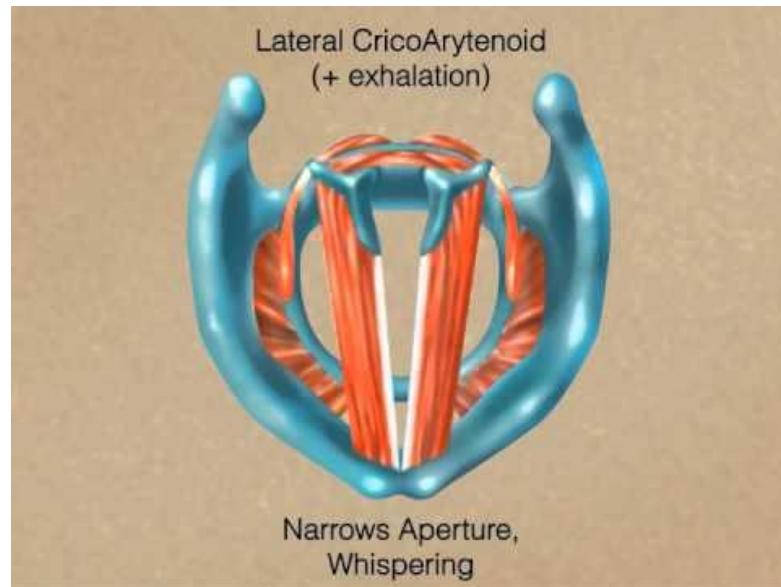
Source and Filter

Human speech: source

- **Voiced:** Vocal folds vibrating
- **Unvoiced:** vocals fold held close but not vibrating



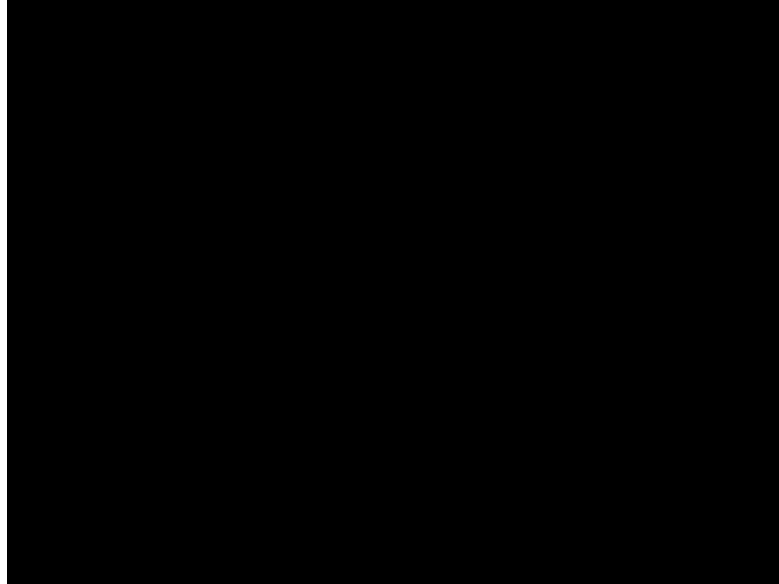
<https://youtu.be/BHfGgrVQ2P0>



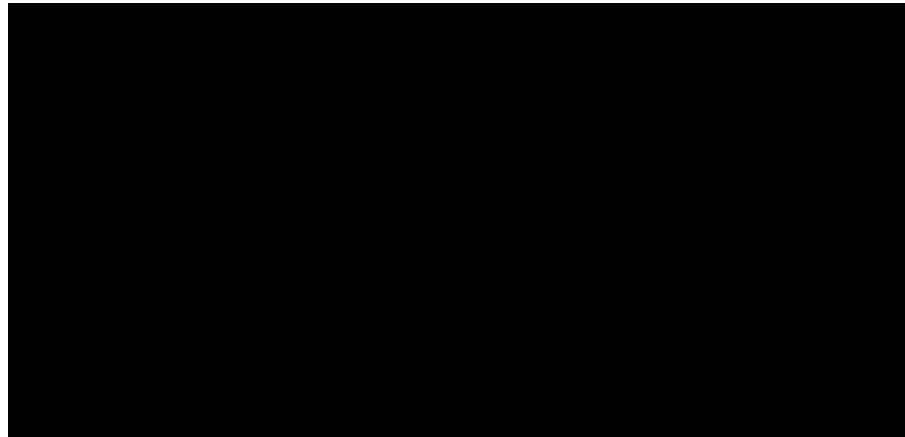
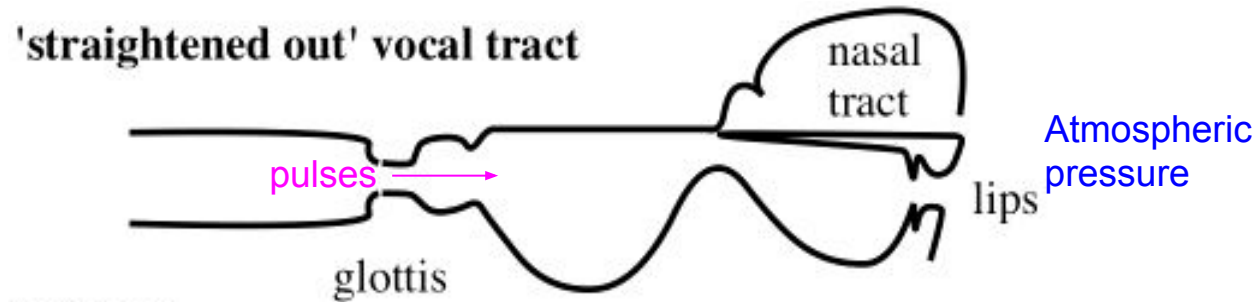
<https://youtu.be/b89RSYCaUBo>

Resonance

A resonant frequency is a natural frequency of vibration determined by the physical parameters of the vibrating object.



Vocal tract as a tube



We model the vocal tract as a tube open at one end (the mouth)

Standing waves: tube open at one end

<https://www.acs.psu.edu/drussell/Demos/StandingWaves/StandingWaves.html>

Standing waves: tube open at one end

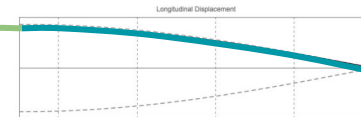
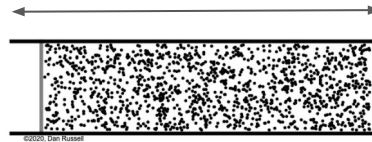
The first resonant frequency can be determined as:

$$R1 = c/\lambda = c/4L$$

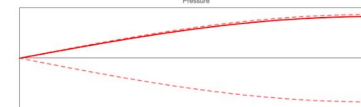
Where c is the speed of sound (343 m/s)

Q: How does tube length change wavelength? How does this change the resonant frequency?

L = length of tube



Air particle displacement



Pressure

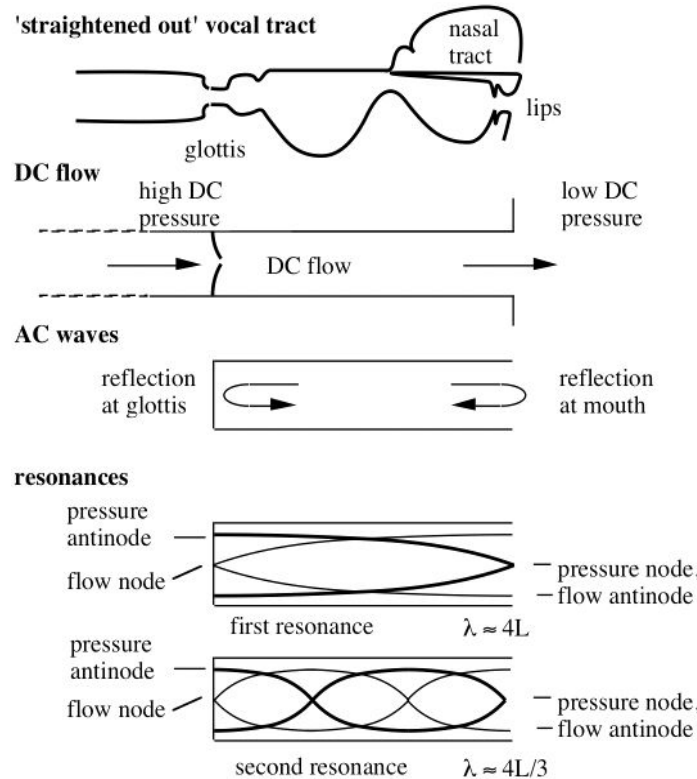
λ = wavelength (metres per cycle)



Vocal tract as a tube

Figure from PhysClips:

<https://newt.phys.unsw.edu.au/jw/voice.html>

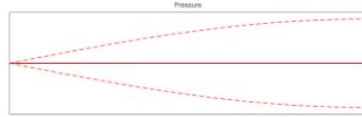
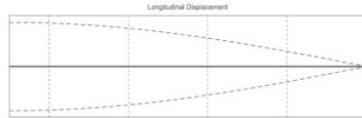
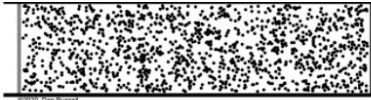


Source: Air through the glottis

Resonances occur for frequencies that get reinforced by the oscillations -> **standing waves**

Oscillation of air particles due to pulses at the source and reflections at the ends of the tube

L = length of tube



$$\lambda = 4L$$

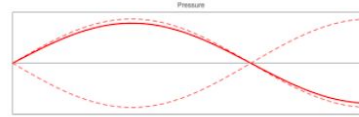
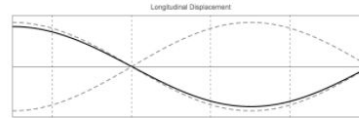
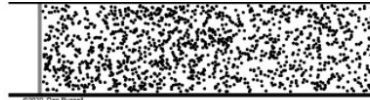
$$R_1 = 343 / 0.15 = 571 \text{ Hz}$$

Calculate the n-th resonance as:

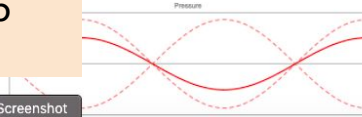
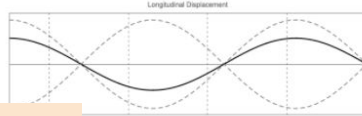
$$R_n = \frac{c(2n - 1)}{4L}$$

$$\lambda = 4L/3$$

$$R_2 = 343 \times 3 / 0.15 = 1715 \text{ Hz}$$

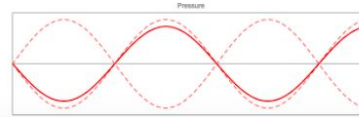
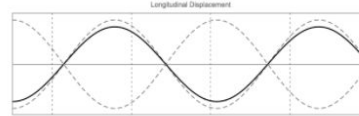


Below left is the third mode (fifth harmonic) and below right is the fourth mode (7th harmonic).



$$\lambda = 4L/5$$

$$R_2 = 343 \times 5 / 0.15 = 2855 \text{ Hz}$$



$$\lambda = 4L/7$$

How about a tube of 15 cm (0.15 m)? What vowel is most like a tube?

Resonances of schwa [ə]

Think of [ə] as a single tube (no constriction)



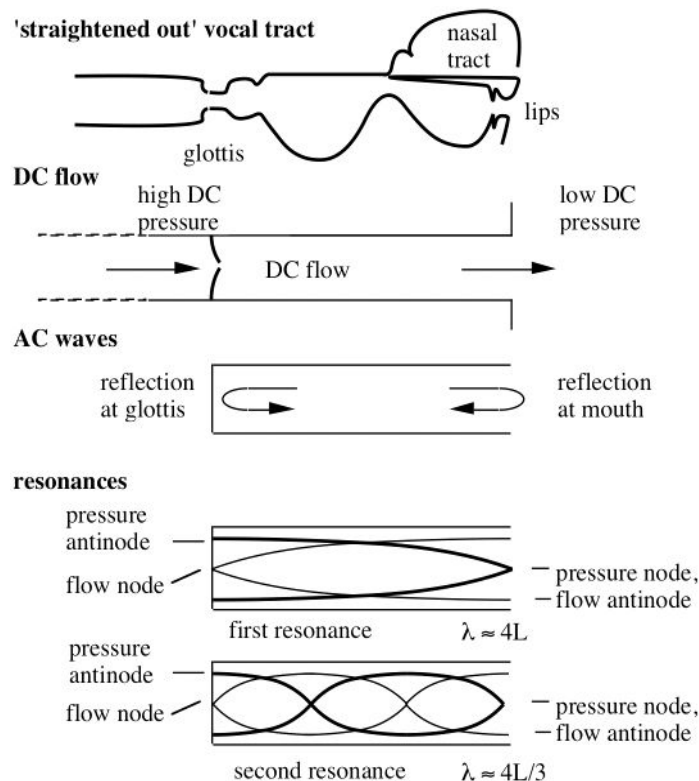
<https://seeingspeech.ac.uk/ipa-charts/?chart=4&datype=1&speaker=1#location=601>

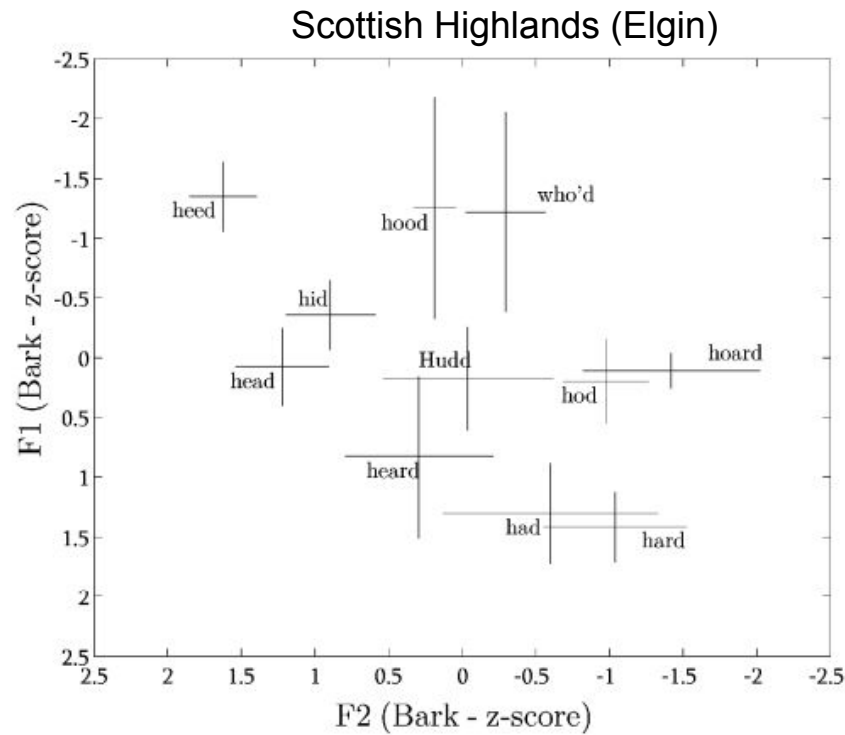
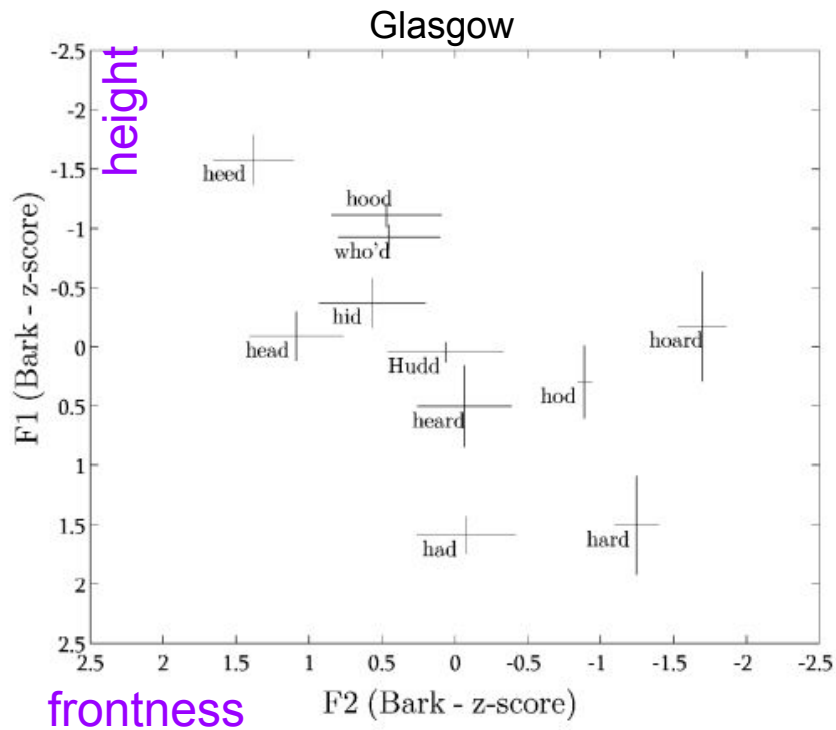
Resonances of schwa [ə]

