PUE 4113 Speech Processing.

Prof. Ciira Maina
ciira.maina@dkut.ac.ke

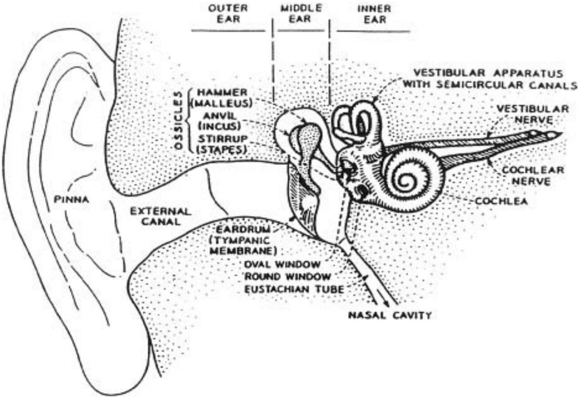
2nd June, 2023

Speech Perception

1. The major components of the auditory perception system
   - Ears
   - Brain
2. The ear transforms sound into vibrations of the basilar membrane
3. Information is extracted by the brain

The Ear

- The outer ear gathers sound and conducts it through the external canal to the middle ear
- The middle ear converts the sound waves to mechanical pressure waves
- The inner ear conducts neural signals to the brain



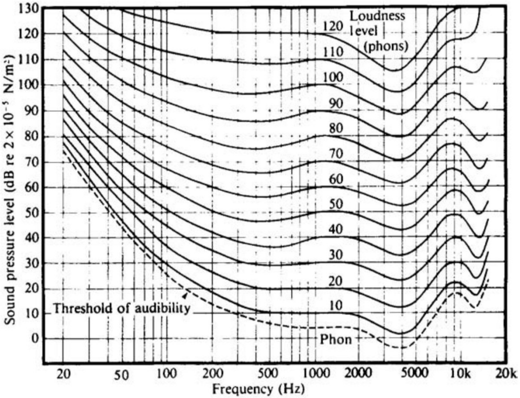
Perceptual vs Physical Quantities

1. There is a distinction between the perceptual qualities of sound and the measurable physical quantities

Physical Quantity	Perceptual Quality
Intensity	Loudness
Fundamental frequency	Pitch
Spectral shape	Timbre
Onset/offset time	Timing
Phase difference in binaural hearing	Location

The Equal Loudness Curve

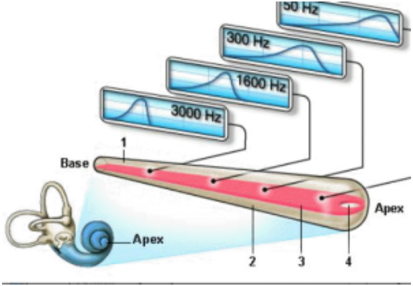
- Loudness is related to the sound pressure level (SPL)
- Perception of loudness is frequency dependant
- Low frequencies must be more intense to be audible



Critical Bands

- The basilar membrane performs spectral analysis on the audio signal
- This spectral analysis is modeled as a filter bank of bandpass filters
- Each bandpass filter has a bandwidth given by

$$\Delta f_c = 25 + 75[1 + 1.4(f_c/1000)^2]^{0.69}$$



1

[1] rcam.fr/docs/AudioSculpt/3.0/co/MaskingEffectIntro.html

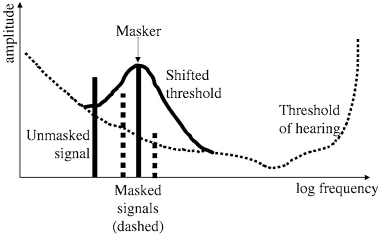
Pitch Perception

- The relationship between pitch and frequency is nonlinear
- 1KHz corresponds to 1000 mels

$$\text{Pitch} \quad \text{in} \quad \text{mels} = 1127 \ln(1 + \frac{f}{700})$$

Masking

- Masking occurs when one sound makes a sound of nearby frequency inaudible
- An intense sound increases the threshold of audibility for nearby frequencies



Introduction to Phonetics

Phonetics

- A phoneme is a minimal unit of speech sound that help distinguish words
- The number of phonemes varies from language to language. Usually between 32 and 64.
- Consider Kenyan and American English[2]
- The The Carnegie Mellon University Pronouncing Dictionary `http://www.speech.cs.cmu.edu/cgi-bin/cmudict?in=desert`

Phoneme	Example	Translation
AA	odd	AA D
AE	at	AE T
AH	hut	HH AH T
AO	ought	AO T
AW	cow	K AW
AY	hide	HH AY D

[2] Gakuru, M. (2009). Development of a Kenyan English Text To Speech System: A Method of Developing a TTS for a previously undefined English Dialect. In Tenth Annual Conference of the International Speech Communication Association.

Vowels

- One of the major sound classes along with consonants
- No constrictions or obstructions in the oral cavity
- Variation in tongue placement leads to different vowel sounds
- The vocal folds vibrate at the fundamental frequency $F0$
- The oral cavity resonates at $F1$ and $F2$

Consonants

- Significant constriction or obstruction within the vocal tract
- Some consonants are voiced
- Consonants are classified as

Manner	Sample Phone	Example Words	Mechanism
Plosive	/p/	tat, tap	Closure in oral cavity
Nasal	/m/	team, meet	Closure of nasal cavity
Fricative	/s/	sick, kiss	Turbulent airstream noise
Retroflex liquid	/r/	rat, tar	Vowel-like, tongue high and curled back
Lateral liquid	/l/	lean, kneel	Vowel-like, tongue central, side airstream
Glide	/y/,/w/	yes, well	Vowel-like

Speech Signal Analysis

DFT Review

Review of the DFT in the notebook

Short Time Analysis

- Speech is a slow varying signal
- We process the signal in blocks over which the properties of the signal are assumed stationary
- The entire speech signal is denoted by $x[m]$ a specific block $\hat{n}$ is obtained as follows

$$x_{\hat{n}}[m] = x[m]w[\hat{n} - m] \tag{1}$$

Short Time Analysis

- The window $w[\hat{n} - m]$ is a time shifted window
- This window selects a segment centered at $m = \hat{n}$
- A common window is the Hamming window given by

$$w[m] = \begin{cases} 0.54 + 0.46\cos(\pi m/M) & -M \leq m \leq M \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Short Time Analysis Example

- Two simple applications of short time analysis are energy computation and zero-crossing rate.
- These features are useful in processing speech and have applications such as voice activity detection
- The short time energy is computed as

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} (x[m]w[\hat{n} - m])^2 \tag{3}$$

Short Time Analysis Example

- The zero crossing rate is computed as

$$Z_{\hat{n}} = \sum_{m=-\infty}^{\infty} 0.5|\mathrm{sgn}\{x[m]\} - \mathrm{sgn}\{x[m-1]\}|w[\hat{n}-m] \quad (4)$$

Where

$$\mathrm{sgn}\{x\} = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases} \quad (5)$$

Short Time Fourier Transform (STFT)

- The STFT is defined as

$$X_{\hat{n}}(e^{j\hat{\omega}}) = \sum_{m=-\infty}^{\infty} x[m]w[\hat{n}-m]e^{-j\hat{\omega}m} \tag{6}$$

- To be practical, we evaluate the STFT at a discrete set of frequencies
- In addition, the finite duration window is moved in steps of $R > 1$

Readings

- HAH - Chapter 5-6
- RS - Chapter 4