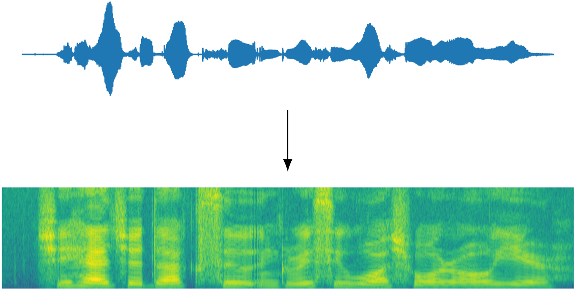
Speech Signal Analysis 2

Hao Tang

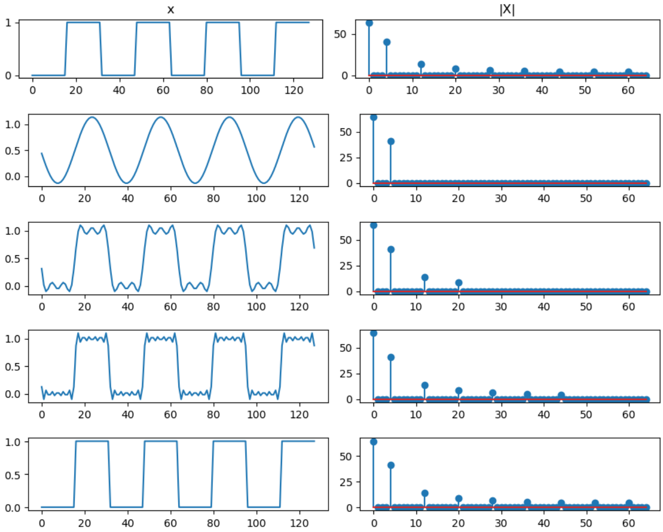
Automatic Speech Recognition—ASR Lecture 3
24 January 2022

Spectrogram

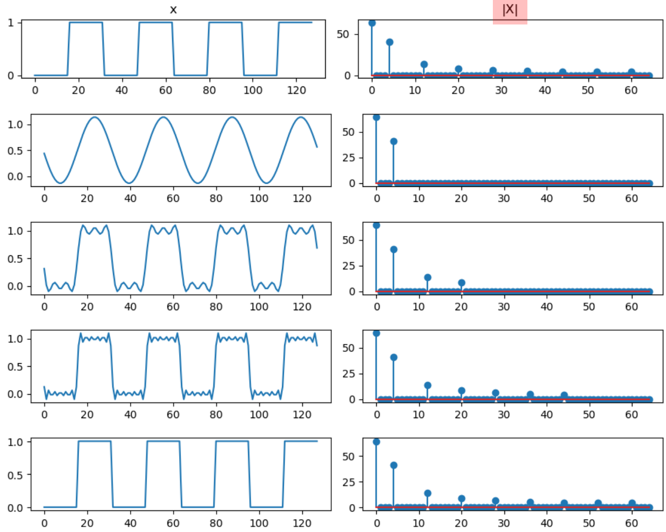


- dithering, removing DC offset, pre-emphasis
- windowing
- Discrete Fourier transform (DFT)
- Short-time Fourier transform (STFT)

Discrete Fourier Transform



Discrete Fourier Transform



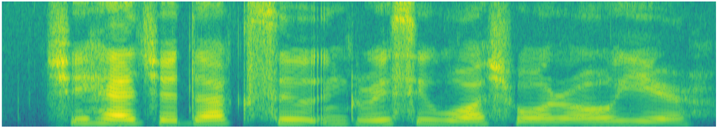
Complex Numbers

$$X[k] = a + bi$$

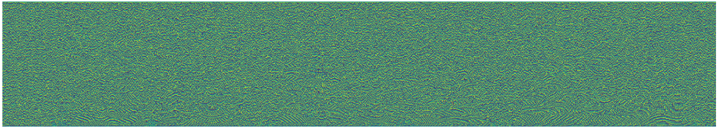
- Real: $\mathfrak{Re}\{X[k]\} = a$
- Imaginary: $\mathfrak{Im}\{X[k]\} = b$
- Magnitude: $|X[k]| = \sqrt{a^2 + b^2}$
- Phase: $\angle X[k] = \arccos \frac{a}{\sqrt{a^2+b^2}}$
- Energy: $|X[k]|^2$

Spectrogram

Magnitude



Phase



- Spectrogram = Magnitude spectrogram = Power spectrogram
- Phase is not as important as magnitude for speech intelligibility.

