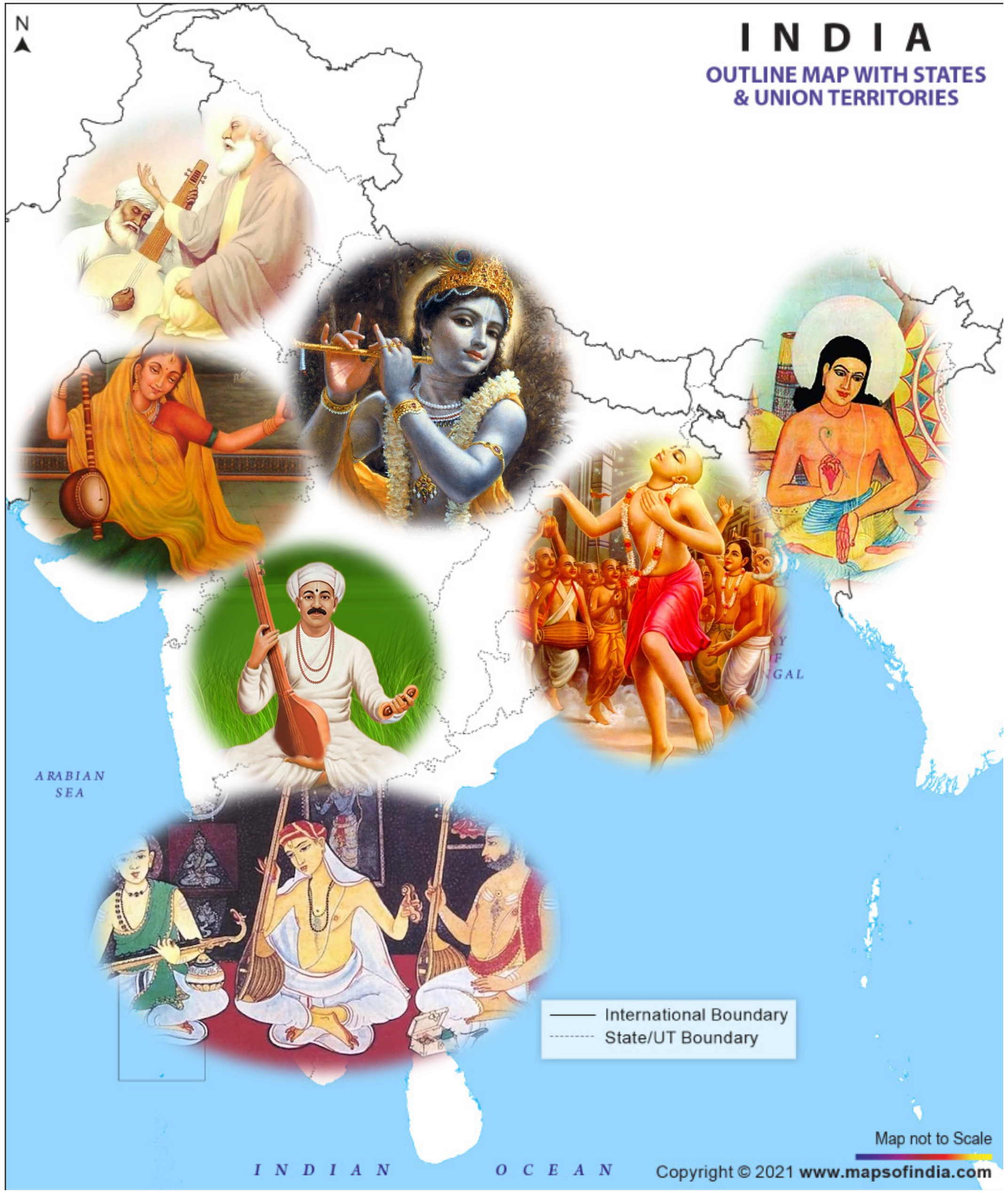
Audio Signal Processing

Vipul Arora
Department of EE, IIT Kanpur

MADHAV lab
Machine Analysis of Data
for Human Audition and Vision

WiSSAP Class Rules

1. Ask questions
2. Ask questions
3. Ask questions
4. Ask questions
5. …

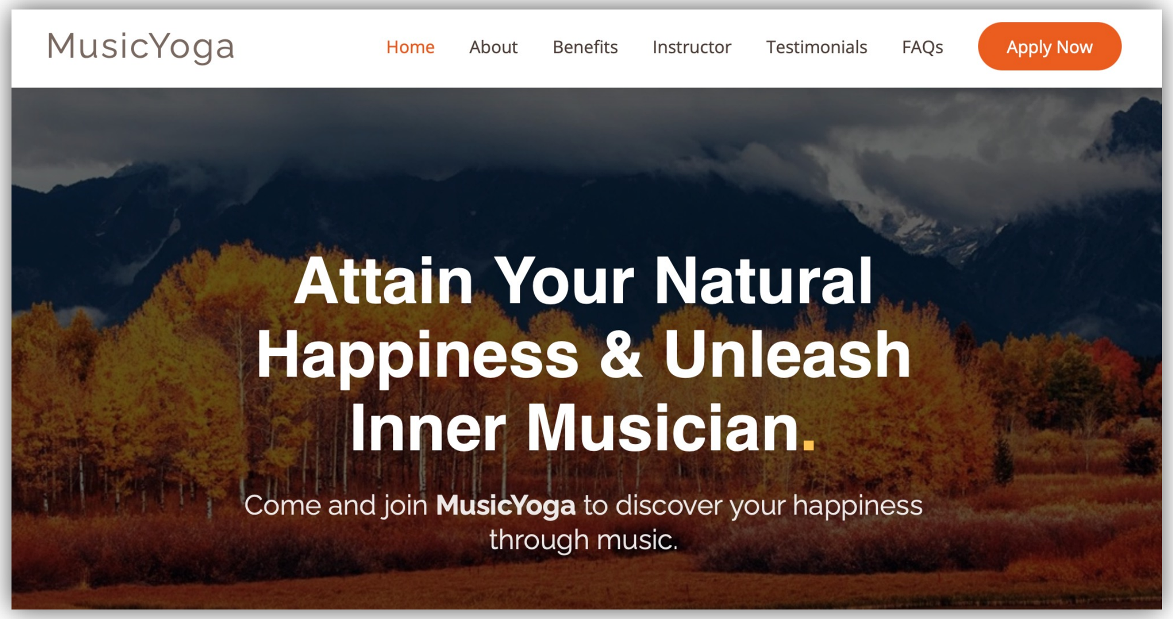


Source: internet

Music in Society

Source: internet

Music and Health



Attain Your Natural Happiness & Unleash Inner Musician.

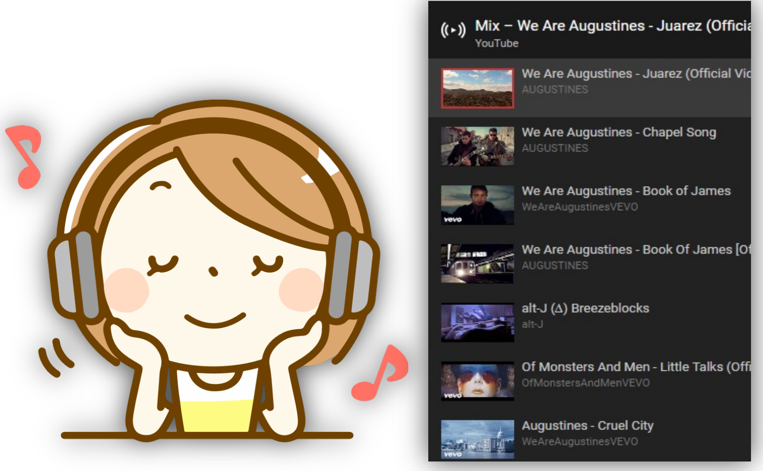
Come and join **MusicYoga** to discover your happiness through music.

Source: internet

Music and Development

Audio Signal Processing and AI

This Photo by Unknown Author is licensed under CC BY-NC-ND



Mix – We Are Augustines - Juarez (Officia
YouTube

We Are Augustines - Juarez (Official Vid
AUGUSTINES

We Are Augustines - Chapel Song
AUGUSTINES

We Are Augustines - Book of James
WeAreAugustinesVEVO

We Are Augustines - Book Of James [Of
AUGUSTINES

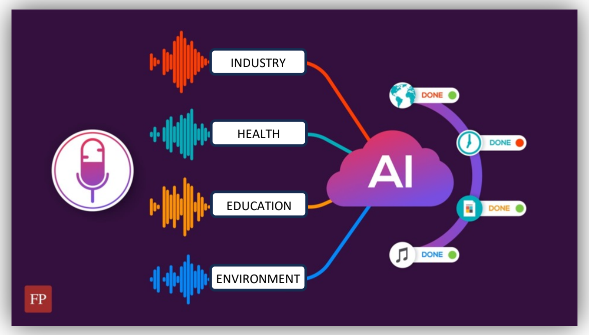
alt-J (∆) Breezeblocks
alt-J

Of Monsters And Men - Little Talks (Offi
OfMonstersAndMenVEVO

Augustines - Cruel City
WeAreAugustinesVEVO

This Photo by Unknown Author is licensed under CC BY

This Photo by Unknown Author is licensed under CC BY-SA-NC



This Photo by Unknown Author is licensed under CC BY-SA-NC

LINGUISTICS ENCYCLOPEDIA
Matthews Syntactic Relations
STRUCTURAL LINGUISTIC
McCawley Everything that Lingu Wanted to Know a
First and Second Language A
An Introduction to Syntactic Theory
MORRIS An Introduction to the Philosoph
MUYSKEN Functional Categories
Sentence Processing in East A
Generative Linguistics
Newmeyer Linguistics: The Cambridge Surv
Newmeyer Linguistics: The Cambridge Surv
Linguistics: The Cambridge Surve

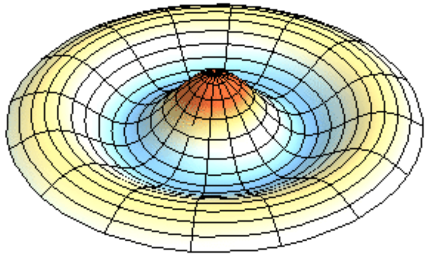
This Photo by Unknown Author is licensed under CC BY-NC

