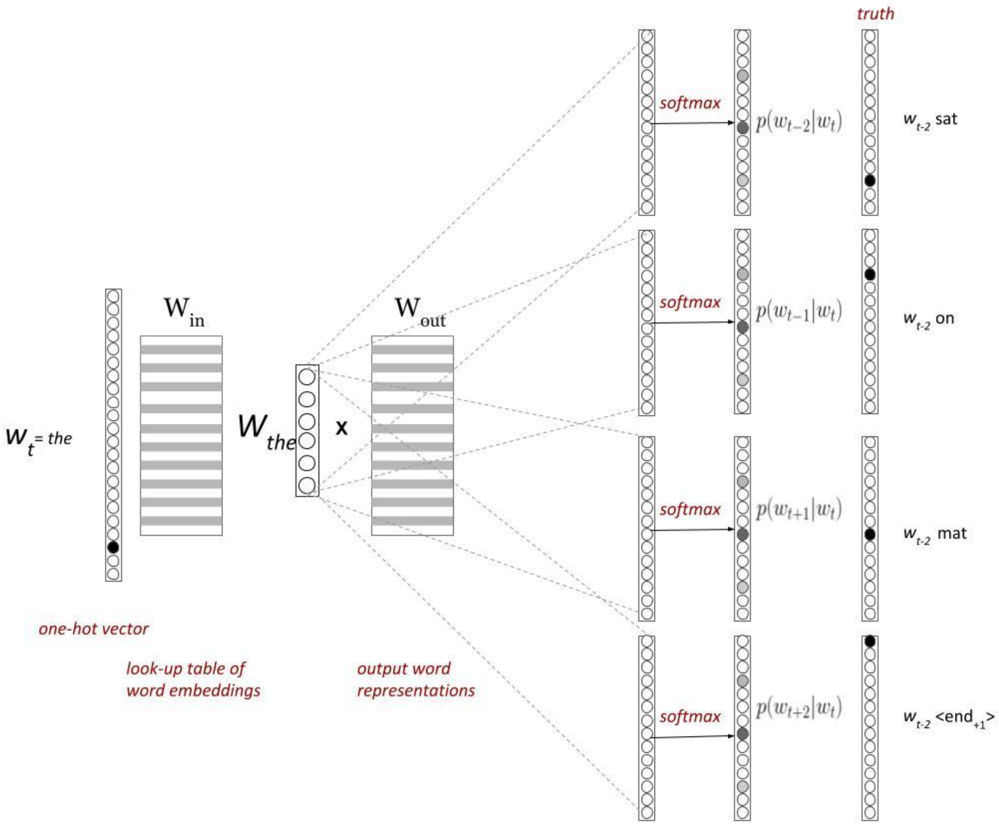
Algorithms for NLP

Automatic Speech Recognition

Yulia Tsvetkov – CMU

Slides: Preethi Jyothi – IIT Bombay,
Dan Klein – UC Berkeley

Skip-gram Prediction

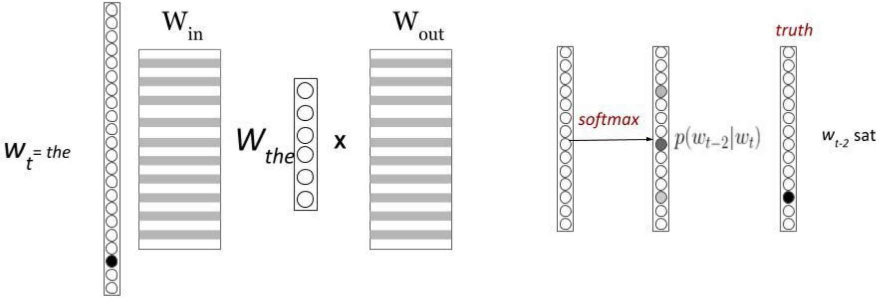


Skip-gram Prediction

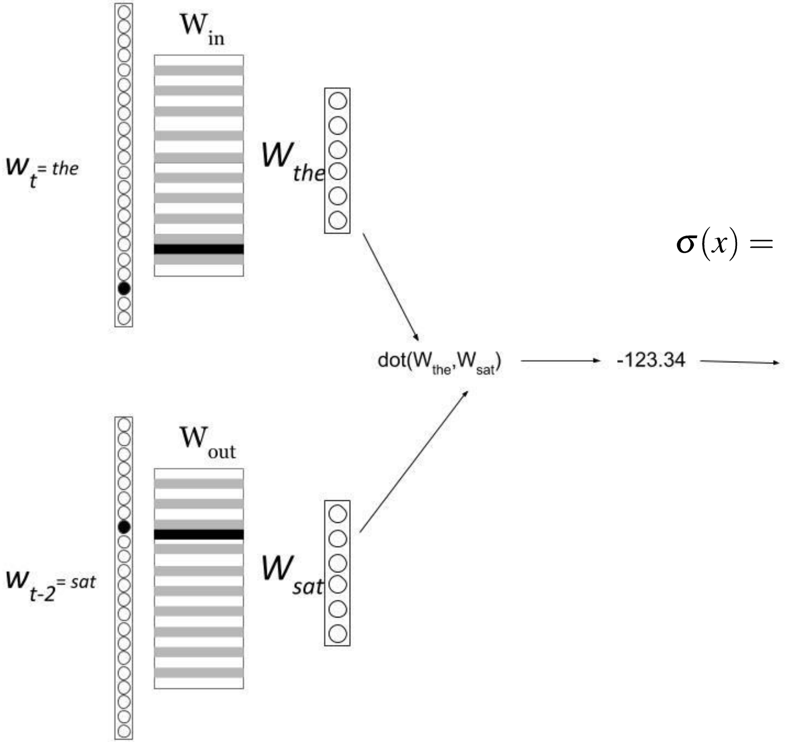
- Training data

w_t , w_{t-2}
w_t , w_{t-1}
w_t , w_{t+1}
w_t , w_{t+2}
...

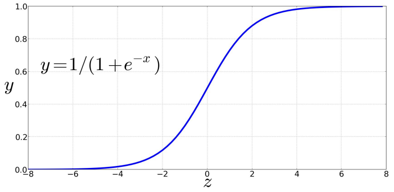
Skip-gram Prediction



How to compute p(+|t,c)?



$$\sigma(x) = \frac{1}{1+e^{-x}}$$



FastText: Motivation

Much'ananayakapushasqakupuniñataqsunamá

Much'a -na -naya -ka -pu -sha -sqa -ku -puni -ña -taq -suna -má

"So they really always have been kissing each other then"

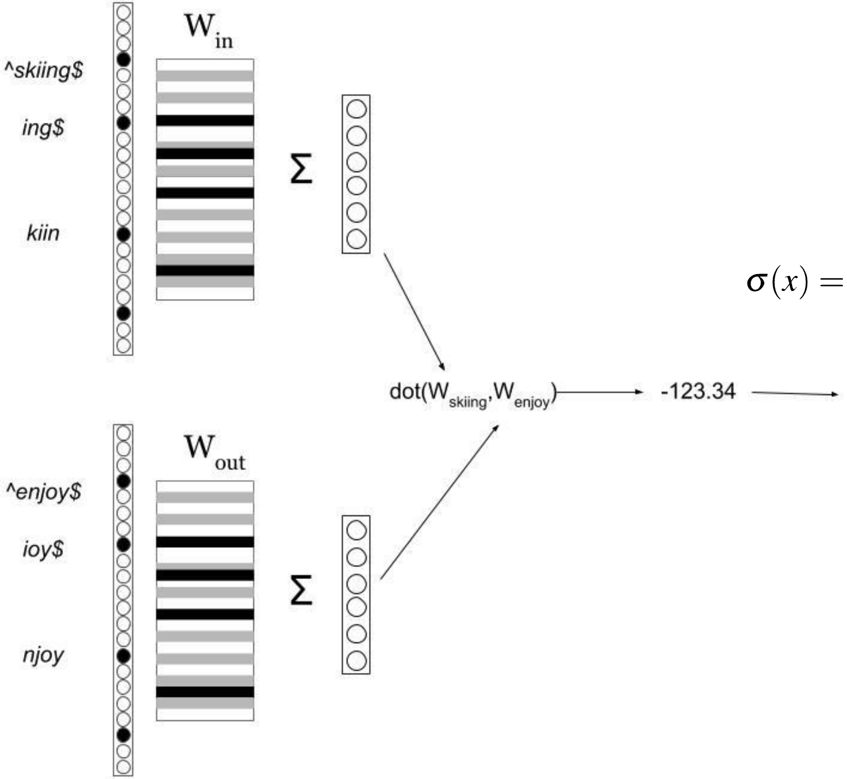
```
Much'a  to kiss
-na     expresses obligation, lost in translation
-naya   expresses desire
-ka     diminutive
-pu     reflexive (kiss *eachother*)
-sha    progressive (kiss*ing*)
-sqa    declaring something the speaker has not personally witnessed
-ku     3rd person plural (they kiss)
-puni   definitive (really*)
-ña     always
-taq    statement of contrast (...then)
-suna   expressing uncertainty (So...)
-má     expressing that the speaker is surprised
```

	Singular+neut	Plural+neut	
Nominative	предложение	предложения	sentence (s)
Genitive	предложения	предложений	(of) sentence (s)
Dative	предложению	предложениям	(to) sentence (s)
Accusative	предложение	предложения	sentence (s)
Instrumental	предложением	предложениями	(by) sentence (s)
Prepositional	предложении	предложениях	(in/at) sentence (s)