Think of [ə] as a single tube (no constriction)

Measure your own vocal tract by measuring your schwa formant frequencies and solve for L

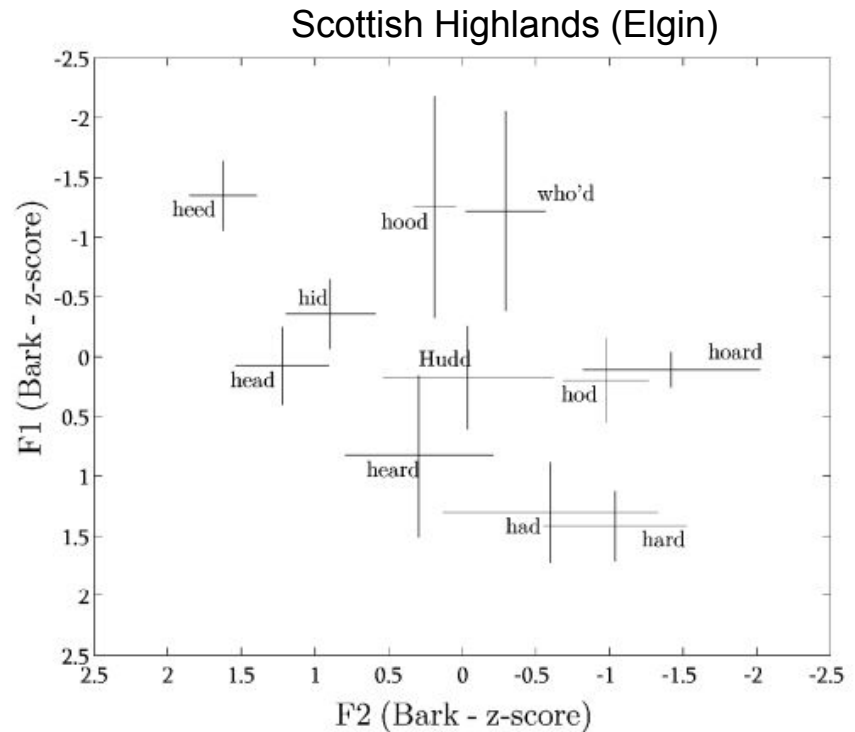
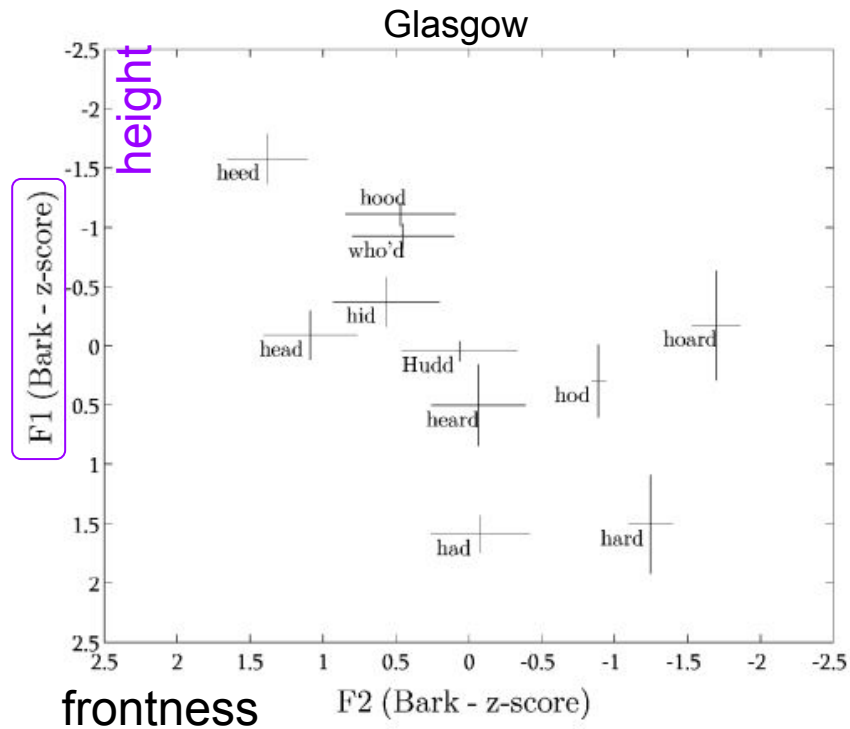
Wavelength $\lambda = 4 * 0.17 \text{ m} = 0.68 \text{ m}$
Quite a lot longer than your vocal tract!





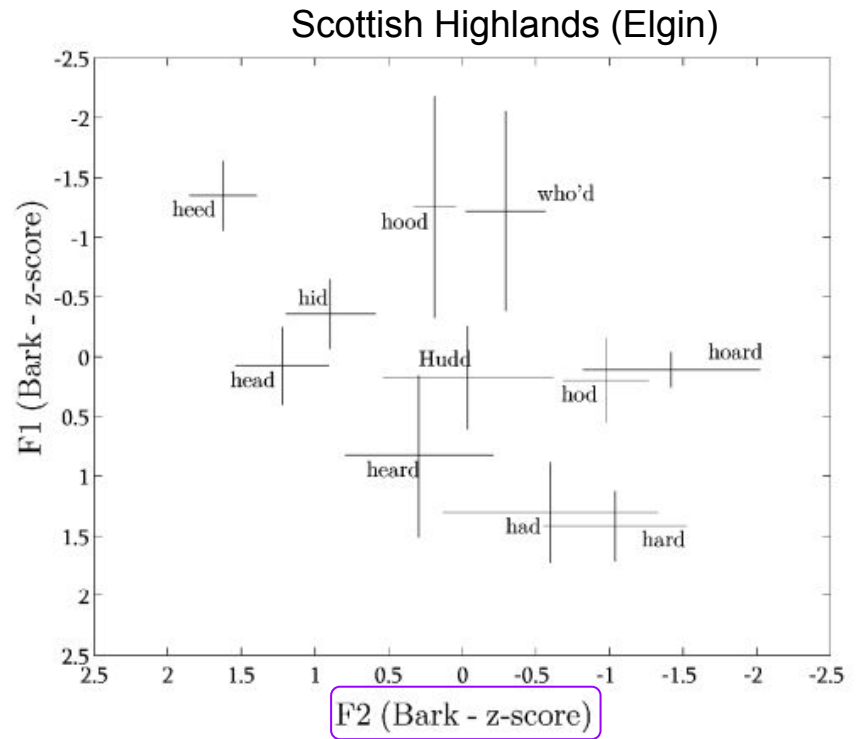
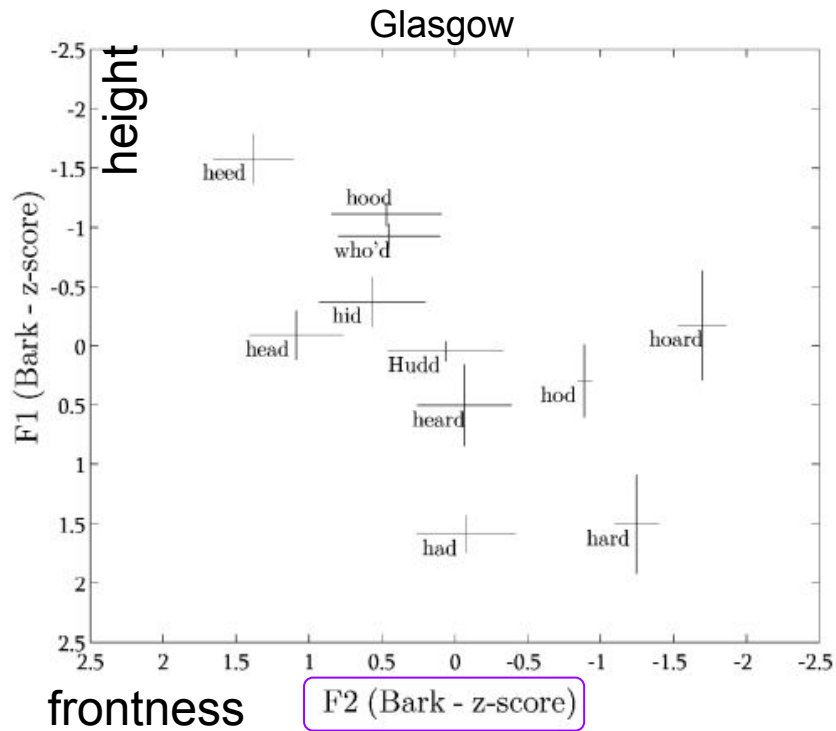
Ferragne, E., & Pellegrino, F. (2010). Formant frequencies of vowels in 13 accents of the British Isles. *Journal of the International Phonetic Association*, 40(1), 1-34. doi:10.1017/S0025100309990247

Formants (acoustic property) correspond to **resonances** (physical property)



Ferragne, E., & Pellegrino, F. (2010). Formant frequencies of vowels in 13 accents of the British Isles. *Journal of the International Phonetic Association*, 40(1), 1-34. doi:10.1017/S0025100309990247

Resonance 1: mouth more open -> lower tongue height -> higher F1



Ferragne, E., & Pellegrino, F. (2010). Formant frequencies of vowels in 13 accents of the British Isles. *Journal of the International Phonetic Association*, 40(1), 1-34. doi:10.1017/S0025100309990247

Resonance 2: tongue constriction further front -> higher F2

Two tube model, e.g. [a]

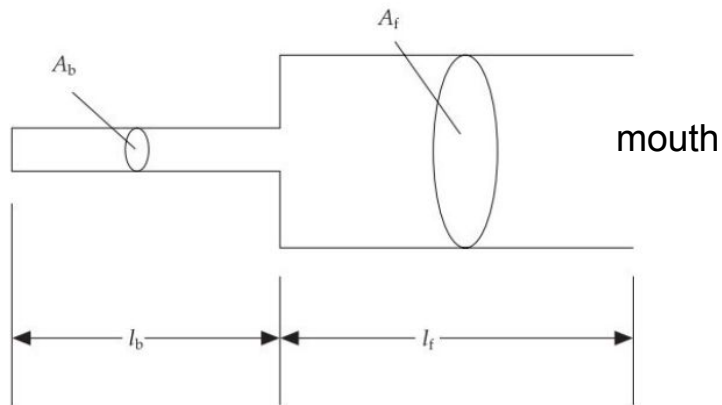


Figure 6.1 Two-tube model of the vocal tract that approximates the shape of the vocal tract for [a].

- Each individual tube has its own resonances
- The tube responsible for F₁ (the first formant) changes as the lengths of the tubes change

Figures from: Johnson, K (2012). Acoustic and Auditory Phonetics, Chapter 6

Details of the tube models are **extension material**

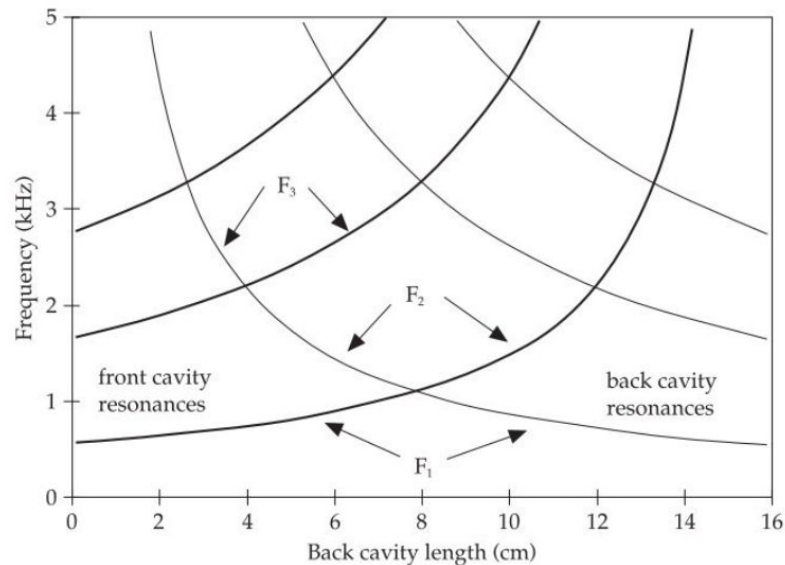
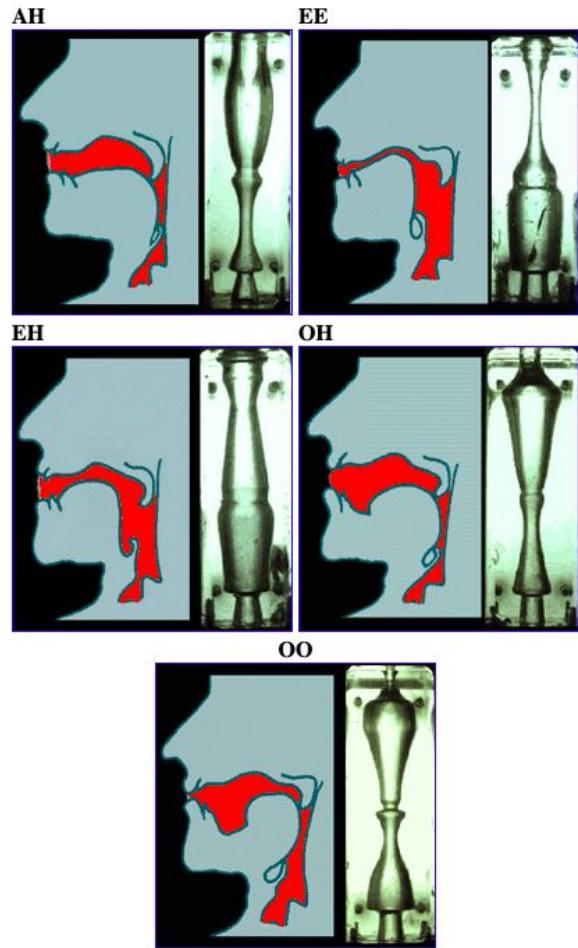


Figure 6.2 Natural resonant frequencies of the back tube (light lines) and front tube (heavy lines) in the tube model shown in figure 6.1 for different lengths of the back cavity. Overall vocal tract length is 16 cm, so the front cavity length is 16 cm minus the back cavity length.

