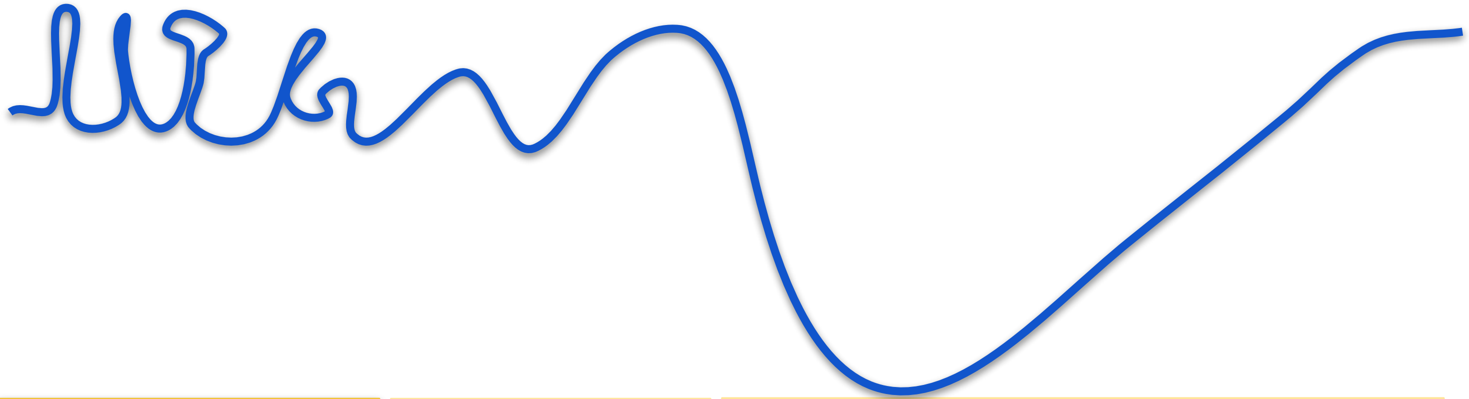
Computing with Signals



DA 623

Jan - May 2024

IIT Guwahati

Instructors: Neeraj Sharma

Lecture-22_23_24

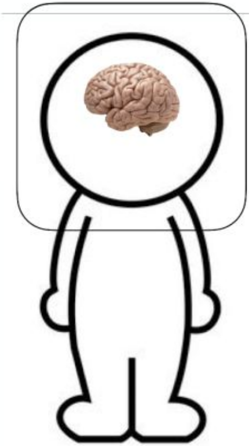
Content

Audio processing

- Features
- Applications

Why features

Imagine you are talking to your friend over phone



Gender

Identity

Cold, fatigue, Age

Emotion

Message

Background noise, reverberation

The speech signal is packed with information!

Why features

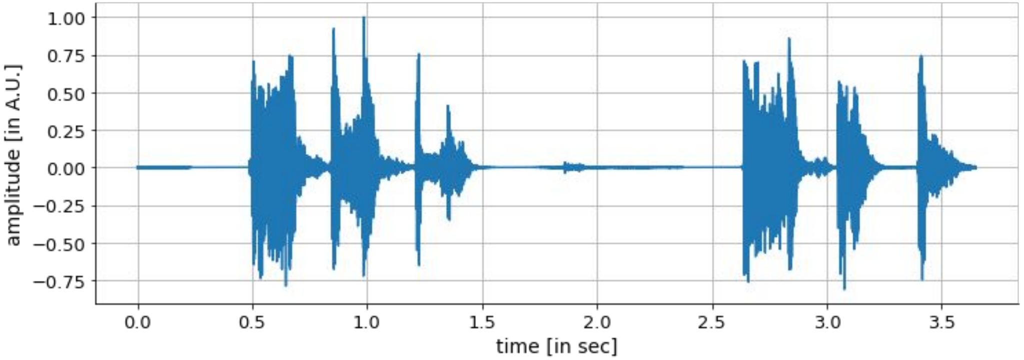
Audio signals are packed with information!

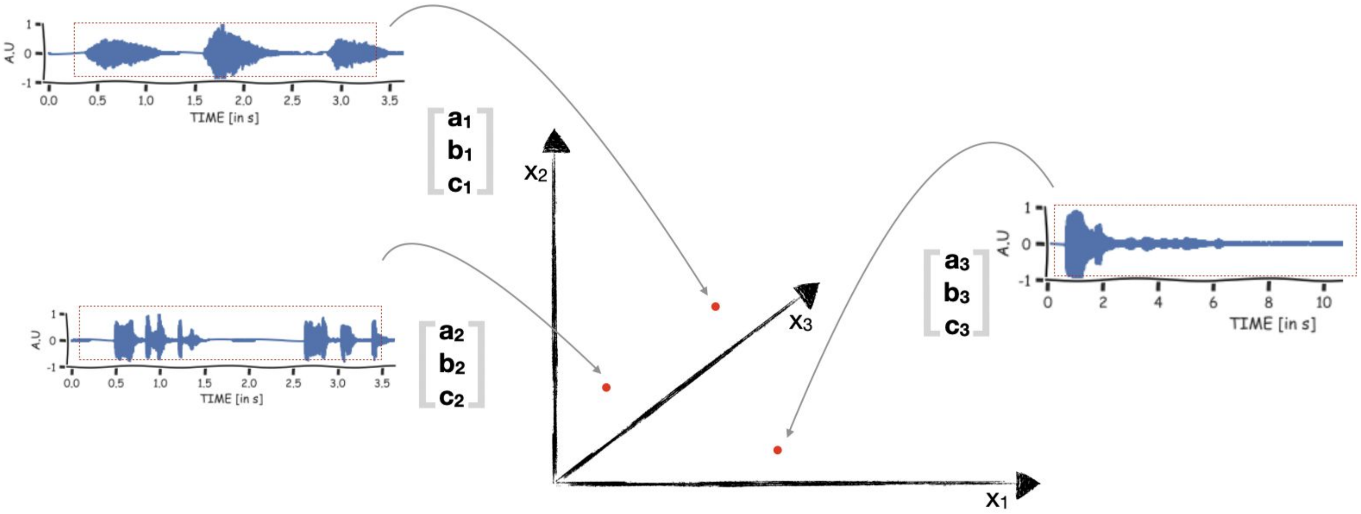
Feature extraction

Process of converting audio data (time-series) into a structured form, allowing us to analyze, understand, and manipulate sound.

Features

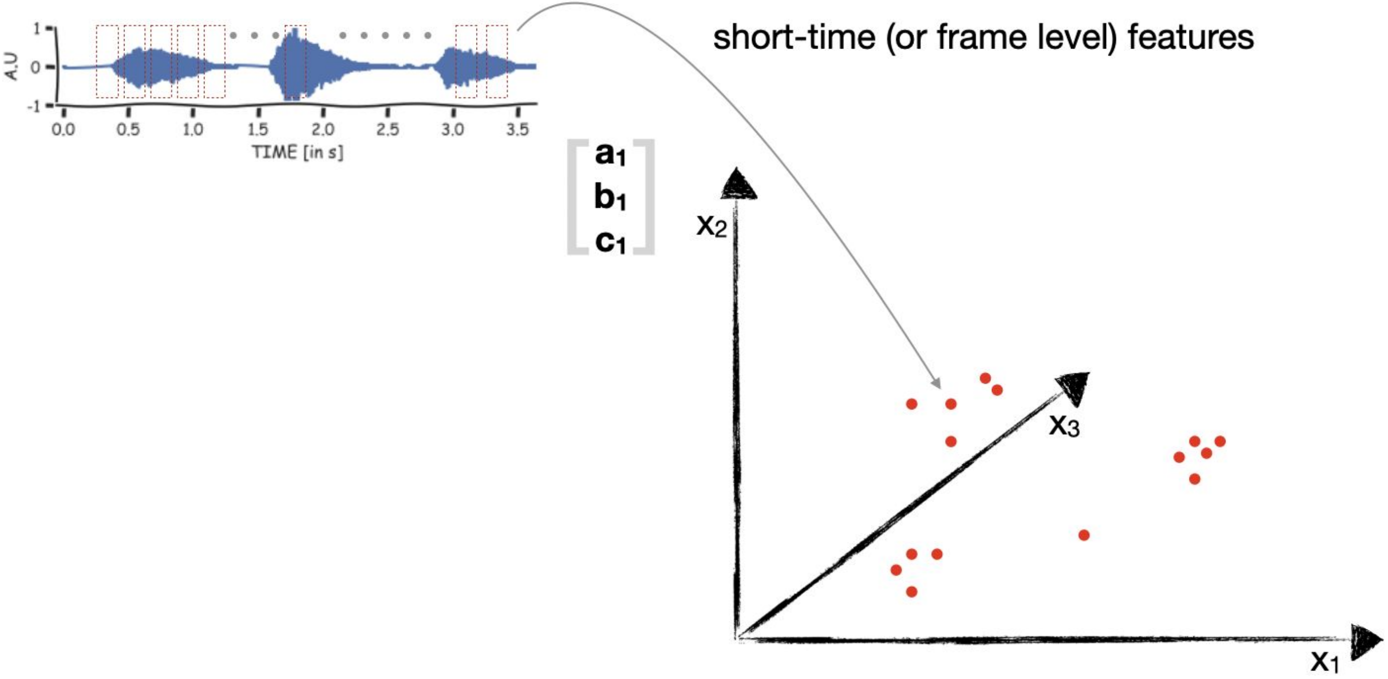
- Long-term features





Features

- Short-term features



short-time (or frame level) features

Units in Spoken Language

- "Cat" can be written as /k/ /æ/ /t/

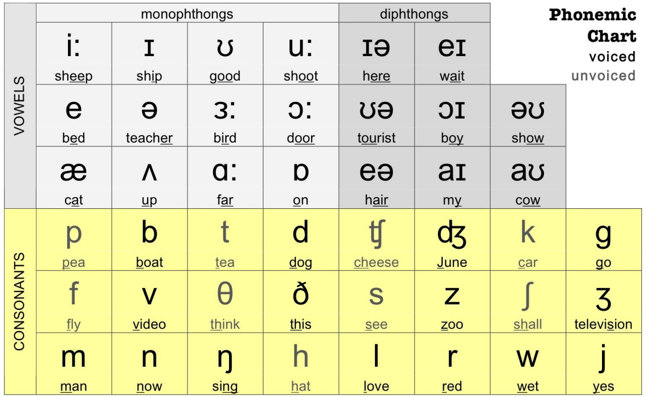
Phoneme

- These are the smallest unit of sound in a language that can change the meaning of a word.
- Phonemes do not have any inherent meaning by themselves but are essential for distinguishing one word from another
- Different languages have different sets of phonemes
- For example, the English language has about 44 phonemes, while other languages may have more or fewer
- Phonemes can be represented by symbols or characters, often enclosed in slashes (/ /) for transcription. For example, in English, the word "cat" consists of three phonemes: /k/ /æ/ /t/.

Units in Spoken Language

International Phonetic Alphabet (IPA)

The International Phonetic Alphabet (IPA) is an alphabetic system of phonetic notation based primarily on the Latin script.



