Speech Recognition (DSAI 456)

Lecture 1

Mohamed Ghalwash
mghalwash@zewailcity.edu.eg

Speech Sounds and Phonetic Transcription

- Represent the pronunciation as a string of phones (speech sounds) which has special alphabets (IPS international phonetics alphabets)
- Mapping between letters of english (orthography) and phones (sound-orthography mapping)

  e.g: 'c' can be mapped to phone 'k' or phone 's'

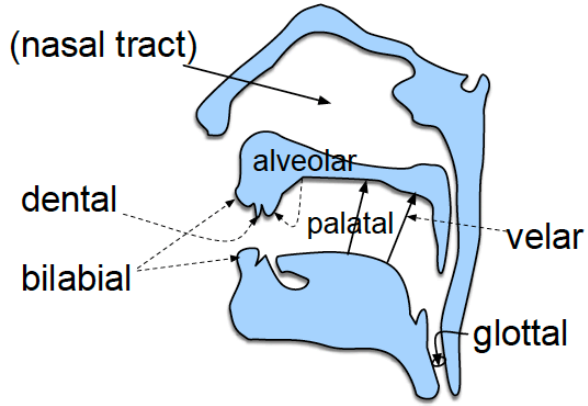
ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription	ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[p]	[p]	parsley	[p aa r s l iy]	[iy]	[i]	lily	[l ih l iy]
[t]	[t]	tea	[t iy]	[ih]	[ɪ]	lily	[l ih l iy]
[k]	[k]	cook	[k uh k]	[ey]	[eɪ]	daisy	[d ey z iy]
[b]	[b]	bay	[b ey]	[eh]	[ɛ]	pen	[p eh n]
[d]	[d]	dill	[d ih l]	[ae]	[æ]	aster	[ae s t axr]
[g]	[g]	garlic	[g aa r l ix k]	[aa]	[ɑ]	poppy	[p aa p iy]
[m]	[m]	mint	[m ih n t]	[ao]	[ɔ]	orchid	[ao r k ix d]
[n]	[n]	nutmeg	[n ah t m eh g]	[uh]	[ʊ]	wood	[w uh d]
[ng]	[ŋ]	baking	[b ey k ix ng]	[ow]	[oʊ]	lotus	[l ow dx ax s]
[f]	[f]	flour	[f l aw axr]	[uw]	[u]	tulip	[t uw l ix p]
[v]	[v]	clove	[k l ow v]	[ah]	[ʌ]	butter	[b ah dx axr]
[th]	[θ]	thick	[th ih k]	[er]	[ɝ]	bird	[b er d]
[dh]	[ð]	those	[dh ow z]	[ay]	[aɪ]	iris	[ay r ix s]
[s]	[s]	soup	[s uw p]	[aw]	[aʊ]	flower	[f l aw axr]
[z]	[z]	eggs	[eh g z]	[oy]	[oɪ]	soil	[s oy l]
[sh]	[ʃ]	squash	[s k w aa sh]	[ax]	[ə]	pita	[p iy t ax]
[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]				

Speech Sounds and Phonetic Transcription

- Time-aligned transcription: mapping between waveform and phones (TIMIT corpus)

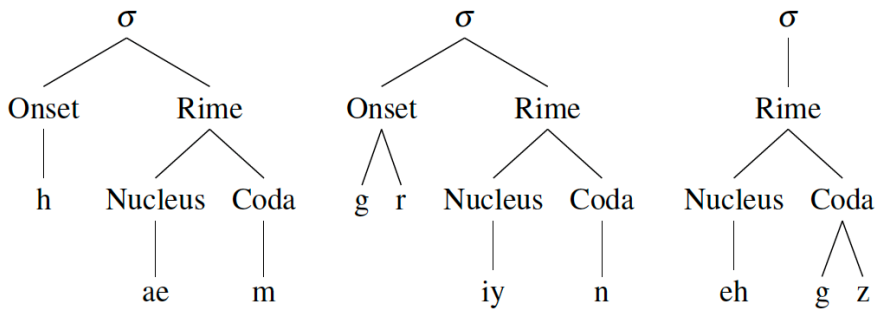
she	had	your	dark	suit	in	greasy	wash	water	all	year
sh iy	hv ae dcl	jh axr	dcl d aa r kcl	s ux q	en	gcl g r iy s ix	w aa sh	q w aa dx axr q	aa l	y ix axr