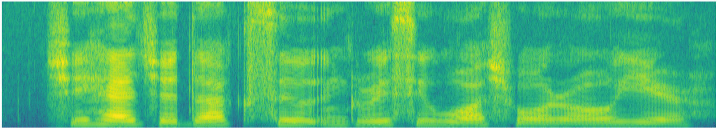
Spectrogram

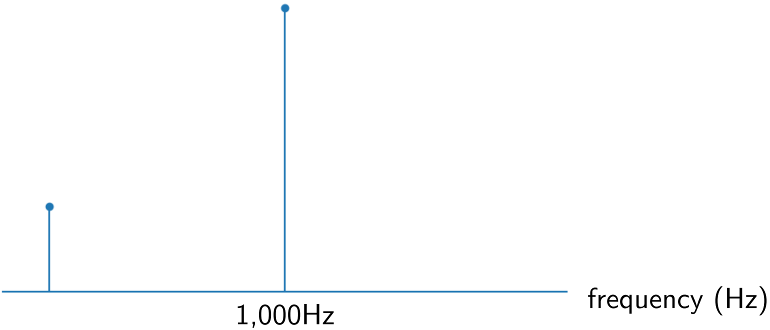
Without log

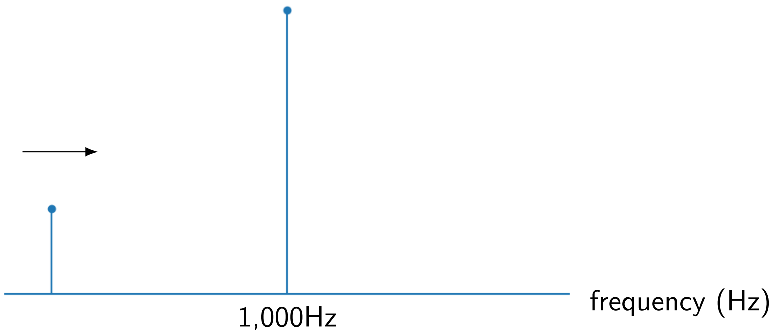
Spectrogram

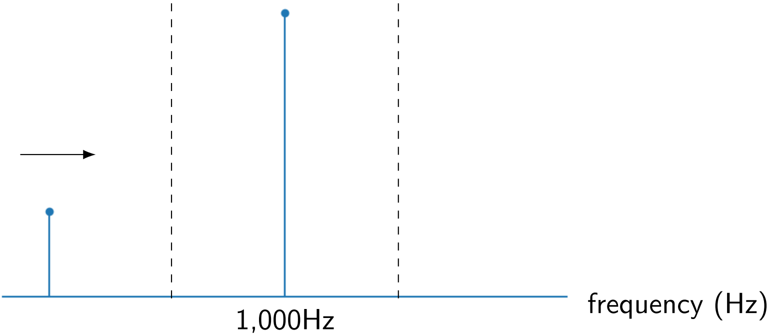
Without log

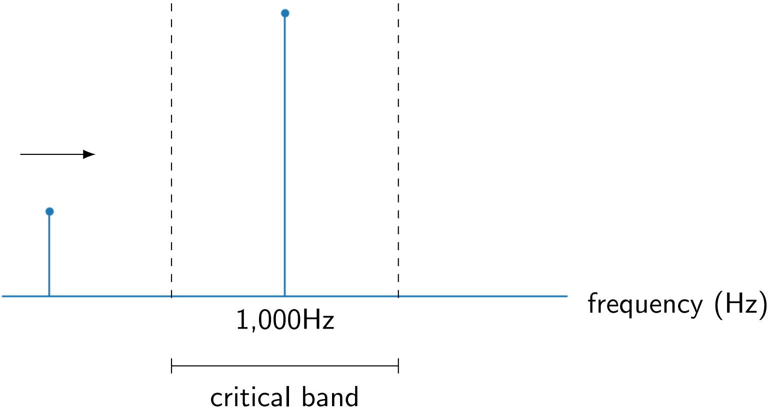
With log




1,000Hz
frequency (Hz)

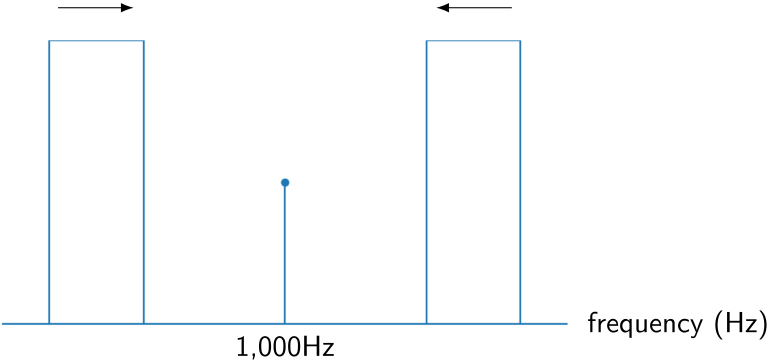

1,000Hz
frequency (Hz)

