Phonetics

Phonology

- **Phonology –** Idea of decomposing speech and words into smaller units
  - Useful for algorithms for **speech recognition** and **speech synthesis** or **text-to-speech.**

- Phonetics - The study of linguistic sounds
  - ✓ how they are produced by the articulators of the human vocal tract,
  - ✓ how they are realized acoustically
  - ✓ how the acoustic realization can be digitized and processed.

Speech Sounds and Phonetic Transcription

- *Phonetics* - The study of the pronunciation of words
- Phone
  - Speech sound
  - Represented with phonetic symbols (similar to alphabets in english)
- Phonetic Transcription:
  - A writing system for representing speech sounds
- Standards to transcribe the sounds of human language

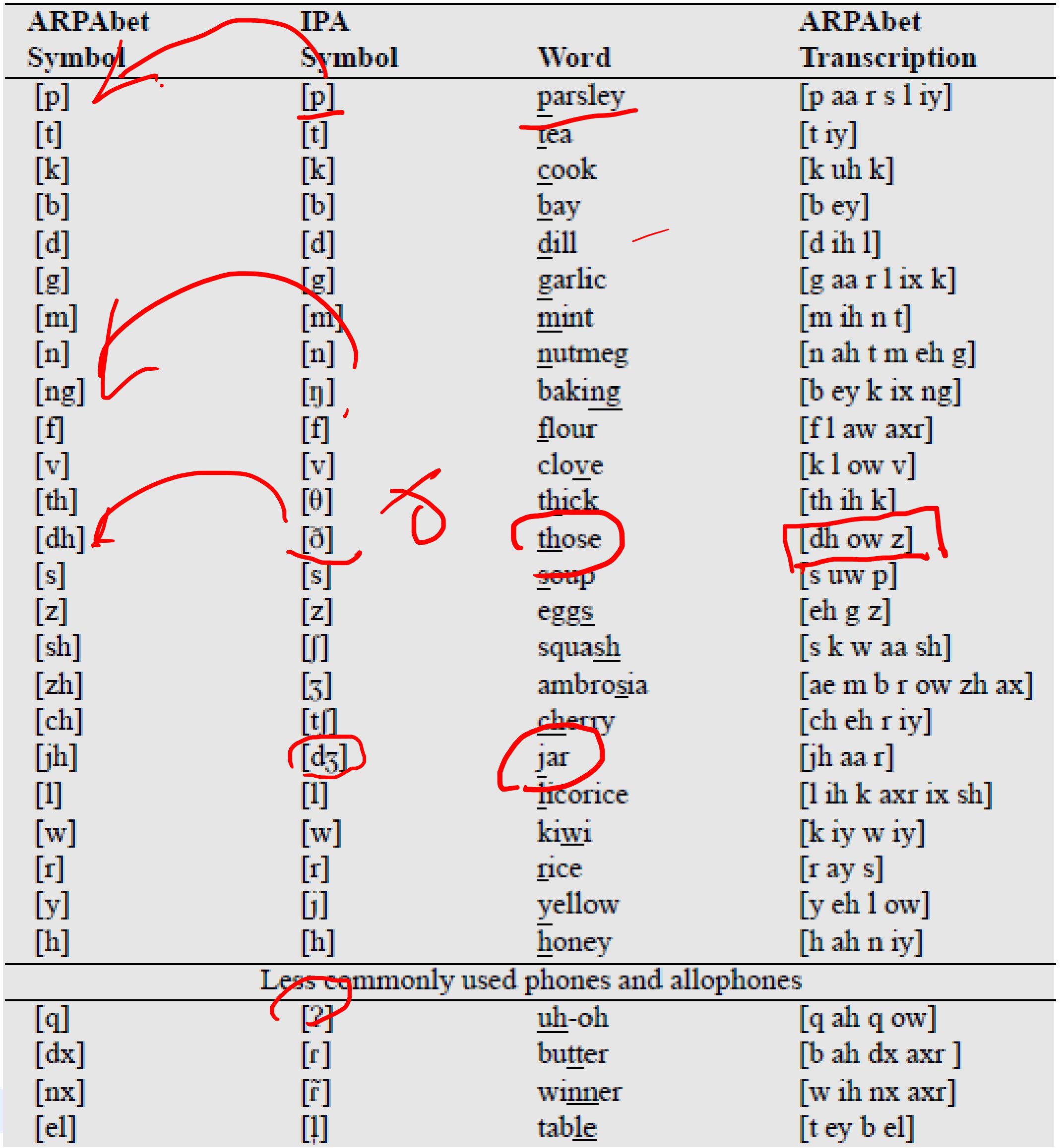
International Phonetic Alphabet (IPA)