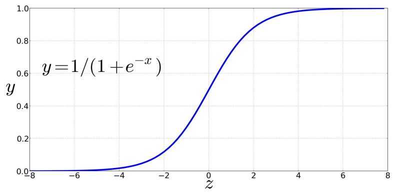
Subword Representation

skiing = {`^skiing$, ^ski, skii, kiin, iing, ing$`}

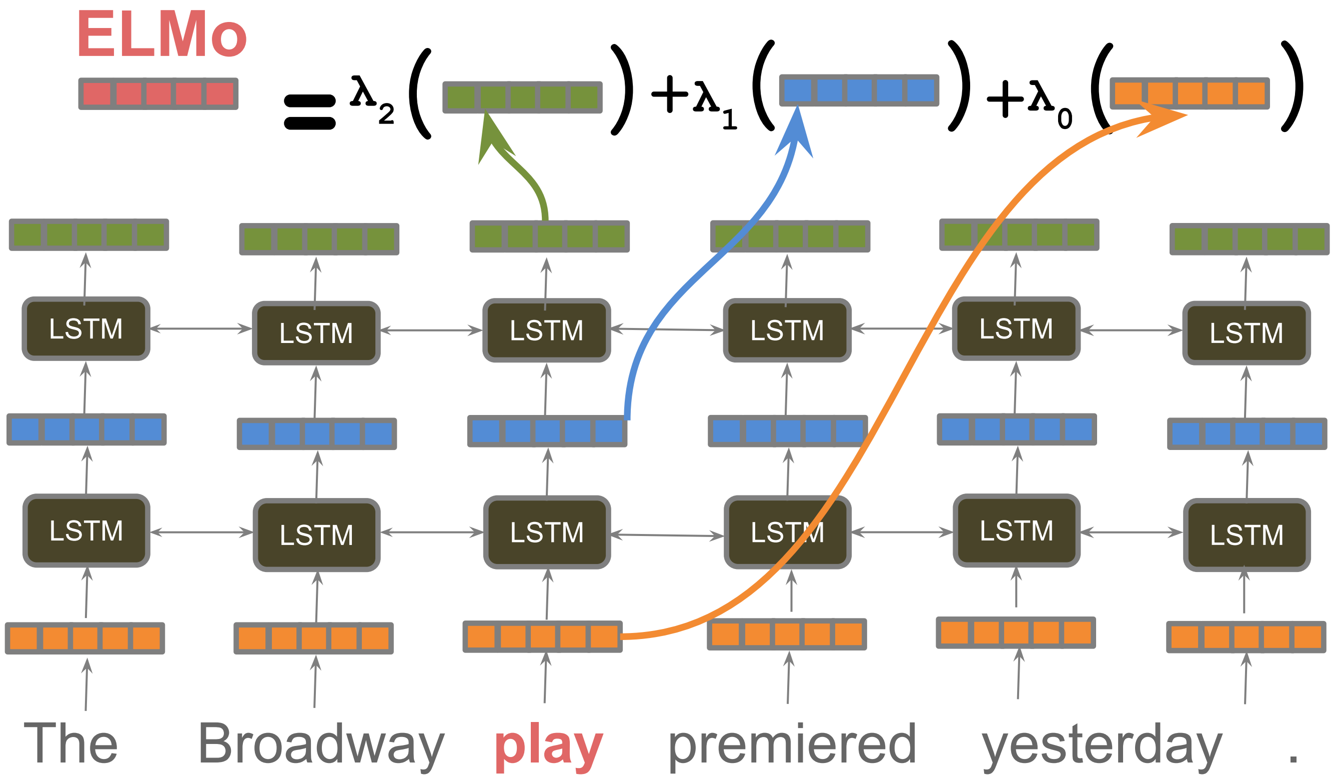
FastText



$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



ELMO



ELMo
= λ2 ( ) + λ1 ( ) + λ0 ( )
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
LSTM
The
Broadway
play
premiered
yesterday
.

Announcements

- HW1 due Sept 24
- HW2 out Oct 2

Automatic Speech Recognition (ASR)

- Automatic speech recognition (or speech-to-text) systems transform speech utterances into their corresponding text form, typically in the form of a word sequence
- Downstream applications of ASR
  - Speech understanding
  - Audio information retrieval
  - Speech translation
  - Keyword search

→ `She sells sea shells`

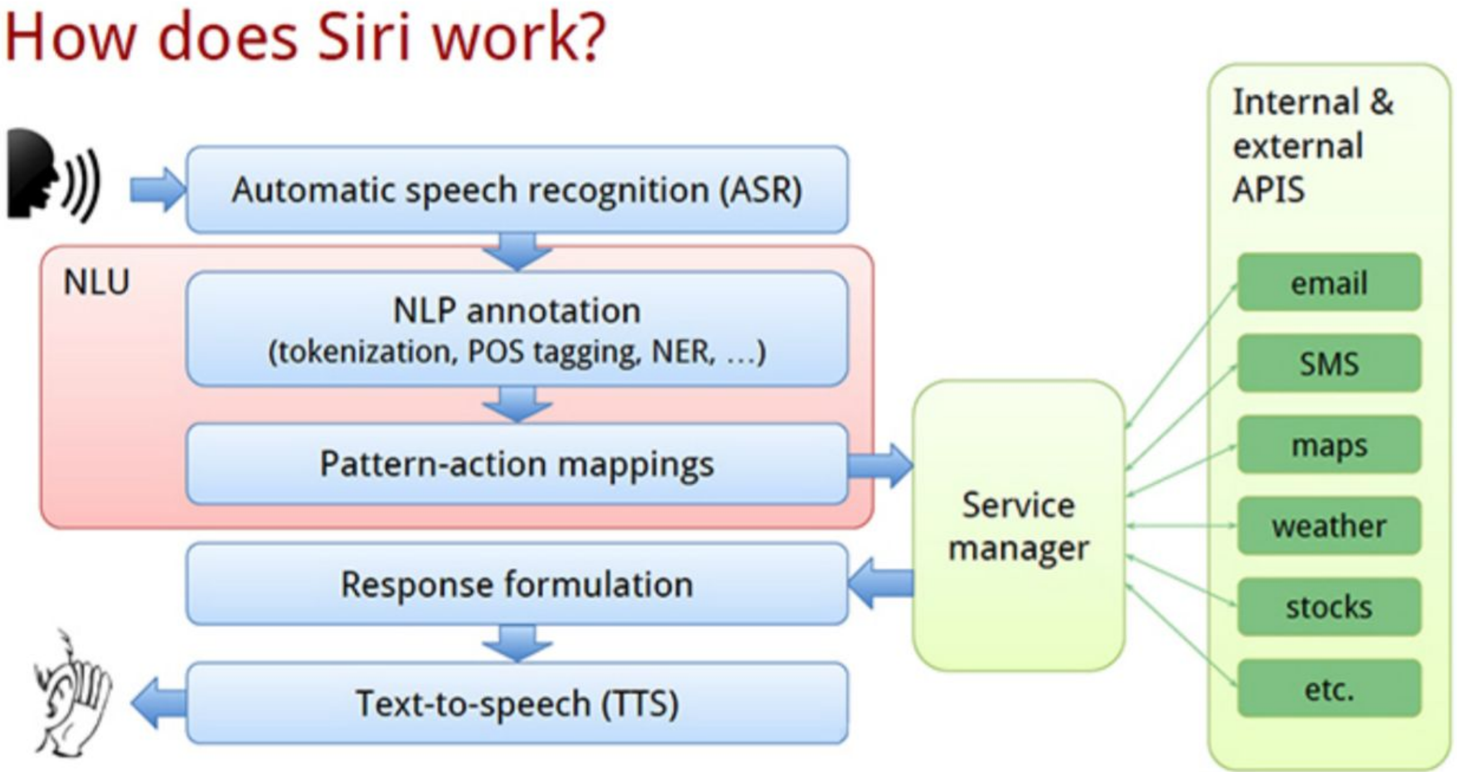
Speech signal

Speech transcript

What ASR is Not

Slide credit: Preethi Jyothi

ASR is the Front Engine



Slide credit: Preethi Jyothi

Why is ASR a Challenging Problem?

