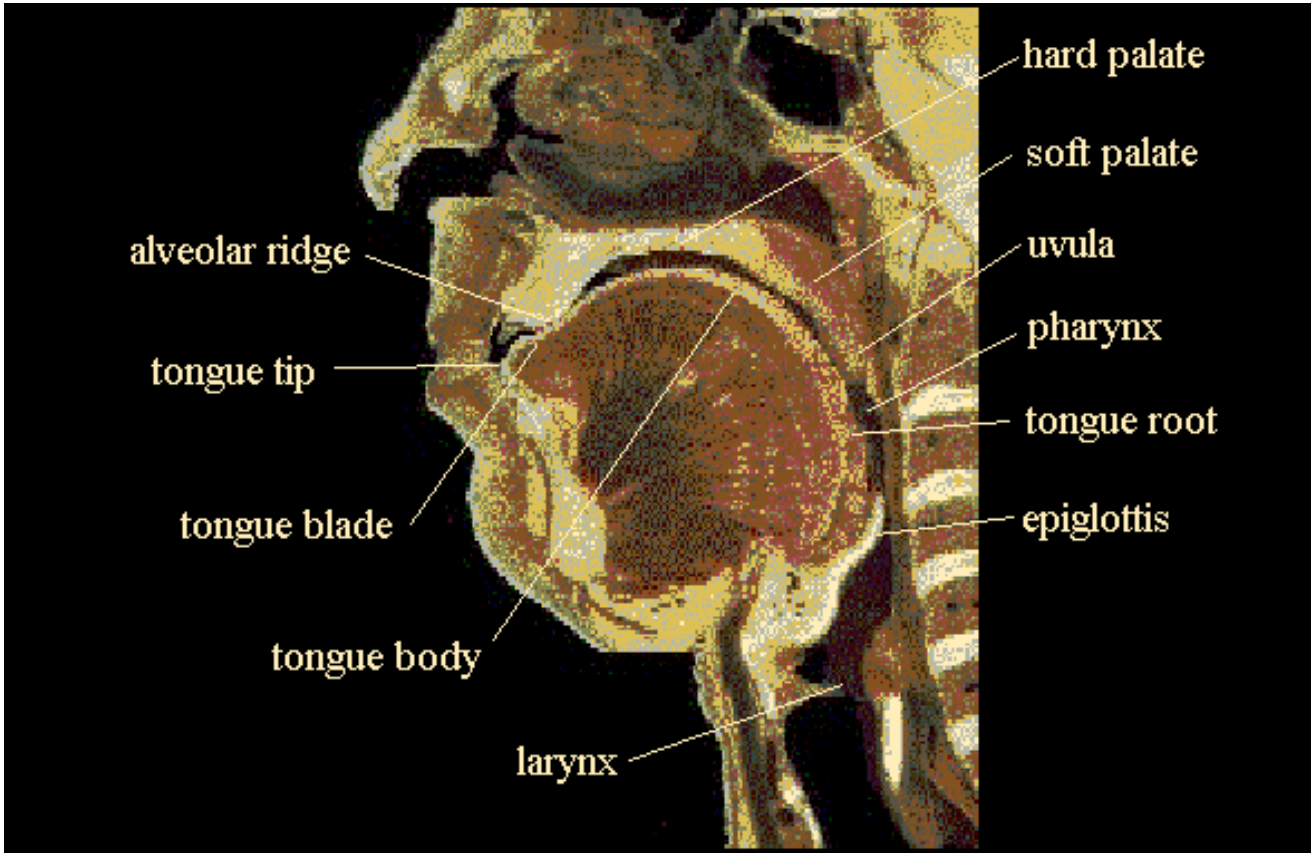
11-737 Multilingual NLP

Speech

Carnegie Mellon University
Language Technologies Institute

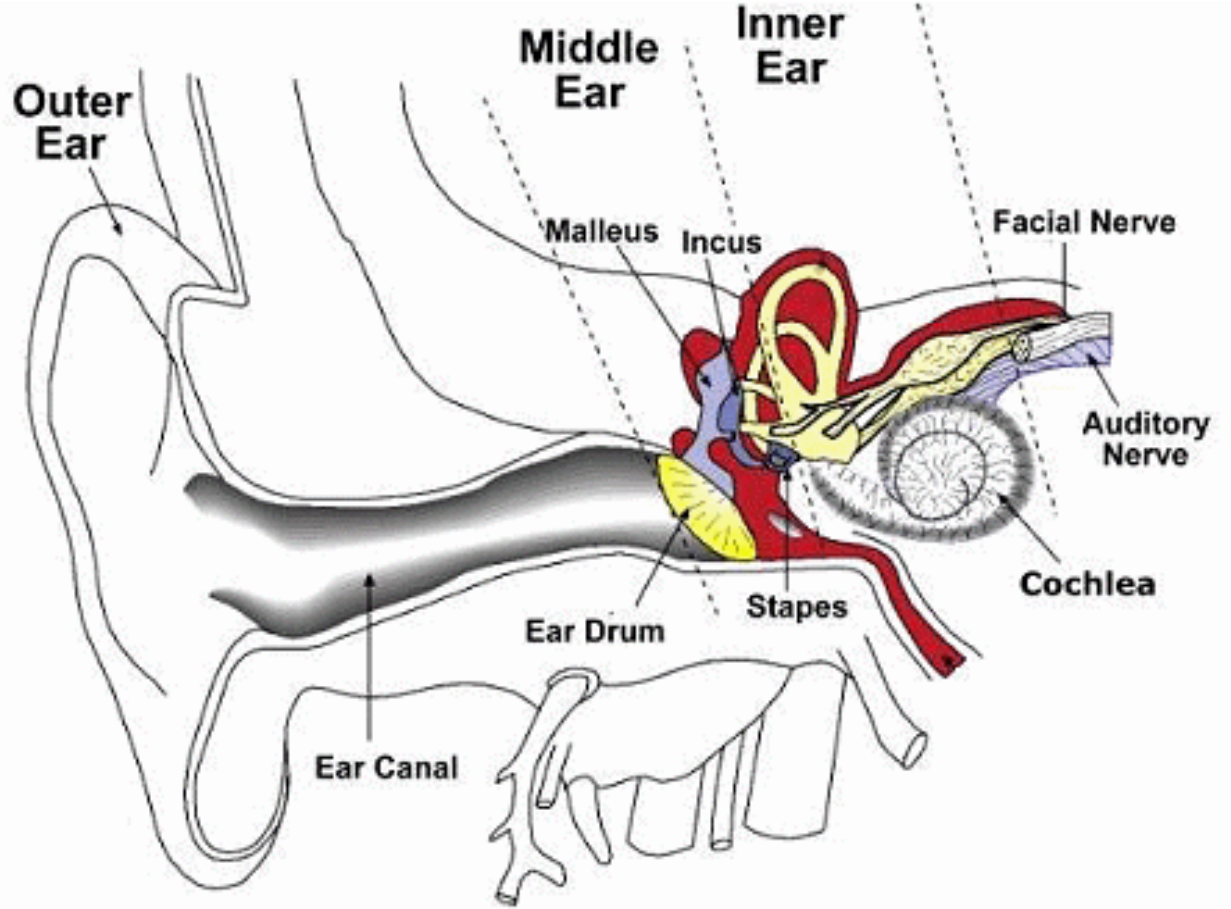
The vocal tract



From Meat to Voice

- Blow air through lungs
  - Vibrate larynx
  - Vocal tract shape defines resonance
  - Obstructions modify sound
    - Tongue, teeth, lips, velum (nasal passage)

The Ear


Outer Ear
Middle Ear
Inner Ear
Malleus
Incus
Facial Nerve
Auditory Nerve
Cochlea
Stapes
Ear Drum
Ear Canal

From sound to brain waves

- Sound waves
  - Vibrate ear drum
  - Cause fluid in cochlear to vibrate
  - Spiral cochlear
    - Vibrate hairs inside cochlear
    - Different frequencies vibrate different hairs
    - Converts time domain to frequency domain