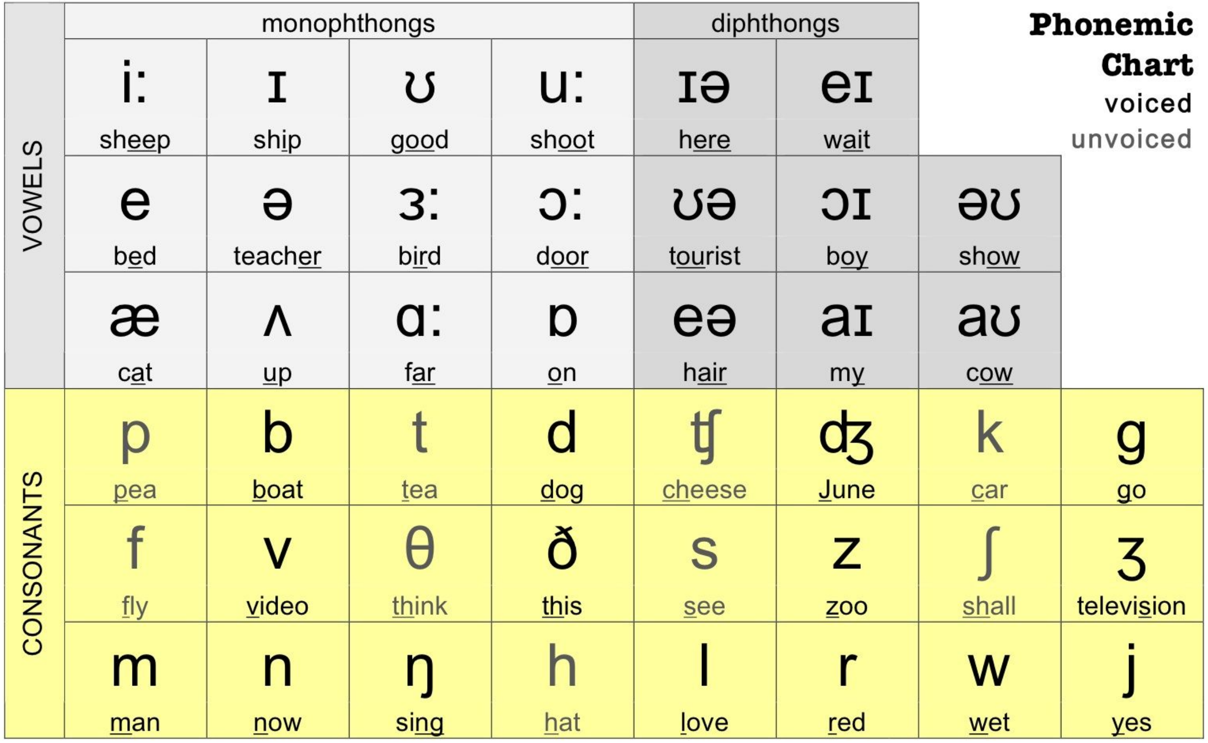
Phonemic Chart
voiced
unvoiced

	monophthongs				diphthongs			
VOWELS	iː sheep	ɪ ship	ʊ good	uː shoot	ɪə here	eɪ wait		
	e bed	ə teacher	ɜː bird	ɔː door	ʊə tourist	ɔɪ boy	əʊ show	
	æ cat	ʌ up	ɑː far	ɒ on	eə hair	aɪ my	aʊ cow	
CONSONANTS	p pea	b boat	t tea	d dog	tʃ cheese	dʒ June	k car	g go
	f fly	v video	θ think	ð this	s see	z zoo	ʃ shall	ʒ television
	m man	n now	ŋ sing	h hat	l love	r red	w wet	j yes

The 44 phonemes of Received Pronunciation based on the popular Adrian Underhill layout

adapted by EnglishClub.com

IPA



Phonemic Chart
voiced
unvoiced

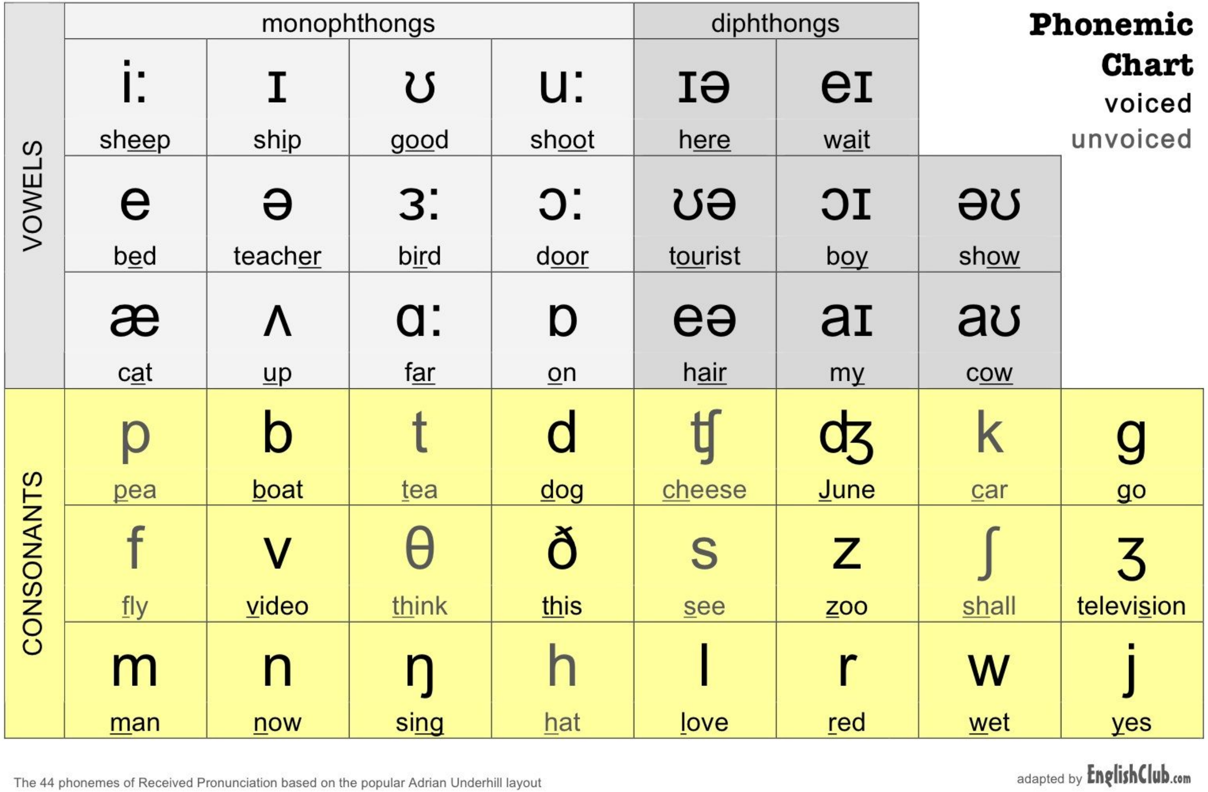
	monophthongs				diphthongs			
VOWELS	iː sheep	ɪ ship	ʊ good	uː shoot	ɪə here	eɪ wait		
	e bed	ə teacher	ɜː bird	ɔː door	ʊə tourist	ɔɪ boy	əʊ show	
	æ cat	ʌ up	ɑː far	ɒ on	eə hair	aɪ my	aʊ cow	
CONSONANTS	p pea	b boat	t tea	d dog	tʃ cheese	dʒ June	k car	g go
	f fly	v video	θ think	ð this	s see	z zoo	ʃ shall	ʒ television
	m man	n now	ŋ sing	h hat	l love	r red	w wet	j yes

The 44 phonemes of Received Pronunciation based on the popular Adrian Underhill layout

adapted by EnglishClub.com

IPA

Can you write your name in IPA?



Phonemic Chart

voiced

unvoiced

	monophthongs				diphthongs			
VOWELS	iː sheep	ɪ ship	ʊ good	uː shoot	ɪə here	eɪ wait		
	e bed	ə teacher	ɜː bird	ɔː door	ʊə tourist	ɔɪ boy	əʊ show	
	æ cat	ʌ up	ɑː far	ɒ on	eə hair	aɪ my	aʊ cow	
CONSONANTS	p pea	b boat	t tea	d dog	ʧ cheese	ʤ June	k car	g go
	f fly	v video	θ think	ð this	s see	z zoo	ʃ shall	ʒ television
	m man	n now	ŋ sing	h hat	l love	r red	w wet	j yes

The 44 phonemes of Received Pronunciation based on the popular Adrian Underhill layout

adapted by EnglishClub.com

Units in Spoken Language

Grapheme

- These are smallest unit of a writing system (script) that represents a phoneme in a specific language.
- These can be letters, characters, or combinations of letters (such as digraphs or trigraphs) that correspond to a single sound (phoneme) or a group of sounds.
- In English, graphemes can be individual letters like "a," "b," or "c," or they can be combinations like "th" or "sh."
- Graphemes can also include punctuation marks and special characters used in writing.

In summary, phonemes are the distinct sounds in spoken language, while graphemes are the written symbols or characters that represent those sounds in a writing system.

Units in Spoken Language

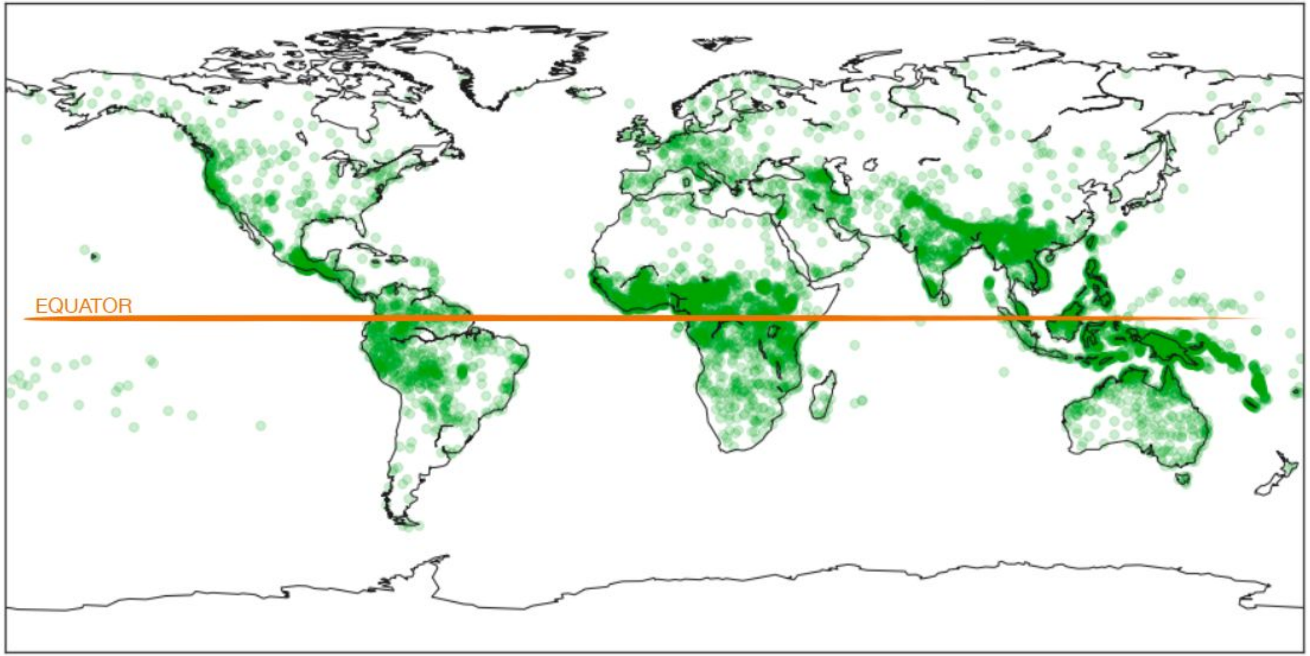
Grapheme

- These are smallest unit of a writing system (script) that represents a phoneme in a specific language.
- These can be letters, characters, or combinations of letters (such as digraphs or trigraphs) that correspond to a single sound (phoneme) or a group of sounds.
- In English, graphemes can be individual letters like "a," "b," or "c," or they can be combinations like "th" or "sh."
- Graphemes can also include punctuation marks and special characters used in writing.

In summary, phonemes are the distinct sounds in spoken language, while graphemes are the written symbols or characters that represent those sounds in a writing system.

7000 plus spoken languages

From Project PHOIBLE 2.0.