- set of principles to transcribe the sounds

ARPAbet

-- ASCII representation of an American-English subset of IPA

-- very common for computational representations of pronunciations



ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[p]	[p]	parsley	[p aa r s l iy]
[t]	[t]	tea	[t iy]
[k]	[k]	cook	[k uh k]
[b]	[b]	bay	[b ey]
[d]	[d]	dill	[d ih l]
[g]	[g]	garlic	[g aa r l ix k]
[m]	[m]	mint	[m ih n t]
[n]	[n]	nutmeg	[n ah t m eh g]
[ng]	[ŋ]	baking	[b ey k ix ng]
[f]	[f]	flour	[f l aw axr]
[v]	[v]	clove	[k l ow v]
[th]	[θ]	thick	[th ih k]
[dh]	[ð]	those	[dh ow z]
[s]	[s]	soup	[s uw p]
[z]	[z]	eggs	[eh g z]
[sh]	[ʃ]	squash	[s k w aa sh]
[zh]	[ʒ]	ambrosia	[ae m b r ow zh ax]
[ch]	[tʃ]	cherry	[ch eh r iy]
[jh]	[dʒ]	jar	[jh aa r]
[l]	[l]	licorice	[l ih k axr ix sh]
[w]	[w]	kiwi	[k iy w iy]
[r]	[r]	rice	[r ay s]
[y]	[j]	yellow	[y eh l ow]
[h]	[h]	honey	[h ah n iy]
Less commonly used phones and allophones			
[q]	[ʔ]	uh-oh	[q ah q ow]
[dx]	[ɾ]	butter	[b ah dx axr]
[nx]	[ɾ̃]	winner	[w ih nx axr]
[el]	[l̩]	table	[t ey b el]

Identity of Speech Sounds

The science of phonetics aims to describe all the sounds of all the world's languages

- Articulatory phonetics: focuses on how the vocal tract produces the sounds of language
- Acoustic phonetics: focuses on the physical properties of the sounds of language
- Auditory phonetics: focuses on how listeners perceive the sounds of language

Introduction to Articulatory Phonetics (Consonants)

LINGUISTICS

Introduction to Articulatory Phonetics (Vowels)

LINGUISTICS

Articulatory Phonetics

Most speech sounds are produced by pushing air through the vocal cords

- ✓Glottis = the opening between the vocal cords
- ✓Larynx = ‘voice box’
- ✓Pharynx = tubular part of the throat above the larynx
- ✓Oral cavity = mouth
- ✓Nasal cavity = nose and the passages connecting it to the throat and sinuses

Consonants: Place of Articulation

- Consonants are sounds produced with some restriction or closure in the vocal tract
- Consonants are classified based in part on where in the vocal tract the airflow is being restricted (the place of articulation)
- The major places of articulation are:

  bilabial, labiodental, interdental, alveolar, palatal, velar,

  uvular, and glottal

Consonants: Place of Articulation

- Bilabials: [p] [b] [m]

  Produced by bringing both lips together

- Labiodentals: [f] [v]

  Produced by touching the bottom lip to the upper teeth

- Interdentals [θ] [ð]

  Produced by putting the tip of the tongue between the teeth

Consonants: Place of Articulation

- Alveolars: [t] [d] [n] [s] [z] [l] [r]

All of these are produced by raising the tongue to the alveolar ridge in some way