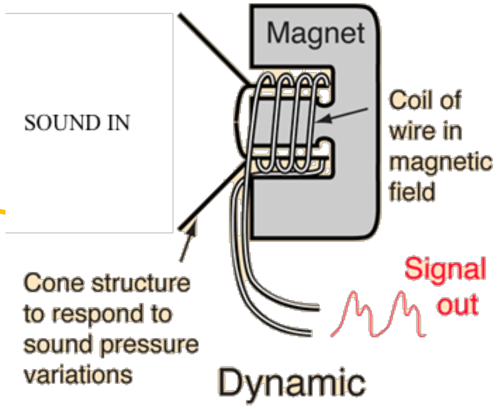
Listen

This Photo by Unknown Author is licensed under CC BY-NC-ND

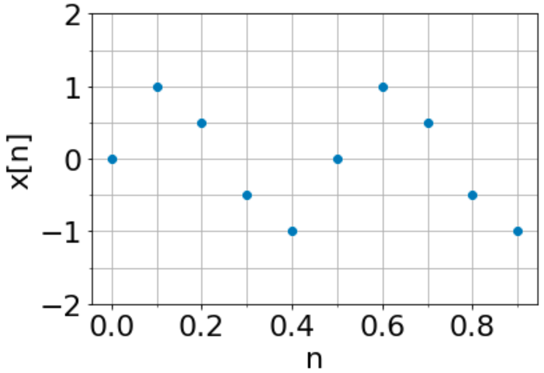
This Photo by Unknown Author is licensed under CC BY-NC

Digital Audio




SOUND IN
Magnet
Coil of wire in magnetic field
Cone structure to respond to sound pressure variations
Signal out
Dynamic

http://hyperphysics.phy-astr.gsu.edu/hbase/Audio/mic.html

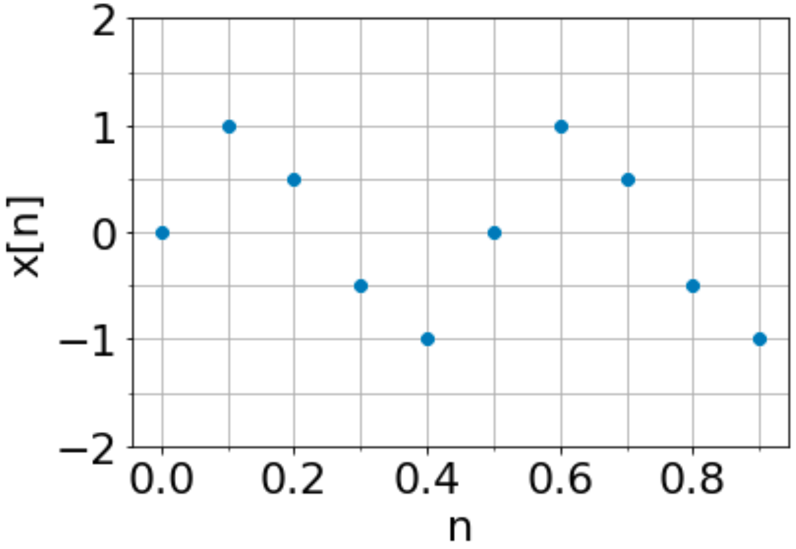


Sampling

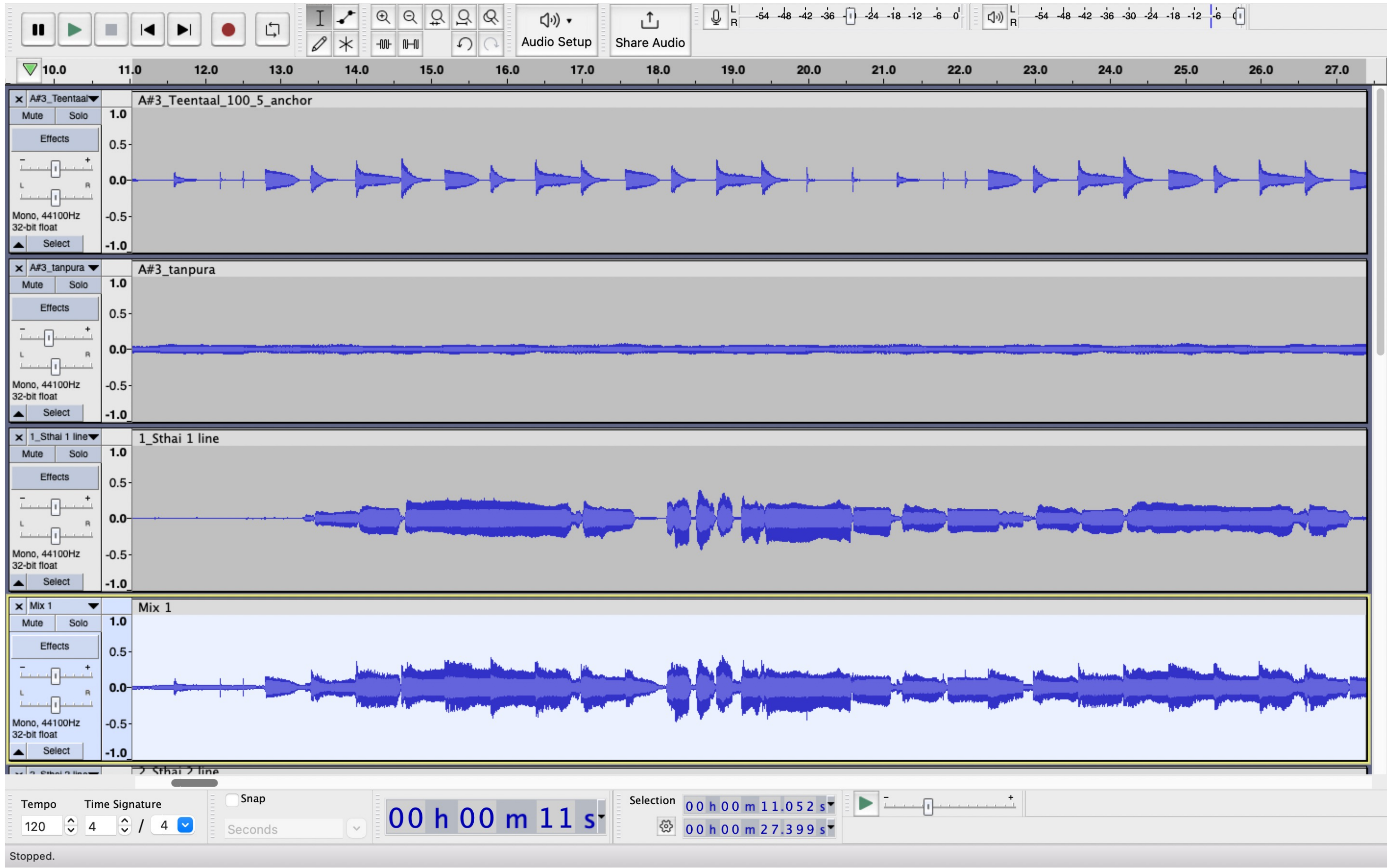
- Nyquist Sampling theorem
- Humans can hear in the range 20Hz to 20kHz
- Popular: 44.1kHz for CD recordings

Quantization

- Converting $x \in \mathbb{R}$ to a digital number
- Q bits per sample => 2^Q possible integer values per sample
- Popular: 16 bits per sample for CD recordings



Waveforms



A#3_Teentaal_100_5_anchor

A#3_tanpura

1_Sthai 1 line

Mix 1

2_Sthai 2 line

Mono, 44100Hz
32-bit float

Tempo 120

Time Signature 4 / 4

Snap

Seconds

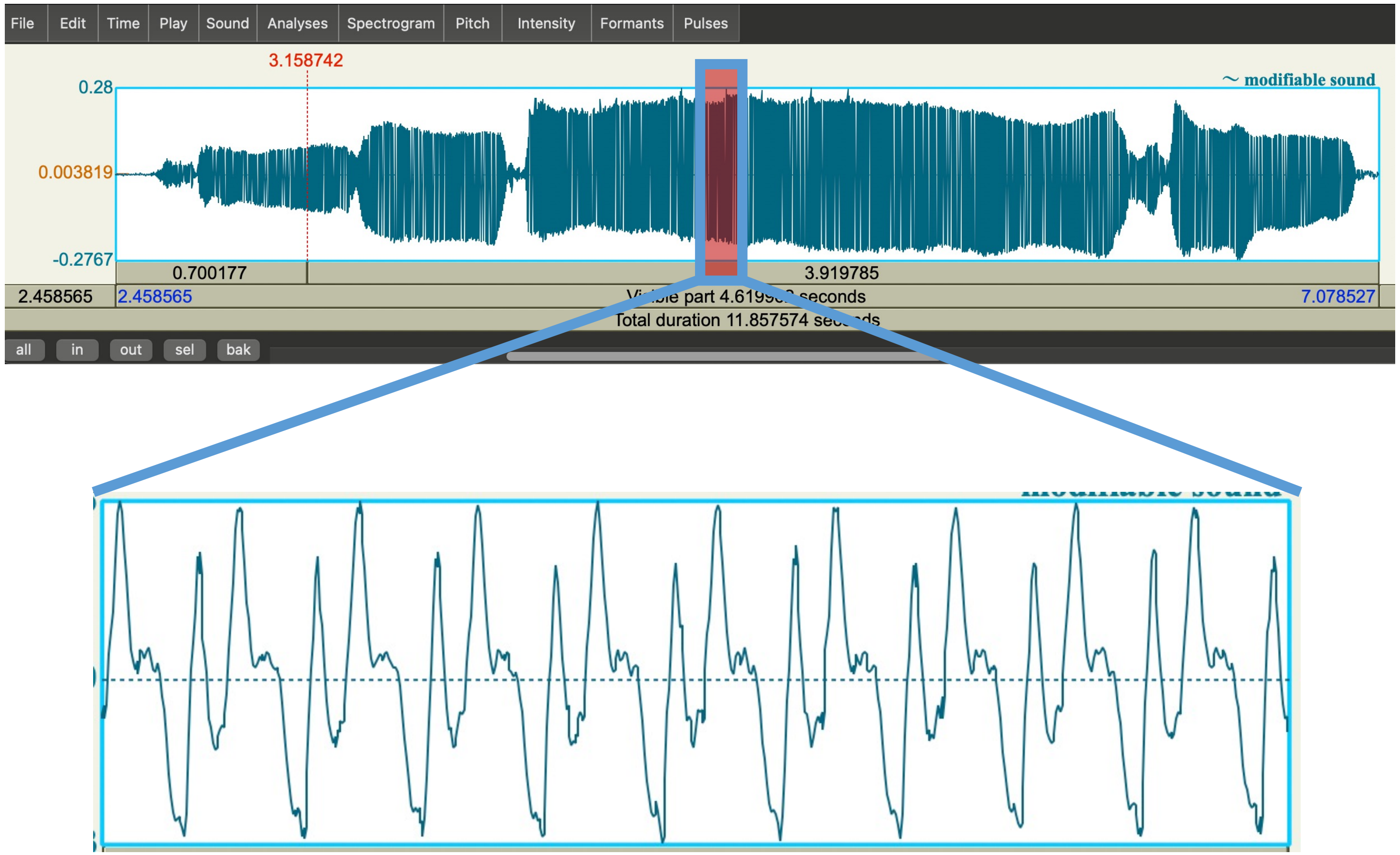
00 h 00 m 11 s

Selection
00 h 00 m 11.052 s
00 h 00 m 27.399 s

Stopped.