- Style:
  - Read speech vs spontaneous (conversational) speech
  - Command & control vs continuous natural speech
- Speaker characteristics:
  - Rate of speech, accent, prosody (stress, intonation), speaker age, pronunciation variability even when the same speaker speaks the same word
- Channel characteristics:
  - Background noise, room acoustics, microphone properties, interfering speakers
- Task specifics:
  - Vocabulary size (very large number of words to be recognized), language-specific complexity, resource limitations

Slide credit: Preethi Jyothi

History of ASR

The very first ASR

RADIO REX (1922)

History of ASR

SHOEBOX (IBM, 1962)

1 word

Freq. detector

1922 1932 1942 1952 1962 1972 1982 1992 2002 2012

Slide credit: Preethi Jyothi

History of ASR

ARPA
ADVANCED RESEARCH PROJECTS AGENCY

HARPY (CMU, 1976)

1 word

Freq. detector

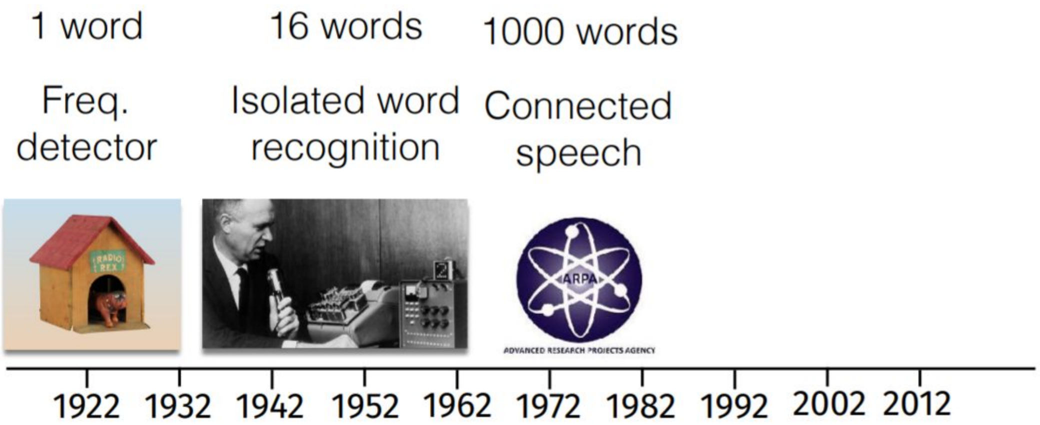
16 words

Isolated word recognition

1922 1932 1942 1952 1962 1972 1982 1992 2002 2012

Slide credit: Preethi Jyothi

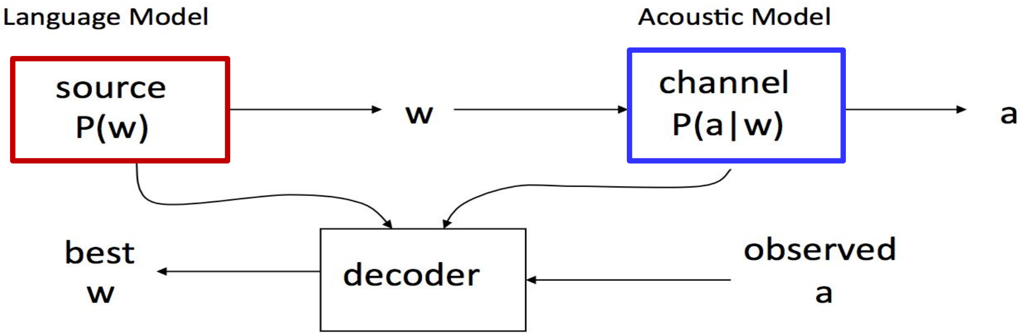
History of ASR



Slide credit: Preethi Jyothi

Statistical ASR : The Noisy Channel Model

~80s



Fred Jelinek
1932-2010

$$w^* = \arg\max_w P(w|a)$$
$$\propto \arg\max_w P(a|w)P(w)$$

Acoustic model

Language model:
Distributions over sequences of words (sentences)

History of ASR

(1980s)

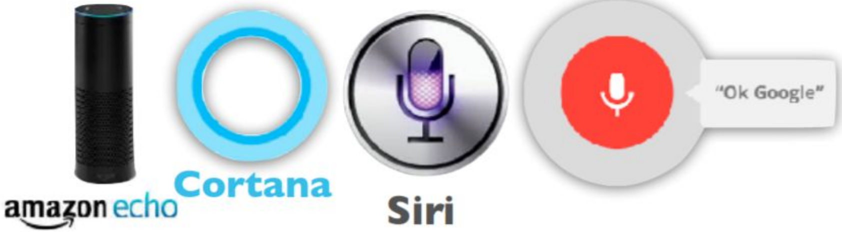
1 word	16 words	1000 words
Freq. detector	Isolated word recognition	Connected speech

ARPA
ADVANCED RESEARCH PROJECTS AGENCY

1922 1932 1942 1952 1962 1972 1982 1992 2002 2012

Slide credit: Preethi Jyothi

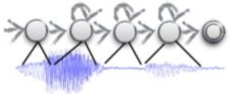
History of ASR



DEEP NEURAL NETWORK BASED SYSTEMS (>2010)

1 word	16 words	1000 words	10K+ words
Freq. detector	Isolated word recognition	Connected speech	LVCSR systems

ARPA
ADVANCED RESEARCH PROJECTS AGENCY



1922 1932 1942 1952 1962 1972 1982 1992 2002 2012

Slide credit: Preethi Jyothi

Evaluating an ASR system

- Word/Phone error rate (ER)
  - uses the Levenshtein distance measure: What are the minimum number of edits (insertions/deletions/substitutions) required to convert W* to W_{ref} ?

$$\mathrm{ER} = \frac{\sum_{j=1}^{N} \mathrm{Ins}_j + \mathrm{Del}_j + \mathrm{Sub}_j}{\sum_{j=1}^{N} \ell_j}$$

- Word Error Rate =

100 (Insertions+Substitutions + Deletions)

Total Word in Correct Transcript

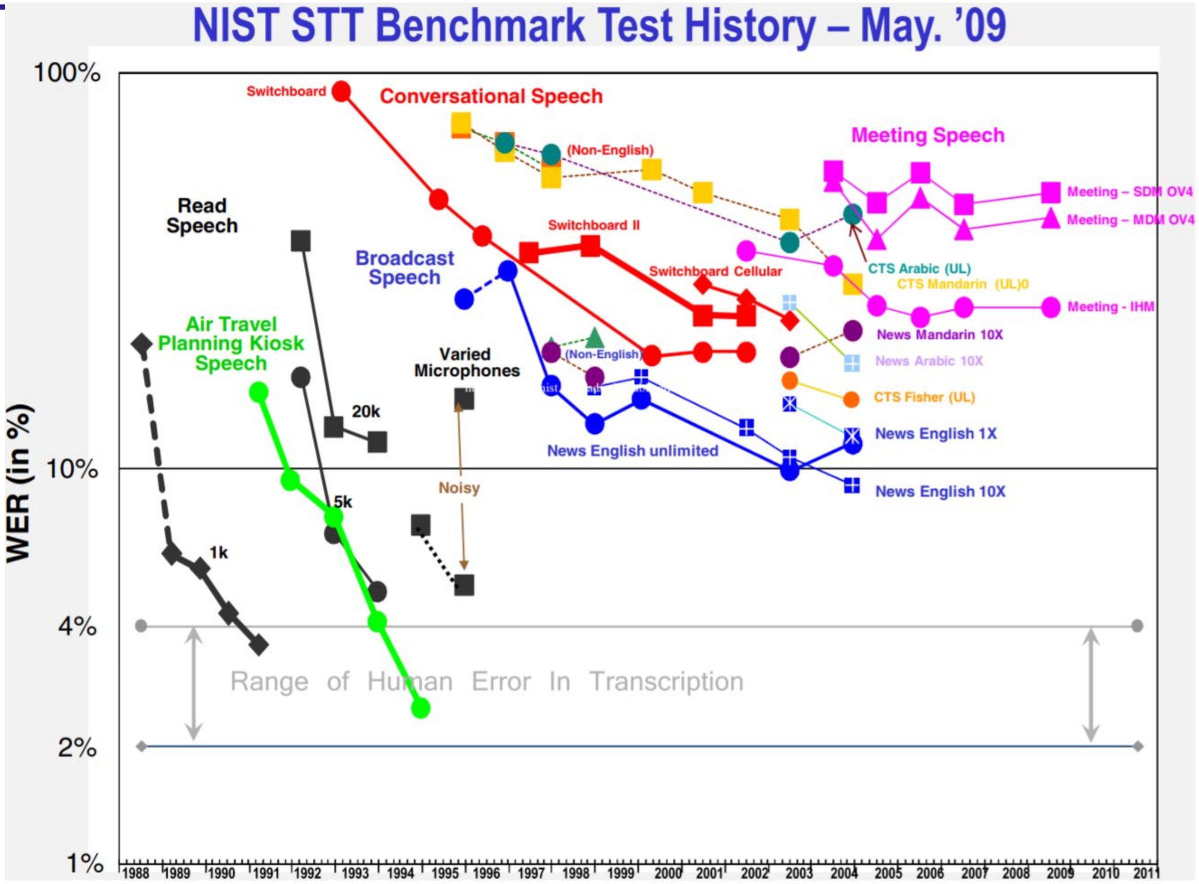
Alignment example:

REF:	portable	****	PHONE	UPSTAIRS	last	night so
HYP:	portable	FORM	OF	STORES	last	night so
Eval		I	S	S		

WER = 100 (1+2+0)/6 = 50%

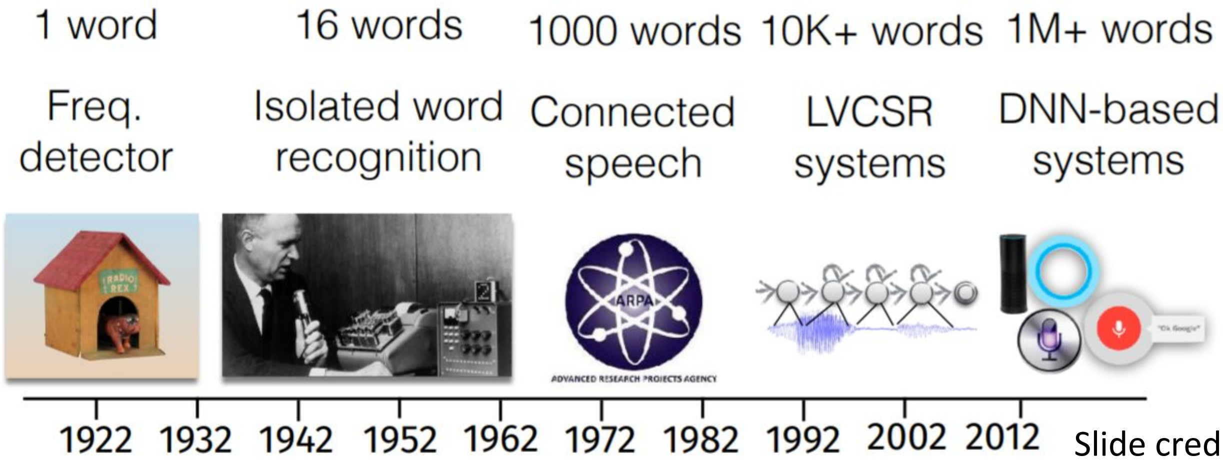
From J&M

NIST ASR Benchmark Test History



http://www.itl.nist.gov/iad/mig/publications/ASRhistory/

What's Next?