- ❑[s, z]: produced with the sides of the front of the tongue raised but the tip lowered to allow air to escape
- ❑[t, d, n]: produced by the tip of the tongue touching the alveolar ridge (or just in front of it)
- ❑[l]: the tongue tip is raised while the rest of the tongue remains down so air can escape over the sides of the tongue (thus [l] is a lateral sound)
- ❑[r]: air escapes through the central part of the mouth; either the tip of the tongue is curled back behind the alveolar ridge or the top of the tongue is bunched up behind the alveolar ridge

Consonants: Place of Articulation

- Palatals: [ʃ] [ʒ] [ʧ] [ʤ][ʝ]
  – Produced by raising the front part of the tongue to the palate
- Velars: [k] [g] [ŋ]
  – Produced by raising the back of the tongue to the soft palate or velum
- Uvulars : [ʀ] [q] [ɢ]
  – Produced by raising the back of the tongue to the uvula
- Glottals: [h] [ʔ]

– Produced by restricting the airflow through the open glottis ([h]) or by stopping the air completely at the glottis (a glottal stop: [ʔ])

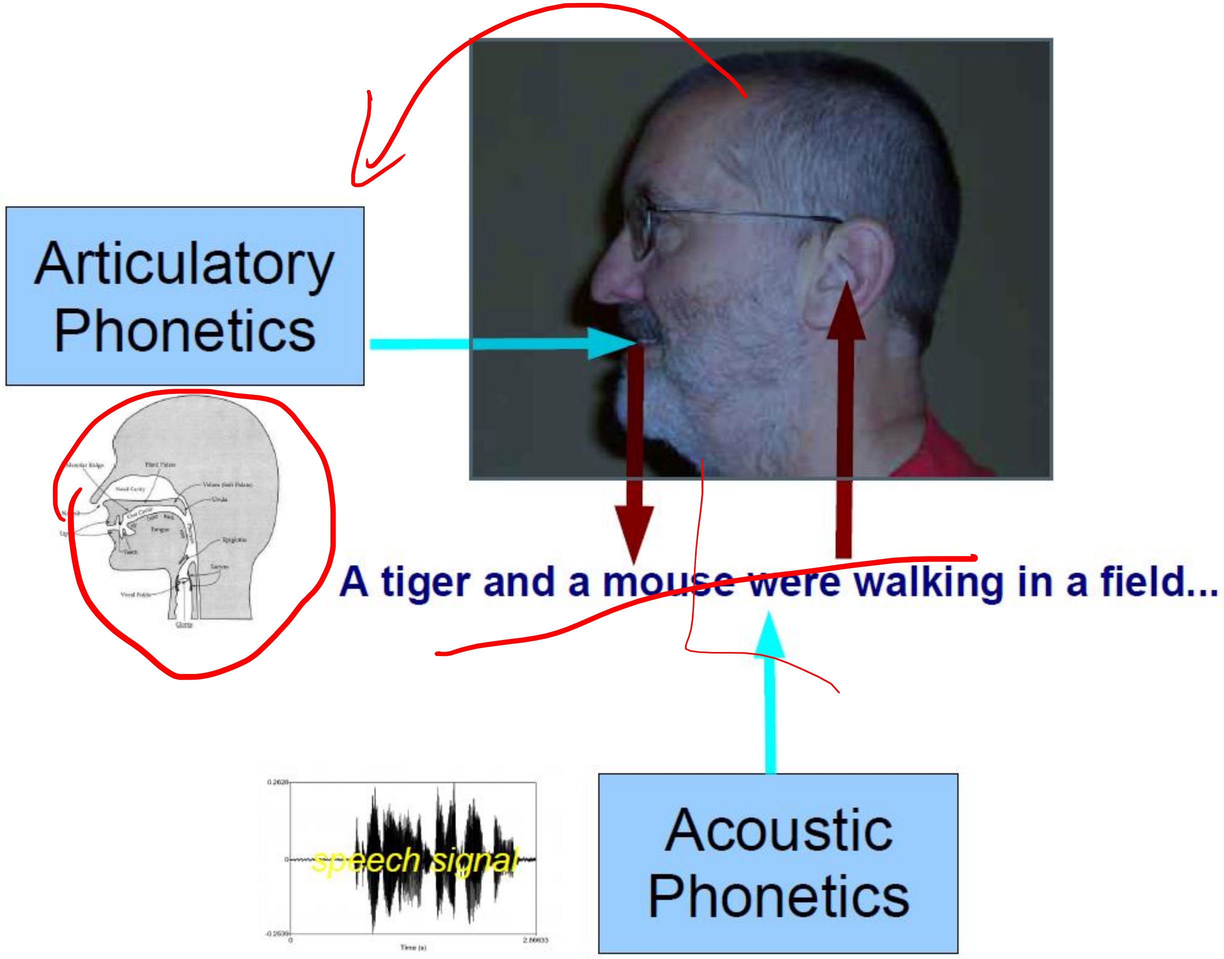
Consonants: Manner of Articulation

- The manner of articulation is the way the airstream is affected as it flows from the lungs and out of the mouth and nose