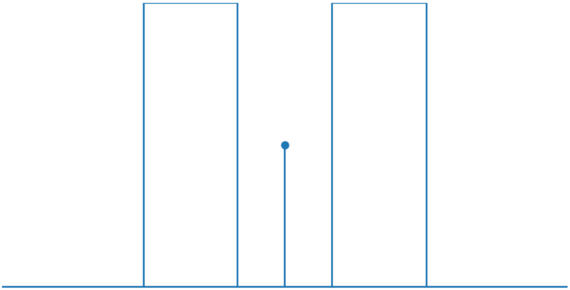

1,000Hz
frequency (Hz)

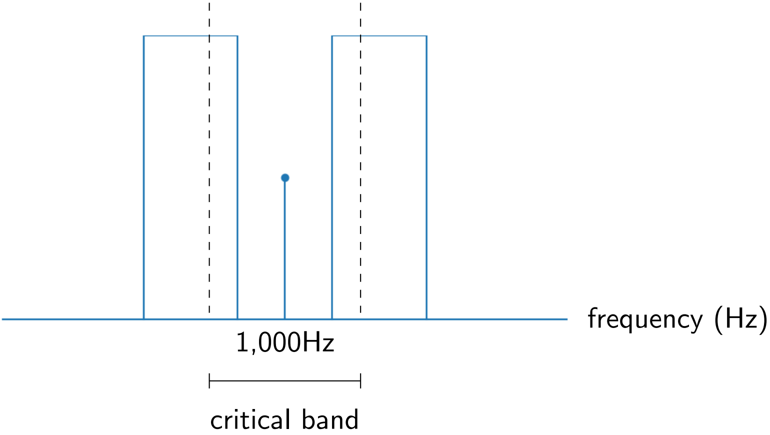

frequency (Hz)
1,000Hz
critical band

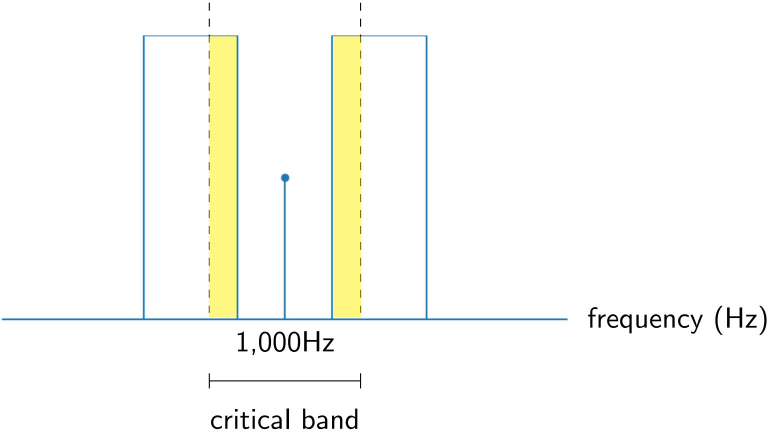
Auditory Masking

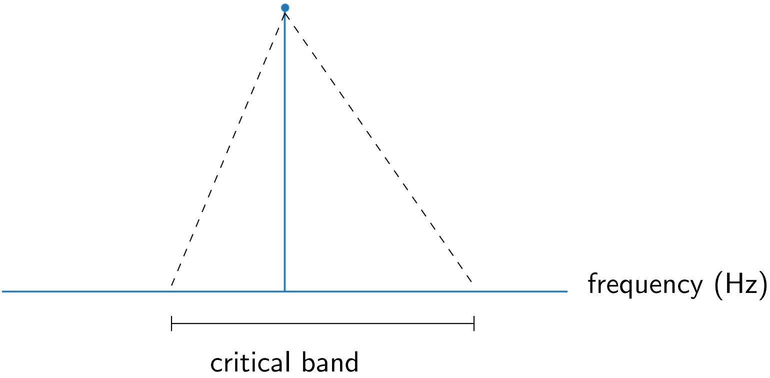
- One sound affects the presense of another sound.
- Both sounds are present, so masking is purely perceptual.
- Masking is a nonlinear effect.
- Many applications take advantage of masking (e.g., MP3).