- Articulatory phonetics
  - consonant vs vowel



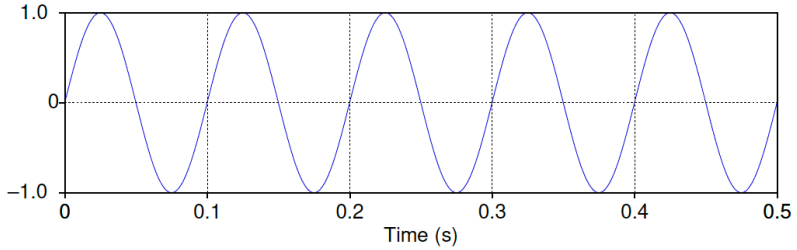
Speech Sounds and Phonetic Transcription

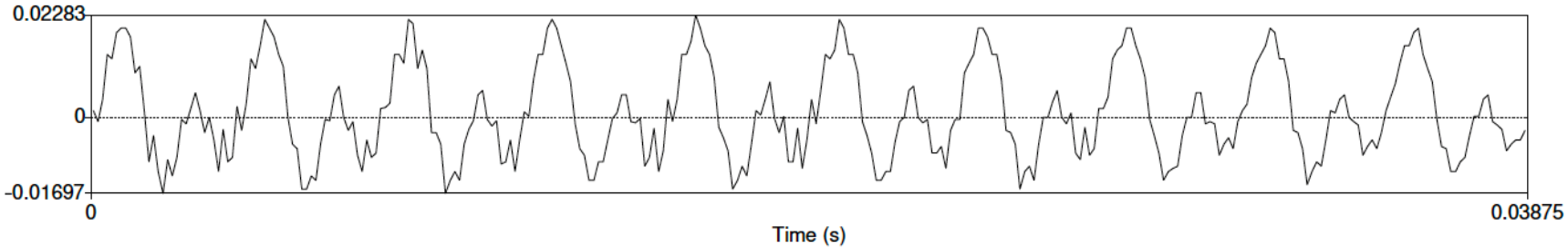
- Syllable: a vowel-like sound together with some of the surrounding consonants that are most closely associated with it
  - `dog` has one syllable (d aa g), `catnip` has two (k ae t) and (n ih p)
  - The vowel at the core of a syllable is called the **nucleus**
  - Initial consonants are called the **onset** (as in strike (s t r ay k))
  - The **coda** is the optional consonant or sequence of consonants following the nucleus
  - The rime, or rhyme, is the nucleus plus coda



Signal Representation

- Signal is a waveform
- Air passes through the glottis -> air pressure -> sound waves
- Waveform can be represented as a combination of `sin` or `cos` functions, e.g. $y = A\ sin(2\pi ft)$
  - A is the **amplitude**: the maximum value on the Y-axis
  - f is the **frequency**: number of cycles per second
- y-axis the amount of air pressure variation
- **Hertz**: number of cycles per second
- The **period** T of the wave is the time it takes for one cycle to complete, i.e. $T = 1/f$