- Voiceless sounds are those produced with the vocal cords apart so the air flows freely through the glottis

- Voiced sounds are those produced when the vocal cords are together and vibrate as air passes through

Acoustic Phonetics and Signal



Articulatory Phonetics

A tiger and a mouse were walking in a field...

Acoustic Phonetics

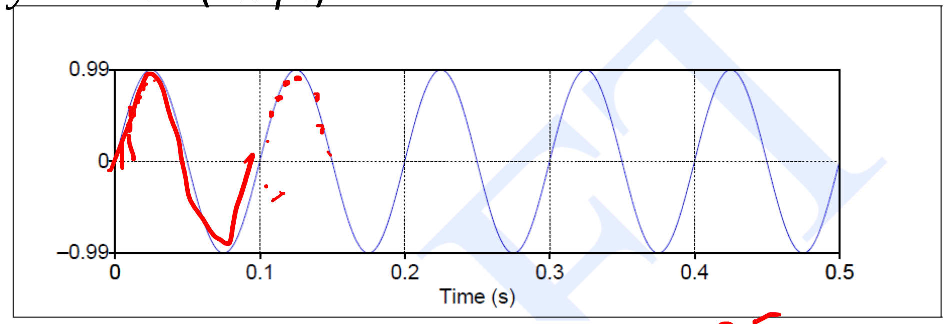
Focuses on the physical properties of the sounds

concerned with investigating the transmission of speech signals through

- gases such as air, other substances (e.g. bone, tissue)
- electronic amplification and storage

Waves

$y = A * sin(2\pi\ f\ t)$



0.5

- Acoustic analysis is based on the sine and cosine functions
- Important characteristics

freq = 10 Hz

- Frequency:
  - ✓ number of times per second that a wave repeats itself, i.e. the number of cycles.
  - ✓ measured in terms of cycles per second usually called Hertz
  - ✓ The signal in Fig repeats itself 5 times in .5 seconds,
    - ✓ Hence, frequency = 10 cycles per second = 10 Hz

$T = \frac{1}{f}$

- Amplitude: is the maximum value on the y-axis
- *Period: The time it takes for one cycle to complete.* $T = \frac{1}{f}$