Slide credit: Preethi Jyothi

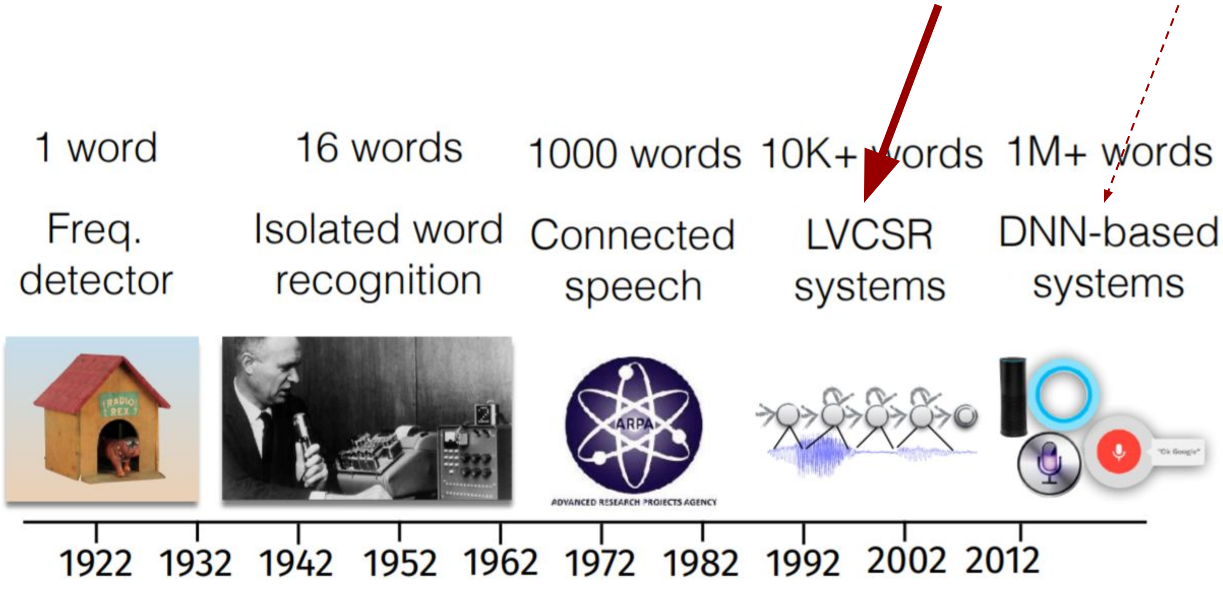
What's Next?

https://www.youtube.com/watch?v=gNx0huL9qsQ

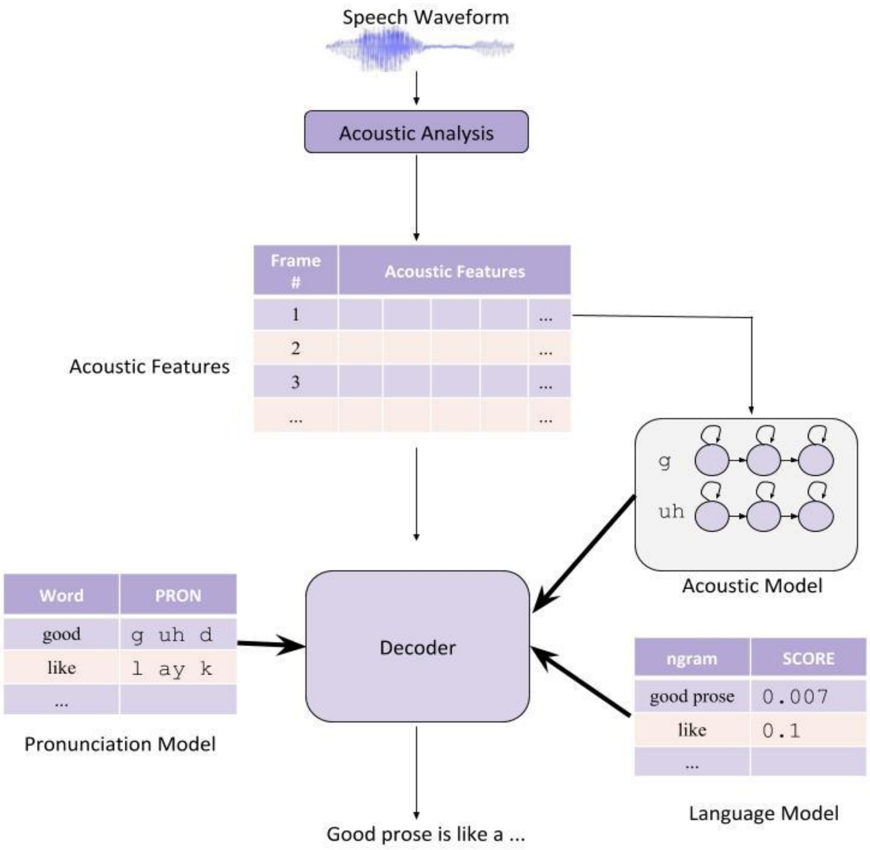
- accented speech
- low-resource
- speaker separation
- short queries
- etc.