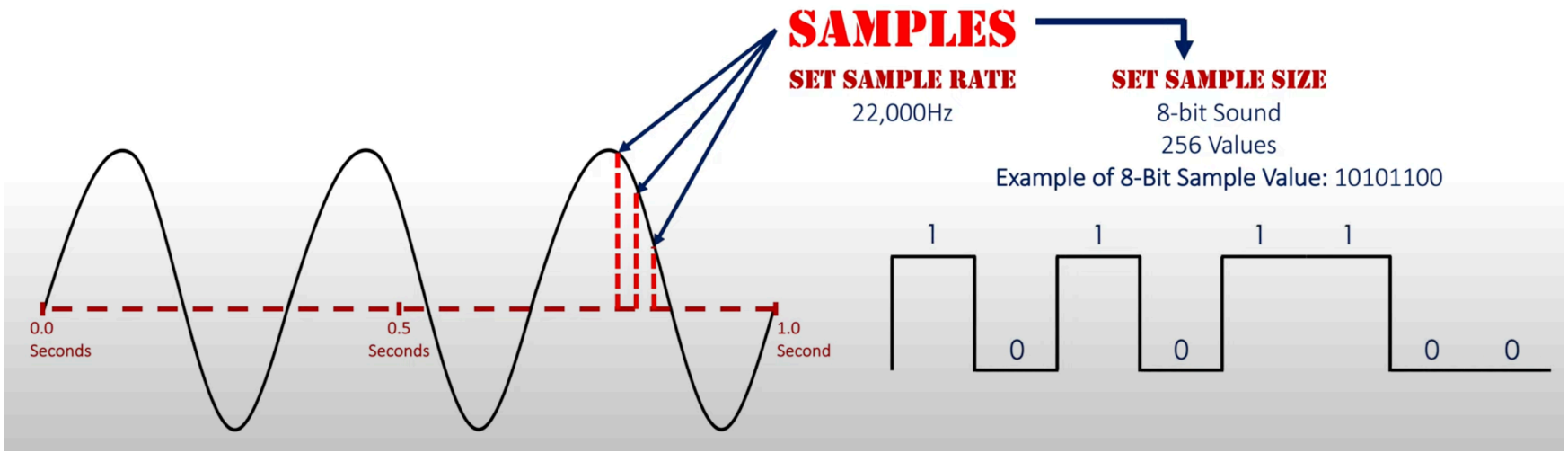


A waveform of the vowel (iy). The y-axis shows the level of air pressure above and below normal atmospheric pressure.

Analog to Digital

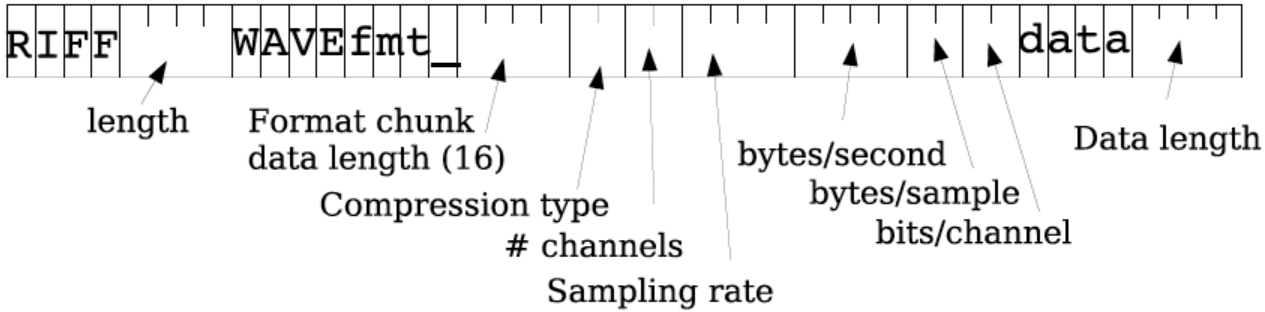
- Sampling (sampling rate is at least twice the frequency)

- Digitization (stored as integers)