Tube models of speech

Physical tubes used to generate vowels

- Plastic models of the vocal tract
- 'Duck call' as source



F0, harmonics, formants

Simplification of reality!



Source

- Fundamental frequency (F0) is driven by the frequency of vocal fold vibrations
- Harmonics are multiples of F0

Filter

- Resonances are driven by the shape of the vocal tract (physical property)
- Formants are peaks in the spectral envelope that correspond to resonances (acoustic property)

Independence of source and filter: You can change F0 without changing the vowel you are saying: harmonics change, formants stay the same

Perceived pitch

- We usually say F_0 is the acoustic correlate of perceived pitch
- But you can still perceive F_0 even if you filter out all frequencies below, e.g. 500 Hz
- The human brain reconstructs F_0 from the harmonics!

Interim Summary

- Resonances of the vocal tract depend on vocal tract constriction and the size of the opening, length of the tube
 - Also depends on the medium the sound wave is travelling through (usually assume it's air)
- Resonances of the vocal tract don't control perceived pitch in the human voice
 - There can be interactions in the physical system though (biomechanics)
- Pitch perception is driven by harmonics, which are determined by F_0 the rate of vocal fold vibration

Q: What's more important for recognizing words? Harmonics or formant structure?

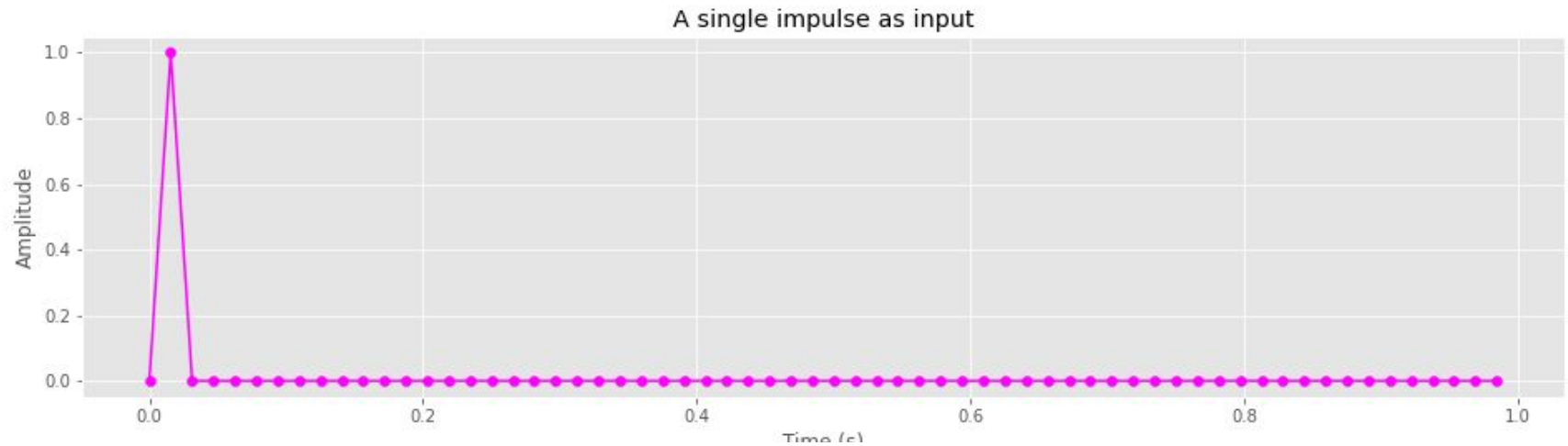
Source and Filter: computationally

Requirements:

- Approximate the source → vocal pulses → **impulse train**
- Approximate the filter → vocal tract → **difference equations and convolution**

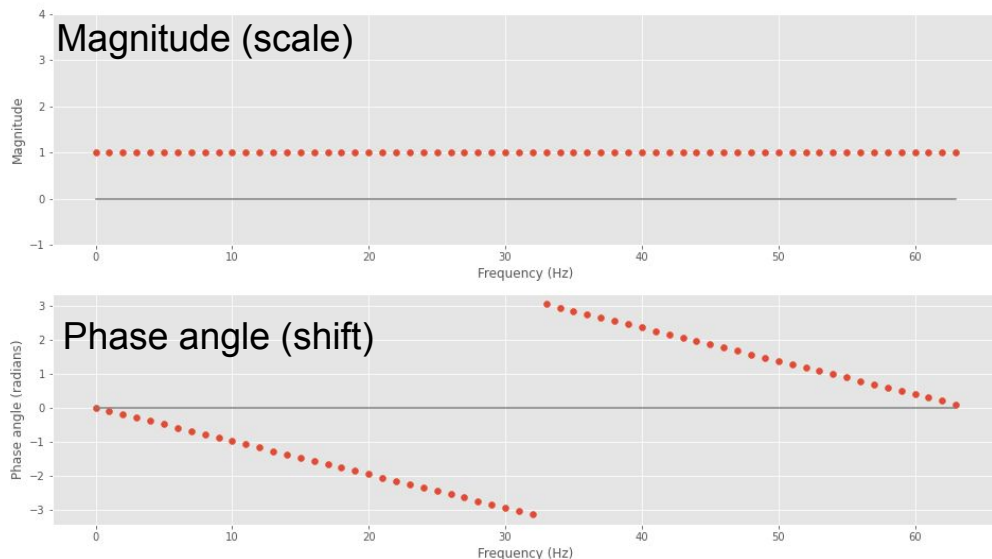
Source: Impulse

- A single non-zero amplitude in time (zero otherwise)
- An infinitely thin rectangular wave



DFT of an Impulse

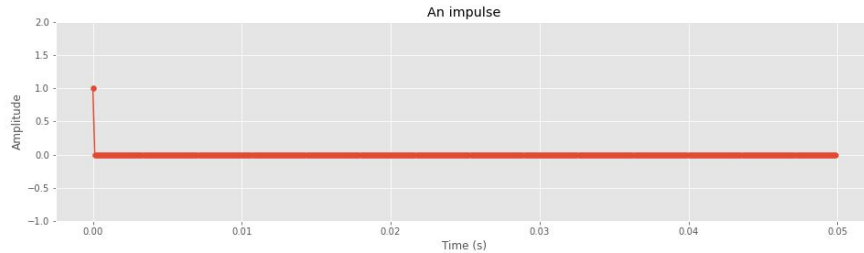
This impulse 'contains' all frequencies



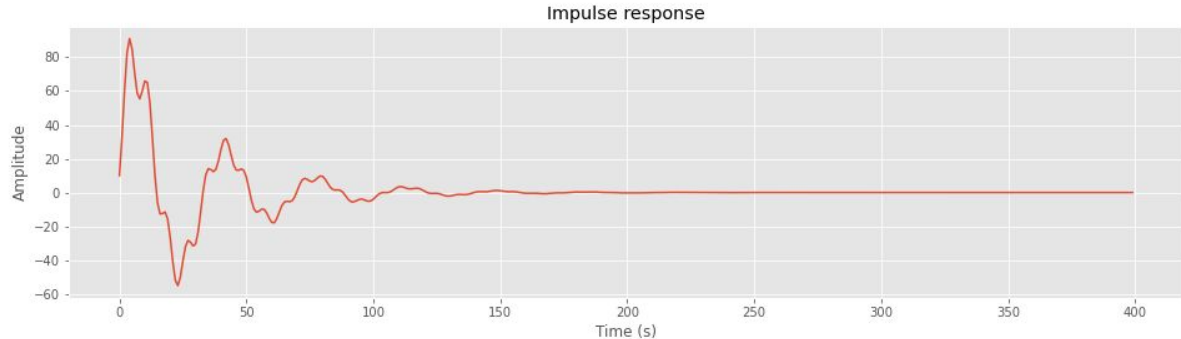
To construct an impulse from sinusoids, you need as many as you can possible get!

Impulse response

- Put a single pulse into a system
- See what frequencies are boosted (i.e. resonances) and which are dampened

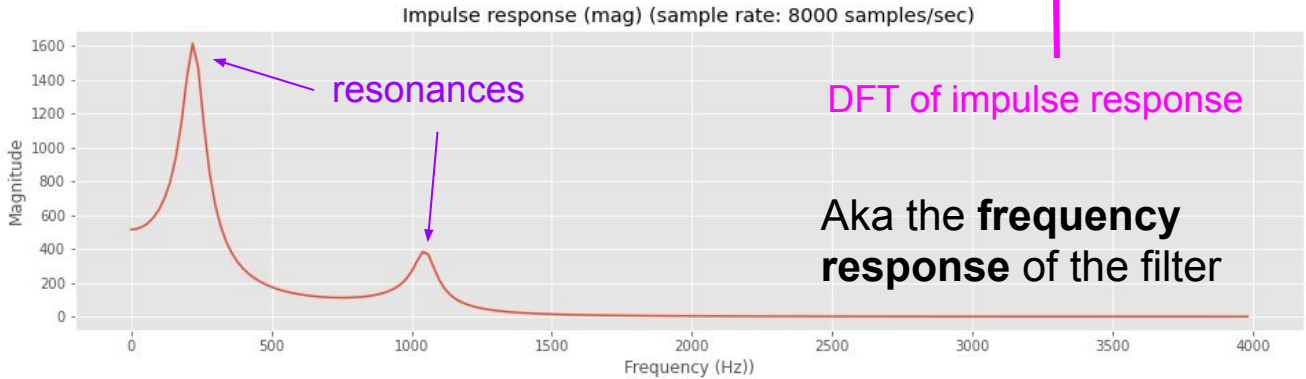
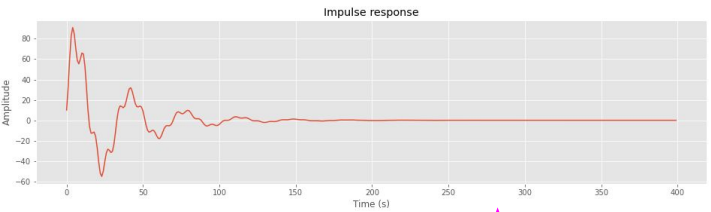
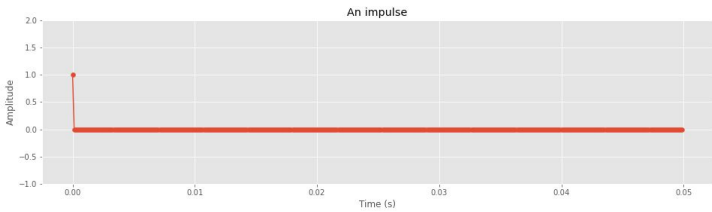


Result of passing the impulse through a specific filter:



Impulse response

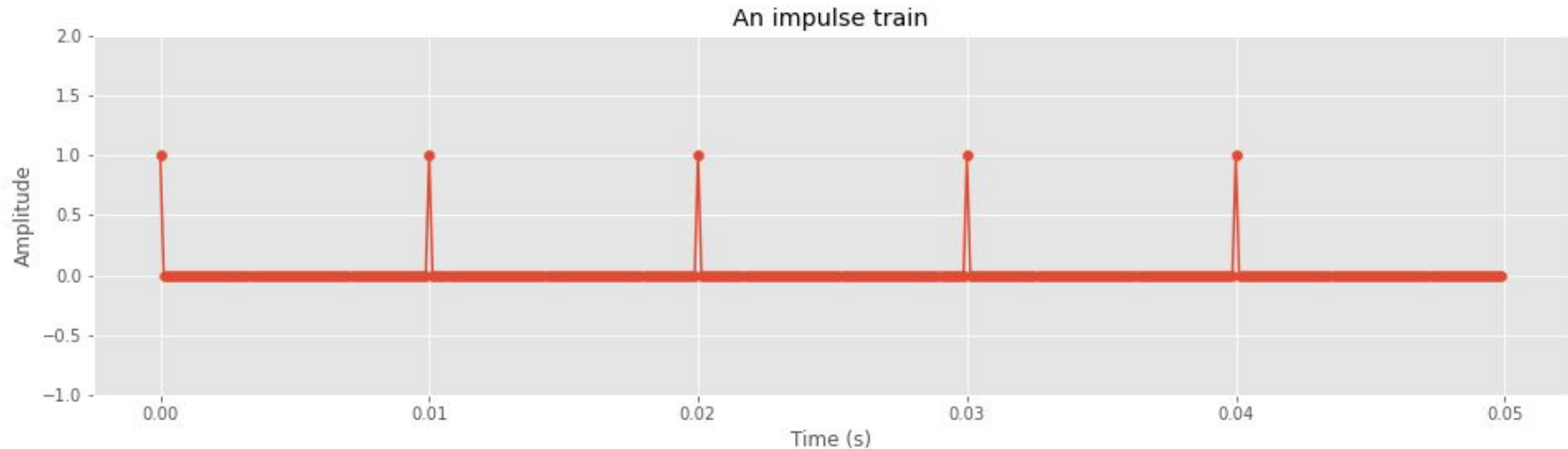
- Put a single pulse into a system
- See what frequencies are boosted (i.e. resonances) and which are dampened



Impulse train

An series of impulses at regular intervals is called an **impulse train**

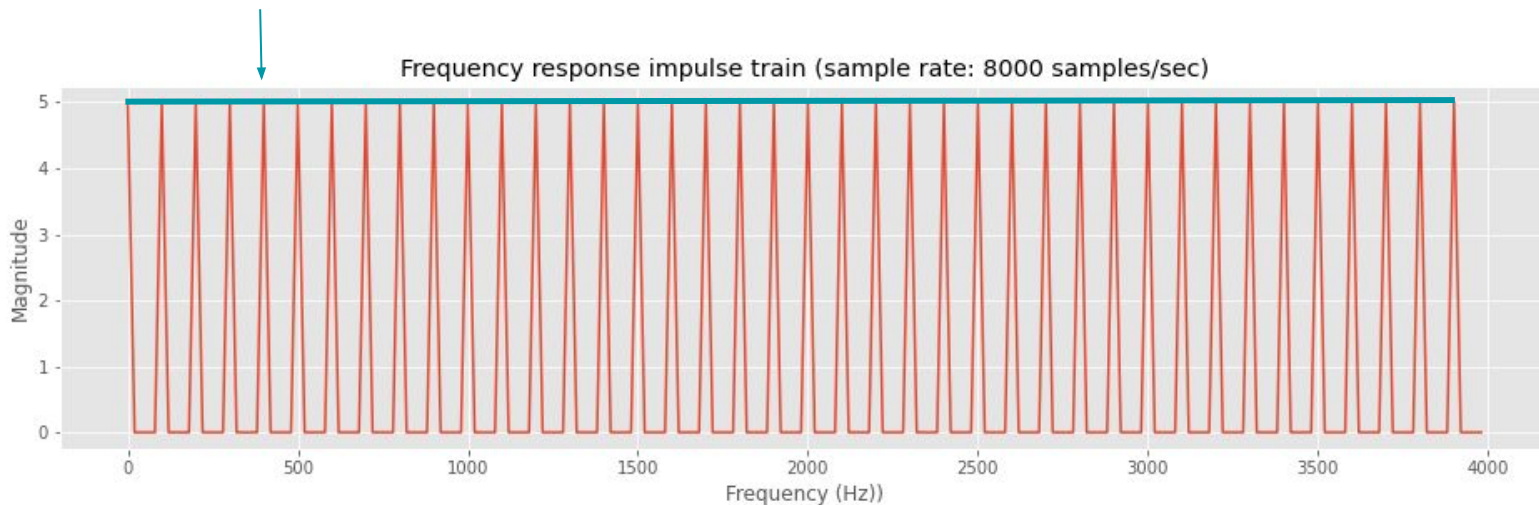
- Period between impulses 0.01 seconds
- So, the fundamental frequency of this impulse train is 100 Hz