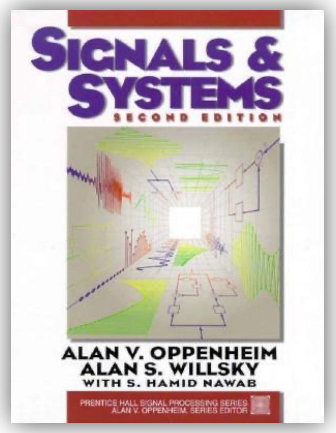
Audacity

Processing … we need mathematical model



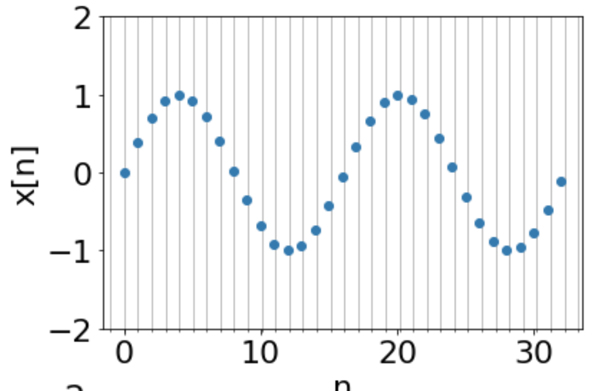
Praat

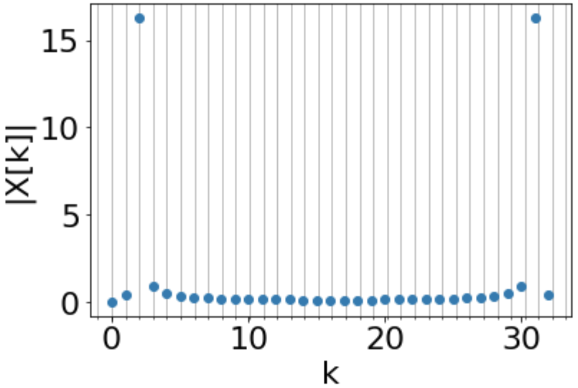
Fourier Transform



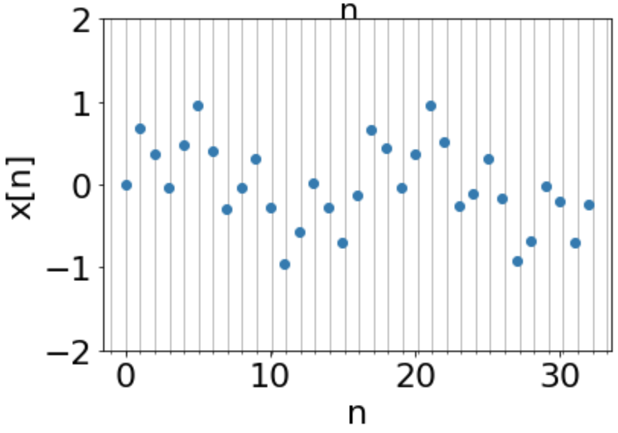
$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j\frac{2\pi}{N}kn} \,; n = 0, \ldots, N-1$$

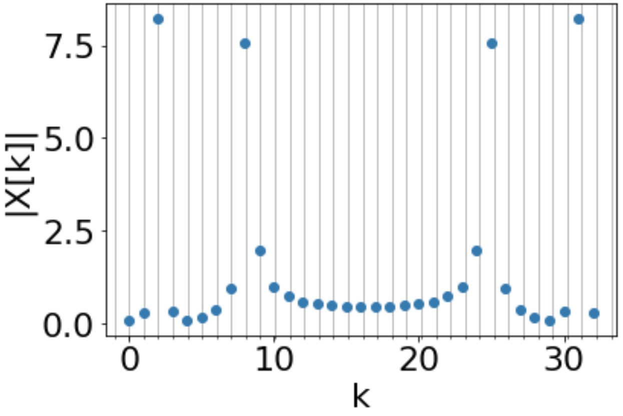
x[n] = sin(2π * 2/32 * n)



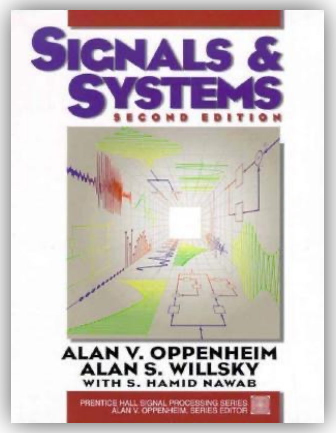


x[n] = 0.5 * sin(2π * 2/32 * n)
+ 0.5 * sin(2π * 8/32 * n)





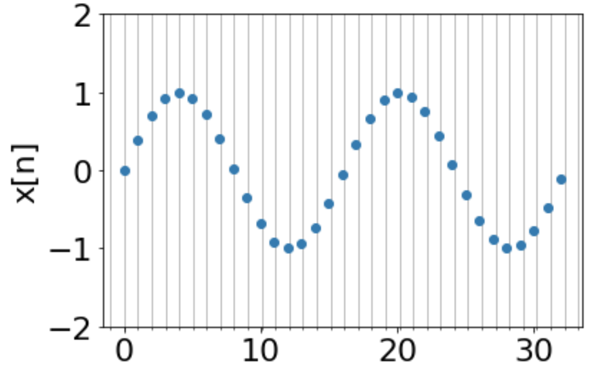
Fourier Transform

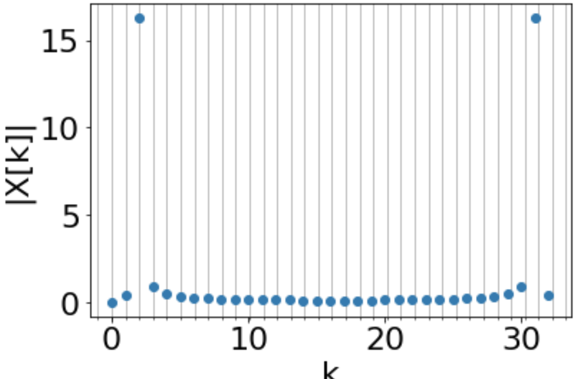


SIGNALS & SYSTEMS
SECOND EDITION
ALAN V. OPPENHEIM
ALAN S. WILLSKY
WITH S. HAMID NAWAB
PRENTICE HALL SIGNAL PROCESSING SERIES
ALAN V. OPPENHEIM, SERIES EDITOR

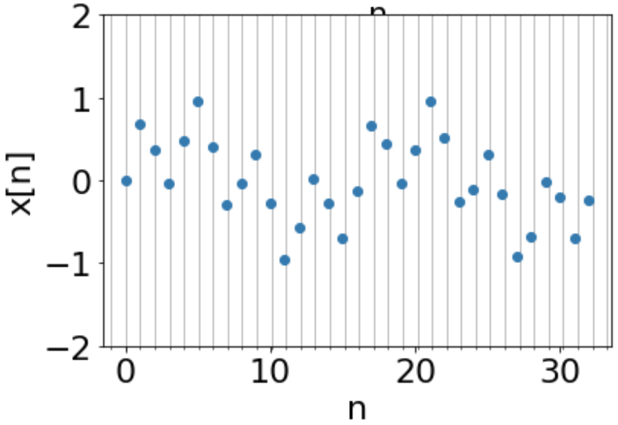
$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} \, ; n = 0, \dots , N-1$$

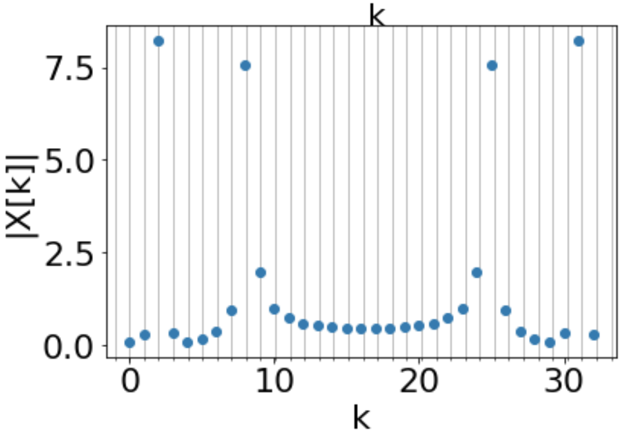
x[n] = sin(2π * 2/32 * n)



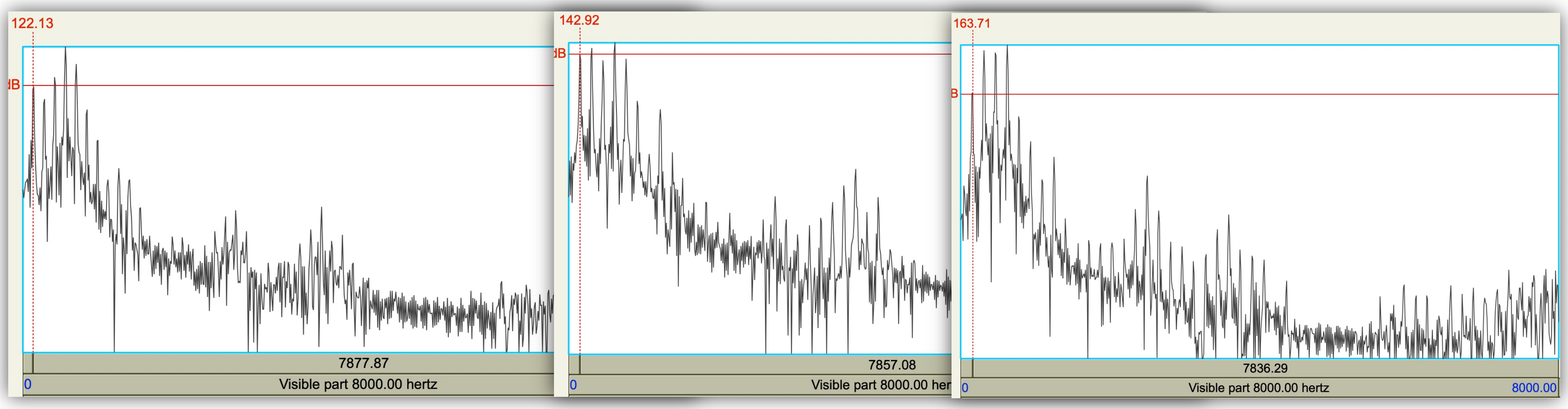


x[n] = 0.5 * sin(2π * 2/32 * n)
+ 0.5 * sin(2π * 8/32 * n)