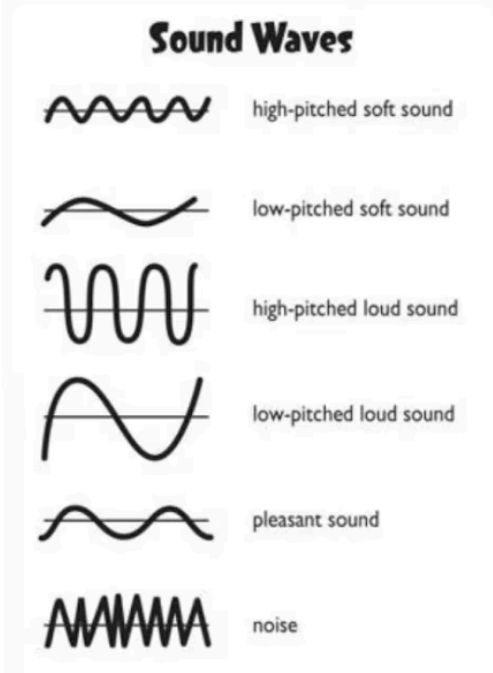
Analog to Digital

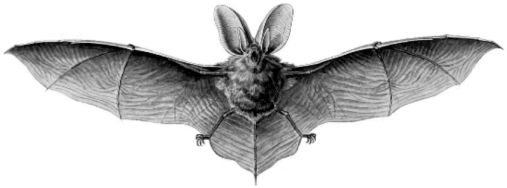
- *Human hearing is more sensitive at small intensities than large ones*
- Compression (linear vs log compression algorithms like μ-law)
- The log represents small values with more faithfulness at the expense of more error on large values
- `.wav` file



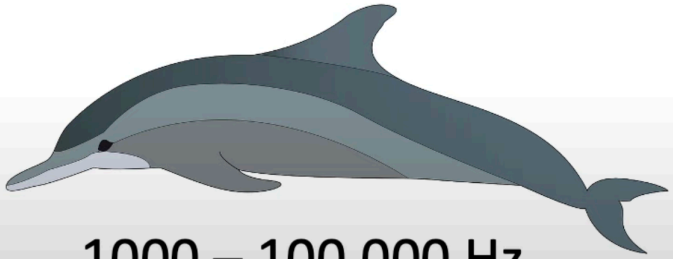
Pitch and Loudness

- Pitch (Hz): pitch is our perception of the frequency of the sound wave, meaning how high or low a sound seems to be
  - High-pitched sounds, like a whistle, are caused by sound waves with a high frequency
  - Low-pitched sounds, like a deep bass note, are caused by sound waves with a low frequency
  - pitch is subjective and linked to how the sound is heard rather than the exact physical frequency
- Loudness (decibel dB) is a measure of intensity I of sound ($I \alpha A^2$)