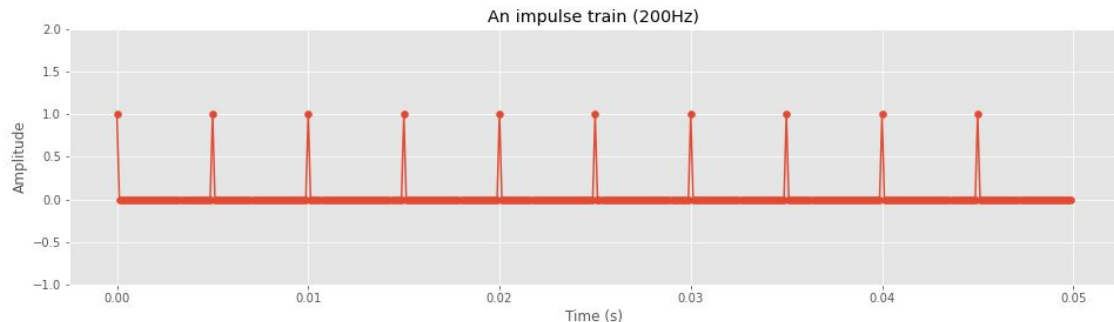
Impulse train: spectrum

We we apply the DFT to the 100 Hz impulse train we see positive magnitudes for integer multiples of 100 Hz in the magnitude spectrum, i.e. **harmonics**

- Like the actual vocal source!
- The **spectral envelope** here is flat: all harmonics equally present



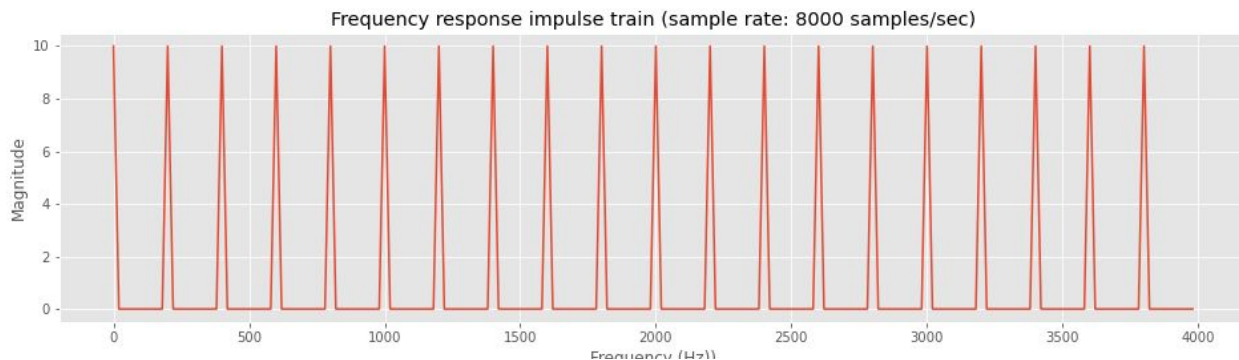
Spectrum of Impulse train (200 Hz)



- Period between impulses is 0.005 seconds
- Smaller period, higher frequency

Harmonics at 200Hz,
400Hz, 600Hz,
800Hz,etc

+ A non-zero 0th
coefficient (bias
term)



Filters

Desired action: shape the spectrum in a specific way

Some common filters:

- Low pass filter: remove all frequencies above a specific frequency
- High pass filter: remove all frequencies below a specific frequency
- Band pass filter: remove all frequencies outside two specified frequencies

Filters applied in the time domain



We use filters to change the frequency characteristics of a waveform

Question: What's the effect of this filter?

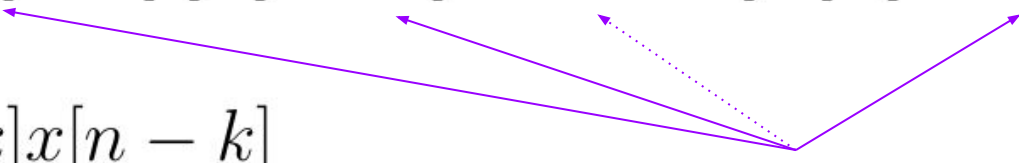
Filters applied in the time domain



We'll look at two types of filters you can apply directly to the waveform:

- Finite Impulse Response (FIR) filters
- Infinite Impulse Response (IIR) filters

Finite Impulse Response (FIR) filters

$$y[n] = b[0]x[n] + b[1]x[n - 1] + \dots + b[K]x[n - K]$$
$$= \sum_{k=0}^K b[k]x[n - k]$$


$x[n], \dots, x[n-K]$ are the previous K samples of the input

A **Finite Impulse Response (FIR)** filter takes the original signal as input and transforms it based on the previous values of the input

Here the filter is expressed as a **difference equation**

FIR Filters

$$\begin{aligned}y[n] &= b[0]x[n] + b[1]x[n - 1] + \dots + b[K]x[n - K] \\ &= \sum_{k=0}^K b[k]x[n - k]\end{aligned}$$

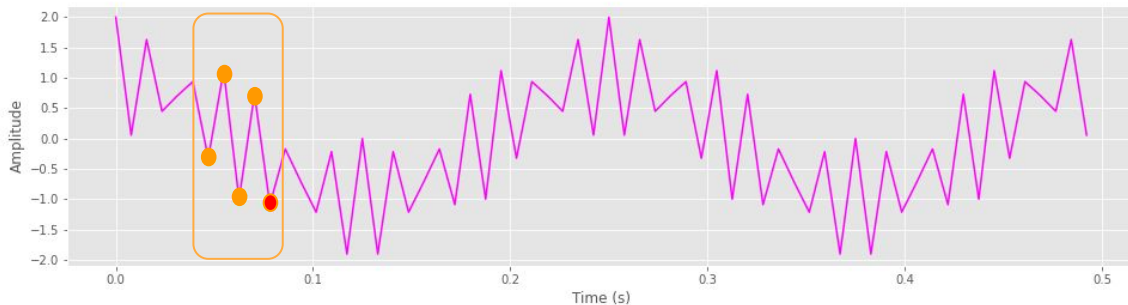
Interpret this as: to calculate the n-th filter output, take a **weighted sum** over the K last elements of the input sequence

Example: Moving average

A moving average is an example of a FIR filter

We define a 5-point moving average filter as:

$$y[n] = \sum_{k=0}^5 \frac{1}{5} x[n - k]$$



So, the 10th output of the filter will be

$$\begin{aligned} y[10] &= \frac{1}{5}x[10] + \frac{1}{5}x[9] + \frac{1}{5}x[8] + \frac{1}{5}x[7] + \frac{1}{5}x[6] \\ &= \frac{x[10] + x[9] + x[8] + x[7] + x[6]}{5} \end{aligned}$$

i.e., the average value of the last 5 points

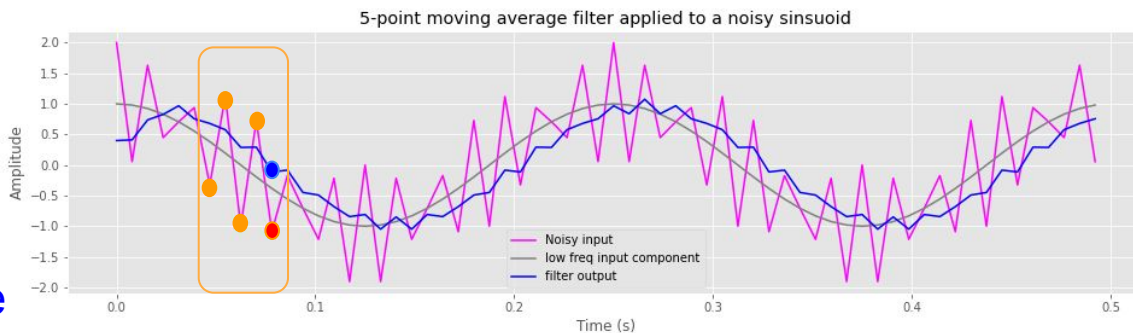
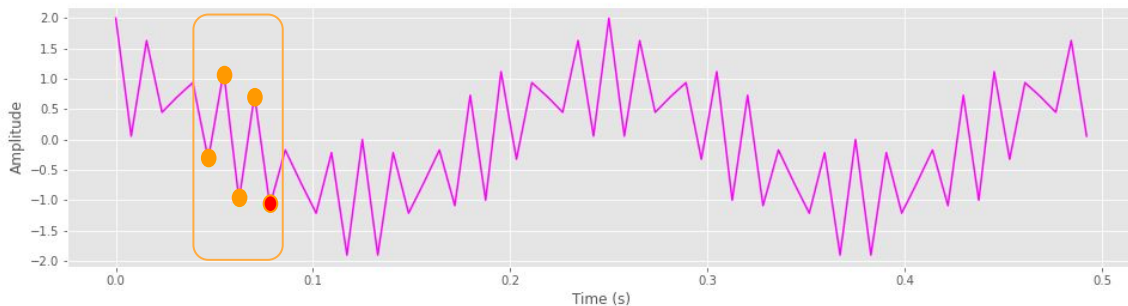
Example: Moving average

A moving average is an example of a FIR filter

We define a 5-point moving average filter as:

$$y[n] = \sum_{k=0}^5 \frac{1}{5} x[n - k]$$

The blue line here represents the output of the filter for each sample in the input sequence



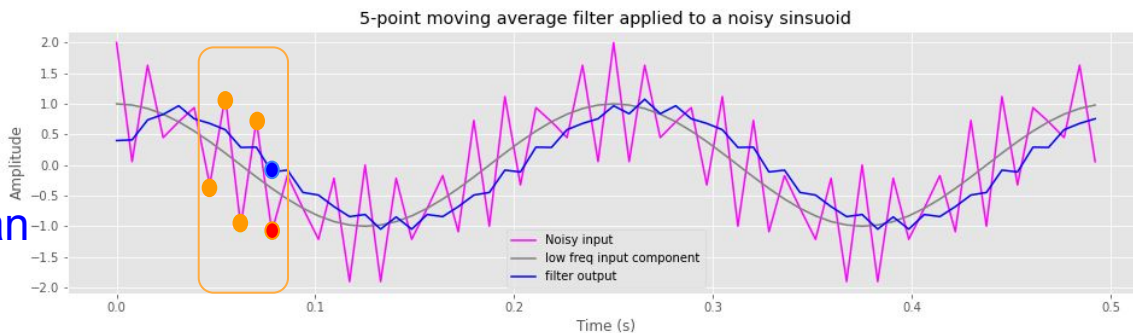
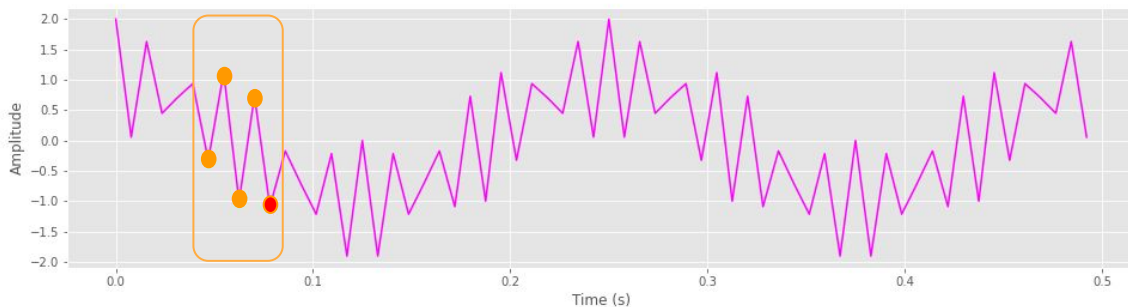
Example: Moving average

A moving average is an example of a FIR filter

We define a 5-point moving average filter as:

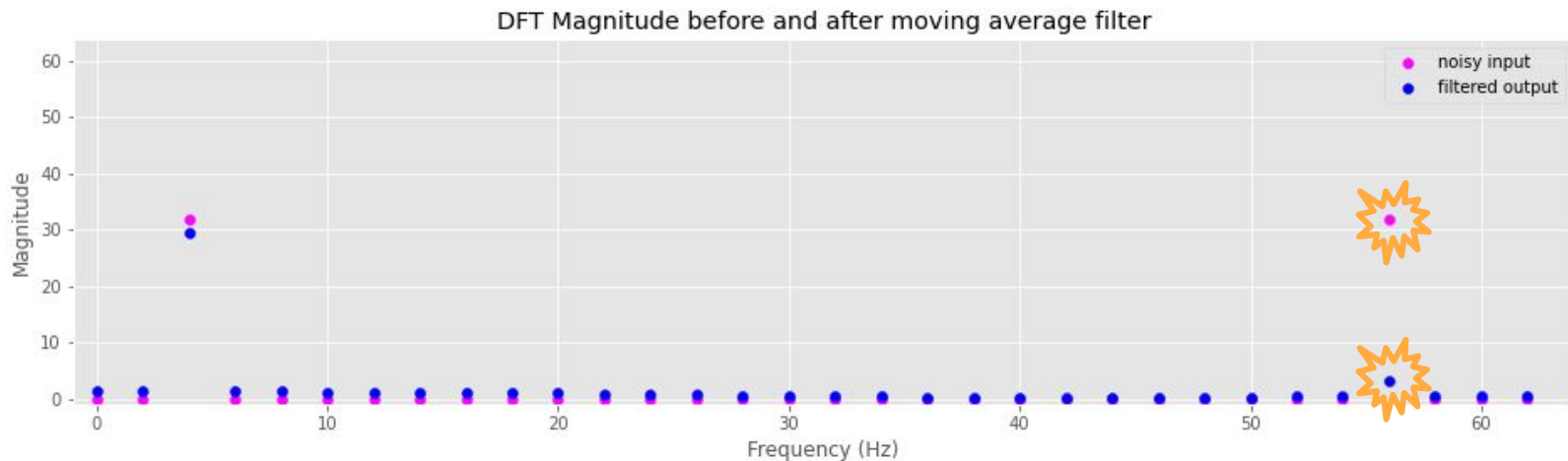
$$y[n] = \sum_{k=0}^5 \frac{1}{5} x[n - k]$$

You can see it's a lot smoother than the original input: the filter has reduced the effect of the higher frequency component in the input