frequency (Hz)
1,000Hz

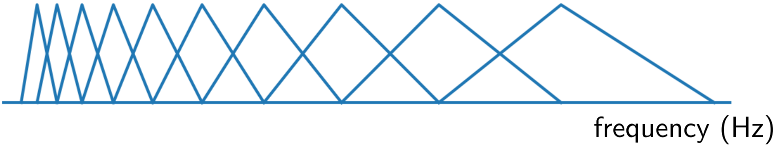

1,000Hz
frequency (Hz)


frequency (Hz)
1,000Hz
critical band


frequency (Hz)
1,000Hz
critical band

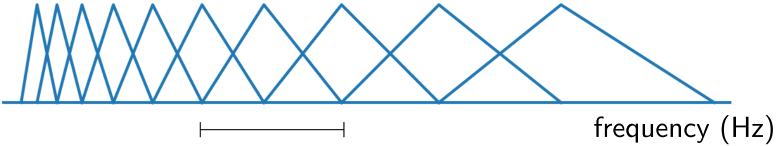

frequency (Hz)
critical band

Auditory Filters



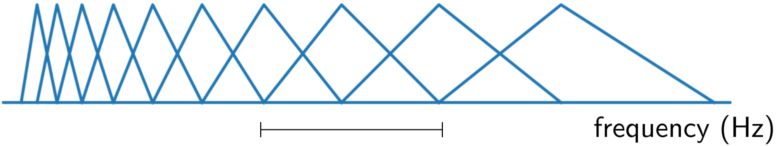
- Triangle-shaped
- Asymmetric
- Sensitive to the amount of energy
- With larger bandwidth at higher frequency

Auditory Filters



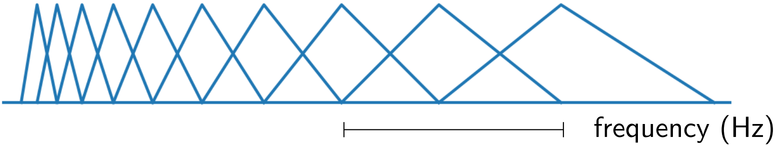
- Triangle-shaped
- Asymmetric
- Sensitive to the amount of energy
- With larger bandwidth at higher frequency

Auditory Filters



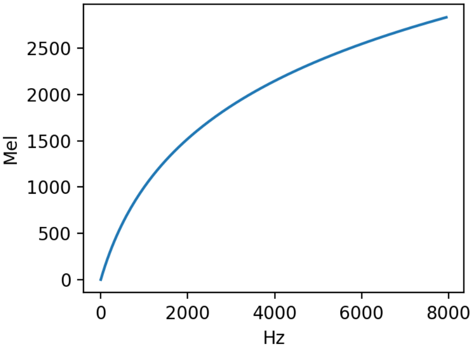
- Triangle-shaped
- Asymmetric
- Sensitive to the amount of energy
- With larger bandwidth at higher frequency

Auditory Filters



- Triangle-shaped
- Asymmetric
- Sensitive to the amount of energy
- With larger bandwidth at higher frequency

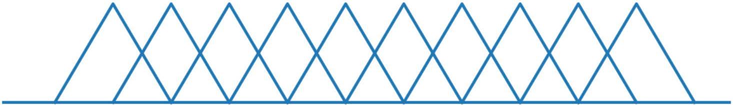
Mel Scale



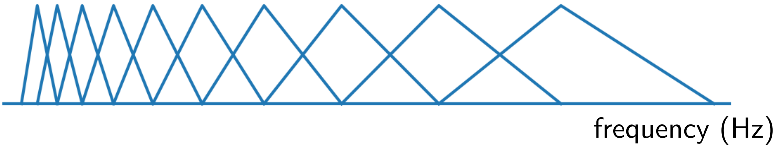
$$m = 1127 \log \left(1 + \frac{f}{700}\right)$$