Reference: https://phoible.org/

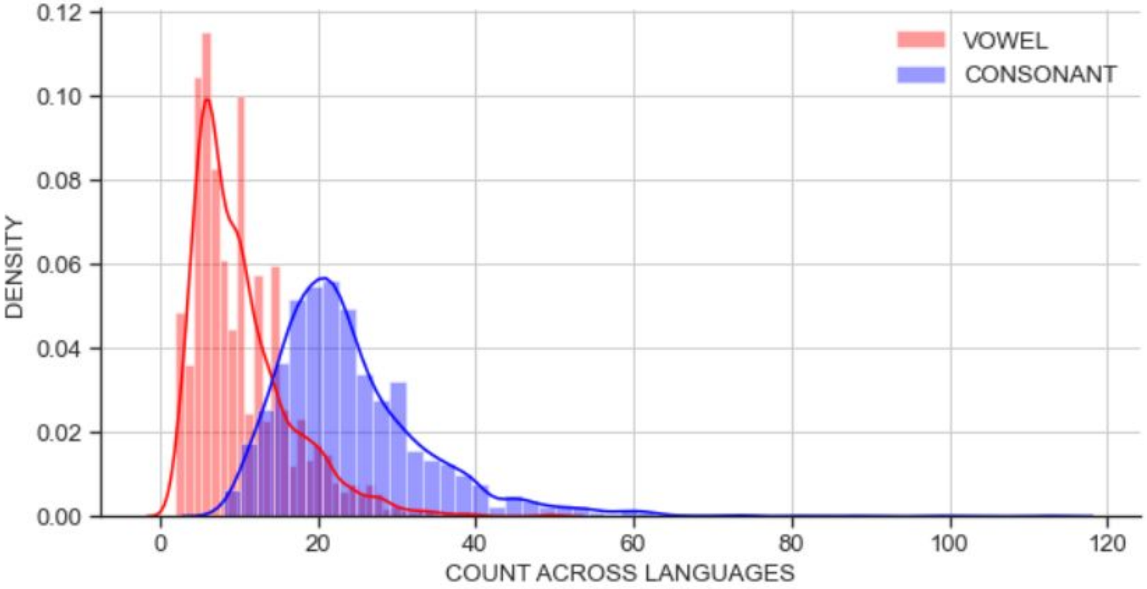
Sounds of a spoken language

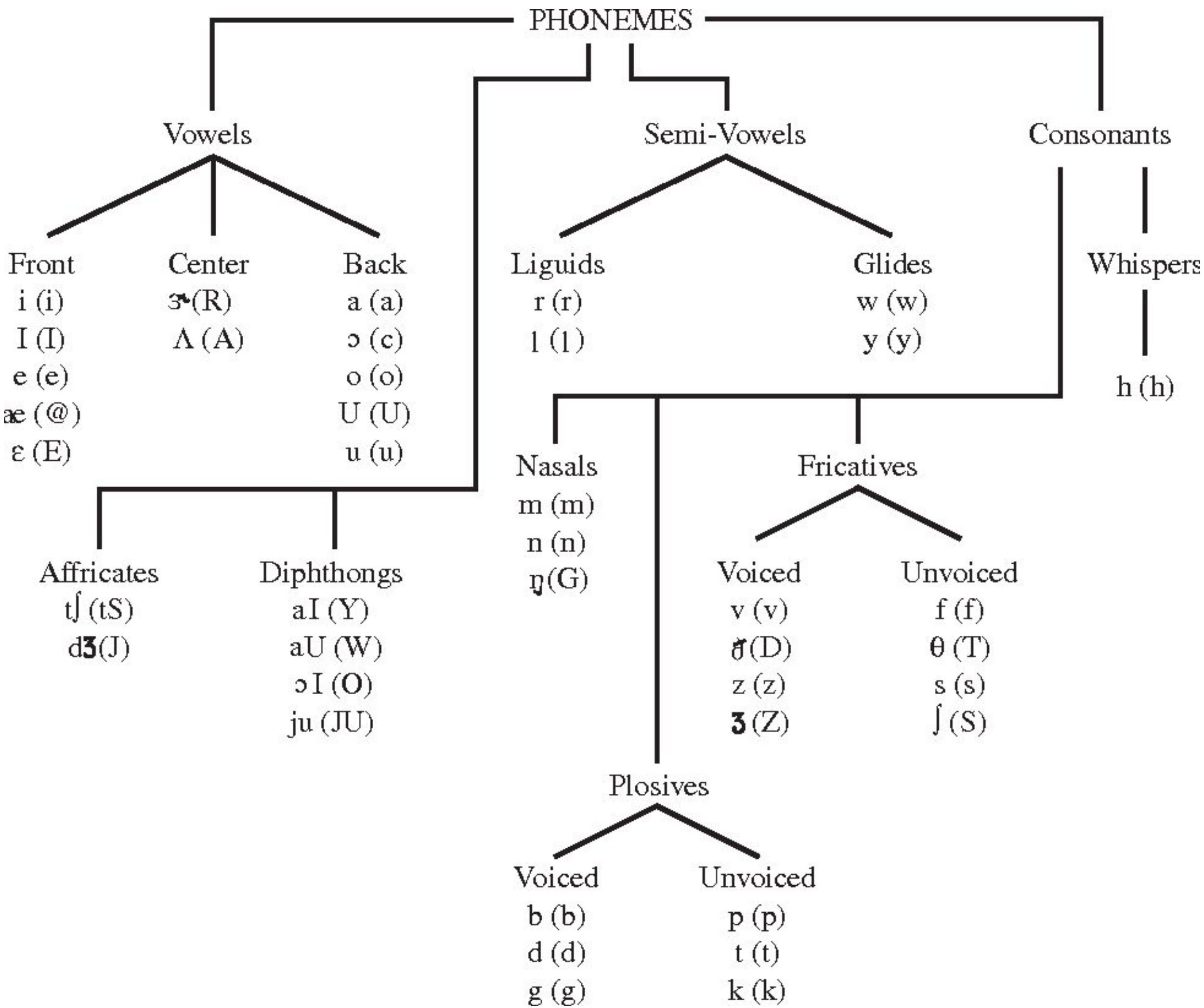
- Phonemes: smallest spoken sound unit

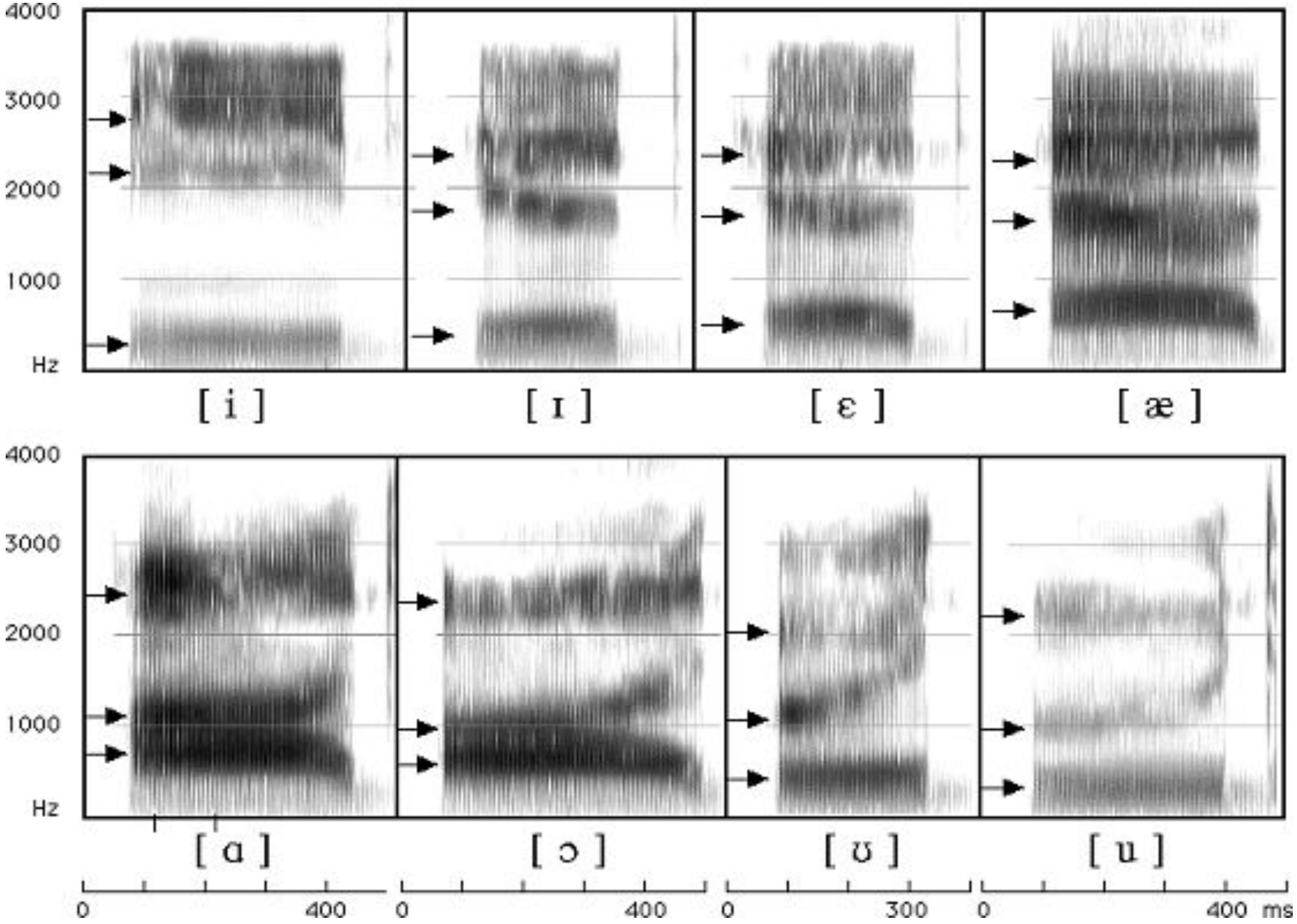
- More consonants than vowels

- Sound distinct

- Concatenation makes words, sentences: Speech




PHONEMES
Vowels
Semi-Vowels
Consonants
Front
i (i)
I (I)
e (e)
æ (@)
ε (E)
Center
ɝ(R)
Λ (A)
Back
a (a)
ɔ (c)
o (o)
U (U)
u (u)
Liguids
r (r)
l (l)
Glides
w (w)
y (y)
Whispers
h (h)
Nasals
m (m)
n (n)
ŋ(G)
Fricatives
Voiced
v (v)
ð(D)
z (z)
ʒ (Z)
Unvoiced
f (f)
θ (T)
s (s)
ʃ (S)
Affricates
tʃ (tS)
dʒ(J)
Diphthongs
aI (Y)
aU (W)
ɔI (O)
ju (JU)
Plosives
Voiced
b (b)
d (d)
g (g)
Unvoiced
p (p)
t (t)
k (k)


4000
3000
2000
1000
Hz
[i]
[ɪ]
[ɛ]
[æ]
4000
3000
2000
1000
Hz
[ɑ]
[ɔ]
[ʊ]
[u]
0
400
0
400
0
300
0
400
ms

Acoustic Features

Statistical Features

Mean: Represents the average value of the signal.

Standard Deviation: Measures the spread or dispersion of the signal values.

Skewness: Describes the asymmetry of the probability distribution of the signal values.

Kurtosis: Measures the "tailedness" of the probability distribution of the signal values.

Acoustic Features

Temporal Features

Zero Crossing Rate: Counts the number of times the signal crosses the zero axis, which can be indicative of the noisiness or percussiveness of a sound.

Root Mean Square Energy (RMSE): Measures the energy of a signal and can indicate loudness.

Temporal Envelope: Represents the slowly varying amplitude envelope of a signal.

Duration: Measures the length of phonemes, words, or segments, contributing to speech rhythm.

Intensity: Represents the loudness of speech segments.

Acoustic Features

Spectral Features

Spectral Centroid: Indicates where the "center of mass" of the spectrum is, often associated with the perceived brightness of sound.

Spectral Bandwidth: Measures the width of the spectrum, providing information about the range of frequencies present.

Spectral Roll-Off: Represents the frequency below which a specified percentage of the total spectral energy lies.

Spectral Flux: Measures changes in the spectral shape over time and is useful for detecting onsets and transients.

Fundamental Frequency (F0): Represents the perceived pitch contour of speech, which can convey intonation and emotion.