2000 – 110,000 Hz



1000 – 100,000 Hz

67 – 45,000 Hz

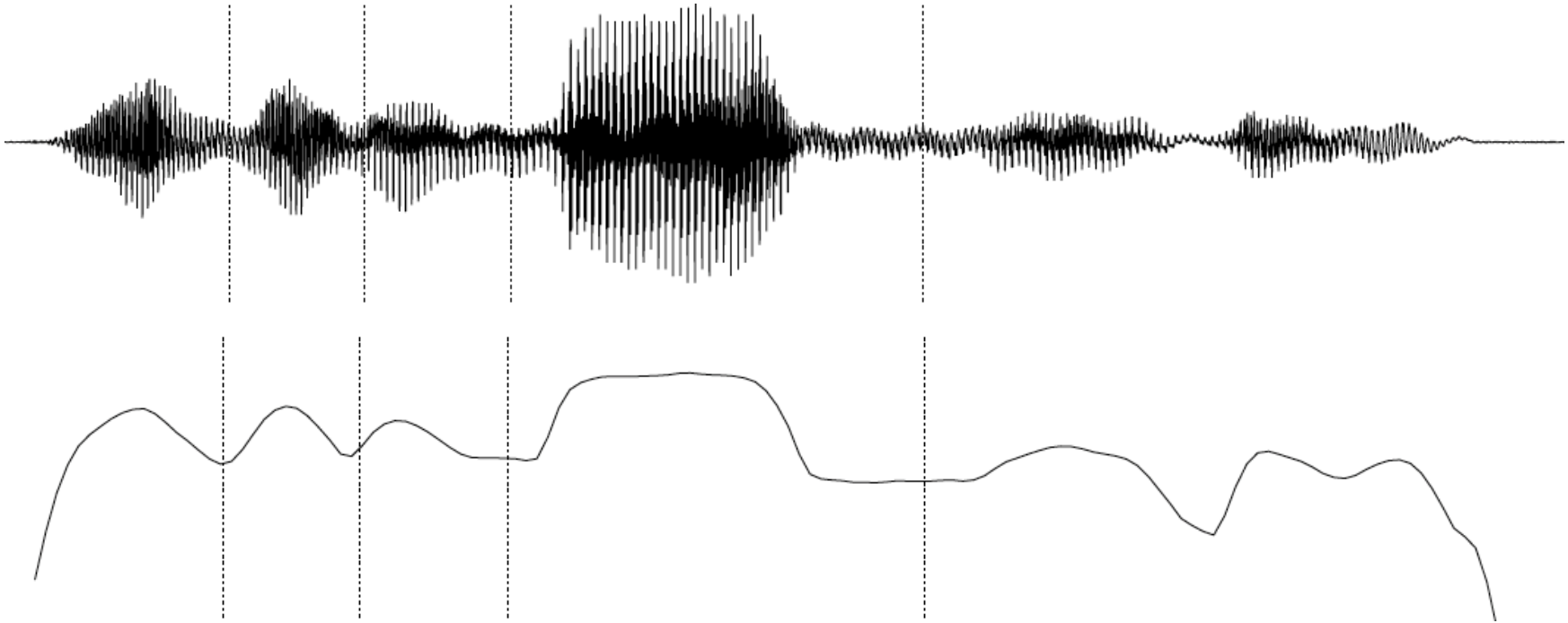
45 – 64,000 Hz

Human can hear sound of frequencies between 20Hz and 20kHz.

Hearing ranges

Features

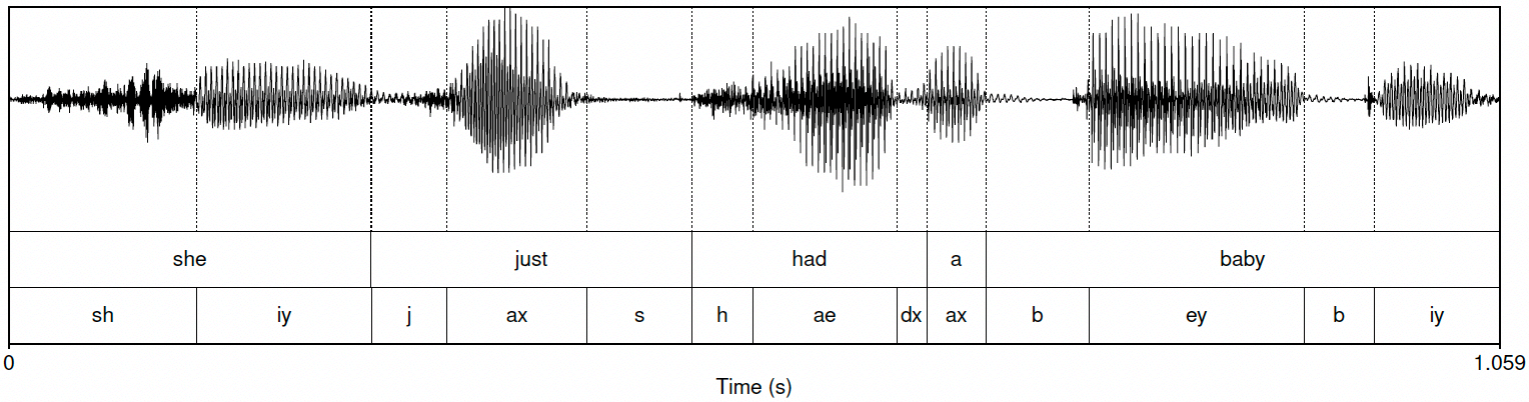
- Intensity



Features

- RMS (root-mean-square) of amplitude ($\sqrt{\frac{1}{n}\sum_{t=1}^{n} a_t^2}$): the average amplitude over some time range, to give us some idea of how great the average displacement of air pressure is
- F0 (fundamental frequency): the lowest frequency of a periodic waveform
- Pitch track: is a time-based representation that shows how the pitch (perceived frequency) of a sound changes over time
- Frame-by-frame, short overlapping time frames

Interpretation of Phones from a Waveform



- Vowels are pretty easy to spot. They are voiced, tend to be long, and are relatively loud
- Each of the six vowels have regular amplitude peaks indicating voicing
- Each major peak corresponding to an opening of the vocal folds
- A stop consonant consists of a closure followed by a release. Can be seen as a period of silence followed by a slight burst of amplitude

Learn More

Slidev · Course Homepage