- 300 Hz vs 310 Hz
- 2000 Hz vs 2010 Hz

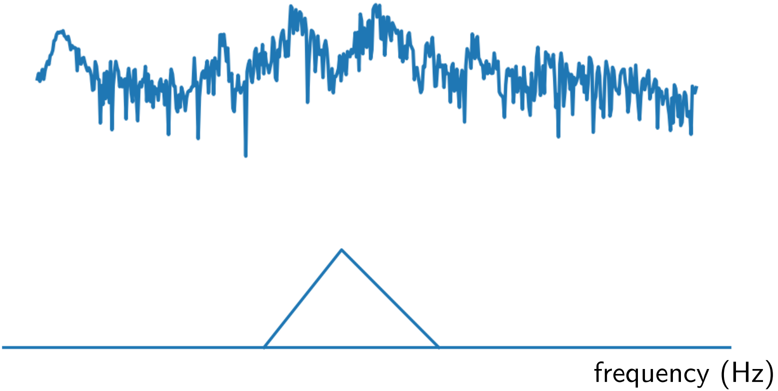
Mel Filters



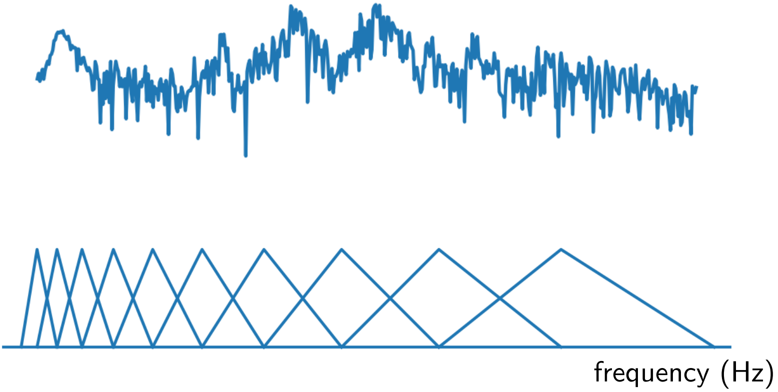
Mel Filters



Mel Filters



Mel Filters



Mel Filters

$$Y[n] = \sum_{k=0}^{T-1} X[k] \cdot H_n[k]$$

- H_n is the n-th Mel filter.
- Mel filters are applied to the magnitude spectrum with dot product.
- The result is an n-dimensional vector for n Mel filters.

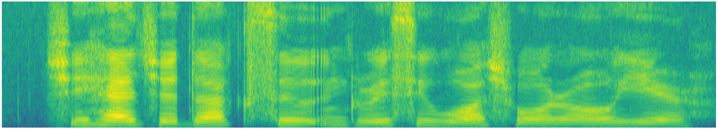
Mel Filters

$$Y[n] = \sum_{k=0}^{T-1} X[k] \cdot H_n[k]$$

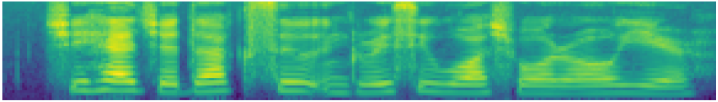
$$Y = \begin{bmatrix} H_1[0] & H_1[1] & \cdots \\ H_2[0] & H_2[1] & \cdots \\ \vdots & \vdots & \\ H_n[0] & H_n[1] & \cdots \end{bmatrix} \begin{bmatrix} X[0] \\ X[1] \\ \vdots \\ X[T-1] \end{bmatrix} = \begin{bmatrix} H_1 \\ H_2 \\ \vdots \\ H_n \end{bmatrix} X = HX$$