Speech Sound Waves

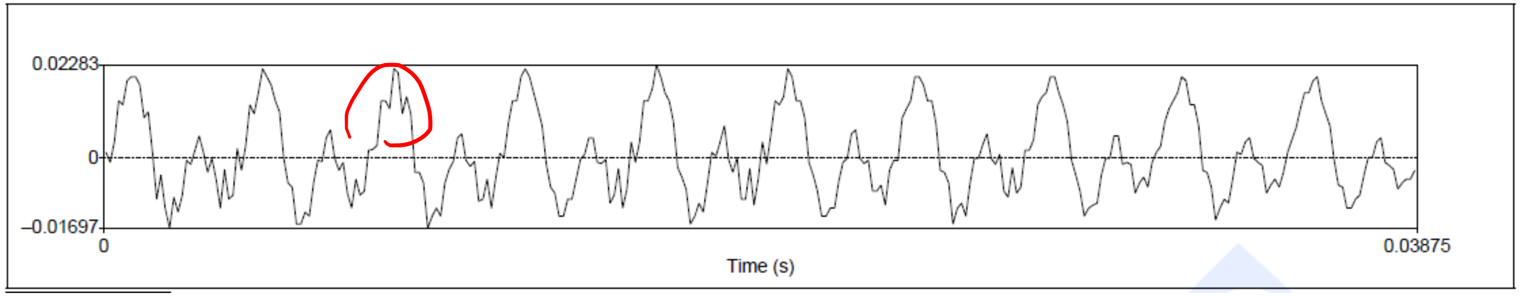
- The input to a speech recognizer, is a complex series of changes in air pressure.
- The changes in air pressure
  - originate with the speaker,
  - caused by the specific way that air passes through the glottis and out the oral or nasal cavities.
- Sound waves are represented by plotting the change in air pressure over time
- The first step in processing speech is to convert the analog representations into a digital signal.
- This process of **analog-to-digital conversion** has two steps: **sampling** and **quantization**.
  - A signal is sampled by measuring its amplitude at a particular time;
  - **sampling rate** is the number of samples taken per second.
  - It is necessary to have at least two samples in each cycle
    - one measuring the positive part of the wave and one measuring the negative part.
  - The maximum frequency wave that can be measured is half the sample rate
  - *Nyquist frequency* - The maximum frequency for a given sampling rate

Quantization

16 4-bit

- 8000Hz sampling rate requires 8000 amplitude measurements to be stored for each second of a speech.
- This is stored as integers (8-bit or 16-bit)
- This process of representing real-valued numbers as integers is called quantization.
    - Telephone speech is often sampled at 8 kHz and stored as 8-bit samples
    - Microphone data is often sampled at 16 kHz and stored as 16-bit samples.
- **Channels -** Two-party conversations, we can store both channels in the same file, or we can store them in separate files.
- **Compression** – represents whether the sample is stored linearly or it is compressed. common compression format used for telephone speech is μ-law

Frequency and Amplitude



- Vocal fords open- air pushing up through lungs – creates high pressure
- Vocal folds closed –no pressure
- Each major peak corresponds to an opening of the vocal folds.
- **Fundamental frequency** – frequency of the vocal fold vibration
- The vertical axis measures the amount of air pressure variation;
  - pressure is force per unit area, measured in Pascals (Pa).
  - Positive value - normal (atmospheric) air pressure
  - Negative value – lower than normal (rarefaction) pressure

Amplitude

- RMS (root-mean-square) amplitude, which squares each number before averaging (making it positive), and then takes the square root at the end.

$$\text{RMS amplitude}_{i=1}^{N} = \sqrt{\sum_{i=1}^{N} \frac{x_i^2}{N}}$$ ✓

- Power of the signal is related to the square of the amplitude

$$\text{Power} = \frac{1}{N}\sum_{i=1}^{n} x[i]^2$$ ✓

- **Intensity** of the sound, which normalizes the power to the human auditory threshold, and is measured in dB.