Link credit: Preethi Jyothi

In our course

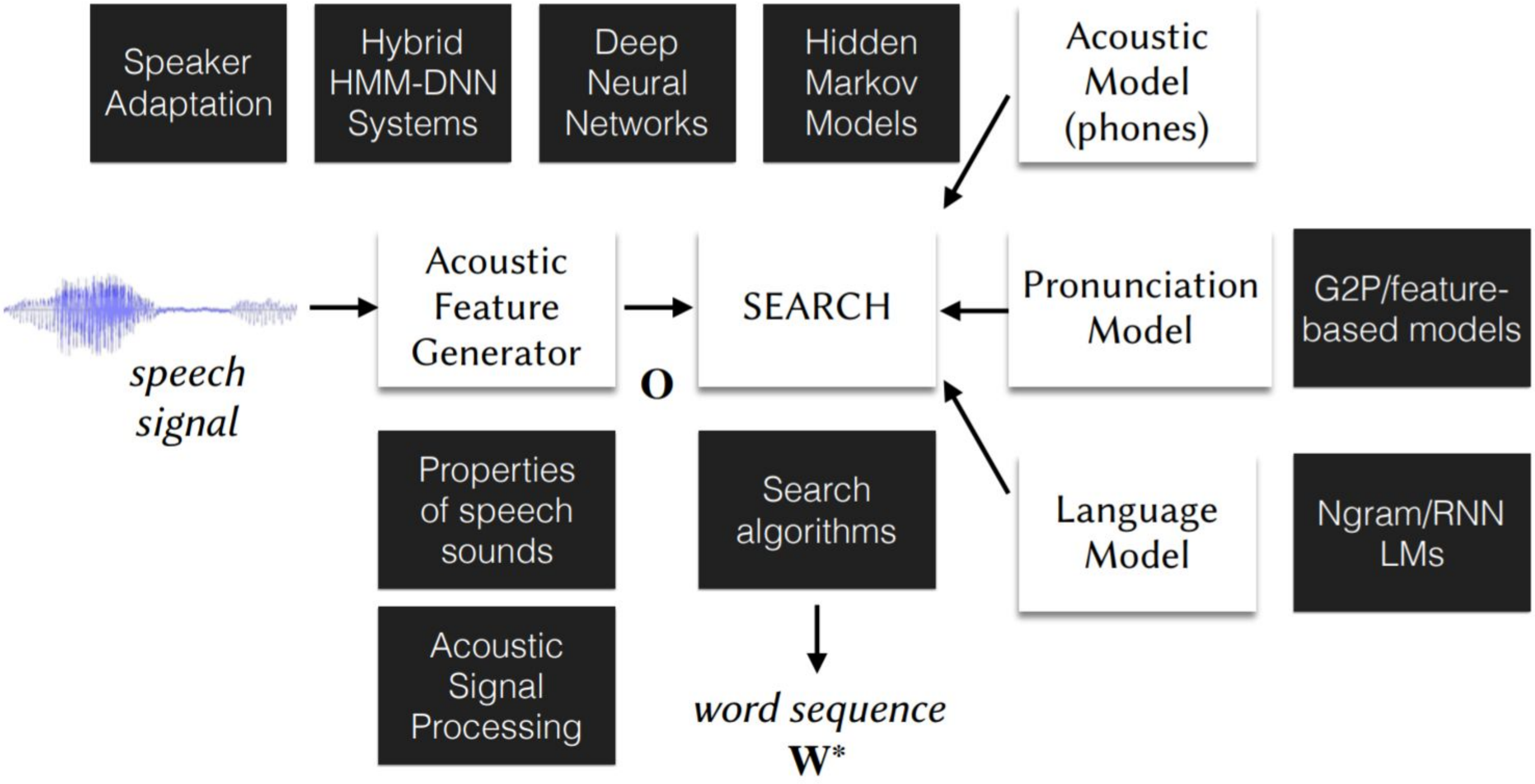


Statistical ASR



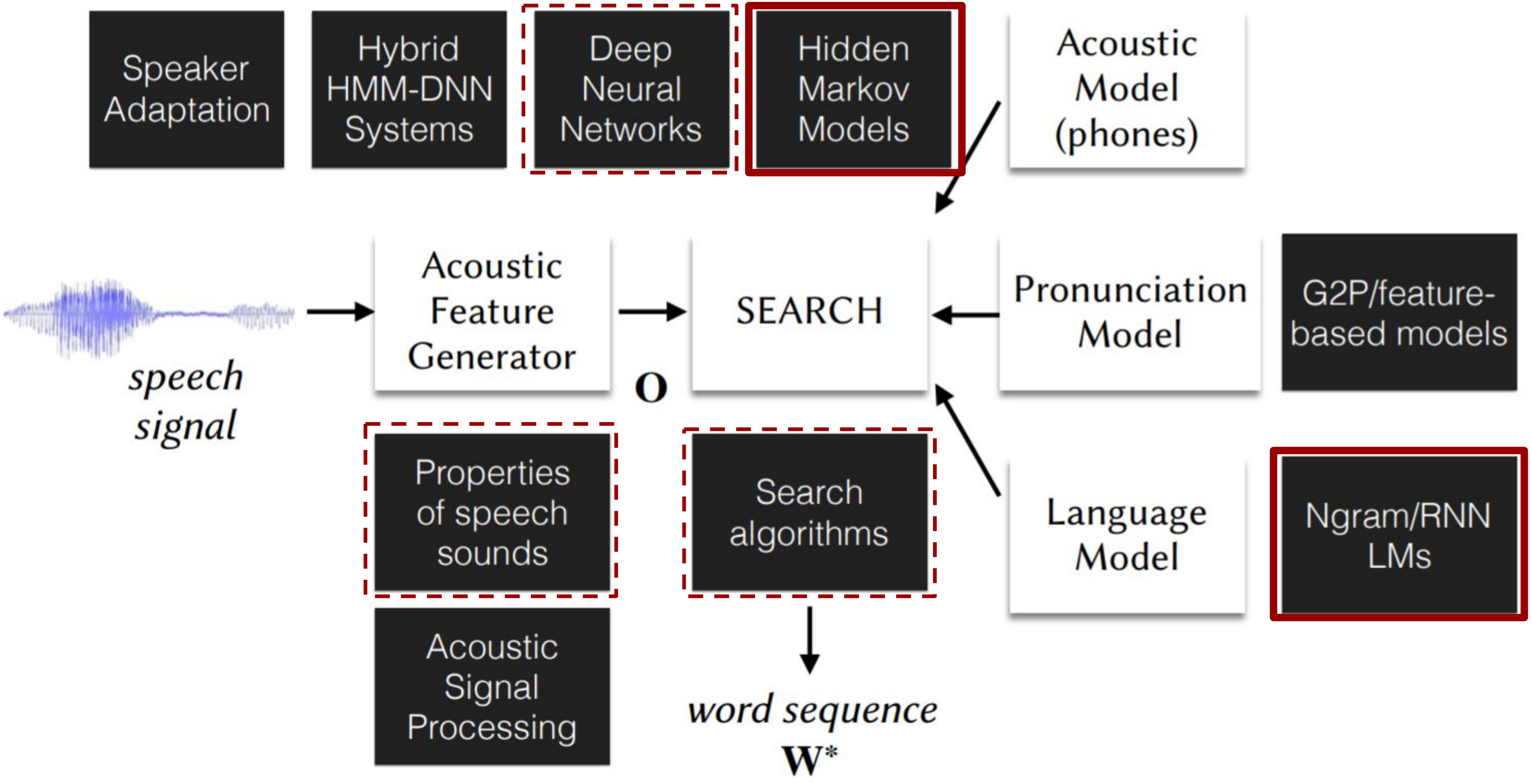
Slide by Preethi Jyothi

ASR Topics



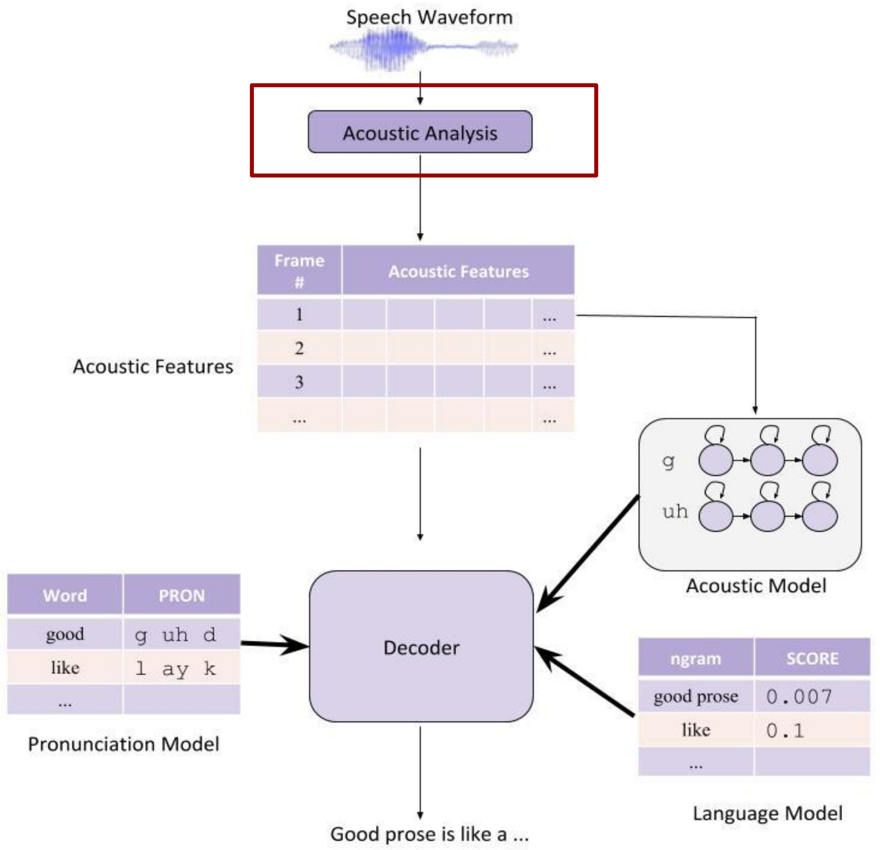
Slide by Preethi Jyothi

In our course



Slide by Preethi Jyothi

Acoustic Analysis



Slide by Preethi Jyothi

What is speech - physical realisation

- Waves of changing air pressure
- Realised through excitation from the vocal cords
- Modulated by the vocal tract, the articulators (tongue, teeth, lips)
- Vowels: open vocal tract
- Consonants are constrictions of vocal tract

- Representation:
  - acoustics
  - linguistics

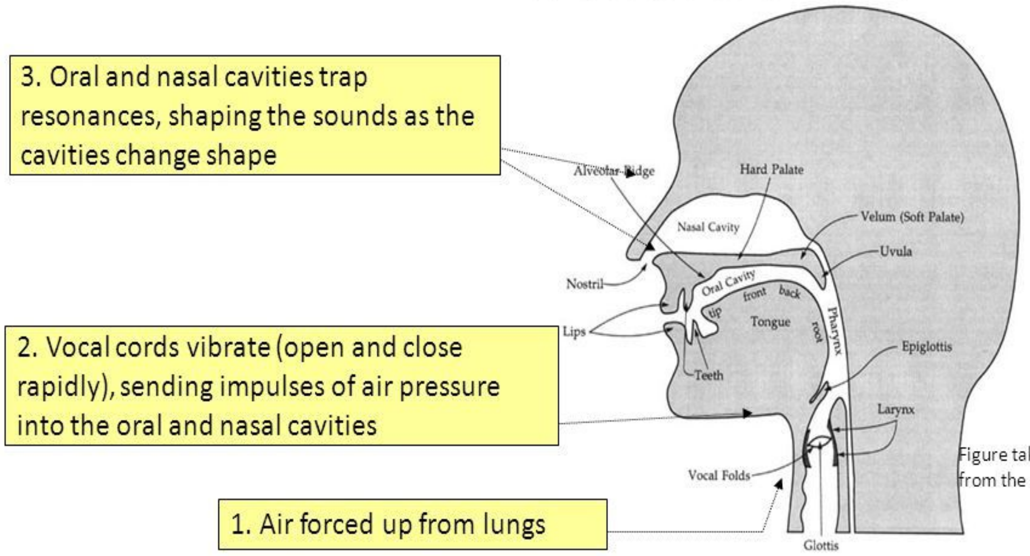
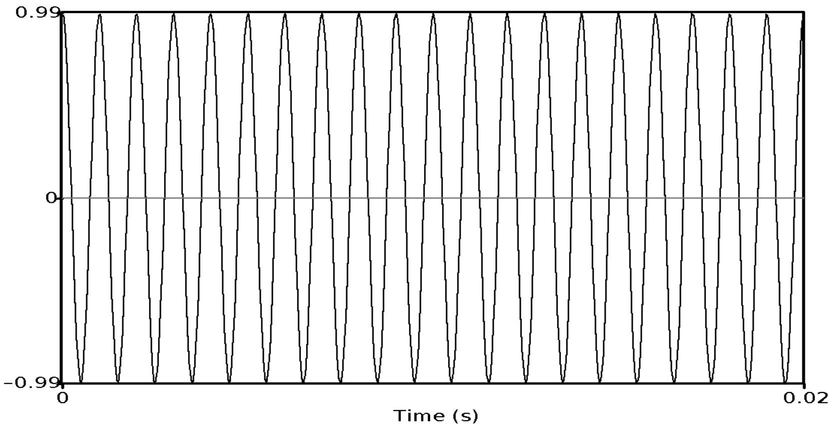