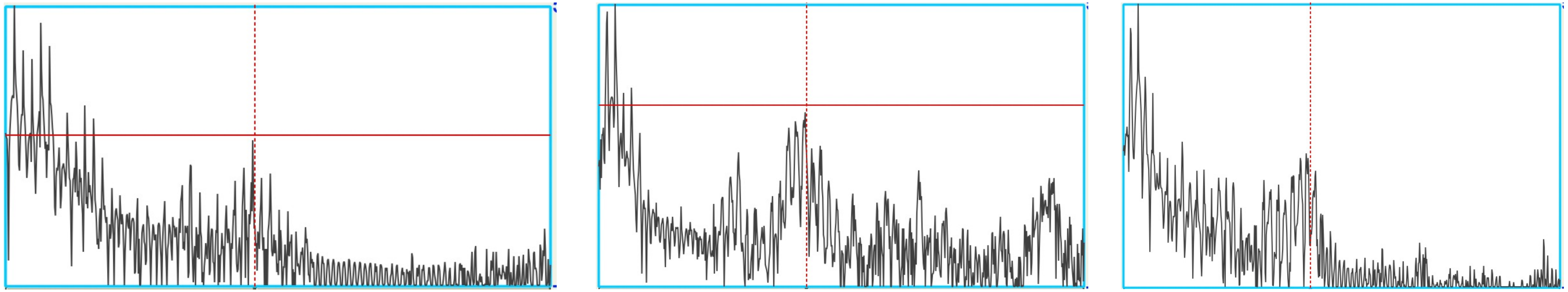


Varying the Pitch

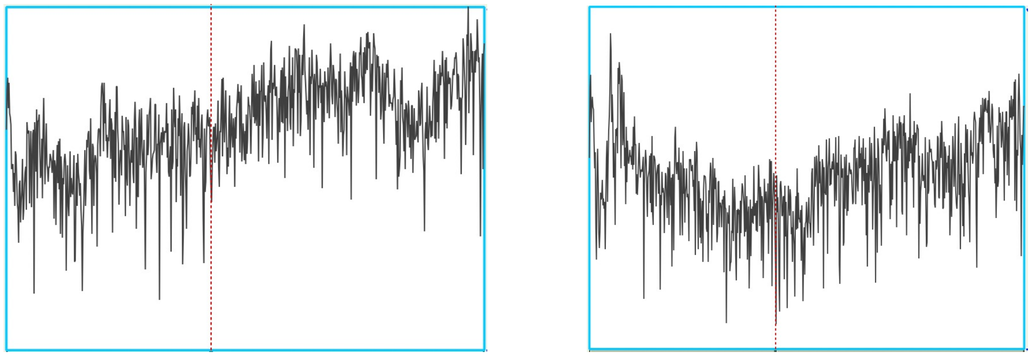


Spectra of speech

- Sounds with periodic waveforms: /a/, /i/, /m/

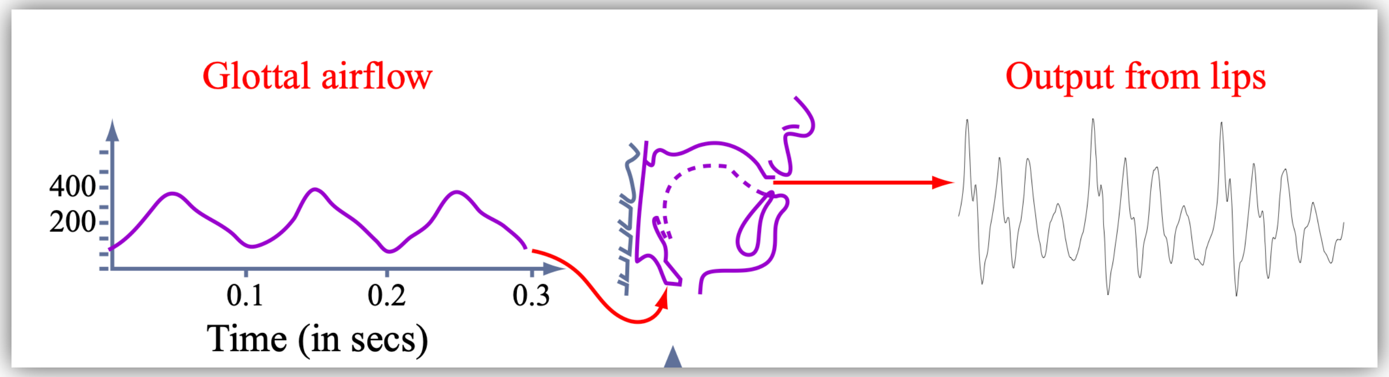


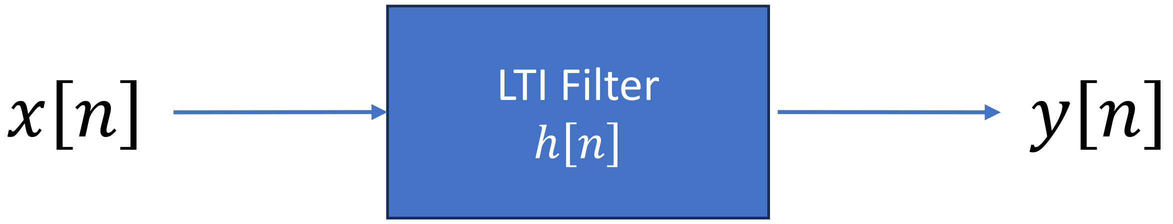
- Sounds with aperiodic waveforms: /s/, /f/



Praat

Filter Theory





$x[n]$

LTI Filter
$h[n]$

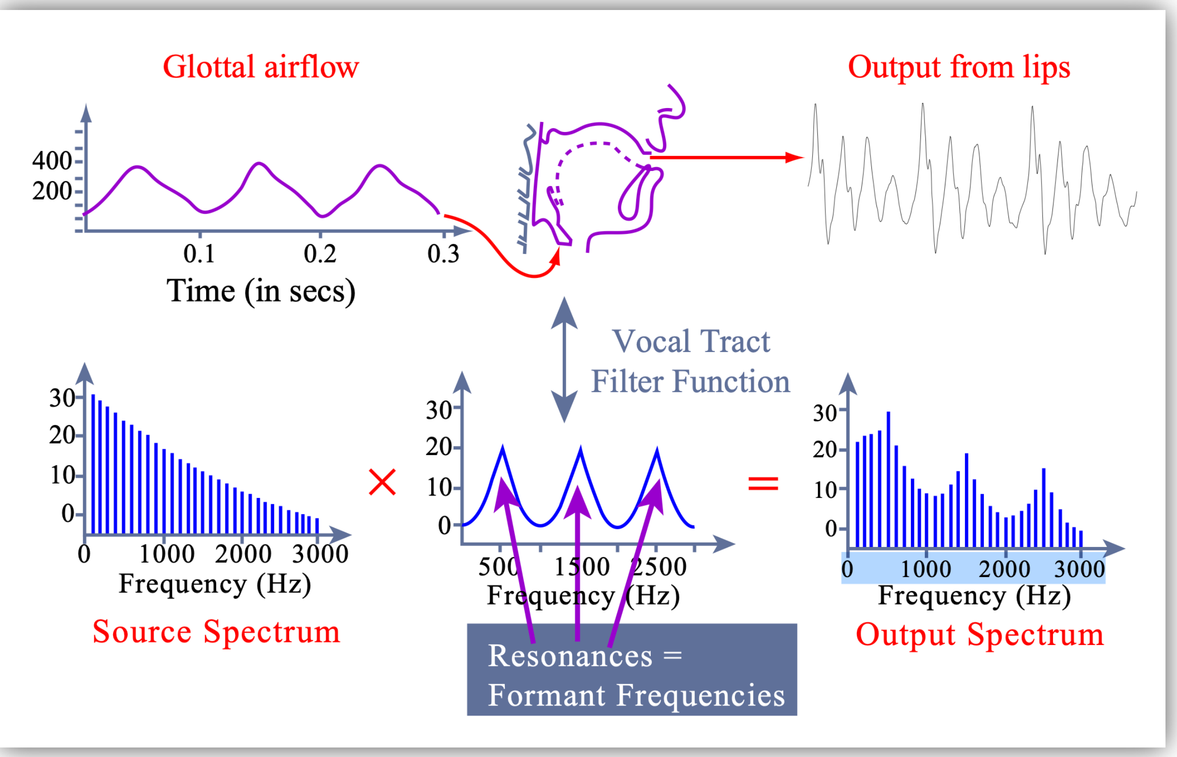
$y[n]$

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

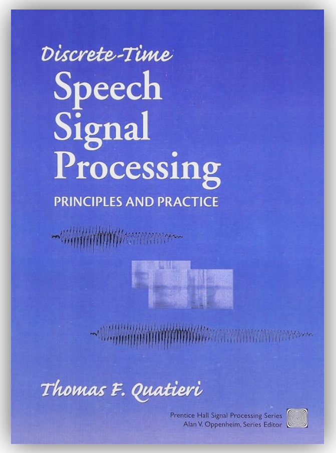
$$Y[k] = X[k]H[k]$$

Source Filter Model

- Linear Time-Invariant (LTI) filters



https://ocw.mit.edu/courses/24-915-linguistic-phonetics-fall-2015/8f4e9d5d8ea634dbbf0e2da8b92675ea_MIT24_915F15_lec4.pdf



Discrete-Time
Speech
Signal
Processing
PRINCIPLES AND PRACTICE

Thomas F. Quatieri

Prentice Hall Signal Processing Series
Alan V. Oppenheim, Series Editor

How do I make real applications with this?