Acoustic Features

Spectro-temporal Features

Spectrogram

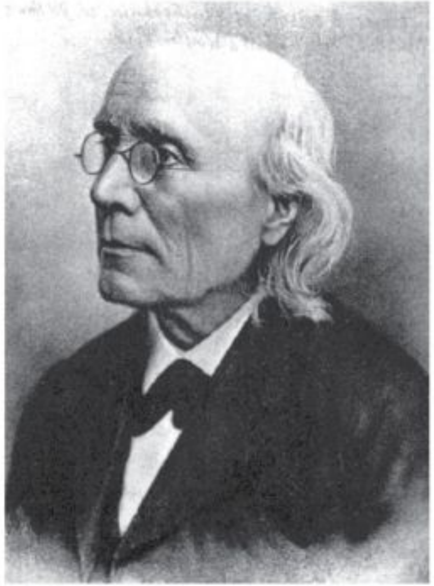
Psychophysical methods & a brief intro to the nervous system

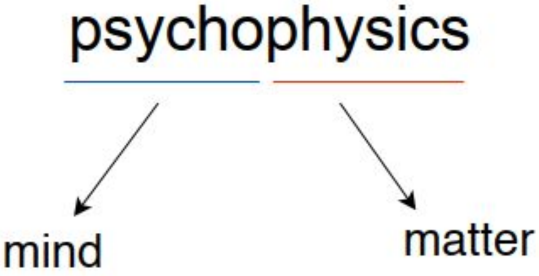
Jonathan Pillow
Perception (PSY 345 / NEU 325)
Princeton University, Spring 2015