Figure taken from the web

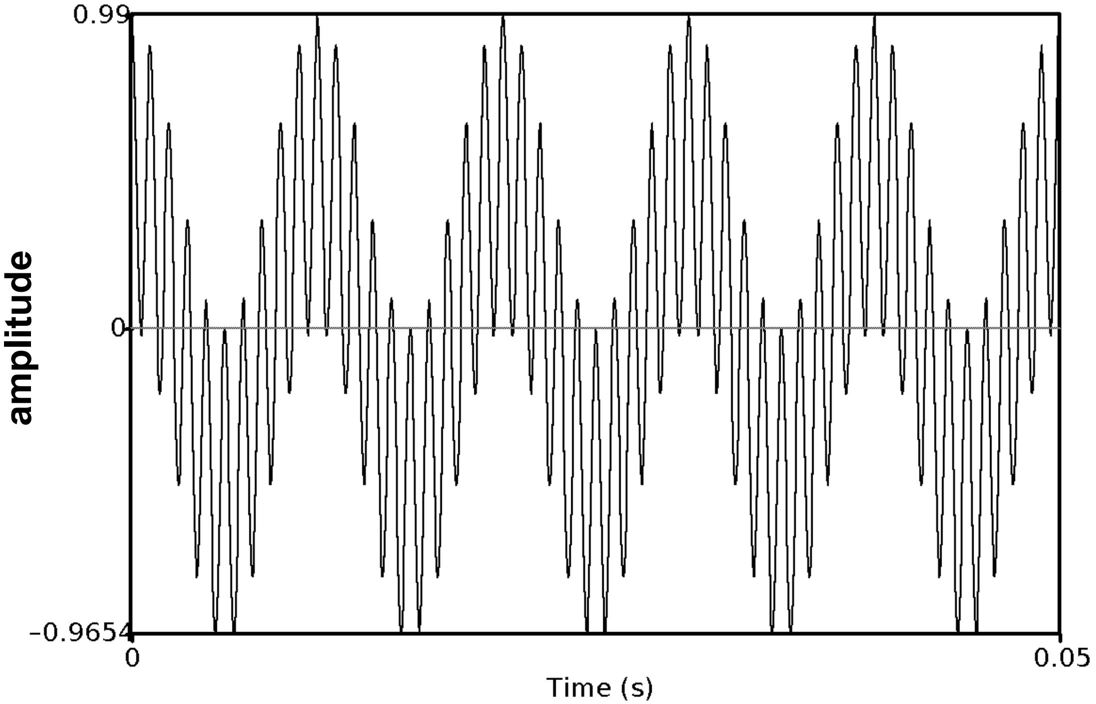
Acoustics

Simple Periodic Waves of Sound



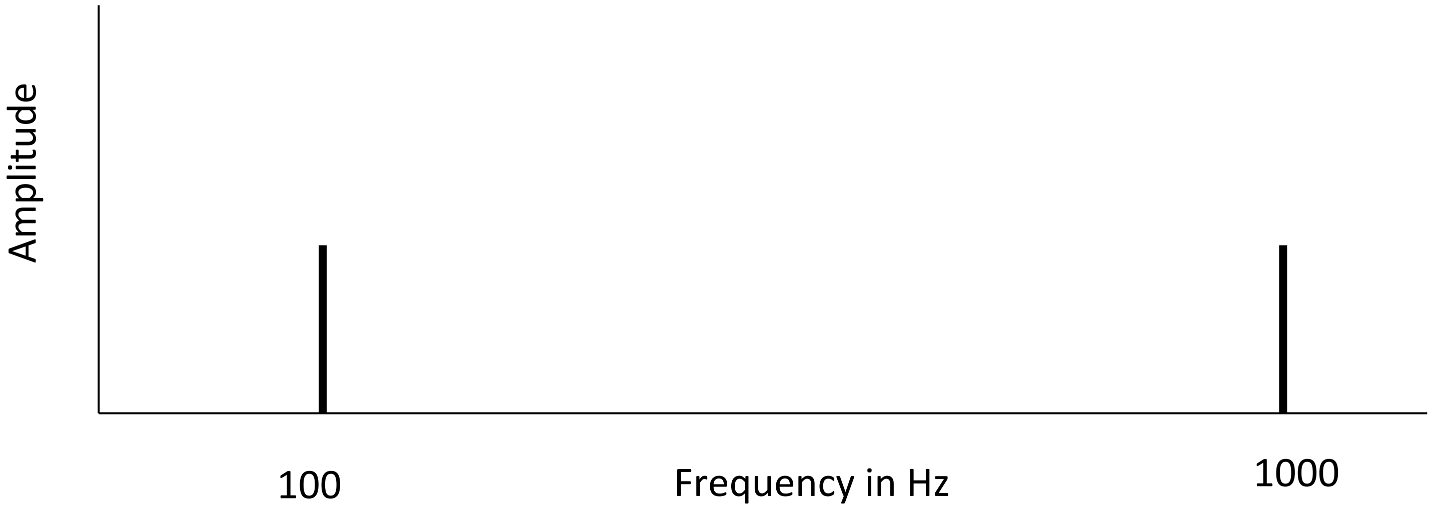
- Y axis: Amplitude = amount of air pressure at that point in time
- X axis: Time
- Frequency = number of cycles per second.
  - 20 cycles in .02 seconds = 1000 cycles/second = 1000 Hz

Complex Waves: 100Hz+1000Hz

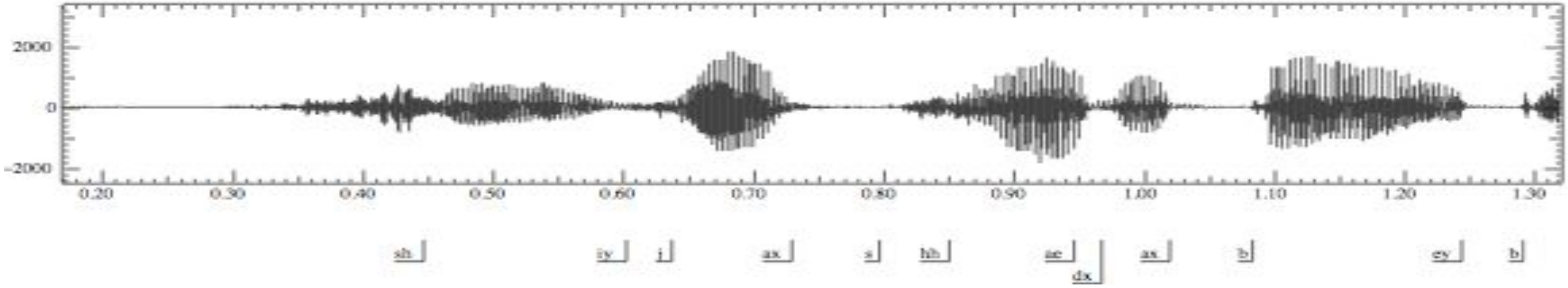


Spectrum

Frequency components (100 and 1000 Hz) on x-axis

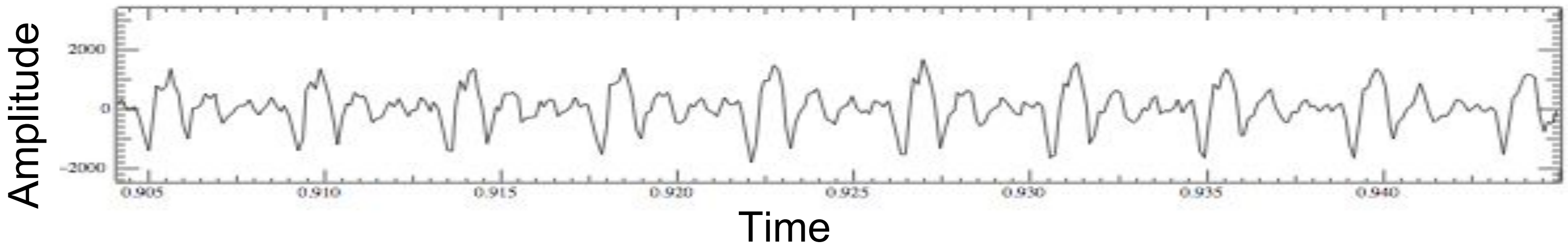


“She just had a baby”