Designing Representations

Representations should be

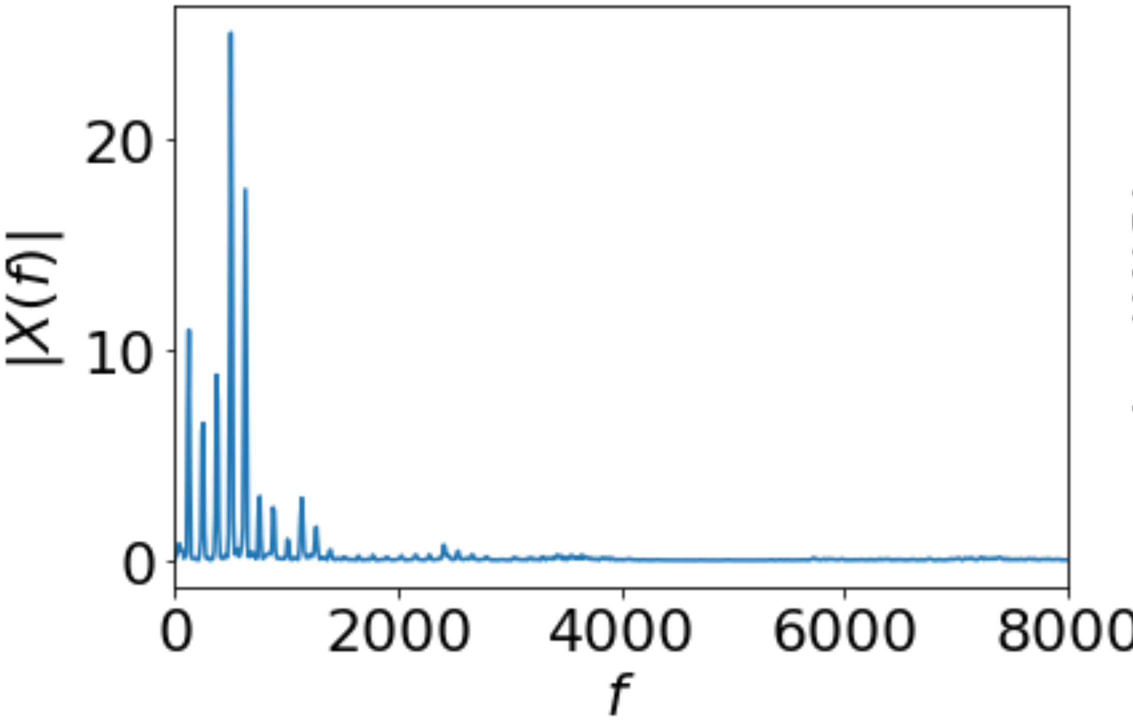
- minimal in size
- distinguishing for what we are interested in
- invariant to what we are not interested in

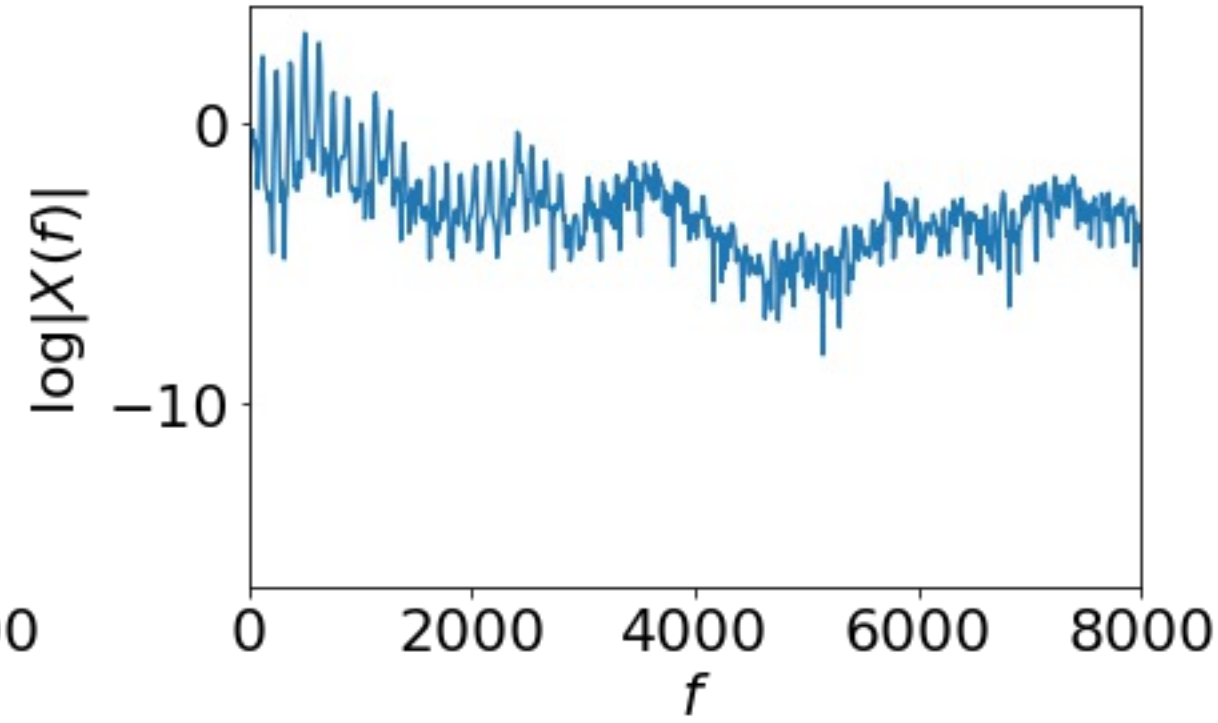
- Design the space so it may have uniform sensitivity (more in Audio Retrieval hands-on by Anup)

Designing Representations

- for Pitch
  - Look at the peaks of spectrum
- for instrument/phoneme
  - Look at the spectral envelope