$$\text{Intensity} = 10\log_{10}\frac{1}{NP_0}\sum_{i=1}^{n} x_i^2$$ ✓

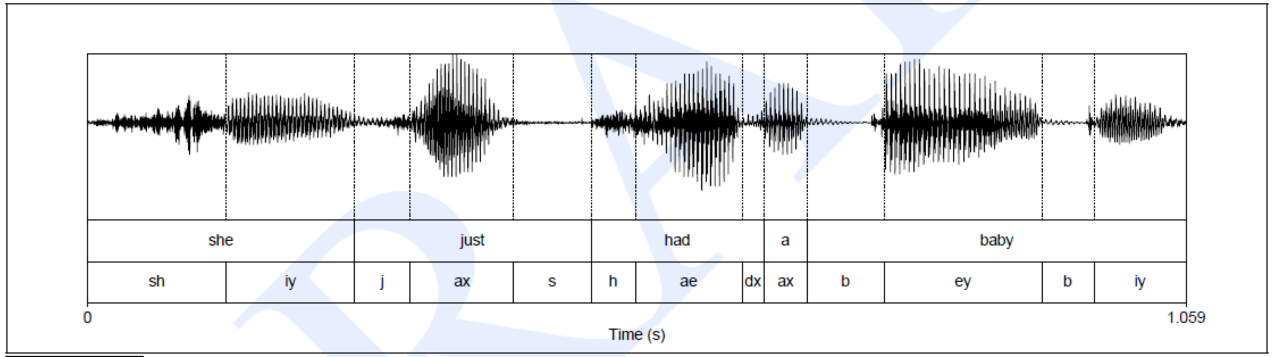
Pitch and Loudness

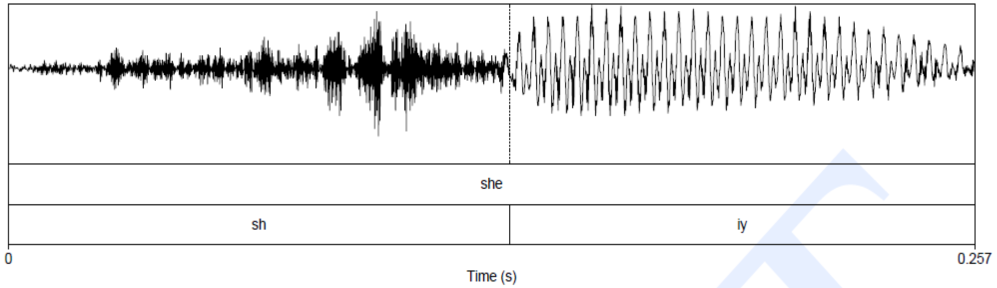
- The **pitch** of a sound - is the mental sensation or perceptual *pitch* correlate of fundamental frequency;
  - A sound has a higher fundamental frequency – higher pitch
  - Human pitch perception is most accurate between 100Hz and 1000Hz
- The mel is a unit of **pitch –** pairs of sounds equidistant in pitch are separated by an equal number of mels.

$$\boxed{m = 1127 \ln(1 + \frac{f}{700})}$$

- **Loudness** - Perceptual correlate of the **power**.
  - Sounds with higher amplitudes are perceived as louder, but again the relationship is not linear.

Interpreting Phones from a Waveform

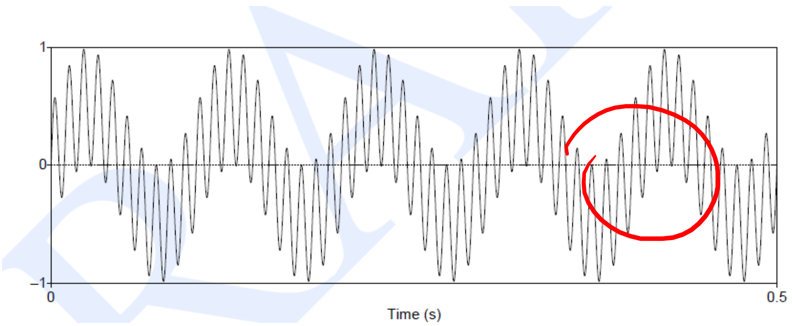


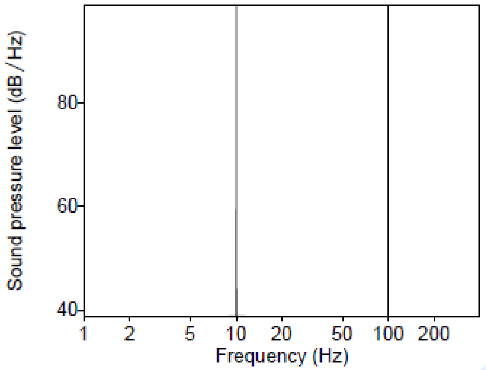


- Vowels can be easily plotted.
- Vowels are voiced, and they tend to be long, and are relatively loud
    - ✓ Length in time – related to the x-axis
    - ✓ Loudness – related to the (square of) amplitude on the y-axis
    - ✓ Voicing - major peaks in amplitude →corresponds to opening of vocal folds
- Stop consonant – closure followed by a release.
    - ✓ Period of silence followed by a slight burst of amplitude - in both [b]'s in baby
    - ✓ fricatives – hissy sounds – have very noisy irregular waveform – in 'she' (in second fig)