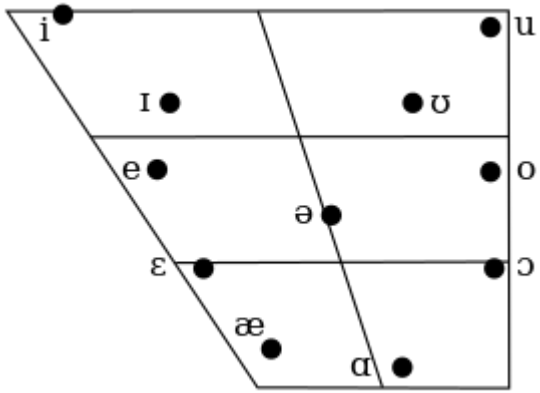
From Grunts to Meaning

- Different noises recognized at different times
- What are the units of noise
  - Phonemes: linguistic segments that when changed give different words

    pat → bat (/p/ to */b/)

- How many sounds are there
  - An infinite amount – or a few score
- International Phonetic Alphabet (IPA)
  - Defines vowels and consonants
  - In (mostly) structured way

Hindi Phonology



		Labial	Dental/ Alveolar	Retroflex	Palatal	Velar	Uvular	Glottal
Nasal		m	n	ɳ	(ɲ)	ŋ		
Plosive/ Affricate	voiceless	p	t	ʈ	tʃ	k	(q)	
	voiceless aspirated	pʰ	tʰ	ʈʰ	tʃʰ	kʰ		
	voiced	b	d	ɖ	dʒ	g		
	voiced aspirated	bʱ	dʱ	ɖʱ	dʒʱ	gʱ		
Flap and trill	plain		r	ɽ				
	voiced aspirated			ɽʱ				
Fricative	voiceless	f	s		ʃ	(x)		ɦ
	voiced	ʋ	z		(ʒ)	(ɣ)		
Approximant			l		j			

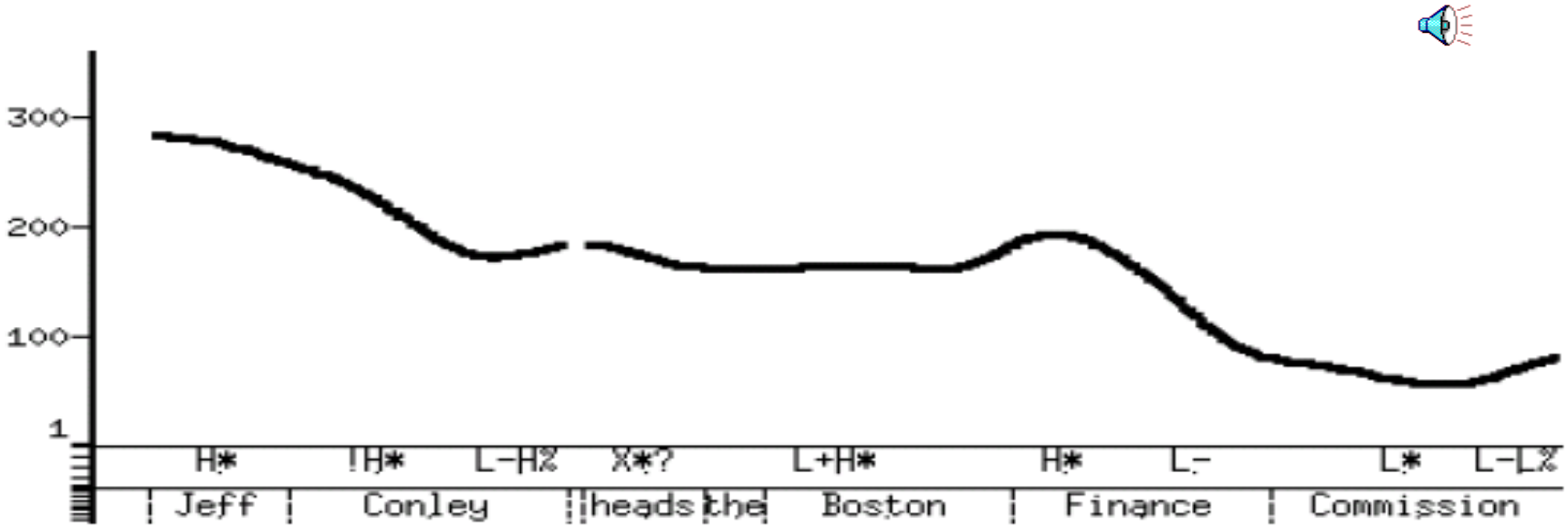
Its not always the same

- English speakers
  - Lots of variation (even in standard dialects, and within speakers)
  - Borrowed phonemes, New phonemes (US English “flap” in “water”)
  - Distinctions: Mary Merry and Marry (same in US, all different in UK)
- Other languages have different choice and distributions
  - Aspirated stops are choice in English(ish), but phonetic in Hindi
- Phonological list of segments might or might not be comprehensive
  - Different dialects, different L1, different age, different styles
  - Maybe want more for ASR/TTS than standard
  - Foreign borrowed phones might be very common