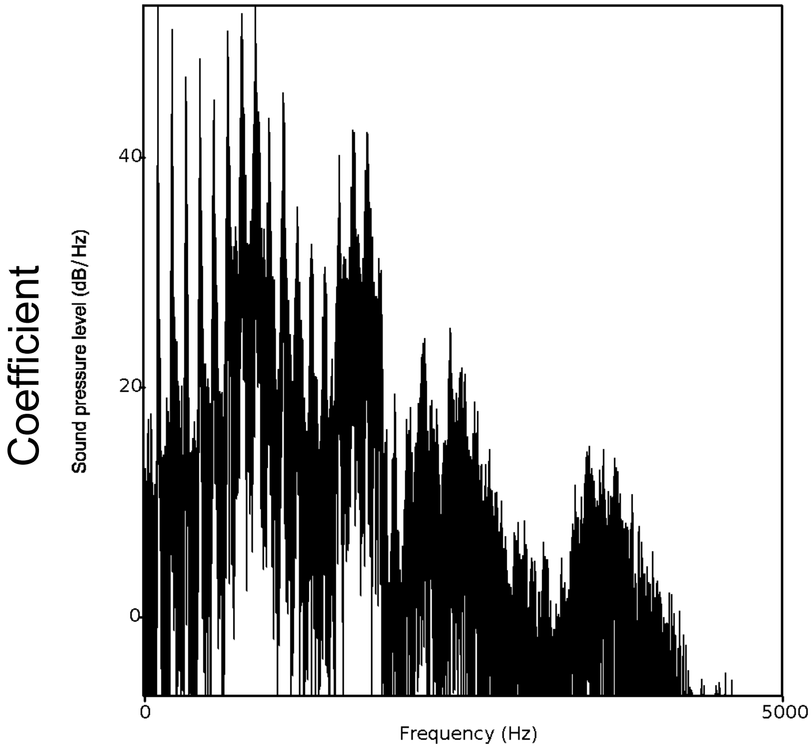
- What can we learn from a wavefile?
  - No gaps between words (!)
  - Vowels are voiced, long, loud
  - Voicing: regular peaks in amplitude
  - When stops closed: no peaks, silence
  - Peaks = voicing: .46 to .58 (vowel [iy], from second .65 to .74 (vowel [ax]) and so on
  - Silence of stop closure (1.06 to 1.08 for first [b], or 1.26 to 1.28 for second [b])
  - Fricatives like [sh]: intense irregular pattern; see .33 to .46

Part of [ae] waveform from "had"

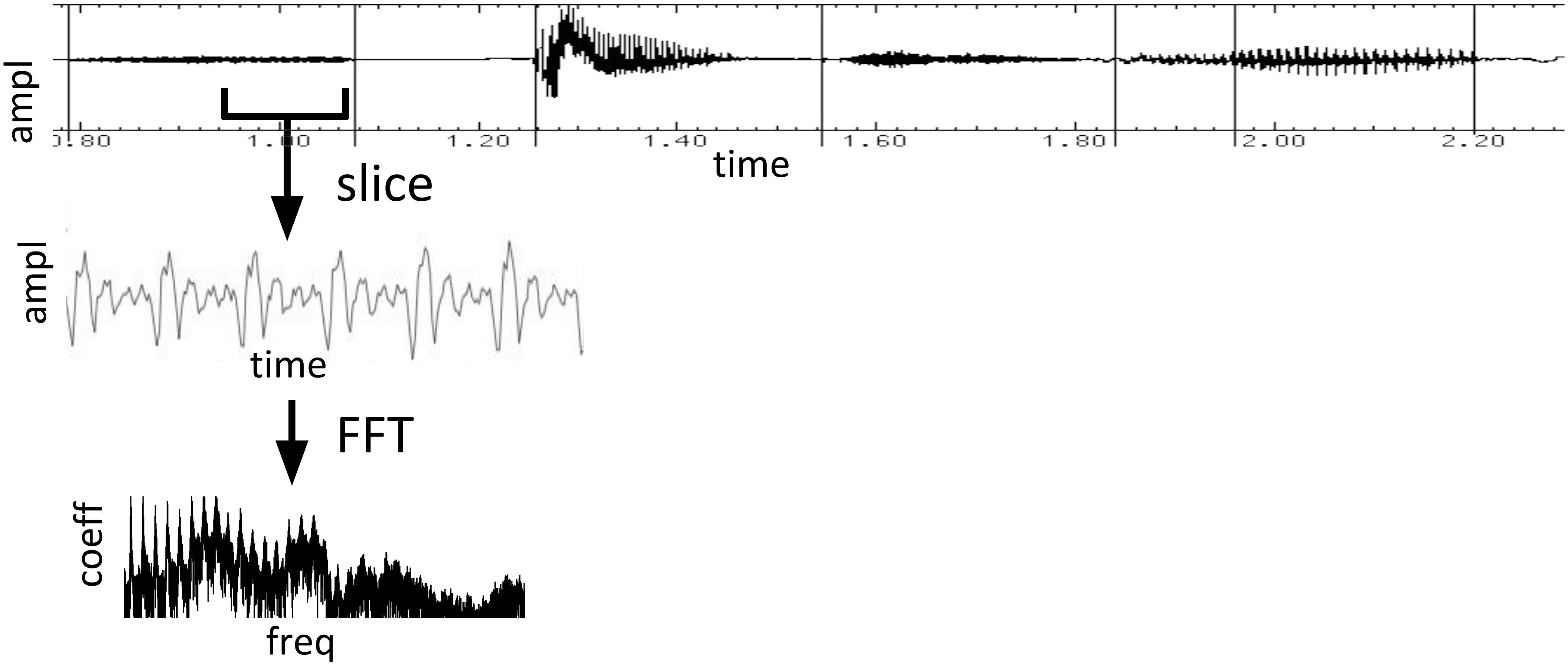


- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- Large wave has frequency of 250 Hz (9 times in .036 seconds)
- Small wave roughly 4 times this, or roughly 1000 Hz
- Two little tiny waves on top of peak of 1000 Hz waves