Mel Spectrograms

linear spectrogram

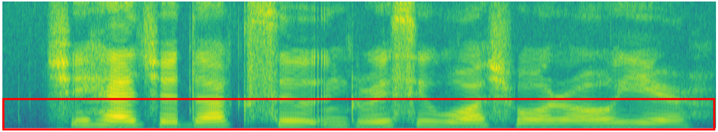


Mel spectrogram

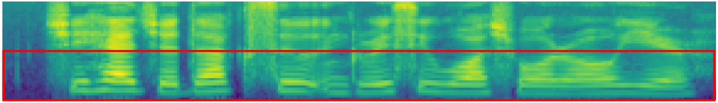


Mel Spectrograms

linear spectrogram

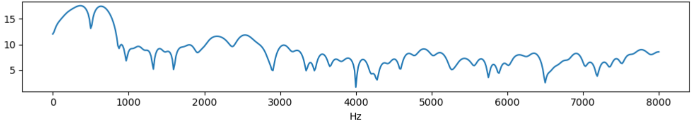


Mel spectrogram

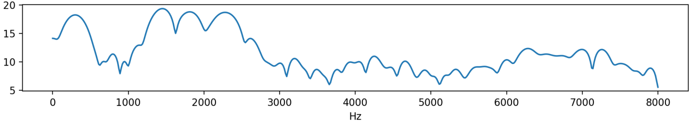


Formants

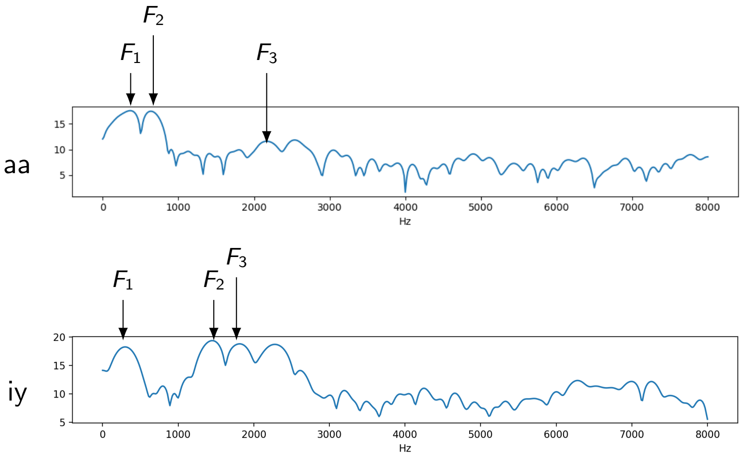
aa



iy

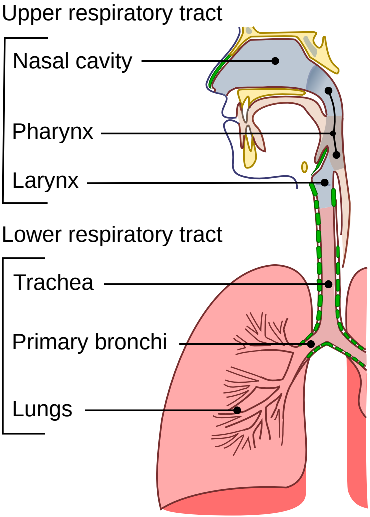


Formants



F_1
F_2
F_3
aa
15
10
5
0
1000
2000
3000
4000
5000
6000
7000
8000
Hz
F_1
F_2
F_3
iy
20
15
10
5
0
1000
2000
3000
4000
5000
6000
7000
8000
Hz

Speech Production



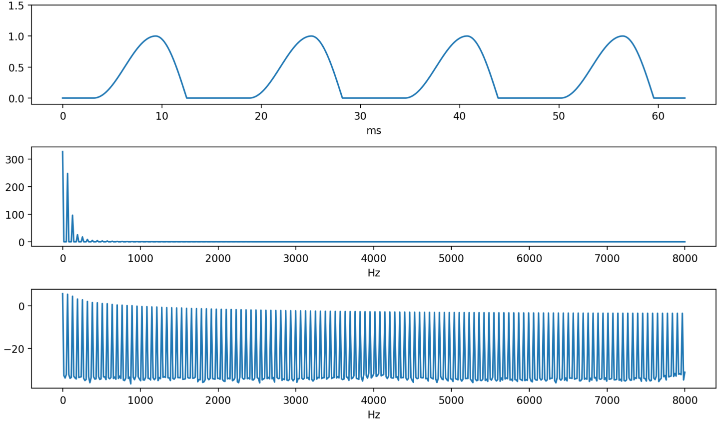
;

Vocal Fold

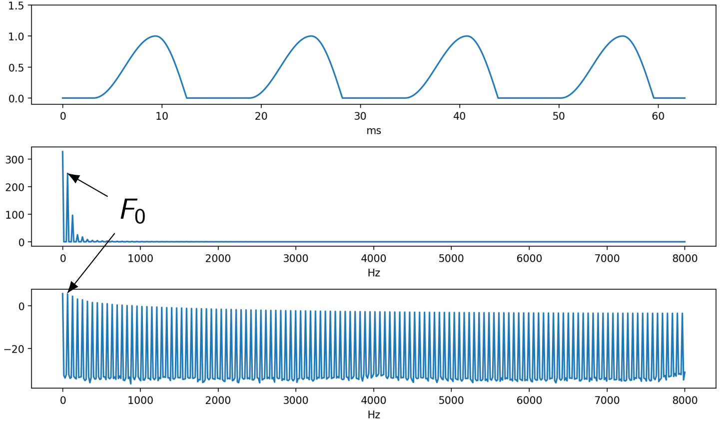
breathing

speaking

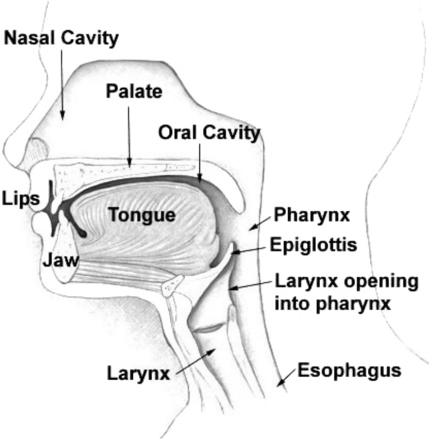
Glottal Pulse



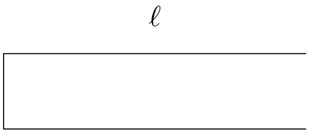
Glottal Pulse

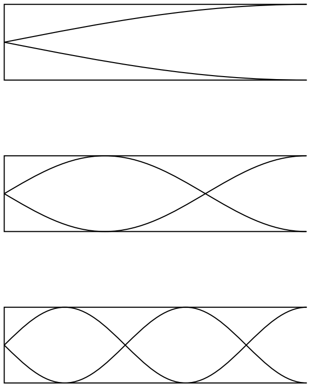


Vocal Tract


Nasal Cavity
Palate
Oral Cavity
Lips
Tongue
Jaw
Pharynx
Epiglottis
Larynx opening
into pharynx
Larynx
Esophagus