Lec. 3

The Dawn of Psychophysics

Gustav Fechner (1801–1887) often considered founder of experimental psychology





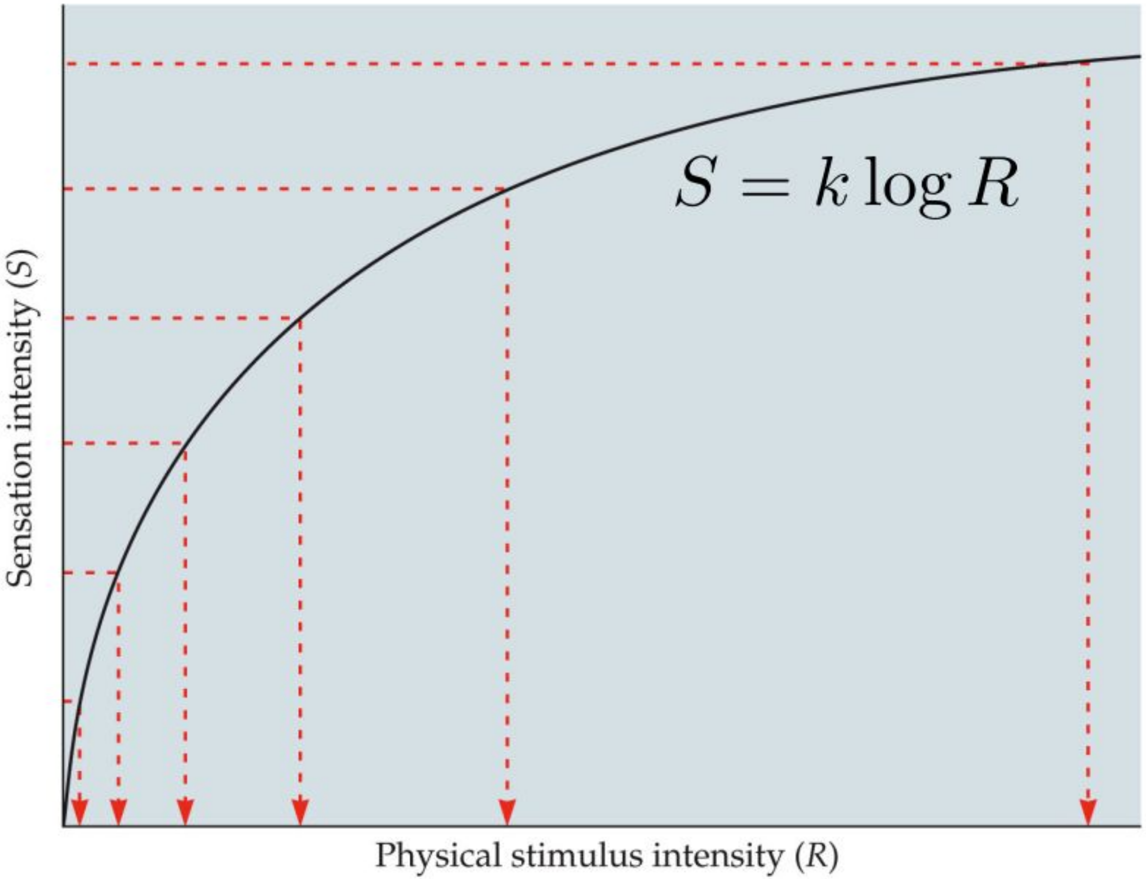
psychophysics

mind

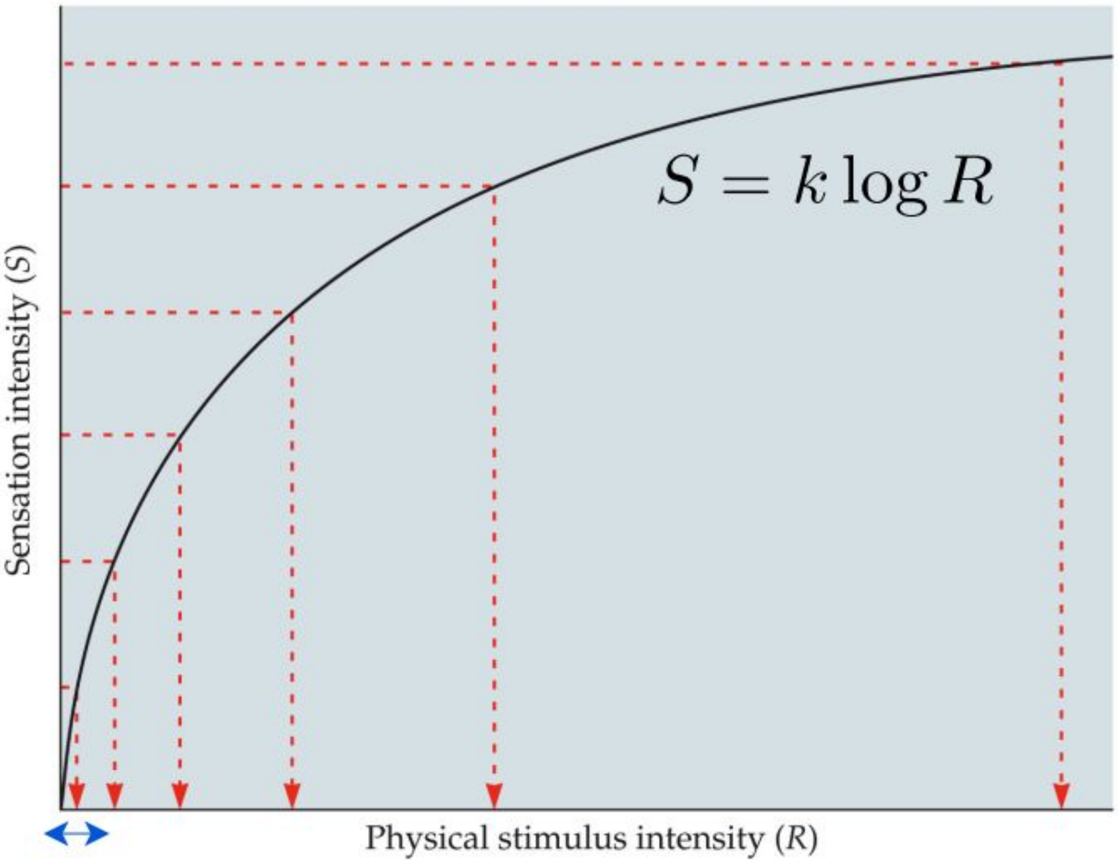
matter

- scientific theory of the relationship between mind and matter

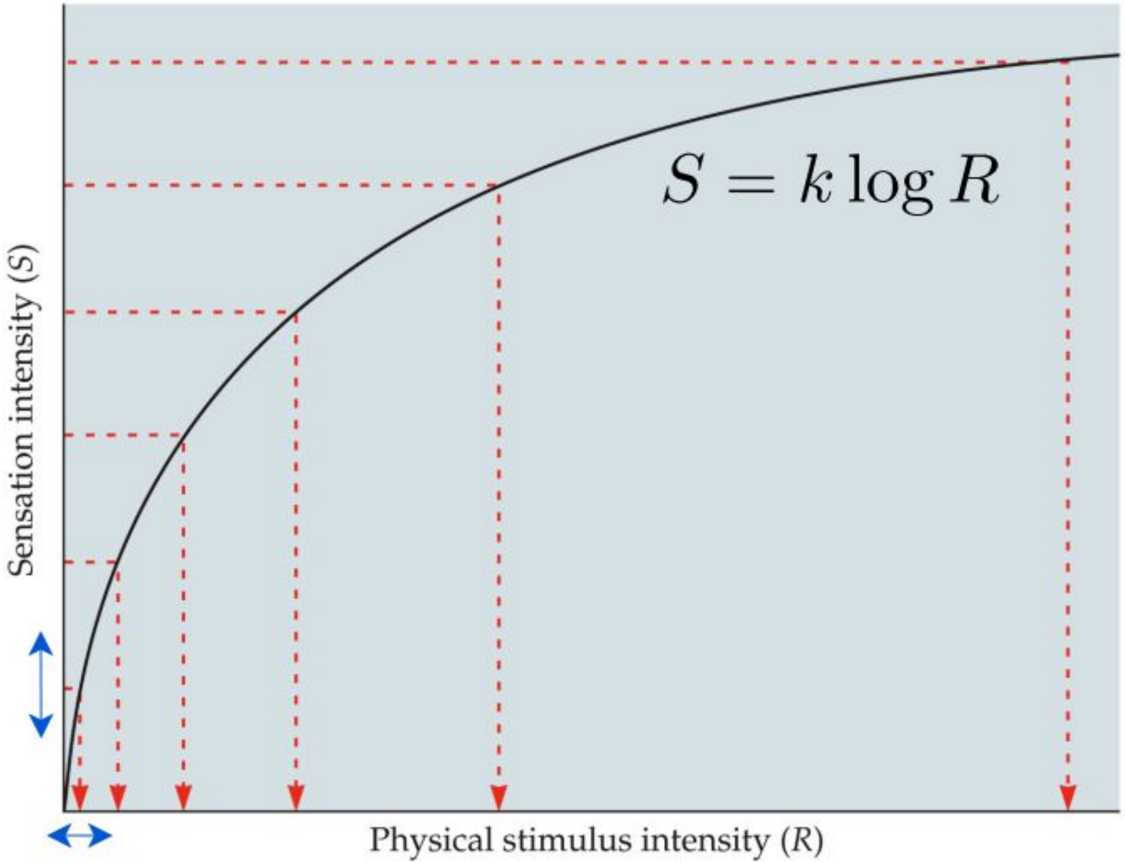
Fechner’s law



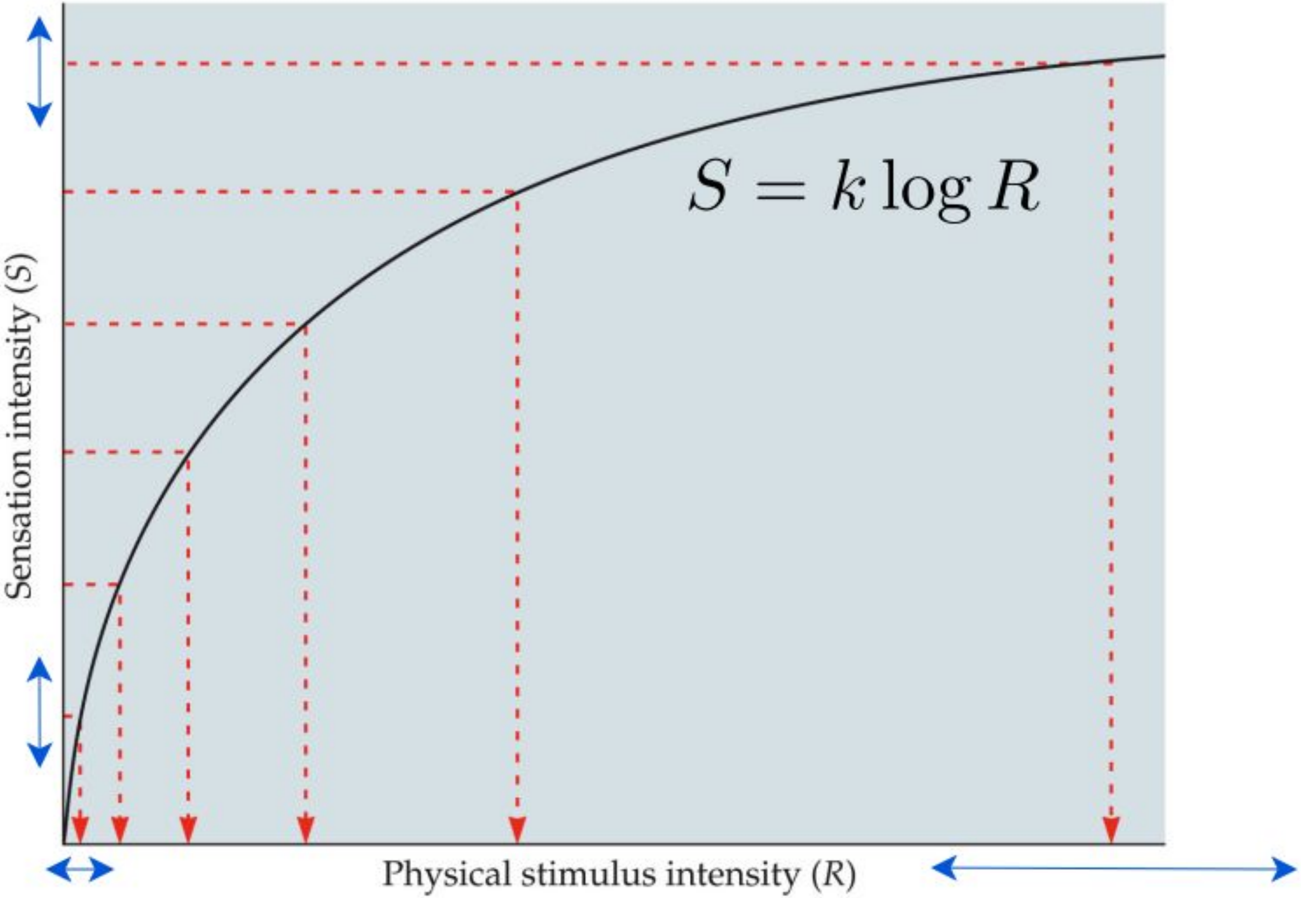
Fechner’s law



Fechner's law



Fechner's law



The Dawn of Psychophysics

Ernst Weber (1795–1878)

"Weber's Law"

- law about how stimulus intensity relates to detectability of stimulus changes
- As stimulus intensity increases, magnitude of change must increase proportionately to remain noticeable

Example:

1 pound change in a 20 pound weight

is just as detectable as

0.2 pound change in a 4 pound weight

The Dawn of Psychophysics

Ernst Weber (1795–1878)

“Weber’s Law”

- law about how stimulus intensity relates to detectability of stimulus changes
- As stimulus intensity increases, magnitude of change must increase proportionately to remain noticeable

Example:

1 pound change in a 20 pound weight $\frac{1}{20} = .05$

is just as detectable as

0.2 pound change in a 4 pound weight $\frac{0.2}{4} = .05$

A little math (don’t freak out)

Fechner’s law:

$$\underset{\substack{\uparrow\\\text{percept}\\\text{intensity}}}{S} = k \log \underset{\substack{\uparrow\\\text{stimulus}\\\text{intensity}}}{R}$$

A little math (don't freak out)

Fechner's law:

$$\underset{\substack{\uparrow \\ \text{percept} \\ \text{intensity}}}{S} = k \log \underset{\substack{\uparrow \\ \text{stimulus} \\ \text{intensity}}}{R}$$

differentiate both sides

A little math (don't freak out)

Fechner's law:

$$\underset{\substack{\uparrow\\ \text{percept}\\ \text{intensity}}}{S} = k \log \underset{\substack{\uparrow\\ \text{stimulus}\\ \text{intensity}}}{R}$$

differentiate both sides

Weber's law:

$$\underset{\substack{\uparrow\\ \text{change in percept}\\ \text{intensity}}}{dS} = k \frac{dR}{R} \longleftarrow \substack{\text{change in}\\ \text{stimulus}\\ \text{intensity}}$$

A little math (don't freak out)

Fechner's law:

$$\underset{\substack{\uparrow \\ \text{percept} \\ \text{intensity}}}{S} = k \log \underset{\substack{\uparrow \\ \text{stimulus} \\ \text{intensity}}}{R}$$

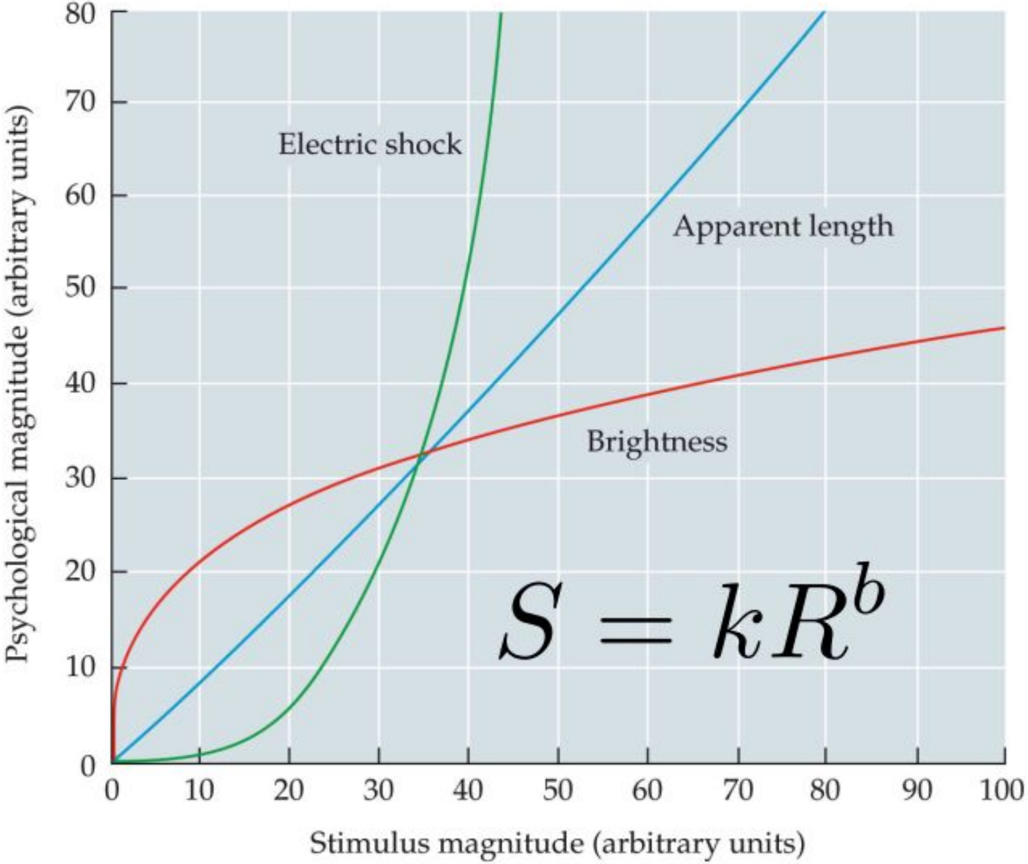
differentiate both sides

Weber's law:

$$\underset{\substack{\uparrow \\ \text{change in percept} \\ \text{intensity}}}{dS} = k \frac{dR}{R} \longleftarrow \text{change in stimulus intensity}$$

So detectability ("how much the percept changes") is determined by the ratio of stimulus change dR to stimulus intensity R.

Stevens’ Power Law



$$S = kR^b$$

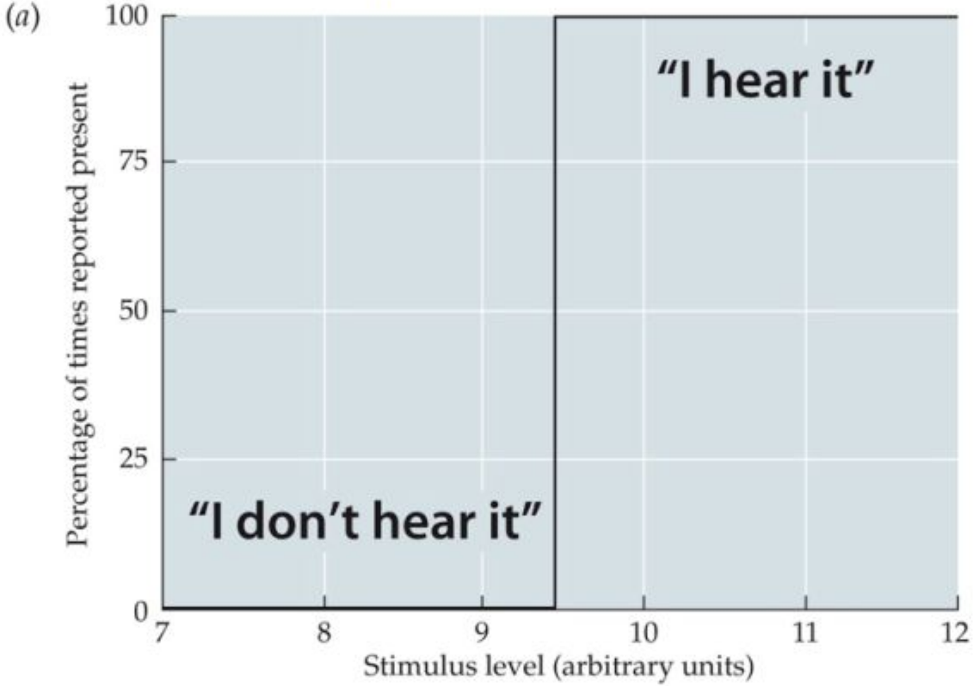
Psychophysics

- detection (yes/no)
- discrimination (e.g., bigger than)
- estimation (report the stimulus exactly)

All provide indirect measure of internal mental state!

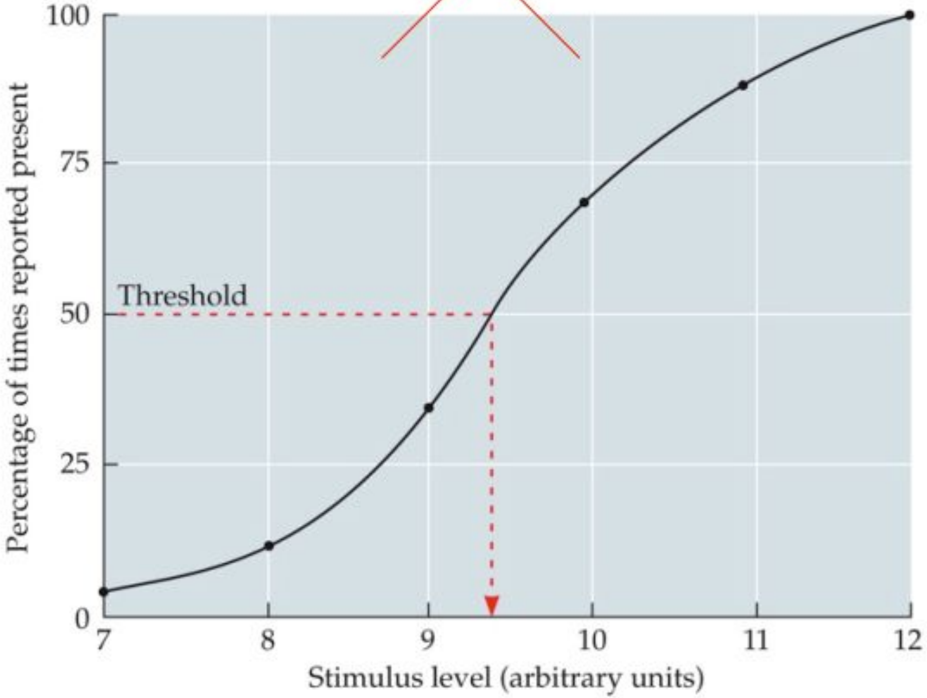
Detection

perfect threshold



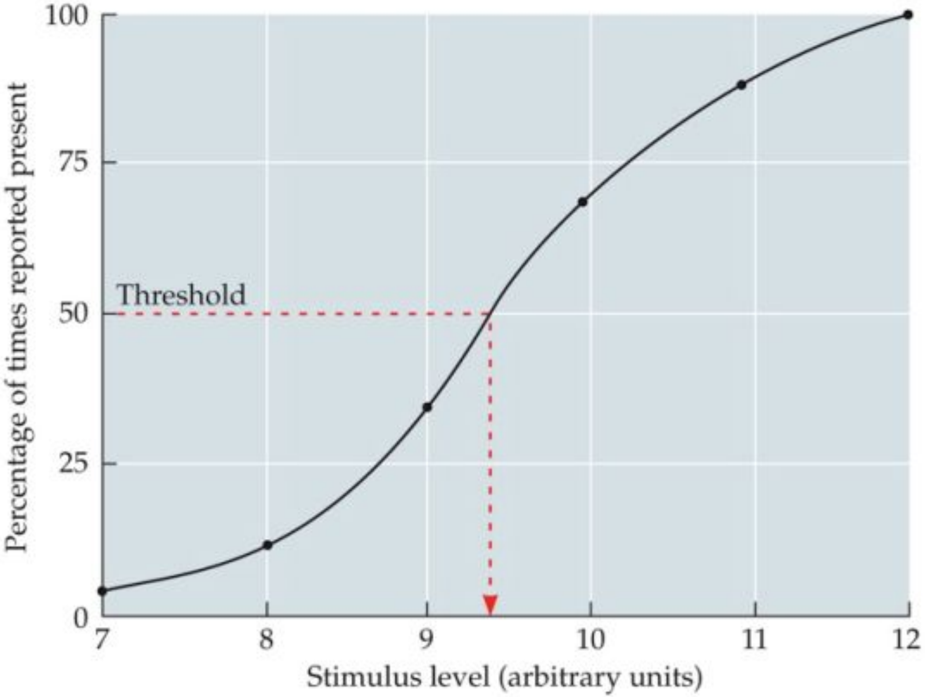
Detection

~~perfect threshold~~ noise



psychometric function

- relates physical quantity to the probability of detecting it



Signal detection theory: A psychophysical theory that quantifies the response of an observer to the presentation of a signal in the presence of noise