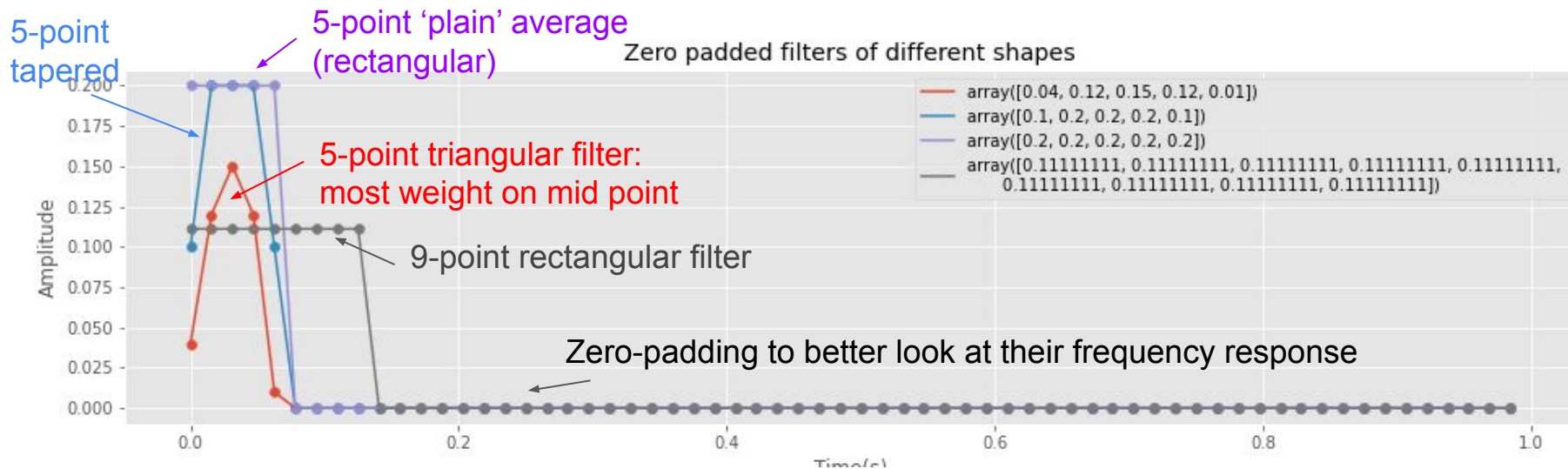
Moving average spectrum

We can see the reduction in the high frequency component in the (magnitude) spectrum of the smoothed (filtered) waveform



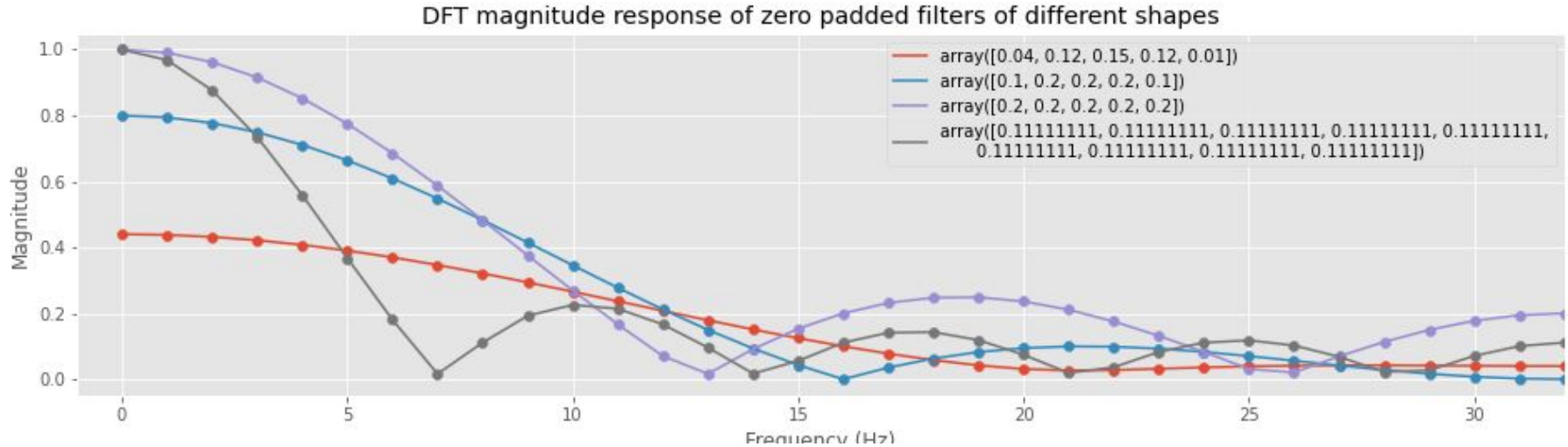
FIR filters as a weighted sums

We can set the coefficients (i.e. 'weights') to different values and visualize them by plotting their values.



FIR filters as a weighted sums

We can also apply the DFT to the sequence of coefficients to see what their effect will be on different frequencies (zero padding to make the effect clearer).



All low pass filters, but with different frequency cut-offs and effects in the higher frequencies

FIR filters as a weighted sums

- These FIR filters all do essentially do the same thing: dampen high frequencies → low pass filter.
- Coefficient (i.e., weight) choice can change which frequencies are dampened
- FIR filters are generally used for smoothing/noise removal and windowing, but not what we want for approximating vocal tract resonances
- They are **finite** because if you apply the filter to a single impulse, you get a finite output in time (outputs drop off to zero if you keep applying it) .

This contrasts with...

Infinite Impulse Response (IIR) Filters

Weighted sum of previous **outputs**

$$y[n] = \sum_{k=0}^K b[k]x[n-k] + \sum_{l=1}^L a[l]y[n-l]$$

Weighted sum of previous **inputs**

- Unlike FIR filters, IIR filters also include previous filter outputs $y[n-1]$ in determining the current output $y[n]$
- This can theoretically lead to an infinite signal in the time domain, hence the name!
- The number of previous inputs (K) and previous outputs (L) can differ

IIR filters

- Can specify low pass filters with less coefficients than FIR filters
- Are better at modelling types of resonances we see in speech
- It's NOT obvious what an IIR filter does from just looking at the coefficients

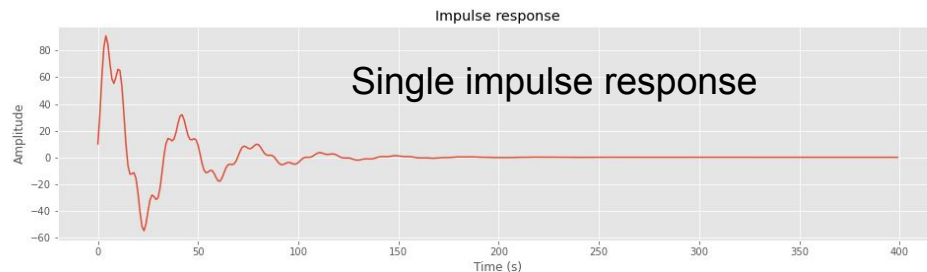
IIR filter example

$$y[n] = x[n] + 3.2y[n - 1] - 4.4y[n - 2] + 3.0y[n - 3] - 0.9y[n - 4]$$

Actual filter coefficients:

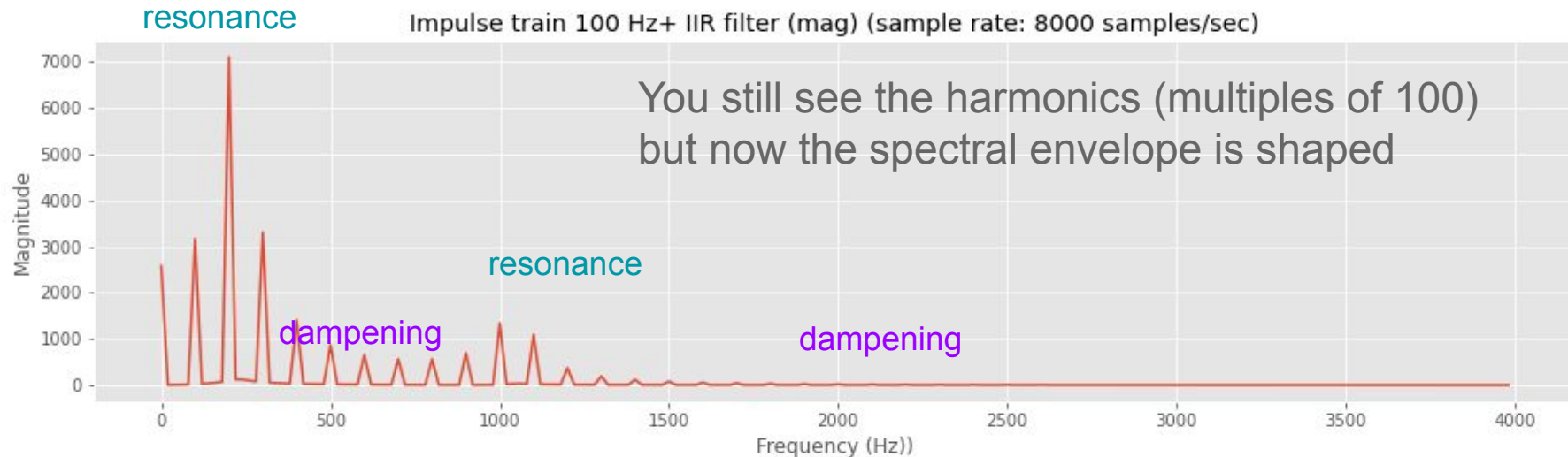
$\mathbf{a} = [3.22666099, -4.3967485,$
 $3.03596532, -0.88529281]$

$\mathbf{b} = [1.0]$



IIR filter + impulse train source

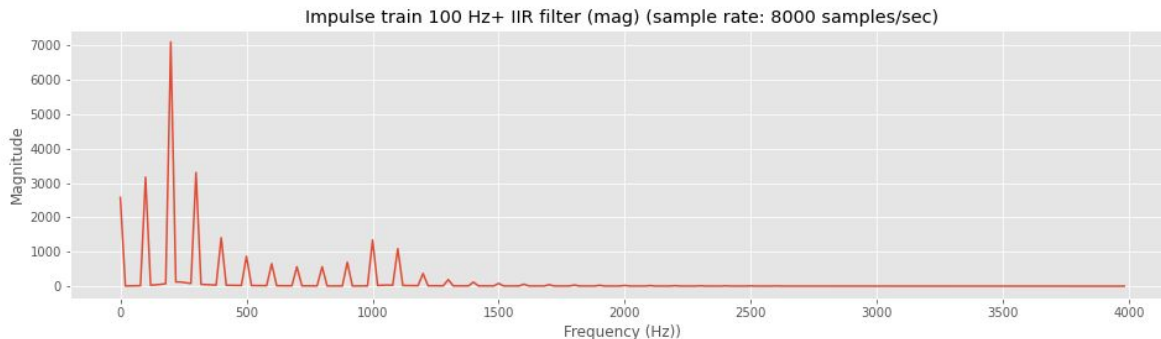
Let's apply this filter to an 100 Hz impulse train:



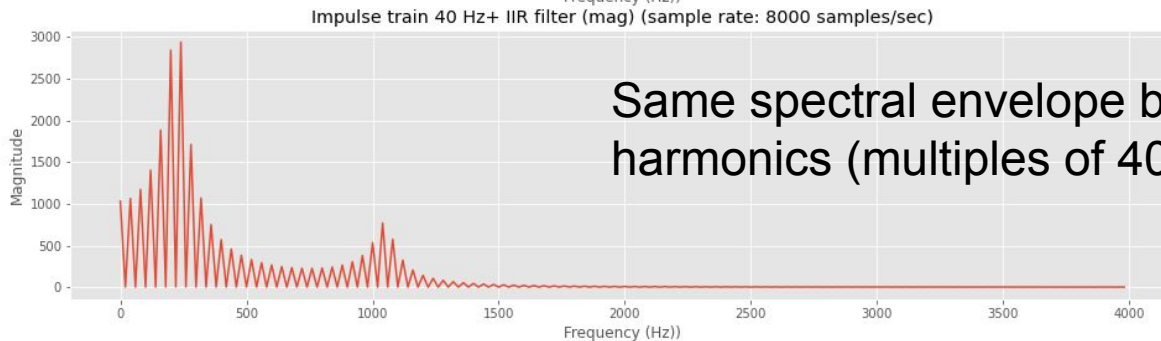
Changing F0 of the impulse train

You can see the spectral envelope is the same but the harmonics are different

100 Hz
impulse
train



40 Hz
impulse
train



Same spectral envelope but more
harmonics (multiples of 40 Hz)

Source-Filter model

$$y[n] = x[n] + \sum_{l=1}^L a[l]y[n-l]$$

Current value of
impulse train

Source

Weighted sum of
previous outputs

Filter

We can now write down a mathematical equation for the source-filter model!

Source-Filter model

$$s[t] = e[t] + \sum_{k=1}^K a[k]s[t - k]$$

“excitation”

Source

“speech signal”

Filter

You’ll probably see it expressed like this but it’s just a change in notation!

Source-Filter model

$$s[t] = e[t] + \sum_{k=1}^K a[k]s[t - k]$$

“excitation”

Source

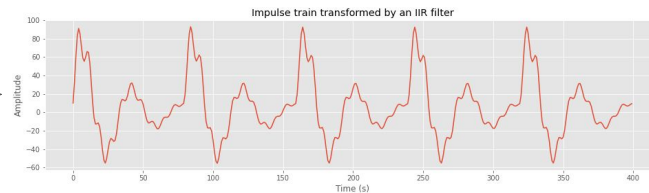
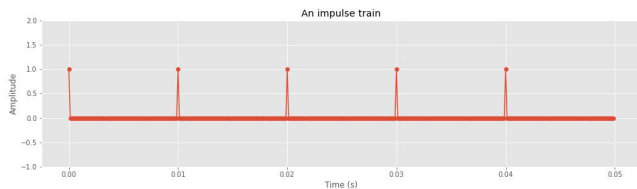
“Speech signal”

Filter

The **excitation** here could be turbulence (e.g. white noise) instead of impulse train

This allows you to create, for example, fricatives.

Source-filter model



To make the signal speech shaped we want to find the right coefficients such that we create the right spectral envelope for the phone, i.e. recreate the effect of the vocal tract as a physical filter

Convolution Theorem

Convolution in the time domain is equivalent to multiplication in the frequency domain

Apply filter window
by window in the
time domain



$$h(k) * x(n) = H(m) \cdot X(m)$$



Alter the spectral
envelope directly in
the frequency domain

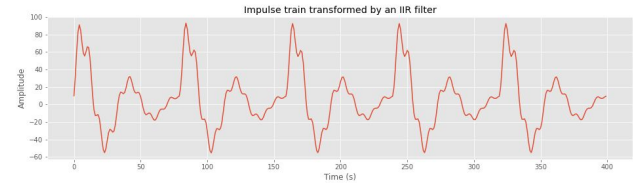
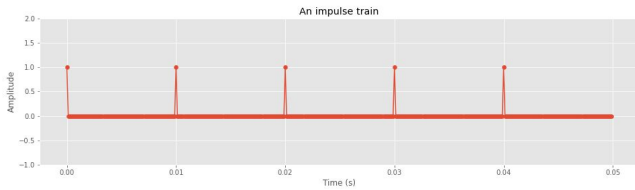
Convolution
In the time domain

Multiplication in the
frequency domain

Convolution in the time domain...