Resonance Frequency of A Tube



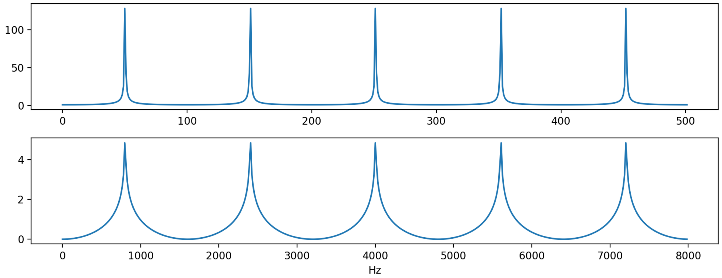


$$f_1 = \frac{v}{4\ell}$$

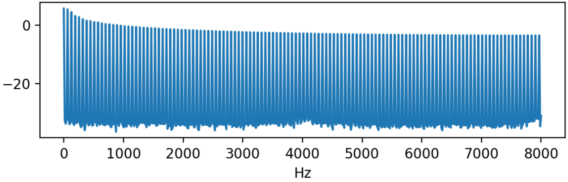
$$f_2 = \frac{3v}{4\ell}$$

$$f_3 = \frac{5v}{4\ell}$$

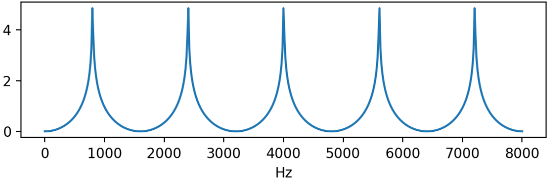
Frequency Response of A Tube



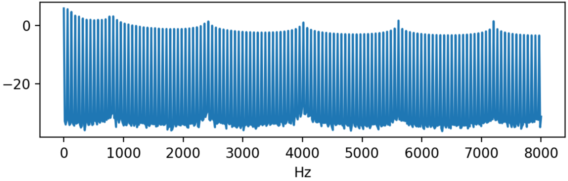
glottal source



vocal tract resonane



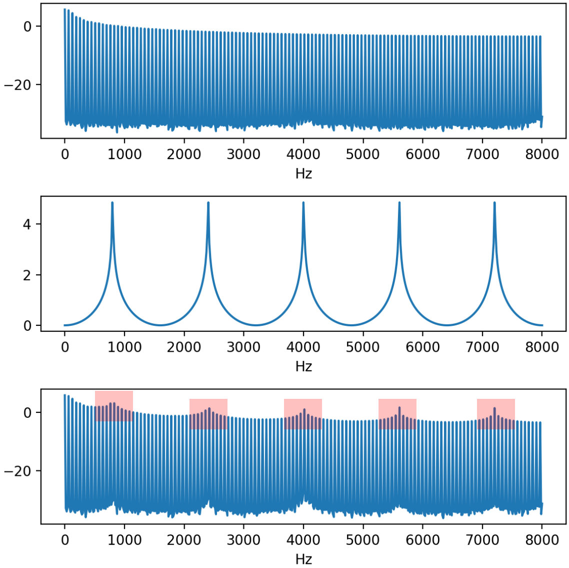
ideal production



glottal source

vocal tract resonane

ideal production



Vowel Production

- Fundamental frequency
  - The first frequency component of the glottal pulse
  - Leading to pitch when perceived
- Harmonics
  - Subsequent frequency components of the glottal pulse
- Formants
  - Resonance frequencies of the vocal tract
  - Leading to the production and perception of certain phones, particularly vowels