Signal detection theory

- **Hit**: Stimulus is presented and observer responds “Yes”
- **Miss**: Stimulus is presented and observer responds “No”
- **False alarm**: Stimulus is not presented and observer responds “Yes”
- **Correct rejection**: Stimulus is not presented and observer responds “No”

Signal Detection Theory Terms to know:

“noise” distribution: values arising when stimulus not present

“signal” distribution: values arising when signal + noise present

Type I error: rate of “false alarms”, or false positives

Type II error: rate of “misses”, or false negatives

psychometric function: describes probability of saying “I heard it” as function of stimulus intensity

Understanding MFCCs

One of the most popular audio feature vector

- Why mel-scale?
- What is mel-scale?
- Reducing dimension of DFT spectrum
- Computing MFCCs

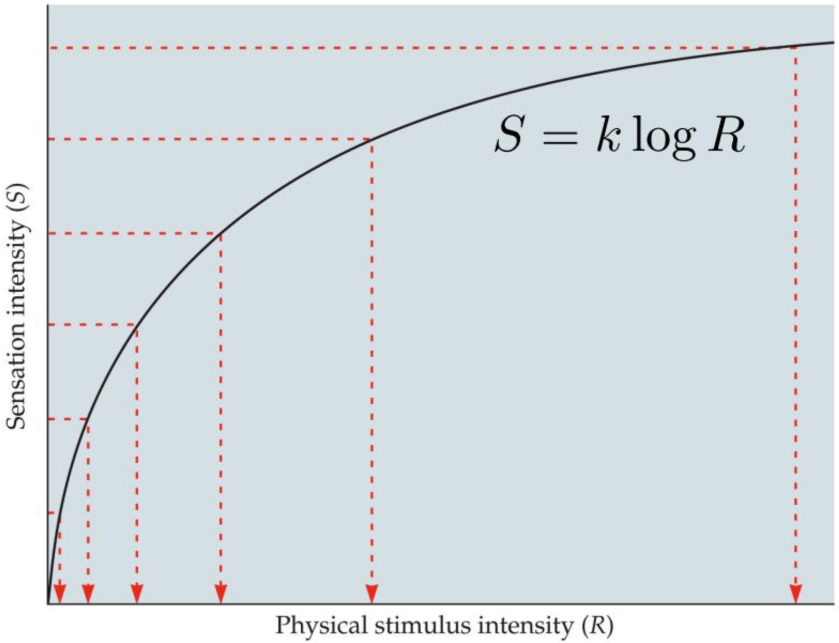
Linear scale for frequency

If you were to analyze the audio signal solely based on its physical frequency content, using a linear scale (Hertz), you might face a problem.

The linear scale would treat all frequency changes equally

- meaning that a change from **100 Hz to 200 Hz** would be considered **as significant as a change** from **1000 Hz to 1100 Hz**

Fechner’s law



However, human hearing doesn't work that way. We perceive frequency changes differently depending on where they occur in the spectrum. We are more sensitive to changes in lower frequencies than higher ones.

Mel-scale

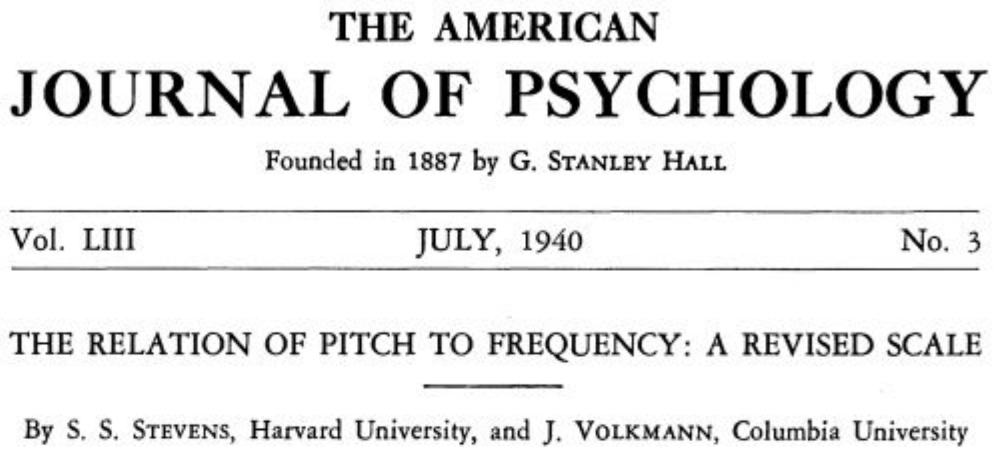
This is where the Mel scale comes into play.

- By converting the frequency representation of the audio signal into the Mel scale, you can ensure that your speech recognition system prioritizes the frequency regions that matter most for human speech perception.

This not only makes your system more efficient but also more accurate in recognizing spoken words, as it aligns with how our ears and brain process sound.

Mel-scale

- It was proposed in the below research paper from 1937



THE AMERICAN

JOURNAL OF PSYCHOLOGY

Founded in 1887 by G. STANLEY HALL

Vol. LIII JULY, 1940 No. 3

THE RELATION OF PITCH TO FREQUENCY: A REVISED SCALE

By S. S. STEVENS, Harvard University, and J. VOLKMANN, Columbia University

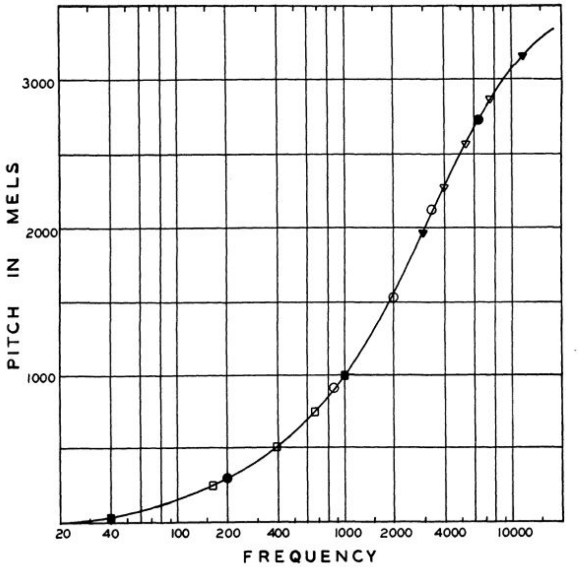
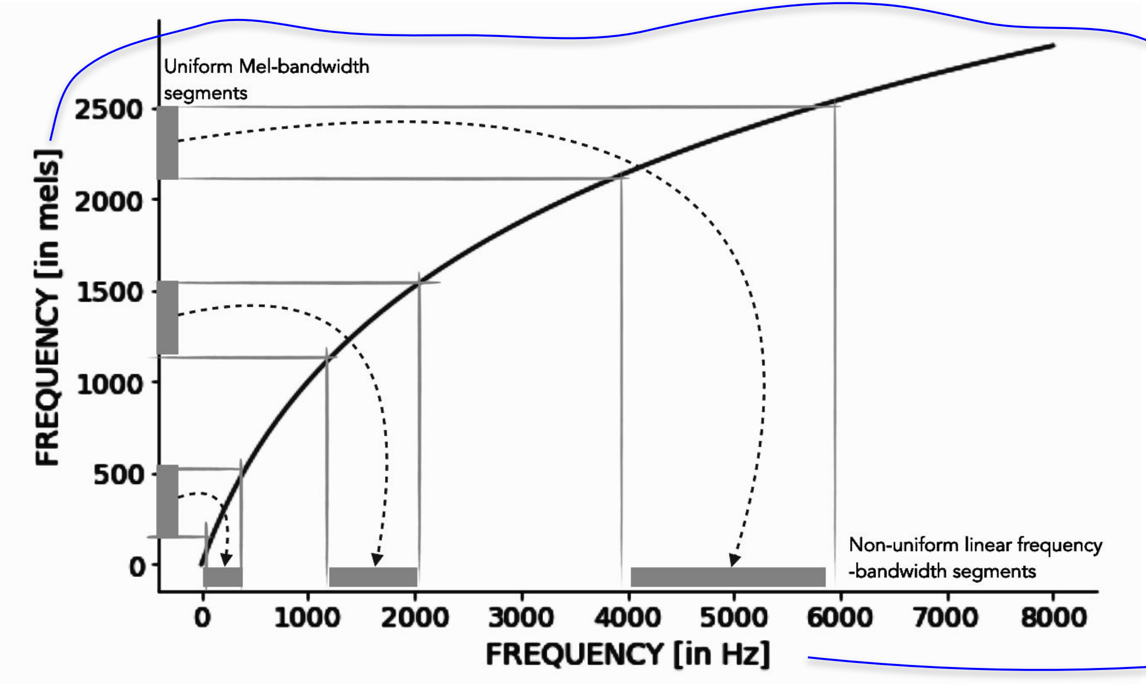


FIG. 2. THE PITCH-SCALE

The curve shows how pitch, scaled in subjective units (ordinates), varies with frequency. The circles, squares and triangles represent data obtained in the experiment on equal sense-distances. The filled figures mark the ends of three frequency-ranges and the hollow figures show the points arrived at when the *Os* divided the ranges into four equal intervals of pitch.

What is mel-scale?

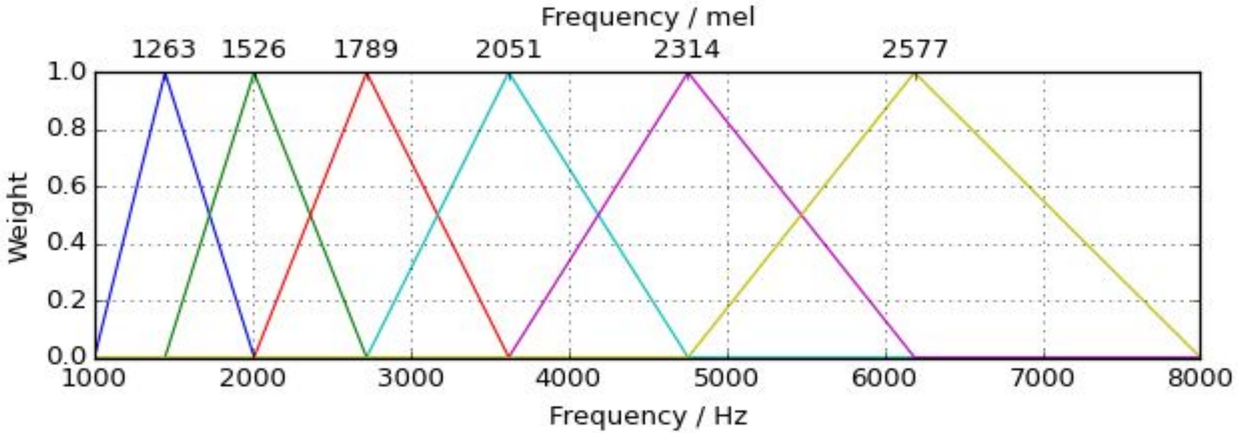
- Helps to map perceived equal frequency difference into linear scale



$$m = 2410 \log_{10}\left(1 + \frac{f}{625}\right)$$

Reducing dimensions of DFT spectrum

Using Filterbank



Filterbank (more generally)

Imagine you're developing a mobile app for musicians and audio enthusiasts.