Amplitude

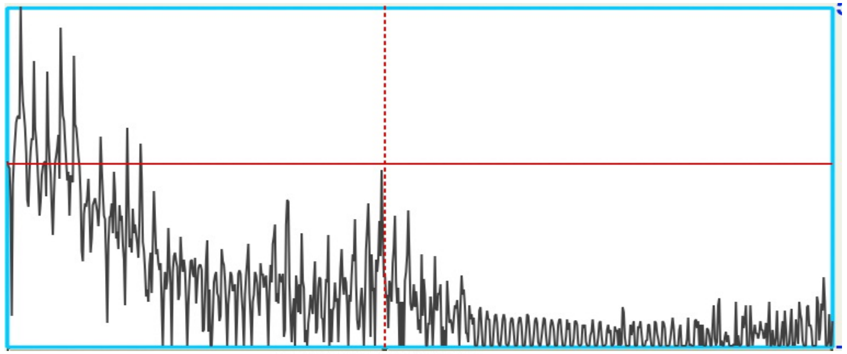


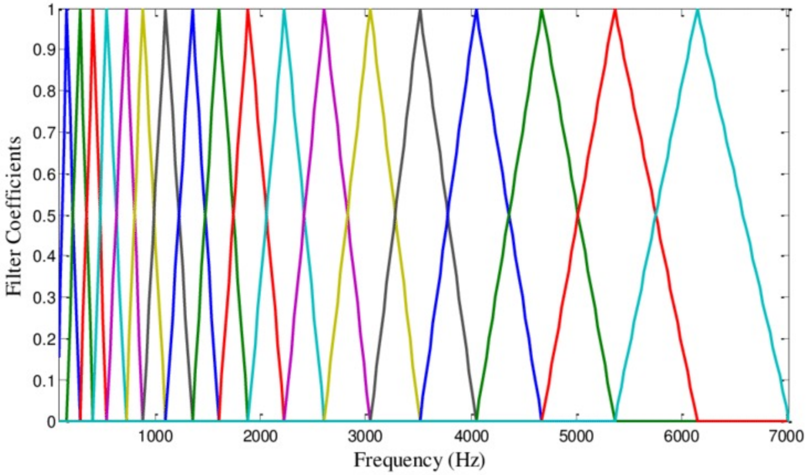


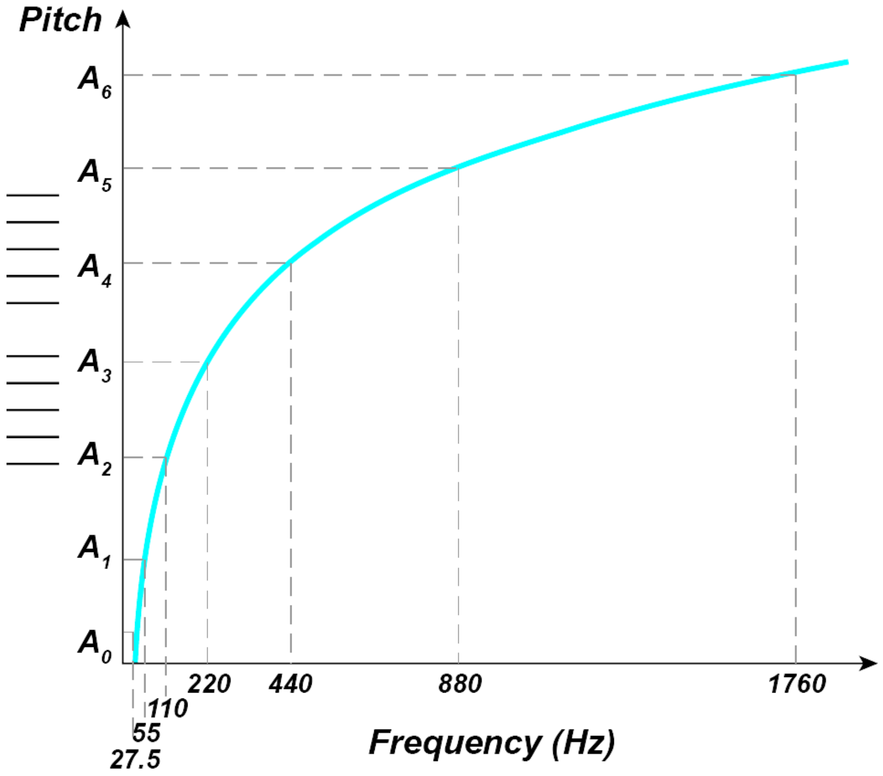
$$A_{dB} = 20\log_{10} A$$

Frequency

- $\tilde{f} \propto \log f$

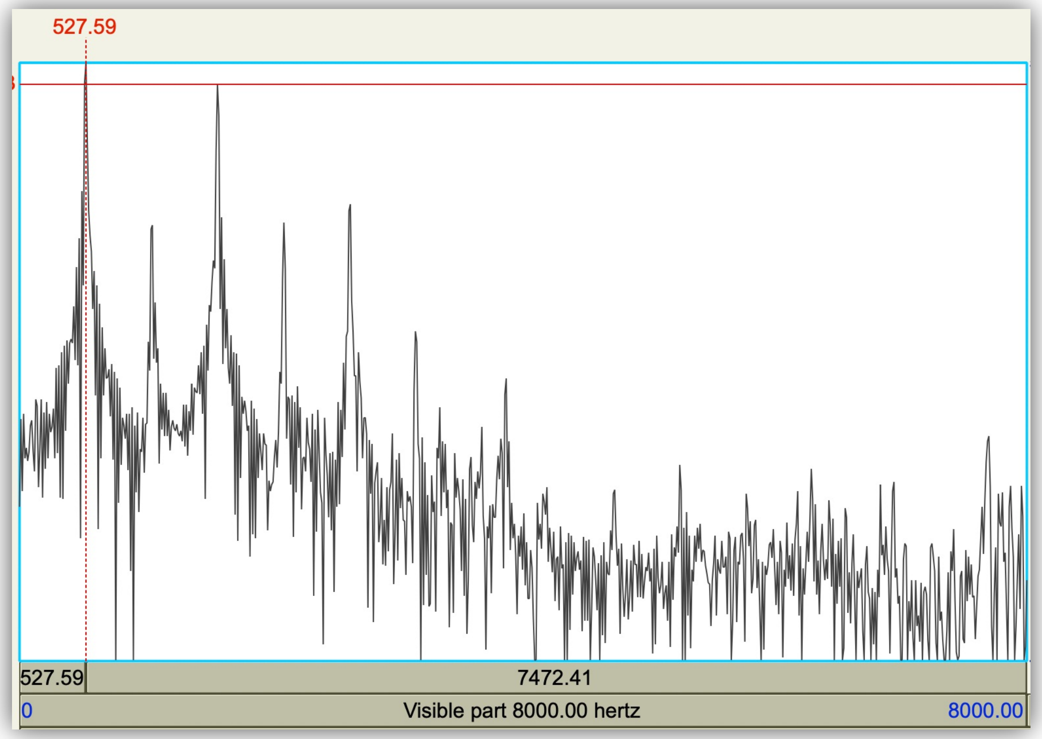






Spectral Envelope

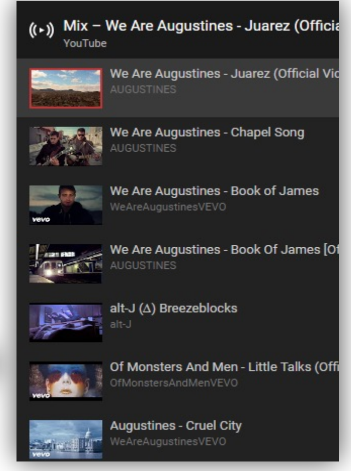
- $|X[\tilde{f}]|_{dB}$
- Mel-frequency, dB amplitude
- Take low frequency components of Fourier transform (DCT) of $|X[\tilde{f}]|_{dB}$



You are ready!!!

This Photo by Unknown Author is licensed under CC BY-SA-NC

This Photo by Unknown Author is licensed under CC BY-NC-ND



Mix – We Are Augustines - Juarez (Officia
YouTube

We Are Augustines - Juarez (Official Vid
AUGUSTINES

We Are Augustines - Chapel Song
AUGUSTINES

We Are Augustines - Book of James
WeAreAugustinesVEVO

We Are Augustines - Book Of James [Of
AUGUSTINES

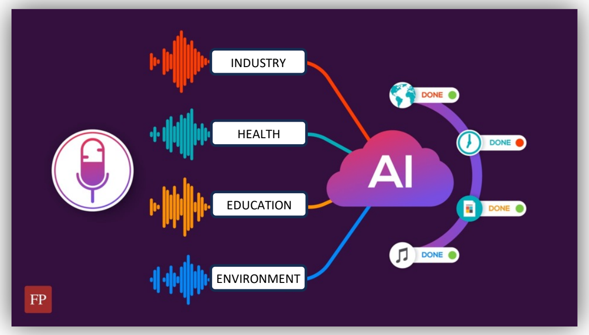
alt-J (∆) Breezeblocks
alt-J

Of Monsters And Men - Little Talks (Offi
OfMonstersAndMenVEVO

Augustines - Cruel City
WeAreAugustinesVEVO

This Photo by Unknown Author is licensed under CC BY

McCawley
Syntactic Relations
Everything that Linguists Wanted to Know A
First and Second Language A
Syntactic Theory
Functional Categories
Sentence Processing in East A
Generative Linguistics
Linguistics: The Cambridge Surv