“Convolution of (filter) h and (input) x ”

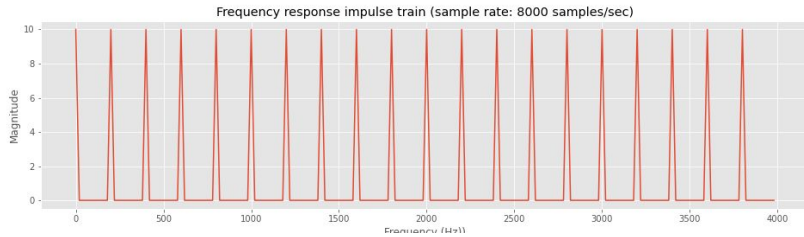
$$h(k) * x(n)$$



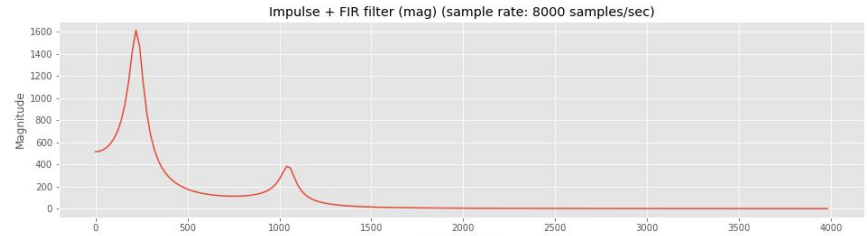
$$y[n] = x[n] + \sum_{l=1}^L a[l]y[n-l]$$

In the time domain: apply the filter to a moving window

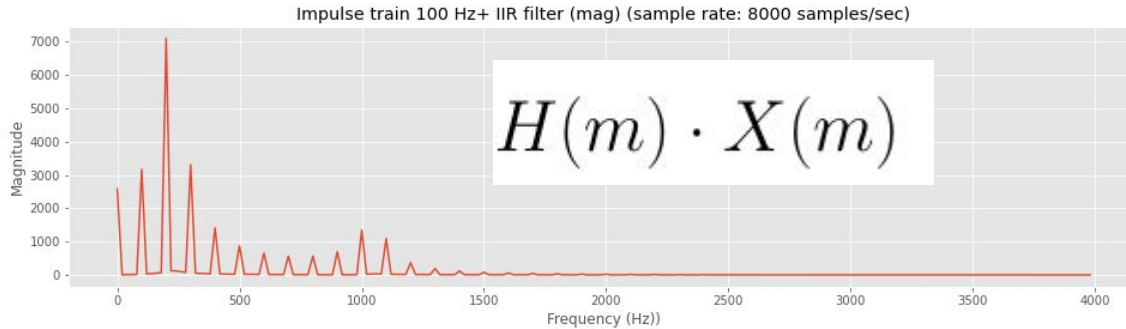
...Multiplication in the frequency domain



X

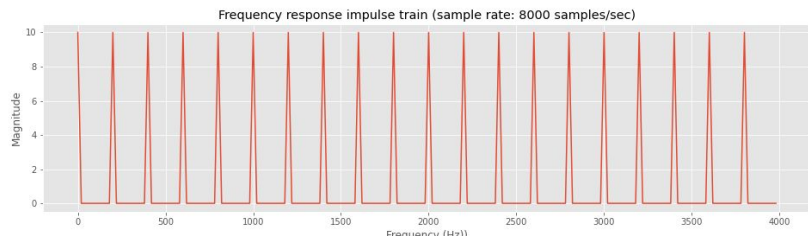


=

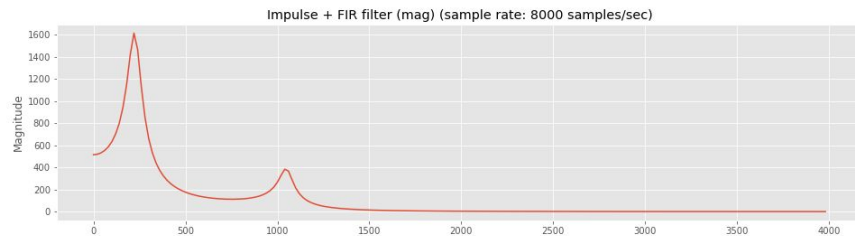


In the frequency domain, we can just multiply the spectrums together at each frequency! Neat!

Magnitude spectrum of the source:
F0 and harmonics

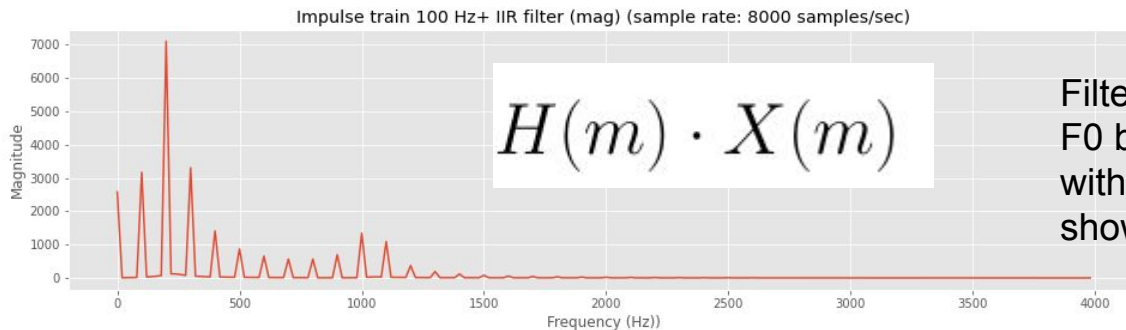


Magnitude spectrum of the filter:
Vocal tract resonance structure
I.e. formants



X

=



Filter applied to source:
F0 based harmonics,
with spectral envelope
showing resonances

Changing the impulse train changes F0 and harmonics

Changing the filter changes the spectral envelope, i.e. resonant/dampened frequencies

Key Points

- We can model speech production in terms of a **source** (vocal folds) and **filter** (vocal tract)
- **Resonances** of the vocal tract depend on vocal tract constriction place, the size of the opening, and vocal tract length
- Resonances of the vocal tract are (more or less) independent of F0
- **Formants** (peaks in the spectral envelope) correlate with resonances of the vocal tract. These change depending on vocal tract constrictions.

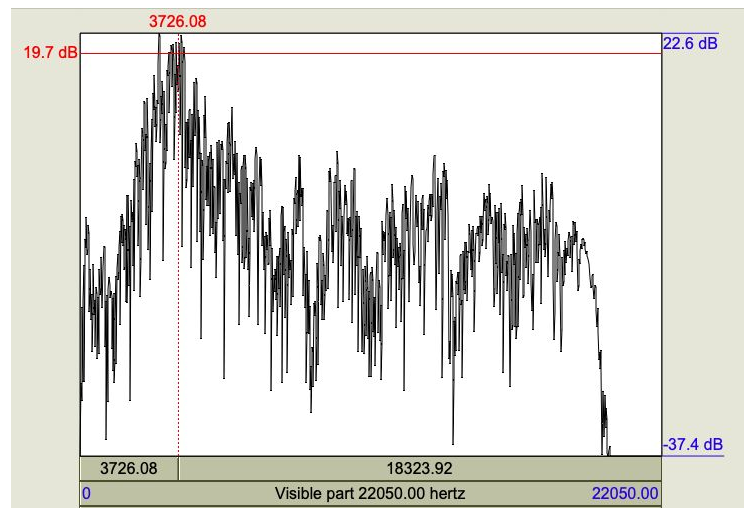
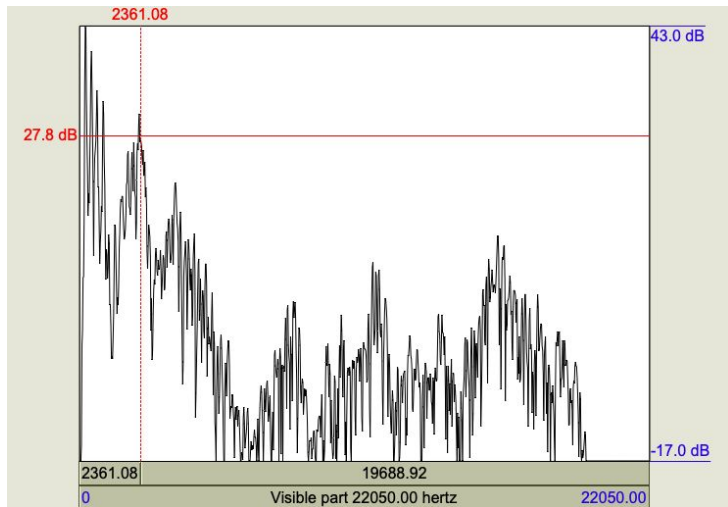
Key Points

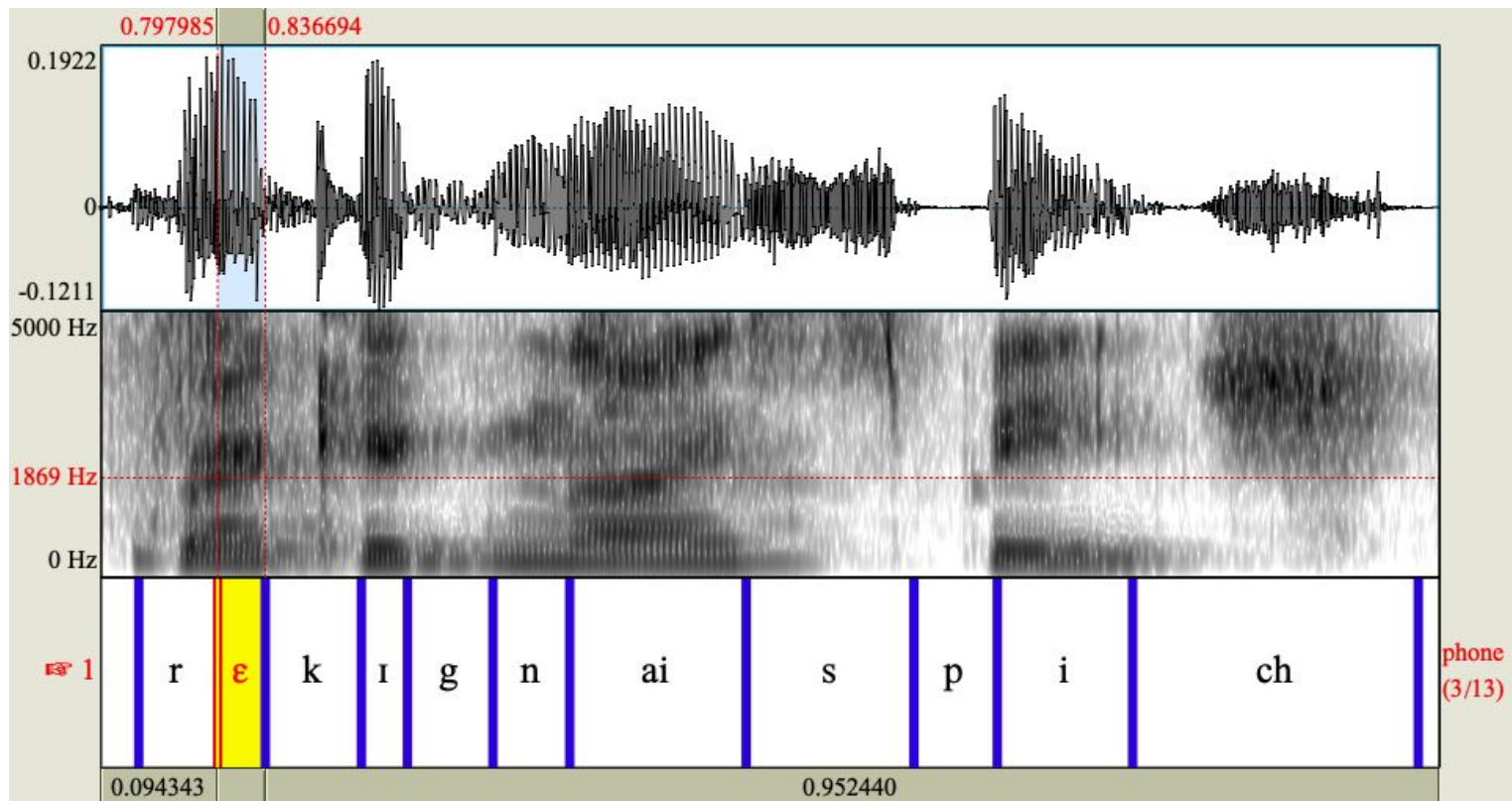
- We can model the source as an **impulse train**, where the frequency of the impulses corresponds to F_0
- The DFT of an impulse train shows **harmonics** (multiples of F_0) with equal magnitude
- We can use **FIR** and **IIR filters** to shape the **spectral envelope** of an impulse train, approximating the vocal tract filter
- We can write down a mathematical version of the source filter model in terms of a **difference equation** (applying an IIR filter to an impulse train)

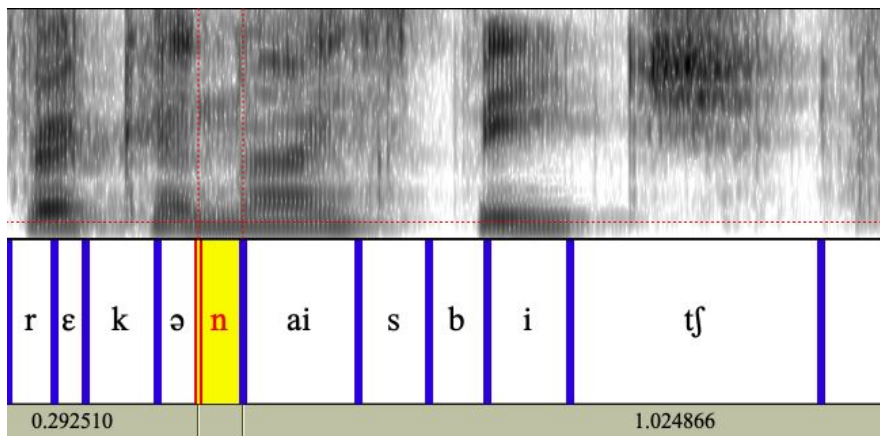
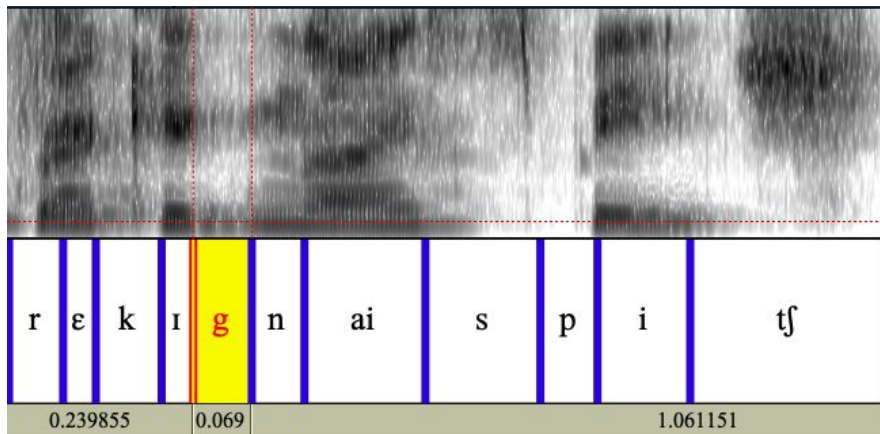
Key Points

- **Convolution** in the time domain is equivalent to multiplication in the frequency domain.
- So if we know the **frequency response** of the filter and the input, we can much more easily see what effect it will have on the input working in the frequency domain.