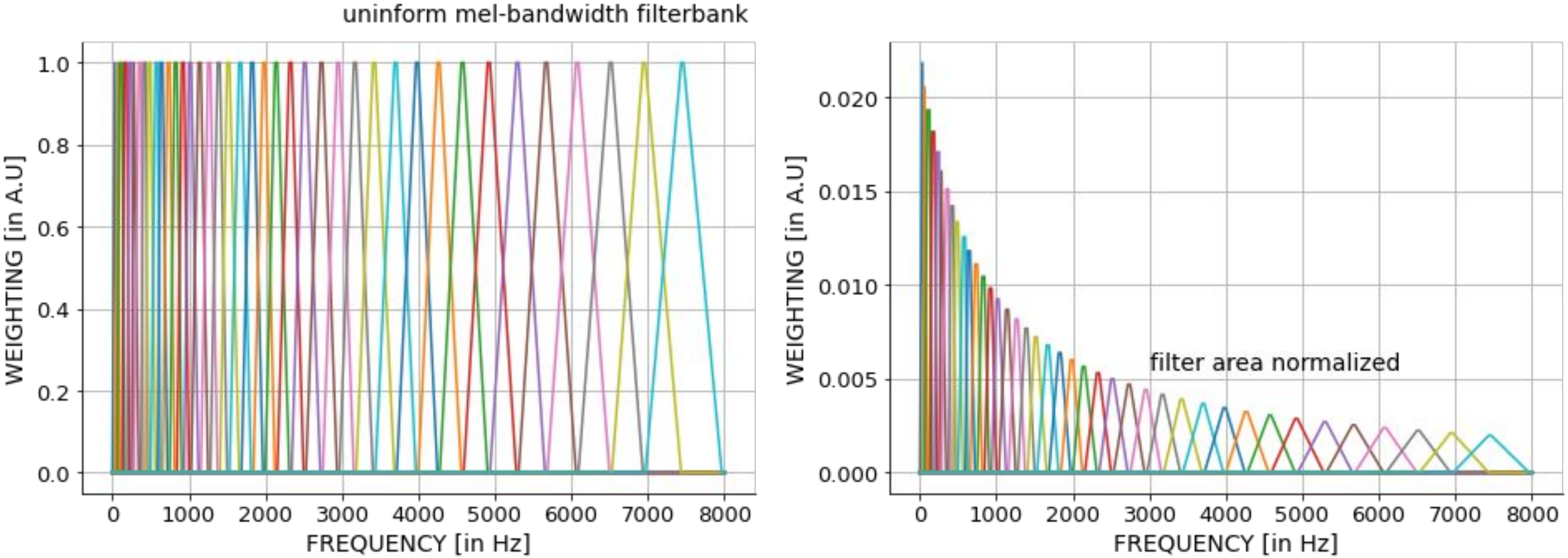
A features you want to include is a real-time audio equalizer that allows users to adjust the frequency balance of the music they're listening to. **This equalizer should let them boost or attenuate specific frequency ranges like bass, midrange, and treble to customize their listening experience.**

To implement this equalizer, you need a way to isolate and manipulate different frequency components in the audio signal. This is where filterbanks come into play.

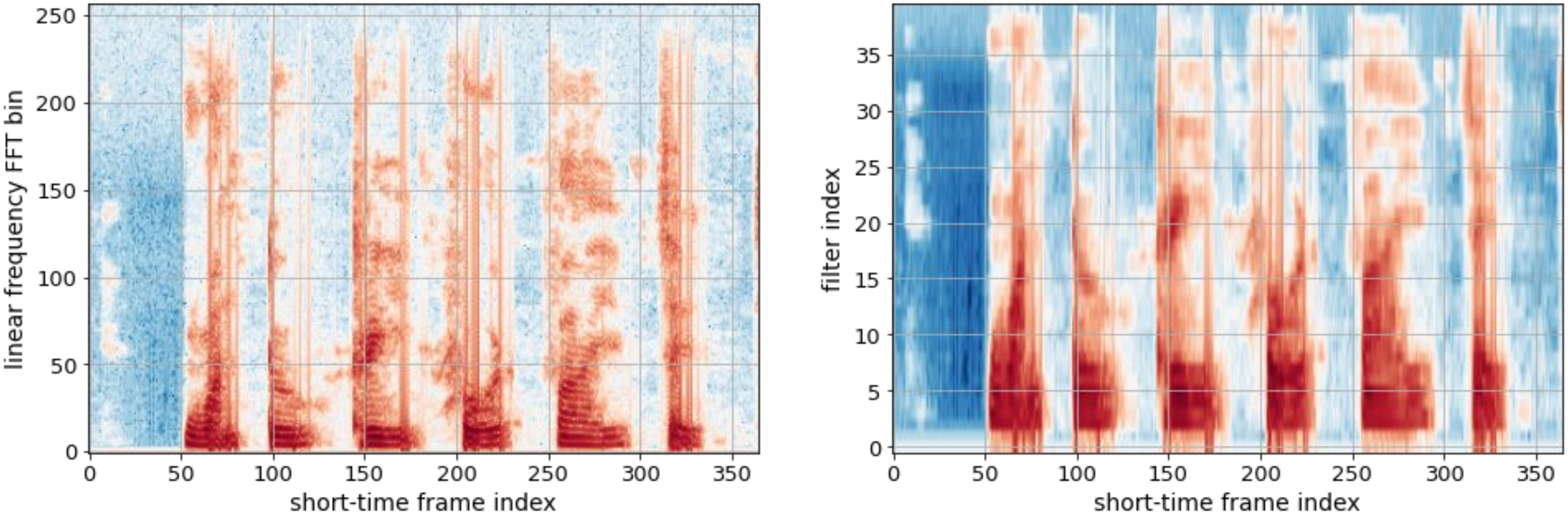
Filterbank
A collection of filters designed to pass specific frequency ranges while attenuating others.
For example, you have filters that emphasize low frequencies (bass), midrange frequencies (vocals and instruments), and high frequencies (treble).

Applications:
Real-time control **Customized audio** **Enhanced audio**

Filterbank in computing MFCC

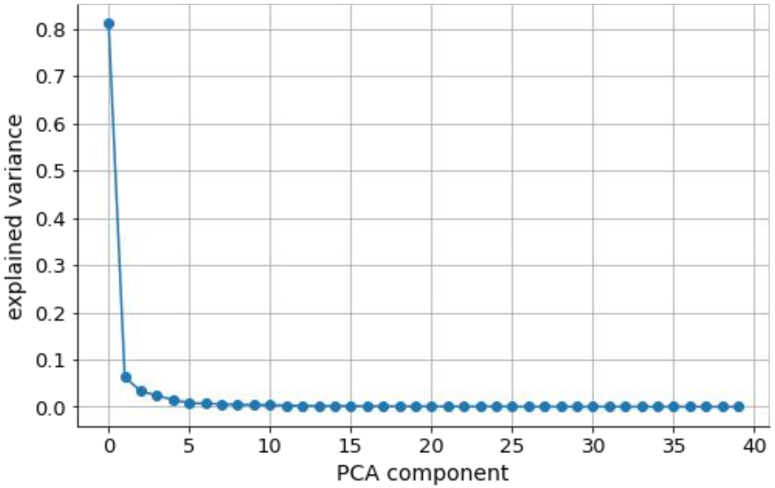


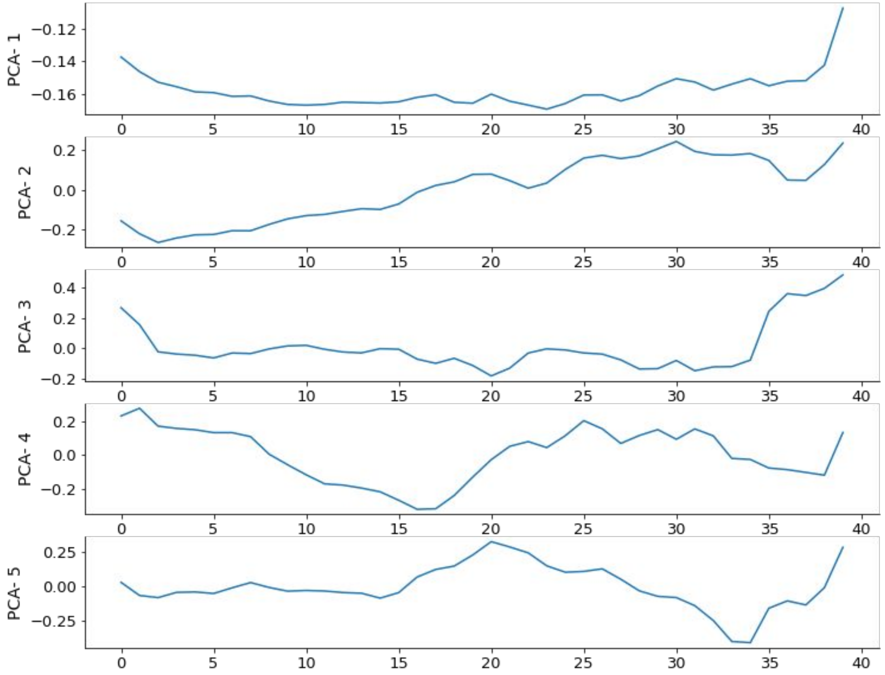
Linear spectrogram to mel-Spectrogram



Reducing dimension further

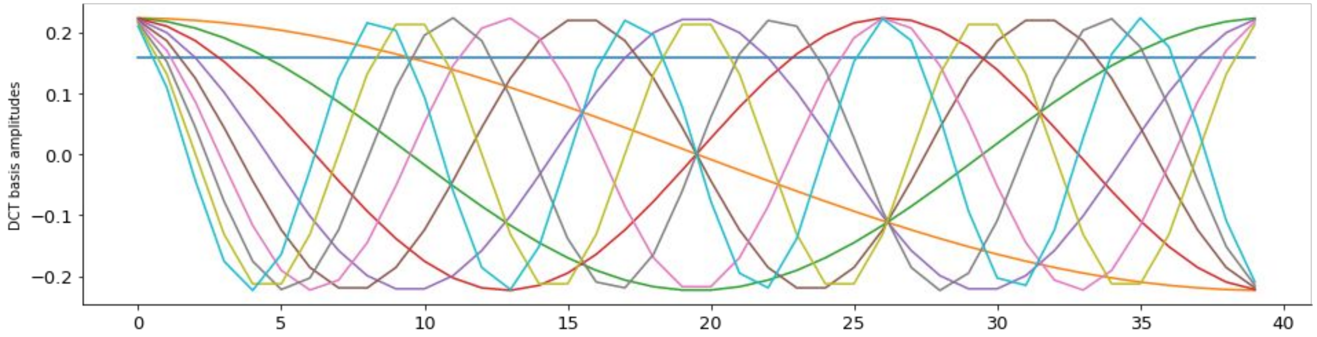
Do a PCA after pooling the columns of the mel-spectrogram