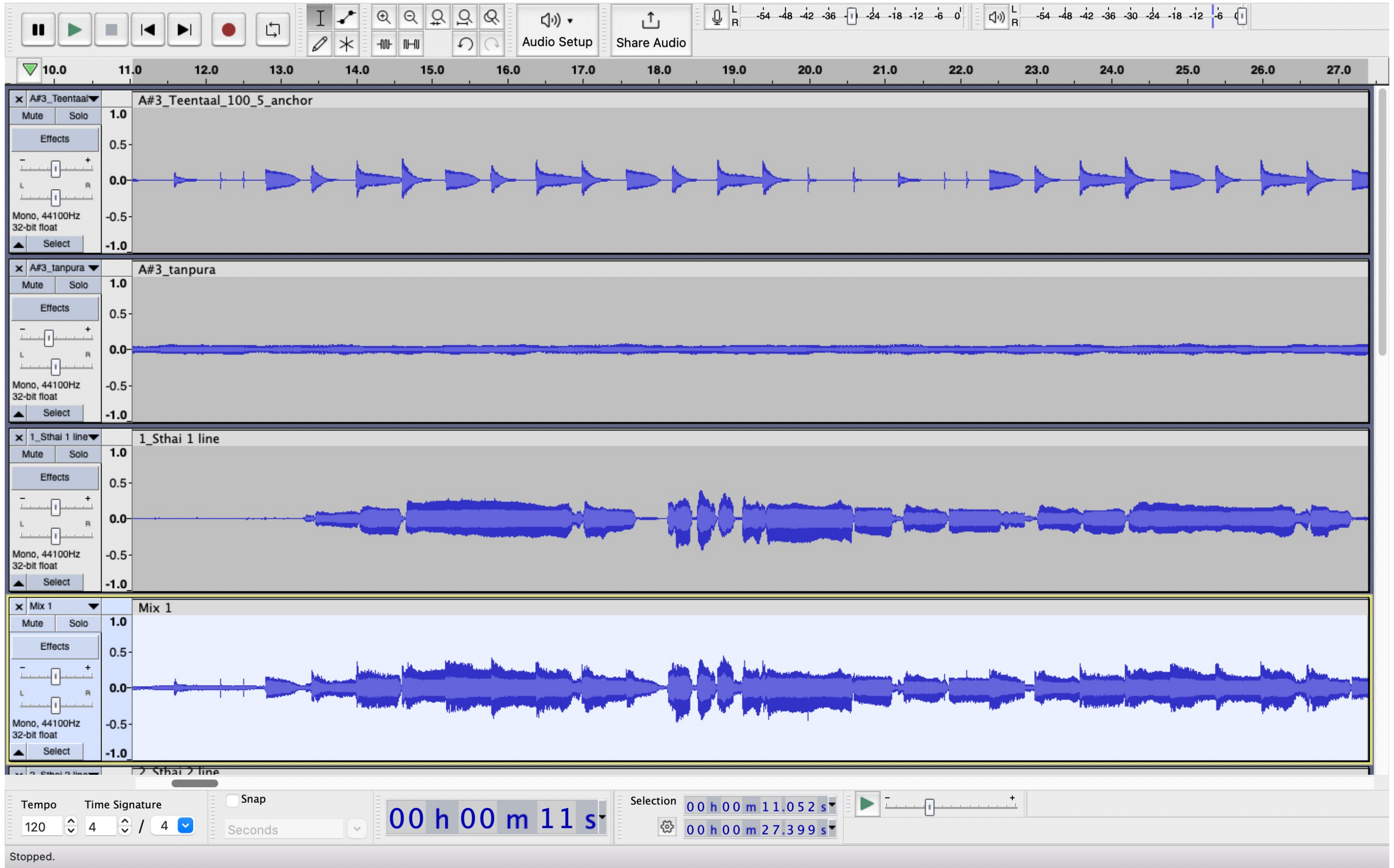


This Photo by Unknown Author is licensed under CC BY-SA-NC

This Photo by Unknown Author is licensed under CC BY-NC

Not yet

- Dynamic behavior
- Time Series Analysis



Tempo 120

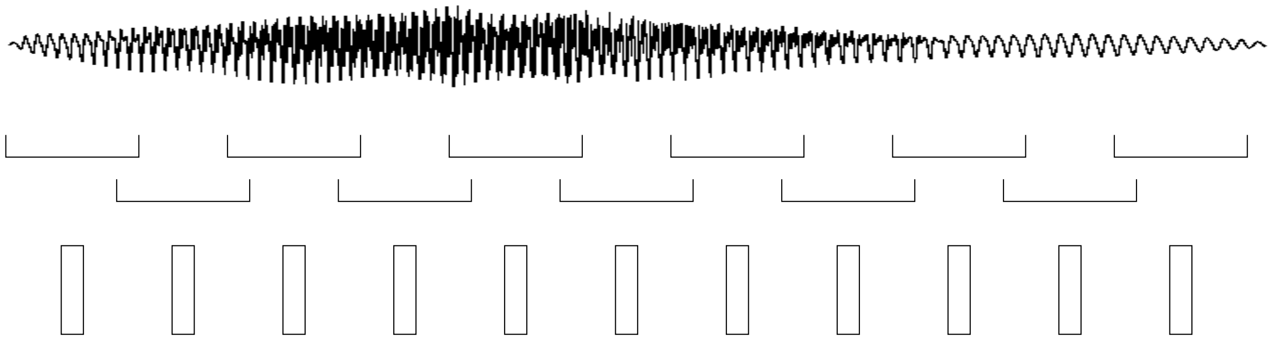
Time Signature 4 / 4

00 h 00 m 11 s

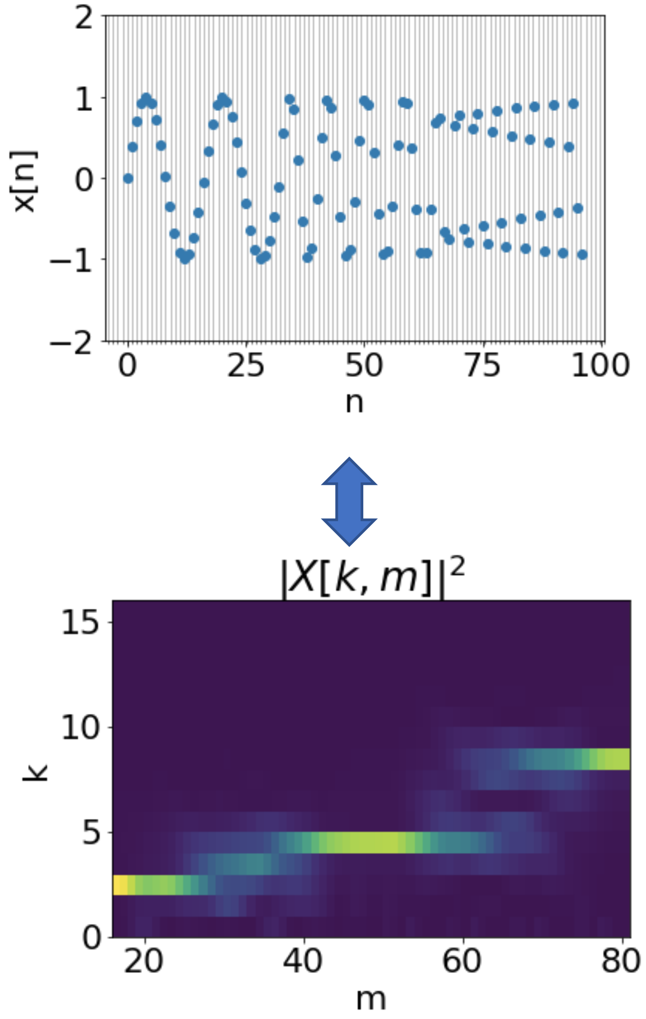
Selection 00 h 00 m 11.052 s

00 h 00 m 27.399 s

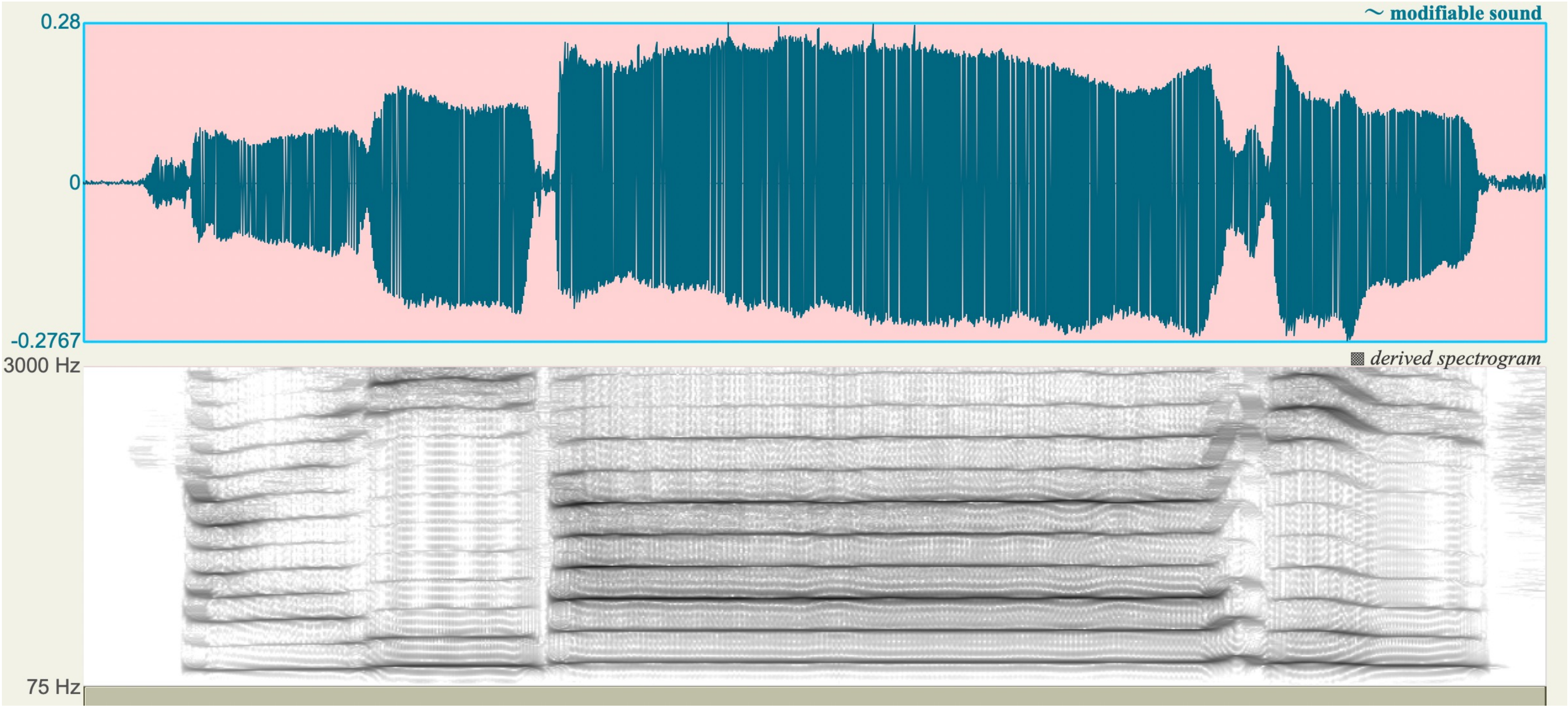
Short Time Fourier Transform



$$X[k,m] = \sum_{n=0}^{N-1} x[n]w[n-mH]e^{-j\frac{2\pi}{N}kn}; \quad k = 0,1,...,N-1$$