Extra slides (out of context)

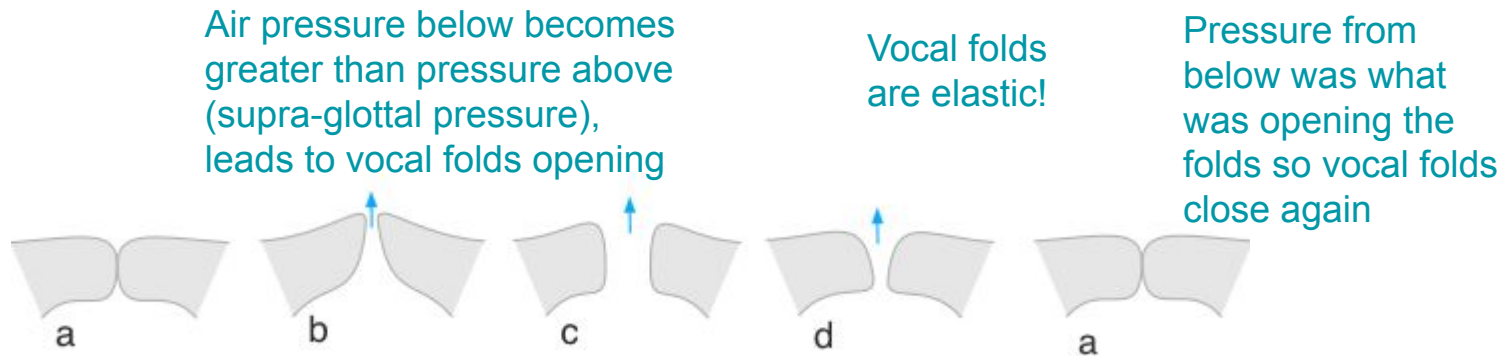






Vocal fold movements

Vocal folds opening and closing drive oscillations in pressure in the vocal tract



Air pressure below becomes greater than pressure above (supra-glottal pressure), leads to vocal folds opening

Vocal folds are elastic!

Pressure from below was what was opening the folds so vocal folds close again

Vocal folds closed (adducted)

Air pressure builds up under the lungs (sub-glottal pressure)

Vocal folds open

Increase in speed through the folds leads to reduction in pressure

Other types of phonation/airstreams

Voice quality

- Whispered
- Breathy
- Creaky
- Harsh

Try making a voiced sound while inhaling!
(e.g. Swedish agreement marker)

Non-pulmonic airstreams

- Glottalic Egressives → Ejectives
(e.g., Navajo)
- Glottalic Ingressive → Implosives
(e.g., Igbo)
- Velaric ingressive → Clicks
(e.g., Xhosa)

MRI/Ultrasound demos:

<https://seeingspeech.ac.uk/ipa-charts/?chart=2&datatype=1&speaker=1>

Bottle resonance: Helmholtz resonator

- Turbulence at the top of the bottle increases air pressure in the bottle → **compression**
- Air springs back after compression → **oscillation**
- Particles oscillating, some frequencies are amplified → **resonance**
- Size of the bottle determines the resonant frequency

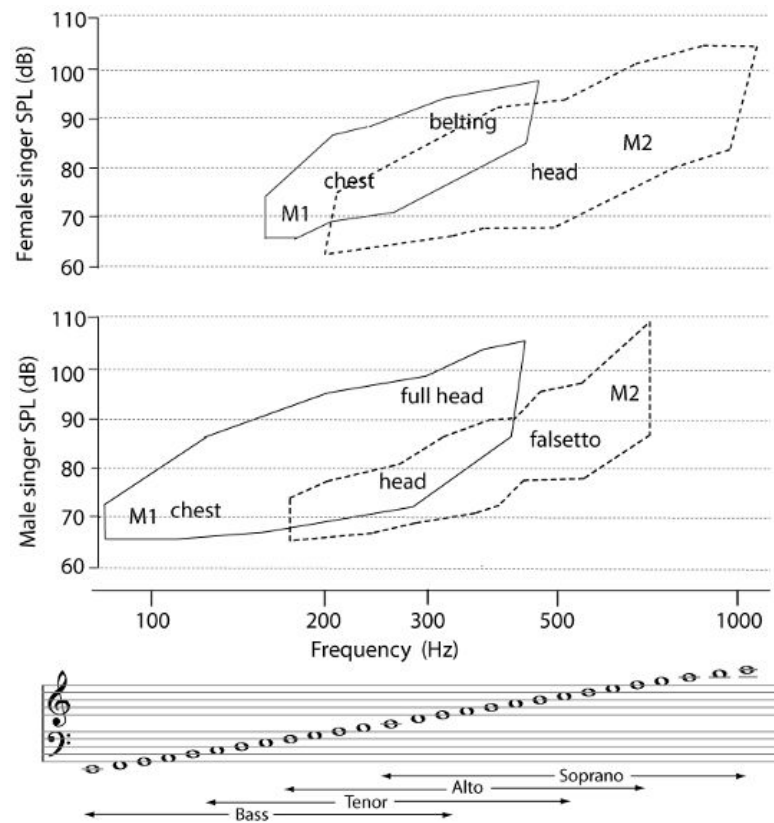


Very different to voiced speech!

Vocal mechanisms

- M0: Creak
- M1: Chest voice
- M2: Head voice ('falsetto')
- M3: Whistle

Usually use 'chest' and 'head' voice for speech, but creaky voice is also common (and contrastive) in some languages



Garnier, M., Henrich, N., Smith, J. and Wolfe, J. (2020) "[The mechanics and acoustics of the singing voice: registers, resonances and the source-filter interaction](#)", in *The Routledge Companion to Interdisciplinary Studies in Singing, Volume I:*

Resonance

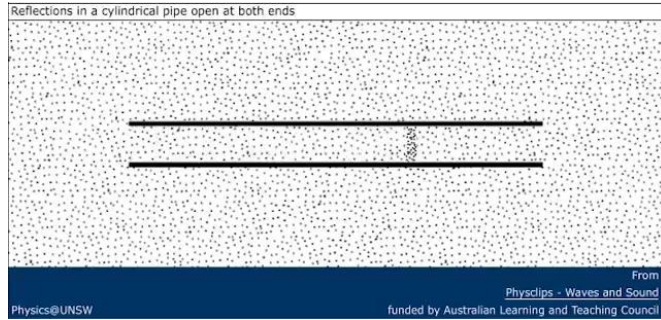
A resonant frequency is a natural frequency of vibration determined by the physical parameters of the vibrating object.



Tacoma bridge
collapse

Waves in tube

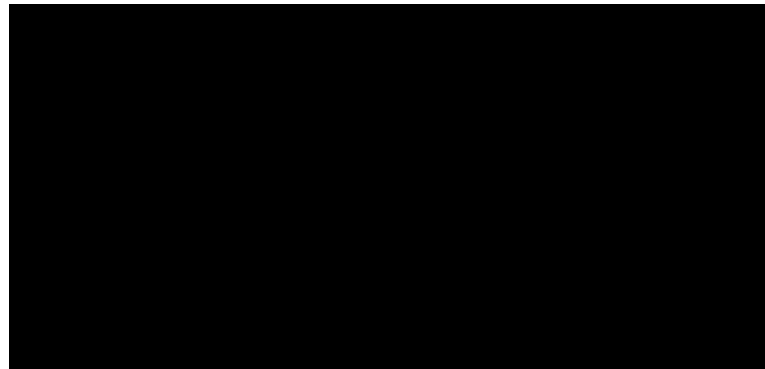
Open at both ends



Closed at both ends



Open at one end



Three tube model, e.g. [i]

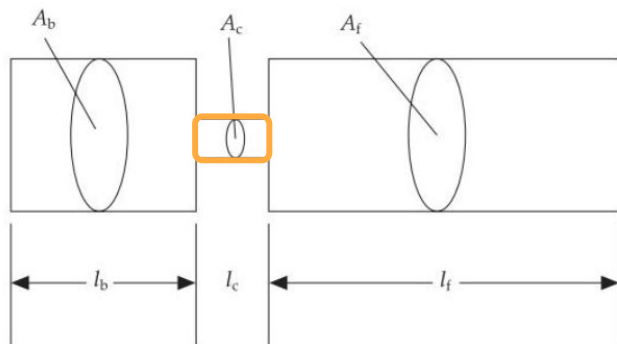


Figure 6.3 Tube model of vocal tract configurations that have a short constriction at some point in the vocal tract.

F1 and F2 change depending of the relative length of the tubes

+ **Helmholtz resonance** due to the small connecting tube

Figures from: Johnson, K (2012). Acoustic and Auditory Phonetics, Chapter 6

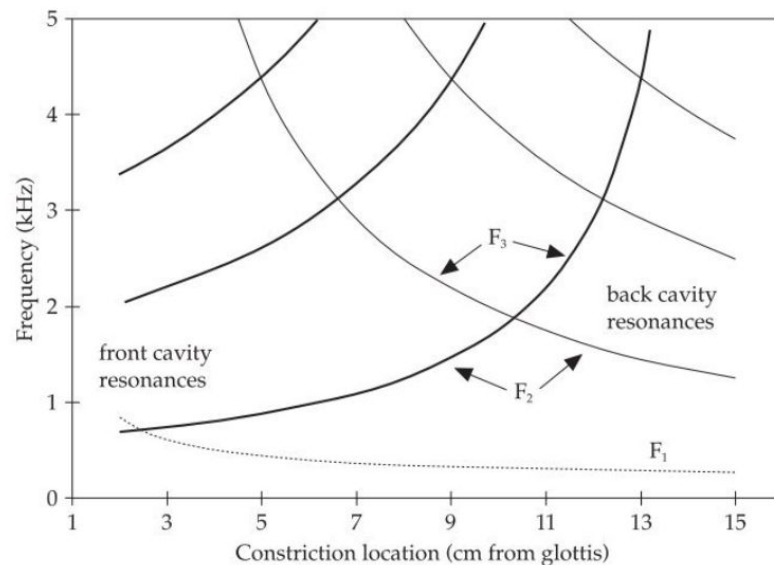


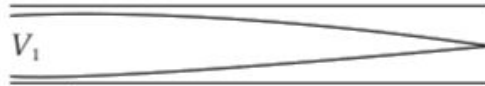
Figure 6.4 Resonant frequencies of the back tube (light lines), front tube (heavy lines) and Helmholtz resonance (dashed line) in the tube model illustrated in figure 6.3. Frequencies are plotted as a function of different back tube lengths (l_b), with the length of the constriction fixed at 2 cm and the total length of the model fixed at 16 cm.

Perturbation Theory

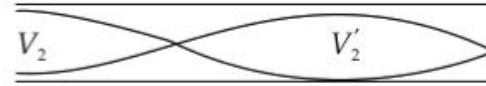
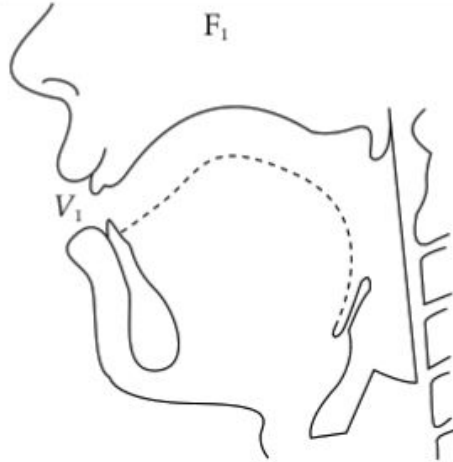
- Claim: Formant value depend on how close the constriction is to velocity nodes and antinodes
- **Velocity Nodes:** Points of maximum pressure (zero velocity)
 - Constriction increases velocity, reduces pressure
 - Raises formant frequency (for the formant related to that node)
 - Increasing velocity increases frequency
- **Velocity Antinodes:** points of zero pressure (max velocity)
 - Constriction increases pressure, reduces velocity
 - Lowers formant frequency (for the formant related to that node)
 - Slowing down the velocity decreases frequency (time it takes to complete a cycle increase)