Spectra and the Frequency Domain

- Most of phonetic features can be extracted from waveform directly.
- However, most computational applications such as speech recognition represent sounds in terms of its **component frequencies**.
- **Fourier analysis** - complex wave can be represented as a sum of many sine waves of different frequencies.
- Summing of 2 sine waves of frequencies 10Hz and 100 Hz





- Two component frequencies can be represented using **Spectrum**.
- The **spectrum** of a signal is a representation of each of its **frequency components** and their **amplitudes.**
- Spectrum is the alternative representation of original waveform.
- Summing of 2 sine waves of frequencies 10Hz and 100 Hz

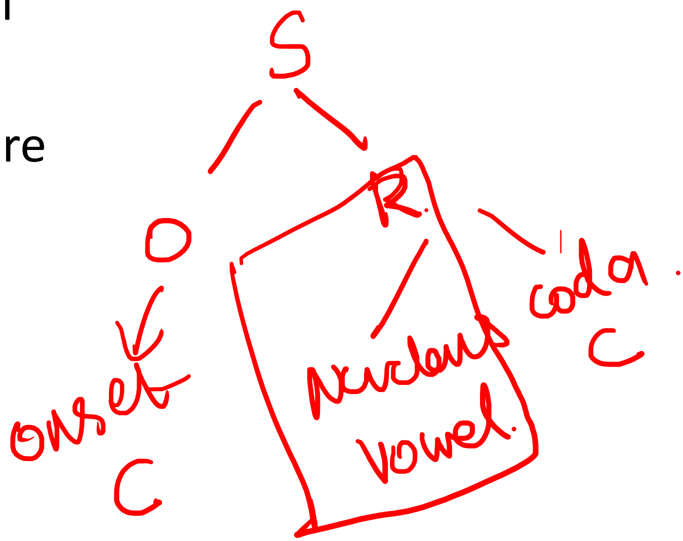
- The **source-filter** model is a way of explaining the acoustics of a sound by modeling how the pulses
- produced by the glottis (the **source**) are shaped by the vocal tract (the **filter**).

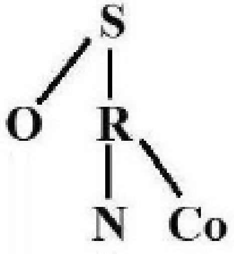
Syllable Structure

En|ter|tain|ing.

En|ter|tai|ning.

- Count of no. of syllables in a word is roughly/intuitively the no. of vocalic segments in a word.
- Thus, presence of a vowel is an obligatory element in the structure of a syllable. This vowel is called “*nucleus”.*
- Basic Configuration: **(C)V(C)**.
- Part of syllable preceding the nucleus is called the *onset*.
- Elements coming after the nucleus are called the *coda*.
- Nucleus and coda together are referred to as the *rhyme*.

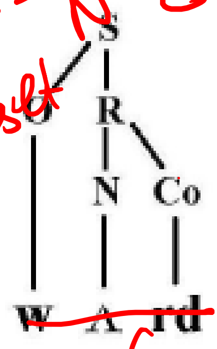




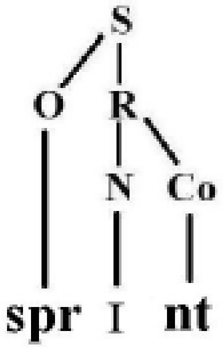
S ≡ Syllable, O ≡ Onset
R ≡ Rhyme, N ≡ Nucleus
Co ≡ Coda

Syllable Structure: Examples

- 'word'