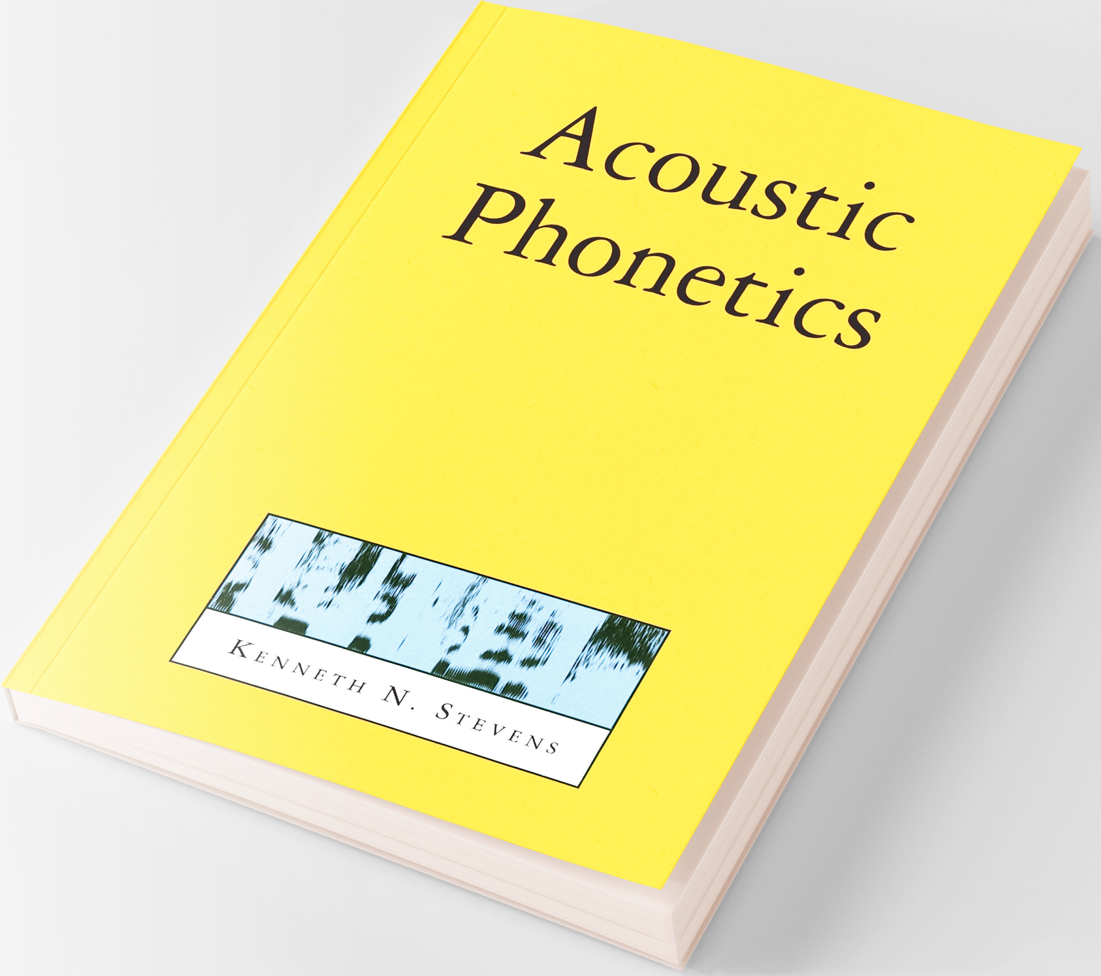


Short Time Fourier Transform



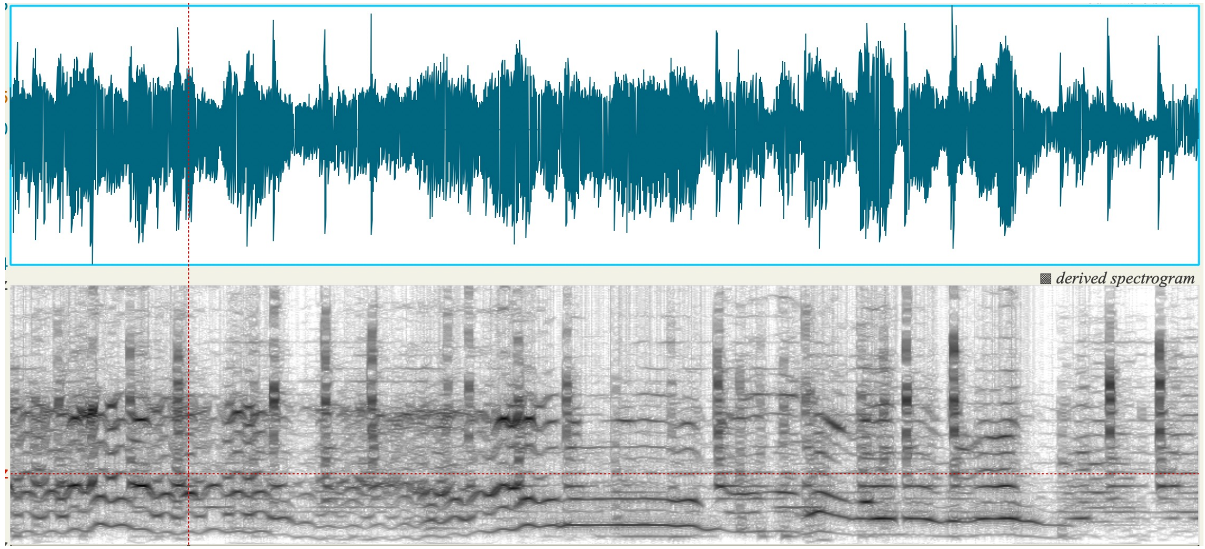
It is possible

But only in ideal situations



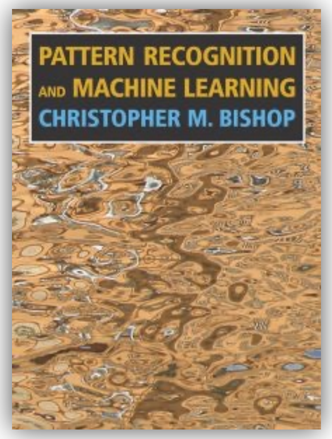
Real-world Variations

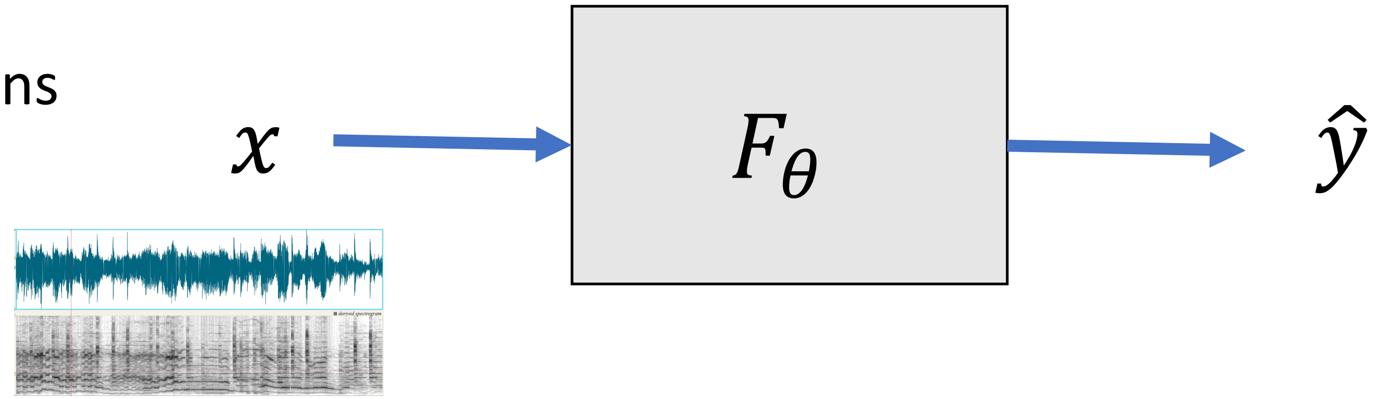
- Context (co-articulation)
- Running speech
- Speakers, instruments
- Languages
- Recording equipment
- Acoustic conditions



Machine Learning

- Parametric models to learn the feature transformations
- Learn the mappings from
  - speech to text
  - audio to audio
  - audio to labels/classes
  - audio to recommendations

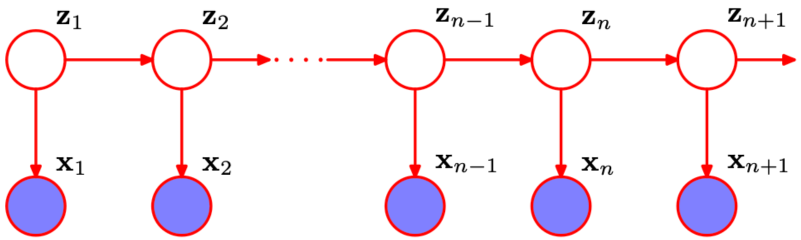


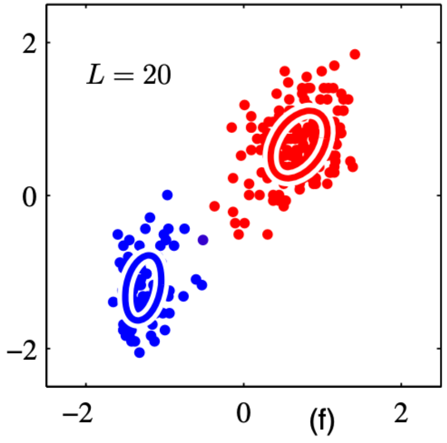


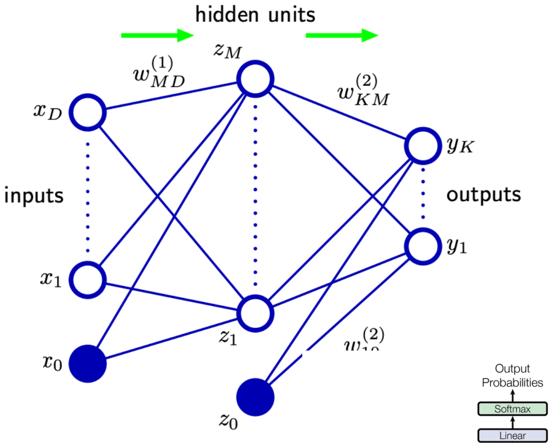
$$\theta = \mathrm{argmin}_\theta\, \mathcal{L}(y, \hat{y}; \theta)$$

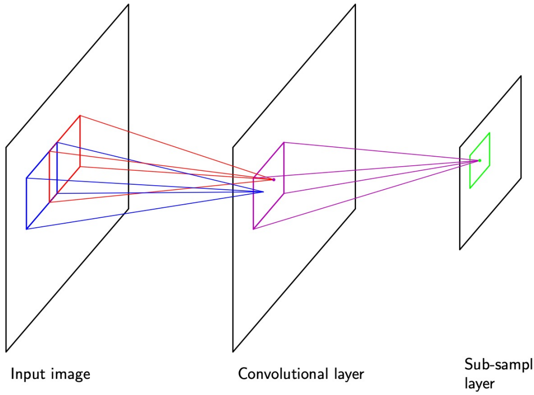
Supervised Learning

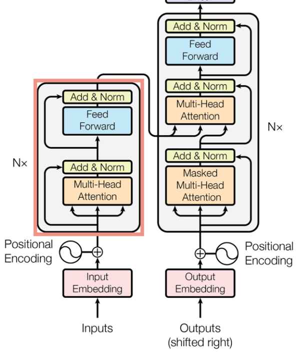
- Gaussian Mixture Model
- Hidden Markov Model
- Multi-Layer Perceptron
- Support Vector Machine
- Convolutional Neural Network
- Recurrent Neural Network
- Transformers





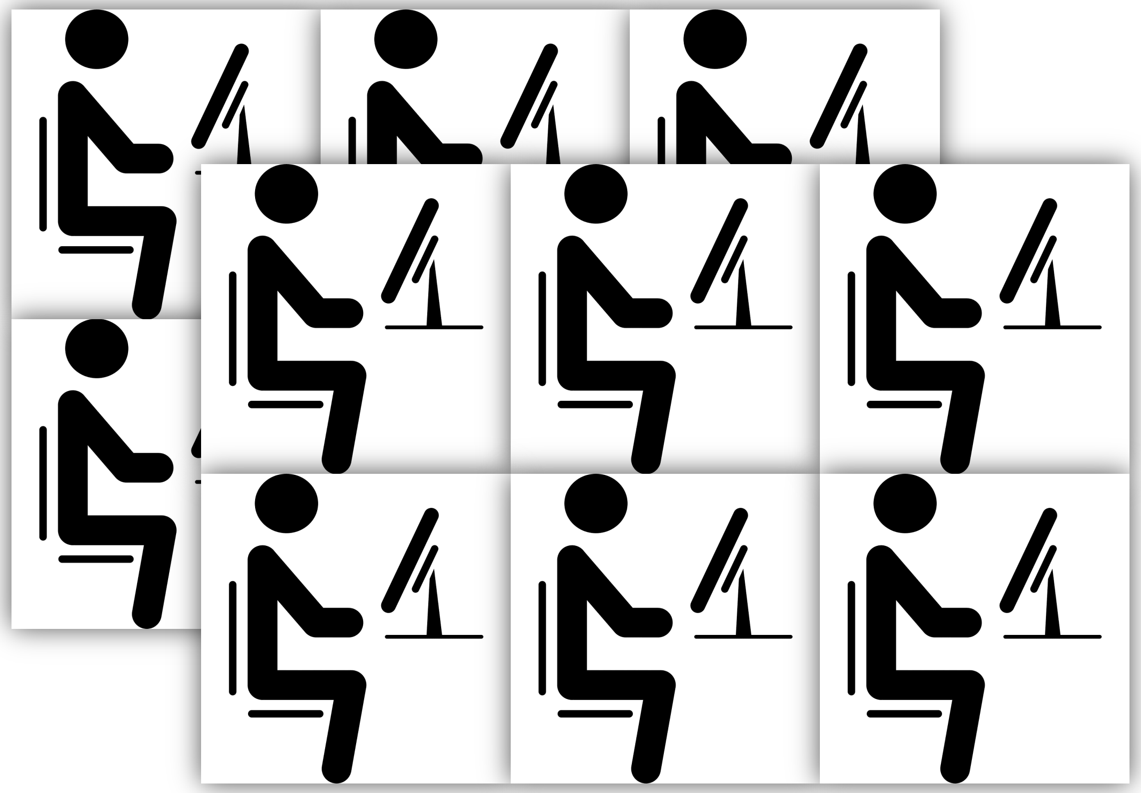






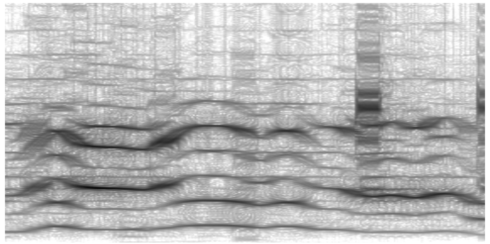
Source: PRML Bishop and https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

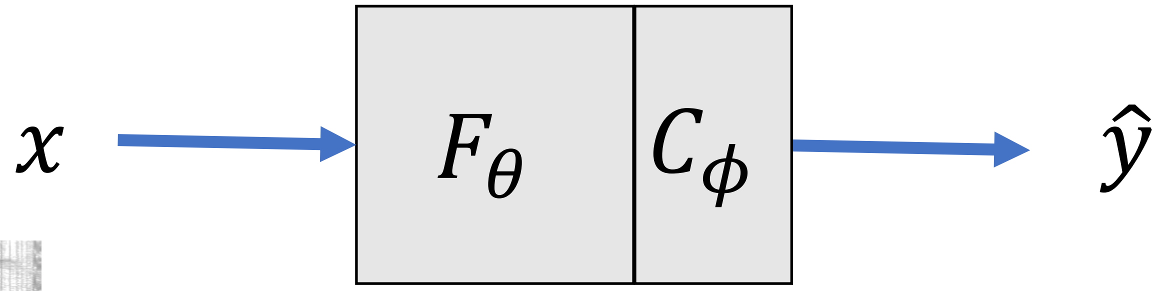
Bottleneck (x, y)

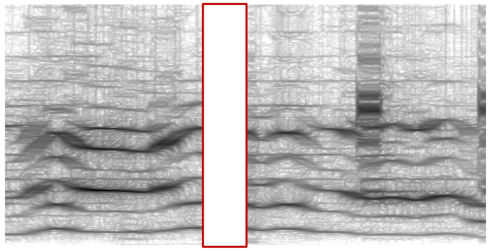


Self-supervised Learning

Use F_θ instead of hand designed representations!







$y =$