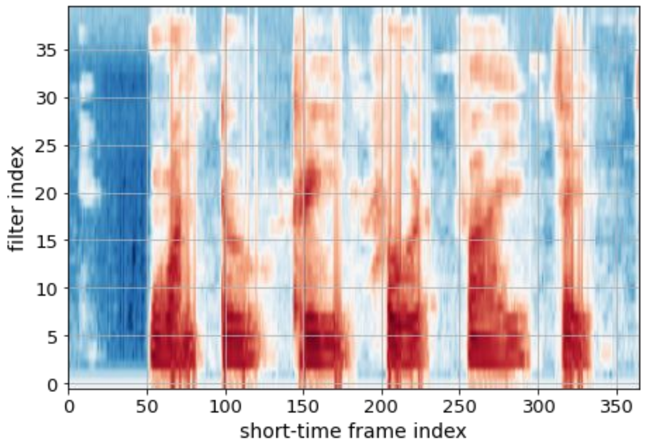


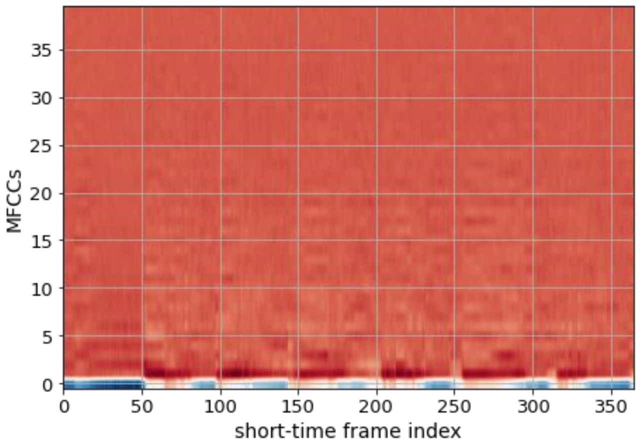
Reducing dimension further

Instead of PCA we do a DCT as the basis of DCT look similar to the PCA basis

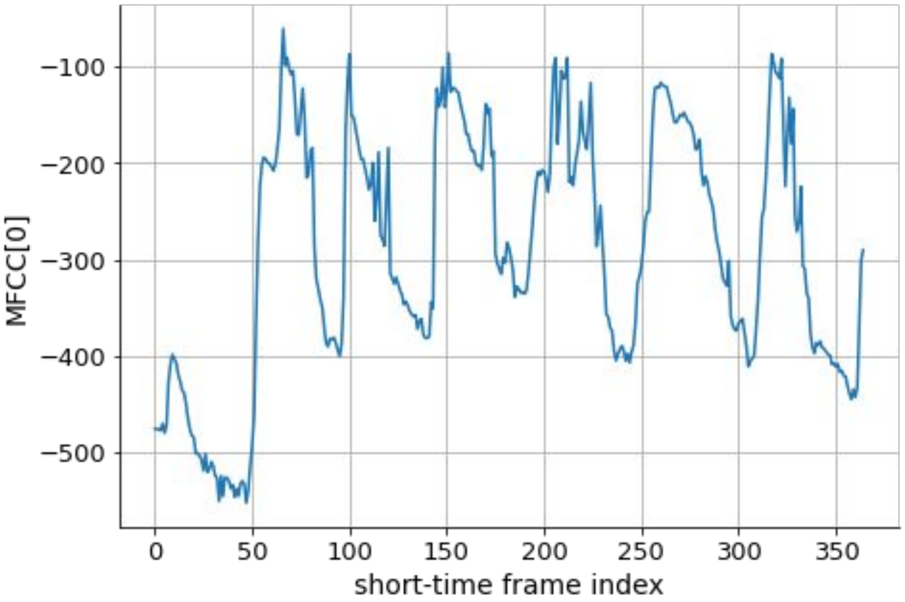


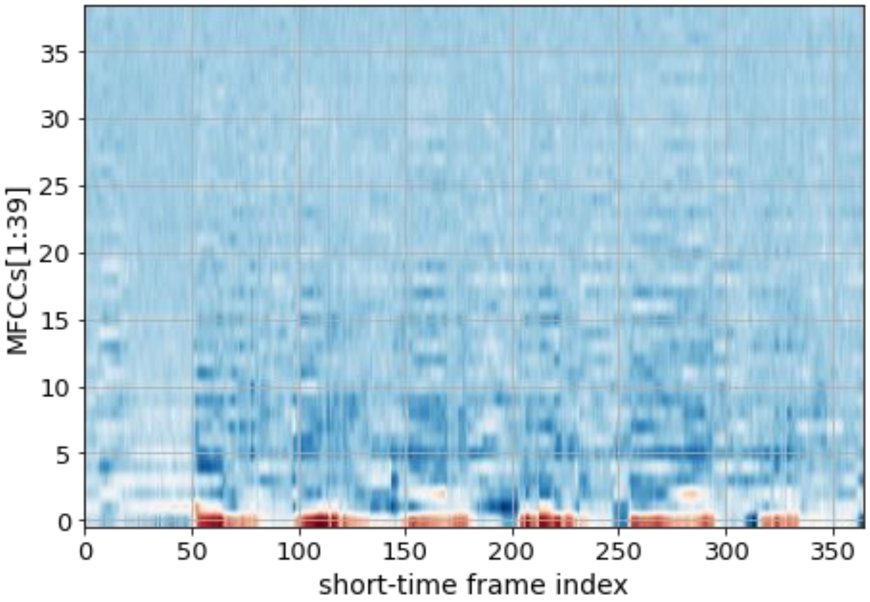
Thus we get - MFCCs





Visualizing one element of MFCC vetcor





Acoustic Features

Spectro-temporal Features

Spectrogram: It provides a time-frequency representation of the signal's energy distribution.

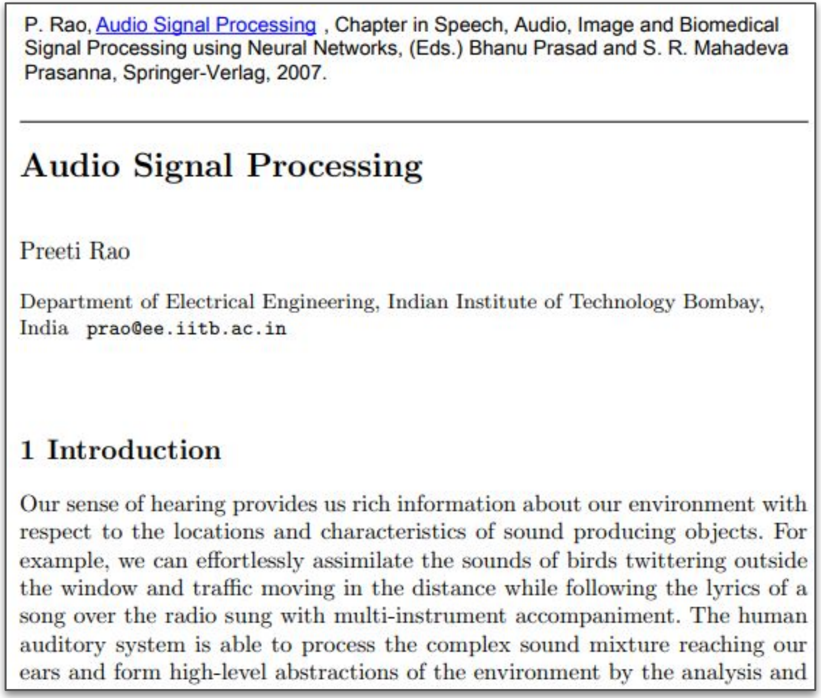
MFCCs (Mel-Frequency Cepstral Coefficients):
Similar to a standard spectrogram, but computed using a mel-frequency scale instead of a linear frequency scale. Mel-frequency spectrograms are commonly used in speech and audio processing, as they better match human auditory perception.

Delta MFCCs: Represent the rate of change of MFCCs over time, useful for modeling dynamic aspects of speech.

Delta-Delta MFCCs: Capture second-order derivatives of MFCCs and provide even more information about dynamics.

Gammatone Features: Gammatone features are designed to mimic the filtering properties of the human auditory system. They are useful for tasks involving auditory scene analysis and sound source localization

References



P. Rao, Audio Signal Processing , Chapter in Speech, Audio, Image and Biomedical Signal Processing using Neural Networks, (Eds.) Bhanu Prasad and S. R. Mahadeva Prasanna, Springer-Verlag, 2007.

Audio Signal Processing

Preeti Rao

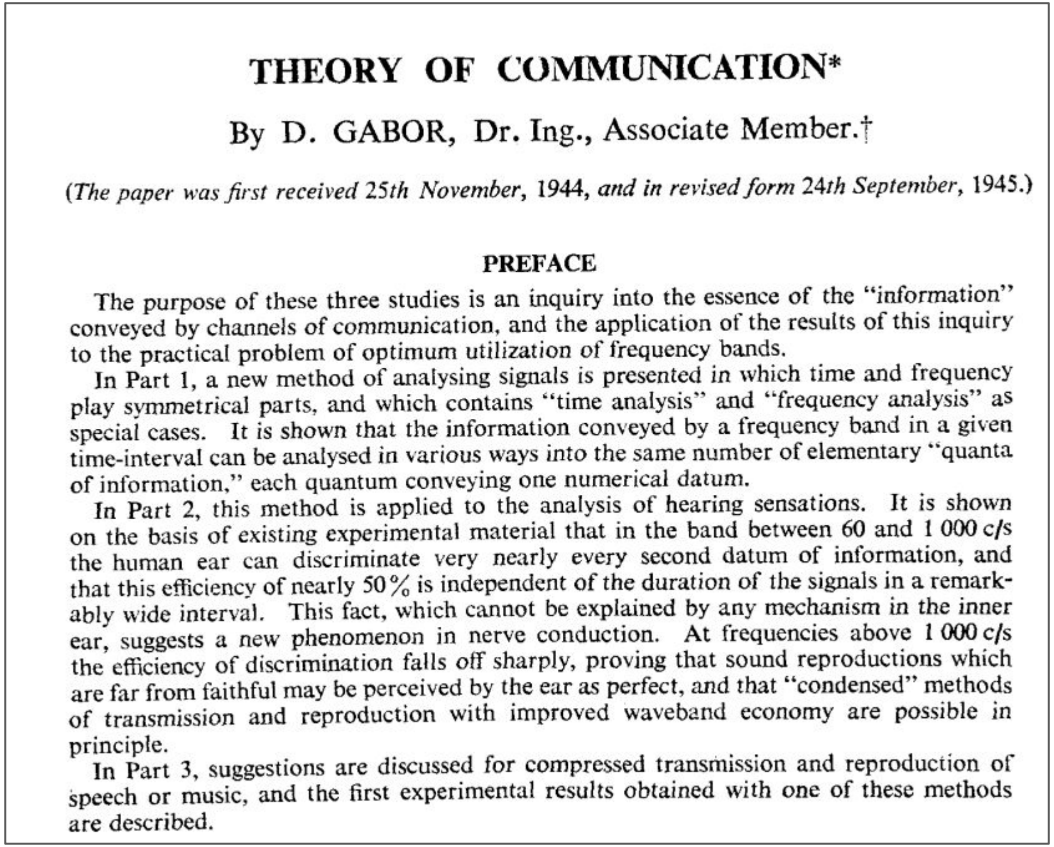
Department of Electrical Engineering, Indian Institute of Technology Bombay, India prao@ee.iitb.ac.in

1 Introduction

Our sense of hearing provides us rich information about our environment with respect to the locations and characteristics of sound producing objects. For example, we can effortlessly assimilate the sounds of birds twittering outside the window and traffic moving in the distance while following the lyrics of a song over the radio sung with multi-instrument accompaniment. The human auditory system is able to process the complex sound mixture reaching our ears and form high-level abstractions of the environment by the analysis and

audio and music processing in Python

Information communication



THEORY OF COMMUNICATION*

By D. GABOR, Dr. Ing., Associate Member.†

(*The paper was first received 25th November*, 1944, *and in revised form* 24*th September*, 1945.)

PREFACE

The purpose of these three studies is an inquiry into the essence of the "information" conveyed by channels of communication, and the application of the results of this inquiry to the practical problem of optimum utilization of frequency bands.

In Part 1, a new method of analysing signals is presented in which time and frequency play symmetrical parts, and which contains "time analysis" and "frequency analysis" as special cases. It is shown that the information conveyed by a frequency band in a given time-interval can be analysed in various ways into the same number of elementary "quanta of information," each quantum conveying one numerical datum.

In Part 2, this method is applied to the analysis of hearing sensations. It is shown on the basis of existing experimental material that in the band between 60 and 1 000 c/s the human ear can discriminate very nearly every second datum of information, and that this efficiency of nearly 50% is independent of the duration of the signals in a remarkably wide interval. This fact, which cannot be explained by any mechanism in the inner ear, suggests a new phenomenon in nerve conduction. At frequencies above 1 000 c/s the efficiency of discrimination falls off sharply, proving that sound reproductions which are far from faithful may be perceived by the ear as perfect, and that "condensed" methods of transmission and reproduction with improved waveband economy are possible in principle.

In Part 3, suggestions are discussed for compressed transmission and reproduction of speech or music, and the first experimental results obtained with one of these methods are described.