Its not just phonemes

- Beyond segments
  - Tones (phonological) ma
  - Lexical accents, Lexical stress
- Other phonemes
  - Tense/lack, Ejective stops, clicks, voices quality may be phonetic
- Prosody
  - Intonation, duration, phrasing and power
  - How people speak not just what people speak

Intonation Contour



Intonation Information

- Large pitch range (female)
- Authoritive since goes down at the end
  - News reader
- Emphasis for Finance H*
- Final has a raise – more information to come

- Female American newsreader from WBUR
- (Boston University Radio)

Syllables and Words

- Syllables
  - {C*} V {C*}
  - Pre-vowel consants: onset
  - Post-vowel consonants: coda
  - Vowel: nucleus
  - Vowel + coda: rhyme
- Words
  - There are no spaces between spoken words
  - Native speakers have strong agreement on (speech) word boundaries

Computer Speech

- Digitized from analog
  - At least 8 thousand times a second
  - Does matter (ish) at more than 16KHz.
  - The older you get the less higher you can hear (and mostly this isn't in speech)
- Microphone quality/placement makes a huge different
  - Close to mouth (telephone) "close-talking"
  - Longer distance (audio source moving) "far field"
  - Multi-microphone array can alleviate this (partly)

Context

- Read speech
- Performed careful speech
  - Broadcast news, studio quality recording
- Quiet office
- Outside
- Around other noises
- Around other people
- Style of speech
  - To machines
  - Human-human
  - To friends
  - Spontaneous conversation
- Drunk with friends, emotional, at busy outside café, far field microphone

Speech is not Text

- Speech is not spoken text
- Text is not written speech
- Many languages are very different between text and speech
  - Different grammar, registers, morphology
- English is different in speech
  - Speech: (more) first person, present tense
  - Text: third person, past tense (if you think Penn Tree Bank is English)
  - ‘I’ is far more common in speech than text
- Historically writing was a different language
  - Now it can sometimes be (MSA vs Arabic dialects, Tamil)
- Social media is closer to written speech
  - Its is much more casual and less edited

Speech Technologies

- Speech Recognition
  - Speech-to-text
- Speech Synthesis
  - Text-to-speech
- Spoken Dialog Systems
  - Interaction through speech (task and non-task oriented)
- Speaker ID, Speaker Recognition
  - Who is talking: passwords, but more often, diarization.
- How people are speaking
  - Expressive speech, recognition and synthesis
  - Emotion (who is angry, who is lying, who is genuine

Speech Data

- Only top economic languages have abundances of speech data
  - Most ASR data (read or transcribed speech)
  - 1000s of hours is good, 100s of hours will do.
- Central repositories
  - LDC/ELRA for major (and some minor languages)
  - Voxforge, openslr (growing public archives)
  - Universities in target geographies
- Found data
  - Youtube,Todou recordings (without transcripts, but maybe subtitles)
  - Audio books, religious texts (Bible, Quran, Linux manuals)
  - CMU Wilderness has 700 languages (20 hours each)

Speech

- Different from text: much more variance
- Not just spoken text (there might not be any text)
- Phones, syllables and words
  - But prosody too
- Channel and context vary the difficulty of the task
  - Quiet office vs loud, busy, outside places
- Data good for high-economic languages
  - Some data for other languages
- Speech processing is better for high literate languages
  - But it is more important for low literate languages

Discussion Points: Speech

- Find examples of differences between written forms and spoken forms
- Find some utterances that do not have a (good) written form
- Find some written forms that don't have clear spoken forms
- Identify the reasons for the mismatch
  - "not a word" (sort of)
  - Grammatical variation
  - Dialect, "not proper"