Perturbation Theory

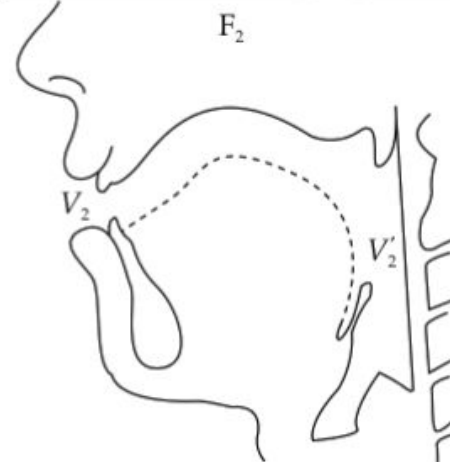
Max
velocity



F_1

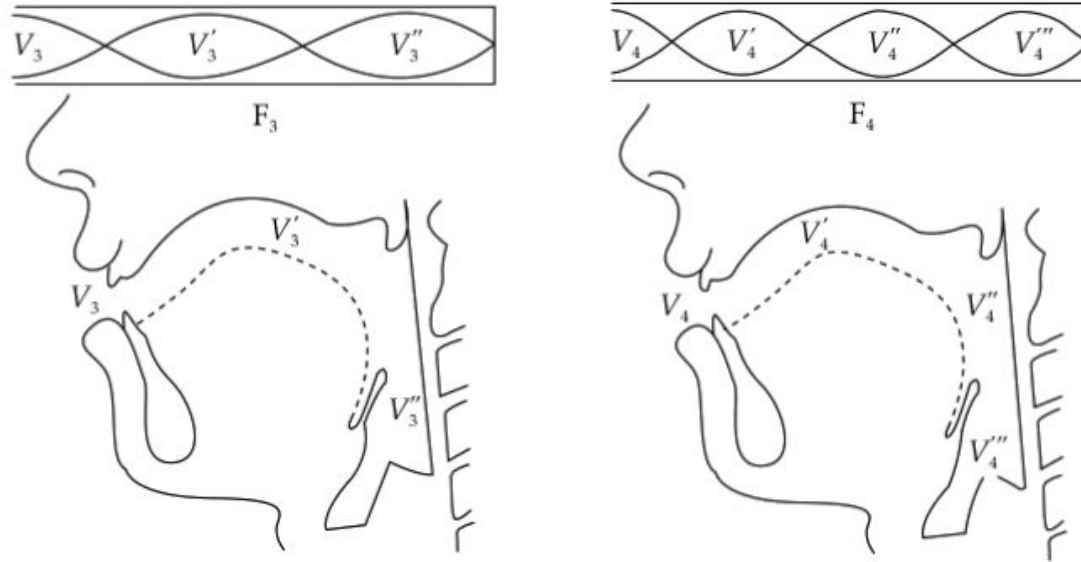


F_2



Velocity **antinodes** for 1st and 2nd resonance

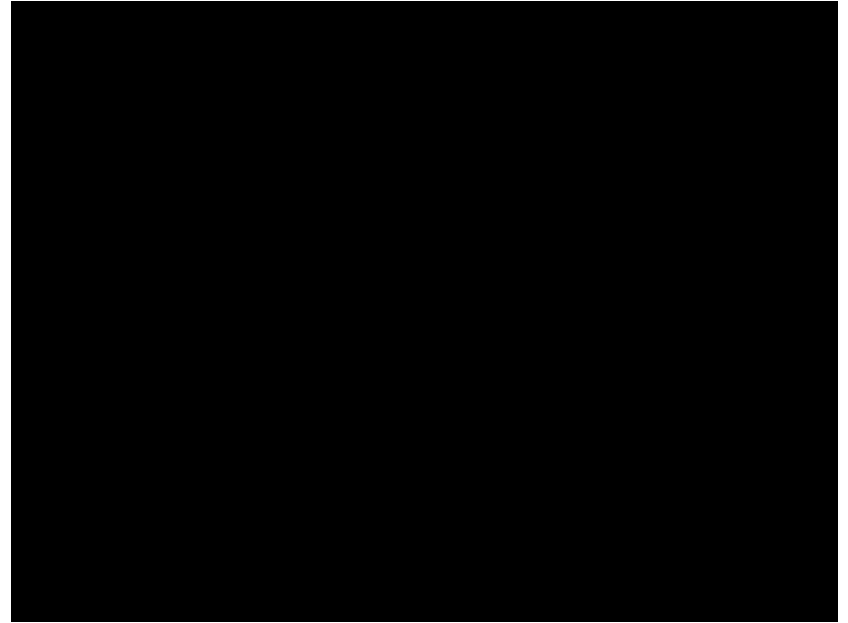
Perturbation Theory



Velocity **antinodes** for 3rd and 4th resonances

Vocal mechanisms

- M0: Creak
- M1: Chest voice
- M2: Head voice ('falsetto')
- M3: Whistle



From PhysClips: <https://newt.phys.unsw.edu.au/jw/voice.htm>

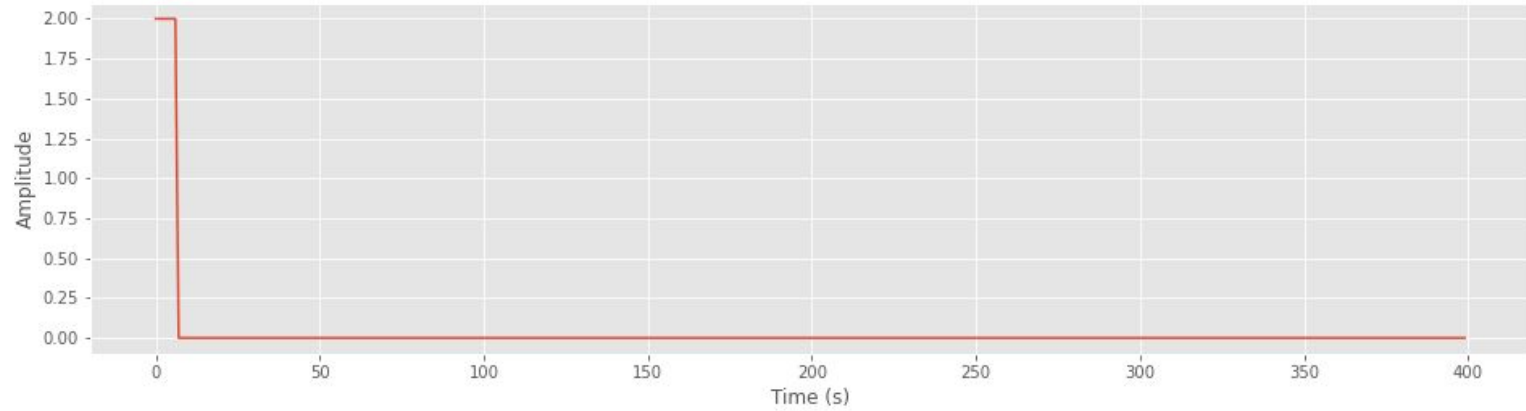
FIR filters

$x[n], \dots, x[n-K]$ are the previous K samples of the input

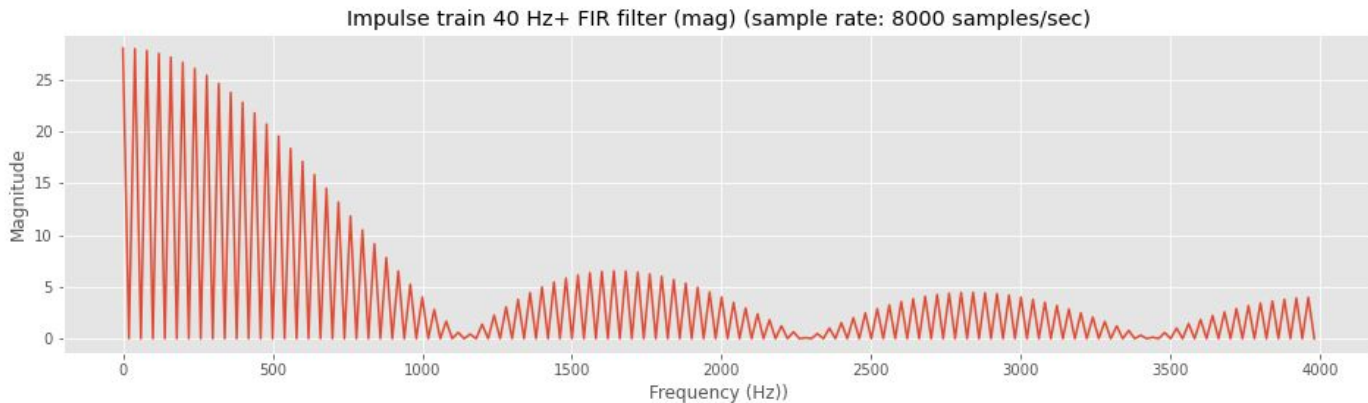
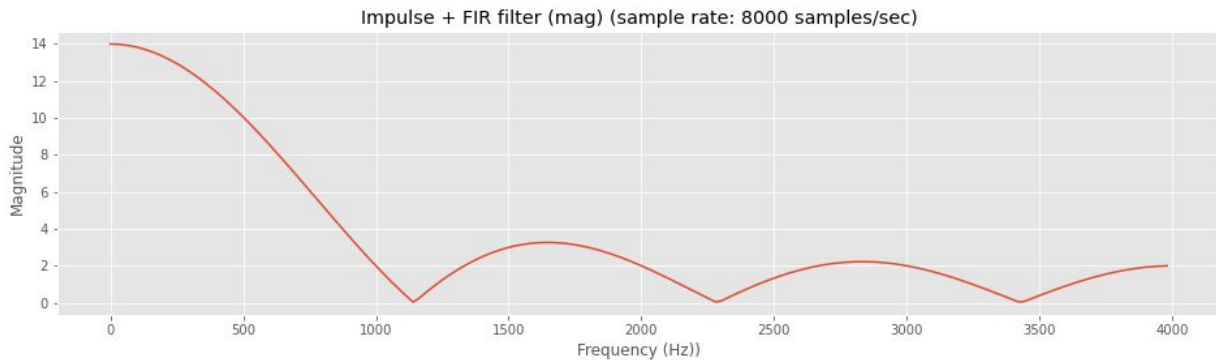
$$y[n] = b[0]x[n] + b[1]x[n-1] + \dots + b[K]x[n-K]$$
$$= \sum_{k=0}^K b[k]x[n-k]$$

- $x[n]$ represents the n th sample of the input
- $y[n]$ represents the n th output of the filter
- $b[0], \dots, b[K]$ represent the **filter coefficients**
 - **Think of them as weights**

Rectangular window



Spectral envelope of a rectangular filter



Convolution Theorem



The way we have been applying the filters with the difference equations is a form of **convolution**: stepping through the sequence, applying the filter at each step.

$$s[t] = e[t] + \sum_{k=1}^K a[k]s[t - k]$$

This means we need to make a lot of calculations to apply a filter in the time domain!

Luckily for us, we have the convolution theorem: Convolution in the time domain is equivalent to multiplication in the frequency domain

Speed matters!

- Normal speech (in air) 
- Speech in helium 

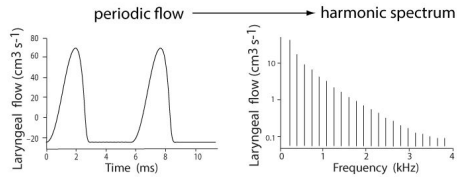
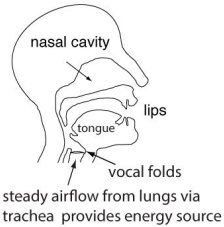
Speed of sound is greater in helium so resonances occur at higher frequencies

Not a difference in pitch, but a difference in **timbre**

Vowels are harder to identify in isolation in helium speech

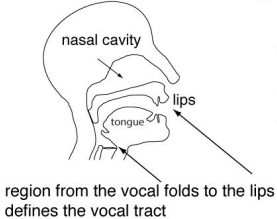
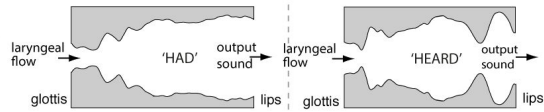
SOURCE

The vocal folds undergo auto-oscillation and produce a pulsed laryngeal flow through the glottis, the oscillating gap between the folds

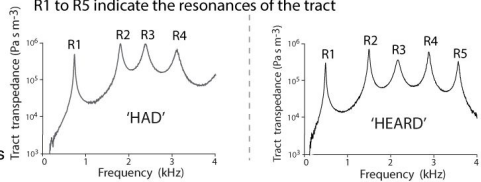


The periodic laryngeal flow then enters the downstream vocal tract
Two different configurations show how the radius varies with distance along the tract. They correspond to the vowels in 'had' and 'heard'.

FILTER

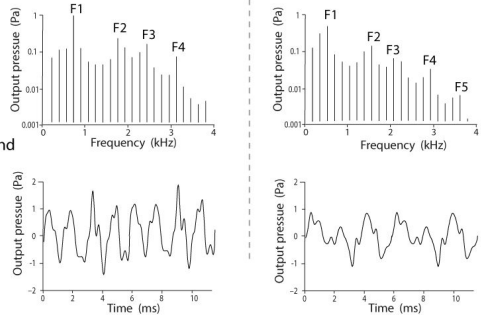
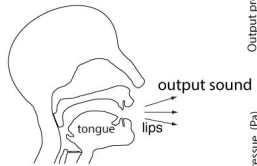


The 2 vocal tract models have the measured transpedances shown below.
R1 to R5 indicate the resonances of the tract



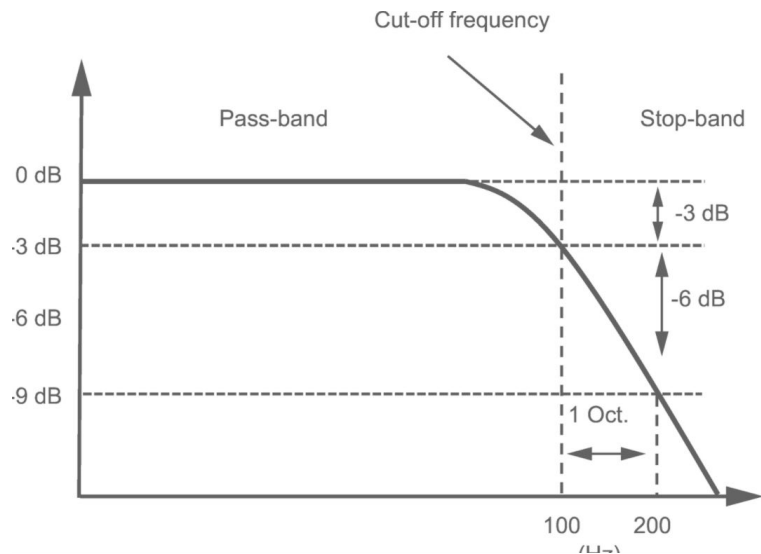
In a linear system the output sounds are the product of the source function and the filter function and will have the pressure spectra and waveforms shown below

OUTPUT SOUND

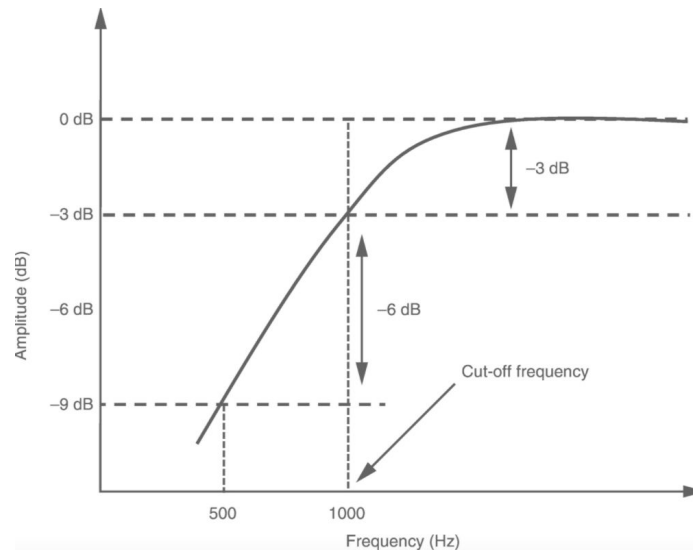


["An experimentally measured Source-Filter model: glottal flow, vocal tract gain and radiated sound from a physical model,"](#) Wolfe, J., Chu, D., Chen, J.-M. and Smith J. (2016) *Acoust. Australia*, **44**, 187–191

Low pass and high pass filters

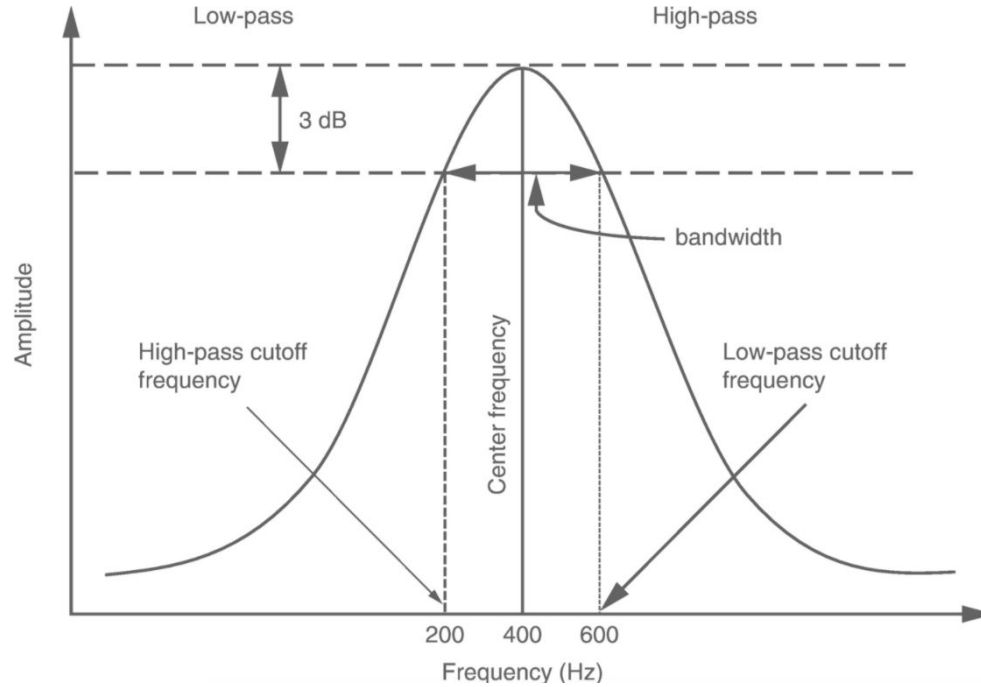


Low pass



High pass

Band-pass filter



An equivalent formulation of the DFT using sines and cosines

Discrete Fourier Transform

A mathematical procedure we can use to determine the frequency content of a discrete signal sequence

Mathematical view: for input $x[n]$ with $n=0, \dots, N-1$ (N inputs)

$$\text{DFT}[k] = \sum_{n=0}^{N-1} x[n] \left[\cos\left(\frac{2\pi n}{N}k\right) - j \sin\left(\frac{2\pi n}{N}k\right) \right]$$

For $k=0, \dots, N-1$ (N analysis frequencies)

Derived from Euler's Formula