- 'sprint'



played → /pleɪd/

C N C→C

onset

k æ t

cat /kæt/

it → /ɪt/

onset N C

Tea → /ti/

O N

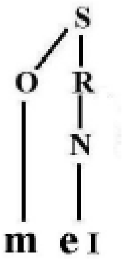
Stand → /stænd/

onset N C

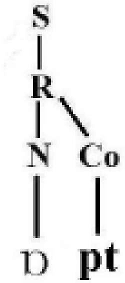
∅

Syllable Structure: Examples

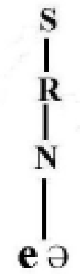
- ‘may’
- ‘opt’
- ‘air’



← No Coda.



← No Onset.



← No Coda, No Onset.

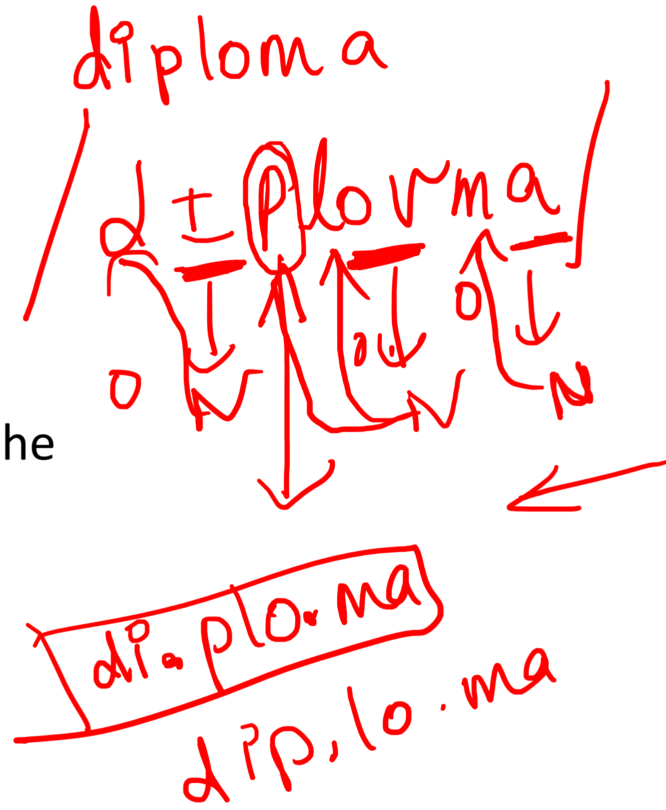
Syllable Structure

- *Open Syllable: ends in vowel*
- *Closed syllable: ends in consonant or consonant cluster*

- *Light Syllable*: A syllable which is open and ends in a short vowel
  - General Description – CV.
  - Example, ‘air’.
- *Heavy Syllable*: Closed syllables or syllables ending in diphthong
  - Example: ‘opt’
  - Example, ‘may’

Syllabification: Determining Syllable Boundaries

- Given a string of syllables (word), what is the coda of one and the onset of another?
- In a sequence such as VCV, where V is any vowel and C is any consonant, is the medial C the coda of the first syllable (VC.V) or the onset of the second syllable (V.CV)?
  - E.g., *ari (अरि; "enemy")*
- To determine the correct groupings, there are some rules, two of them being the most important and significant:
  - Maximal Onset Principle,
  - Sonority Hierarchy



Constraints: Phonotactics

- **Phonotactics**
  - Determines possible comb. of onsets and codas which can occur.
  - Deals with restriction on the permissible combination of phonemes.
  - Defines permissible syllable structure, consonant clusters and vowel sequence by means of phonotactical constraints.
- In general, rules operate around the sonority hierarchy.
- Fricative /s/ is lower on the sonority hierarchy than the lateral /l/, so the combination /sl/ is permitted in onsets and /ls/ is permitted in codas. Opposite is not allowed. ✓
- Thus, ‘*slips*’ and ‘*pulse*’ are possible English words.
- ‘*lsips*’ and ‘*pusl*’ are not possible.