Vowel Production

change in articulators

↓

↓

change in phones

Vowel Production

change in articulators

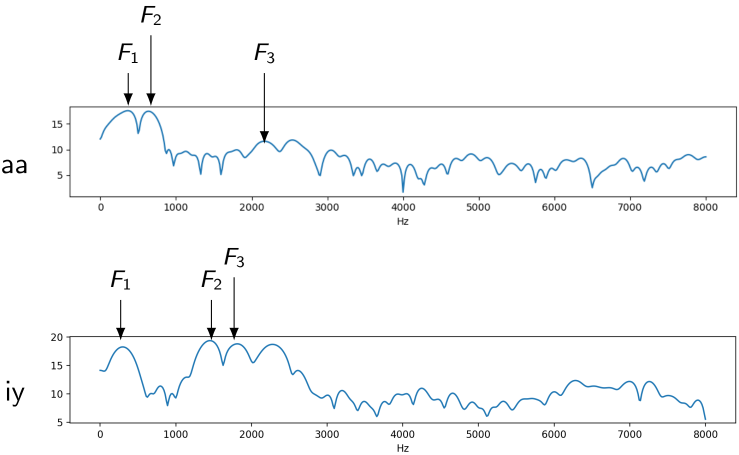
↓

change in resonance frequencies

↓

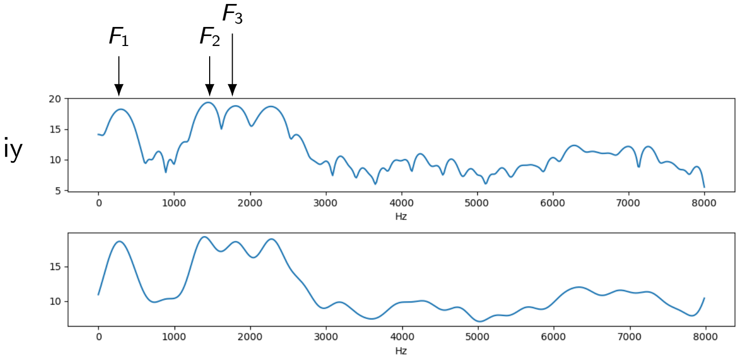
change in phones

Formants

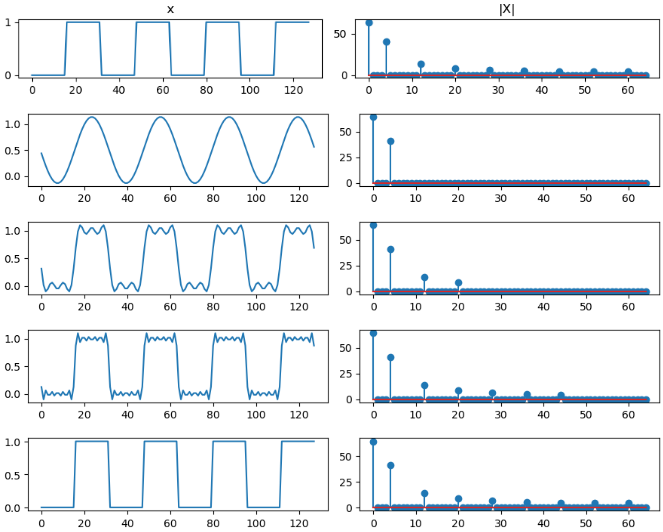


F_1
F_2
F_3
aa
15
10
5
0
1000
2000
3000
4000
5000
6000
7000
8000
Hz
F_1
F_2
F_3
iy
20
15
10
5
0
1000
2000
3000
4000
5000
6000
7000
8000
Hz

Spectral Shape

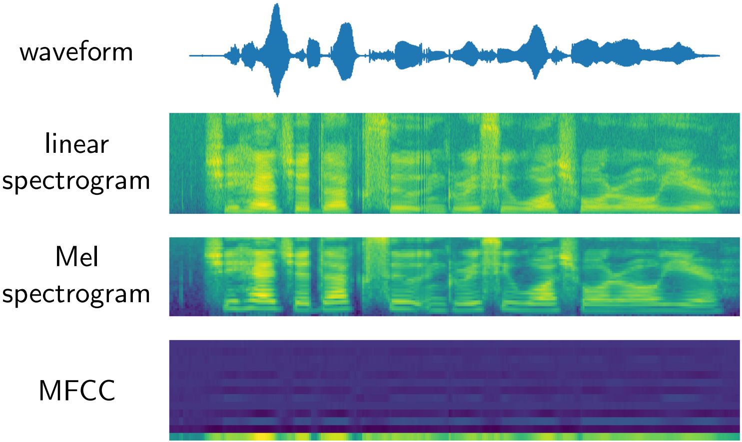


Discrete Fourier Transform



Mel Frequency Cepstral Coefficients (MFCCs)

- Extract Mel spectrogram.
- Apply DFT to every spectrum.
- Truncate the high-frequency components.


waveform
linear spectrogram
Mel spectrogram
MFCC

"All models are wrong, but some are useful."

–George Box, 1978

Summary

- dithering
- removing DC offset
- pre-emphasis
- windowing
- DFT
- Apply Mel filters
- DCT
- Truncate the high-frequency components

Further Reading

- Chapter 3–4, O'Shaughnessy, "Speech Communications: Human and Machine," 2000.