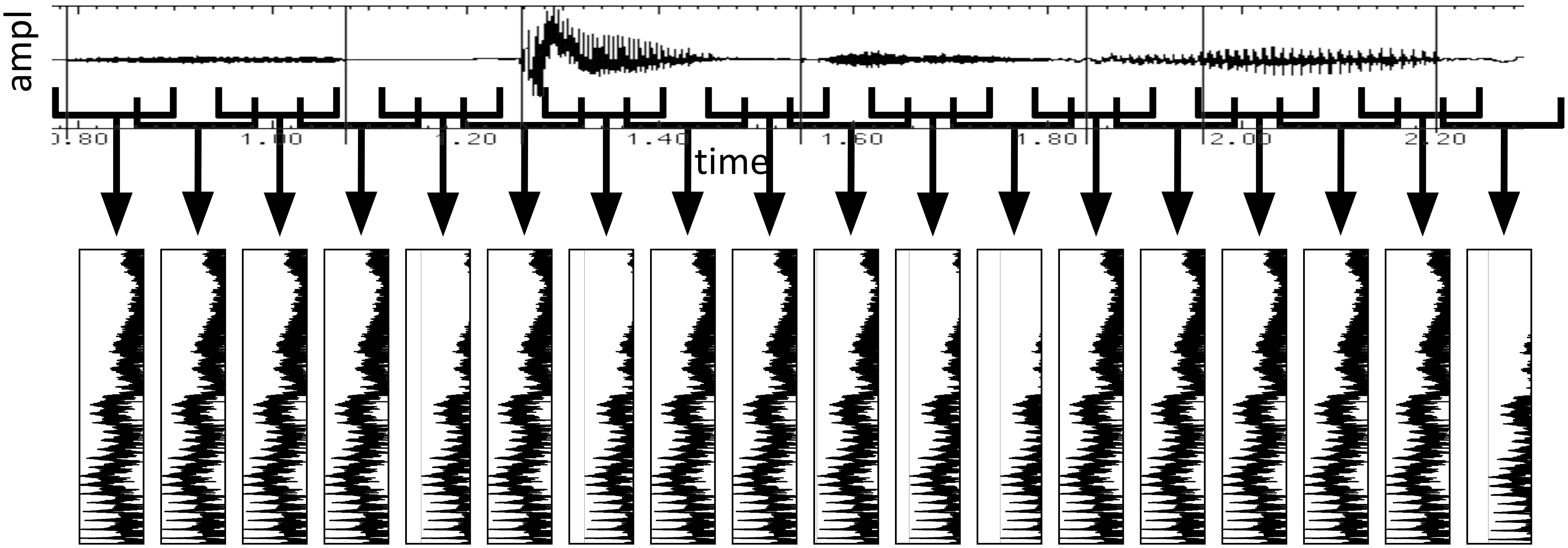
Spectrum of an Actual Speech



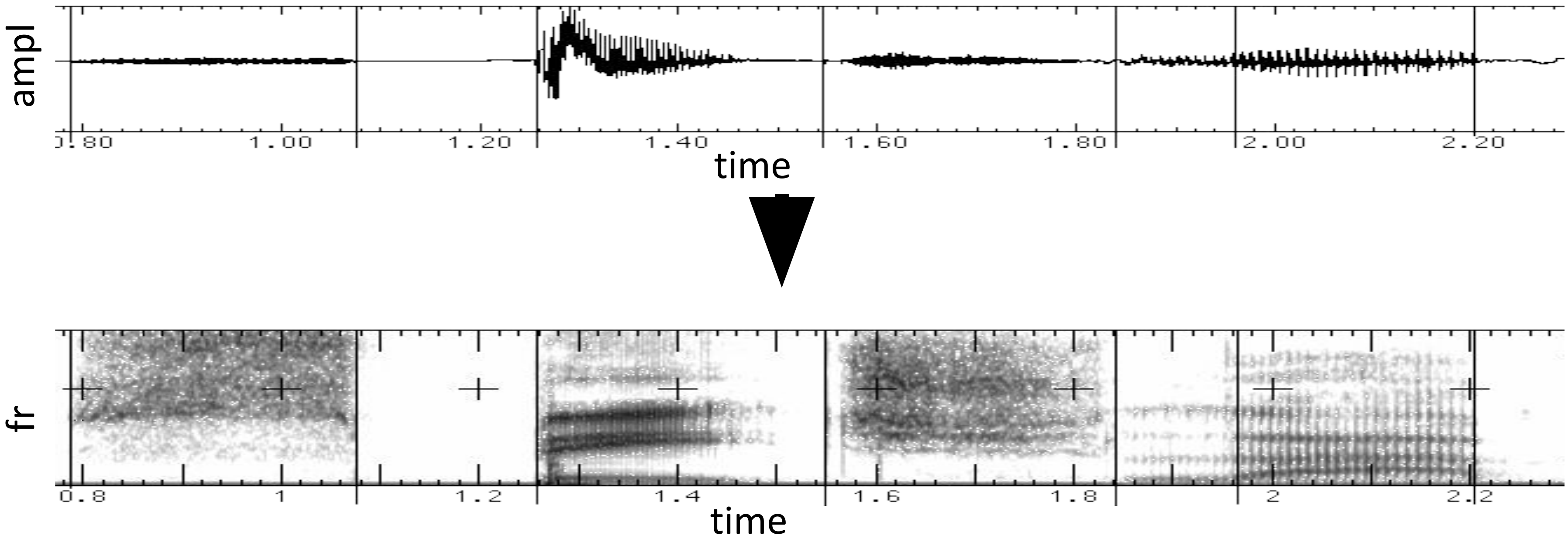
Spectrograms



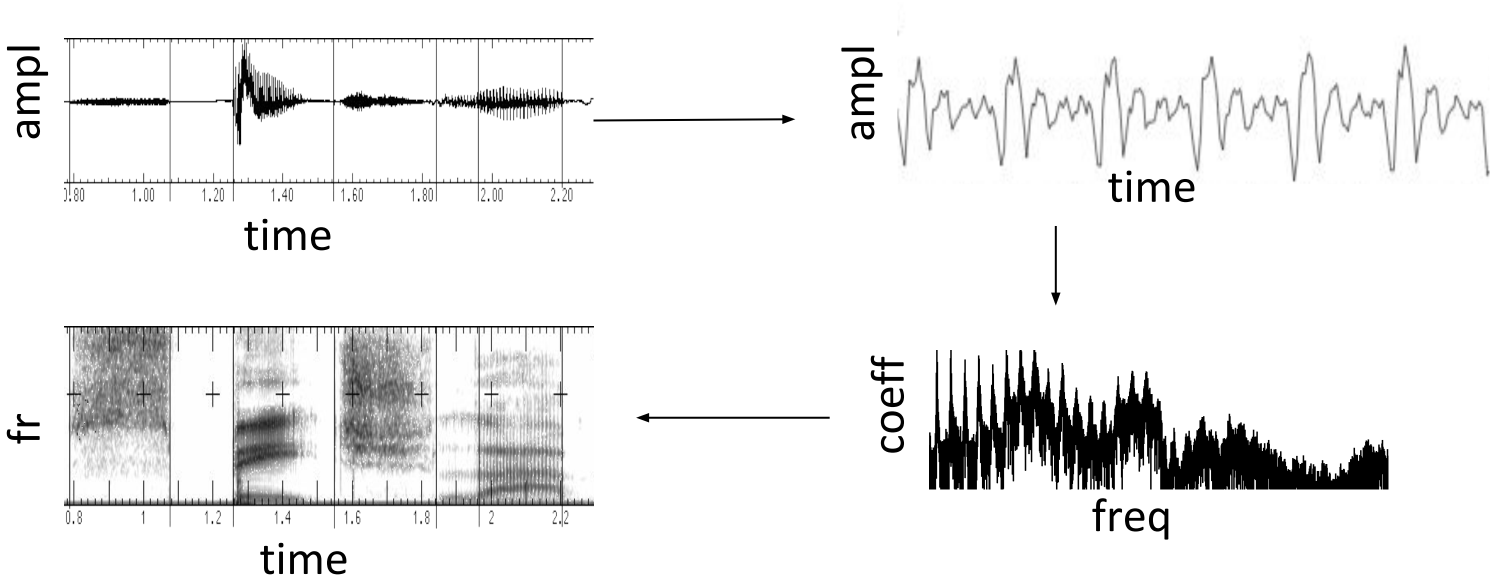
Spectrograms



Spectrograms

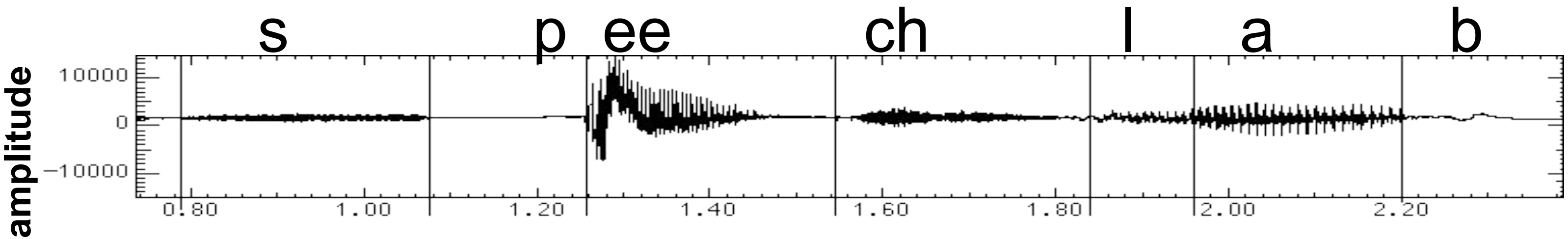


Types of Graphs

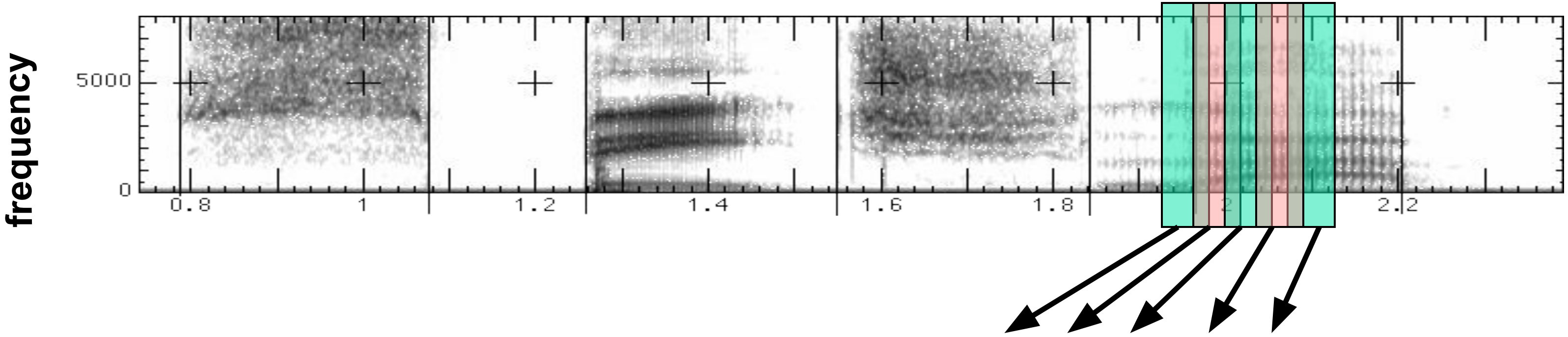


Speech in a Slide

- Frequency gives pitch; amplitude gives volume



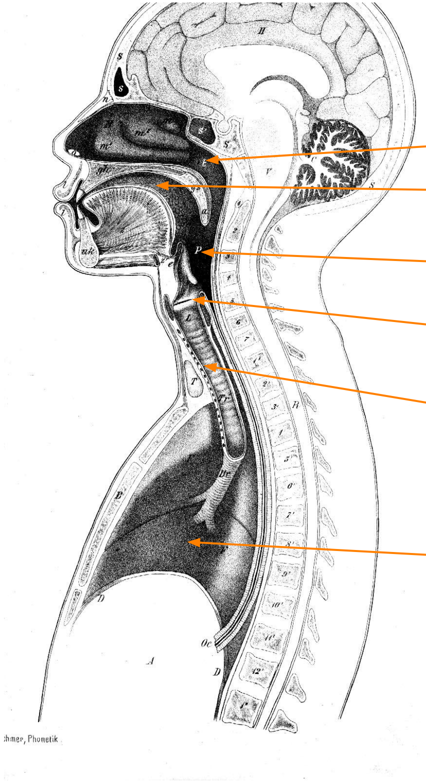
- Frequencies at each time slice processed into observation vectors



……………………………………………$x_{12}x_{13}x_{12}x_{14}x_{14}$………..

Articulation

Articulatory System



Nasal cavity

Oral cavity

Pharynx

Vocal folds (in the larynx)

Trache
a

Lungs

Sagittal section of the vocal tract (Techmer 1880)
Text from Ohala, Sept 2001, from Sharon Rose slide

Space of Phonemes

	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		ᴙ	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ		ꞎ	𝼆	𝼄				
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		𝼈						

- Standard international phonetic alphabet (IPA) chart of consonants

Place

Places of Articulation

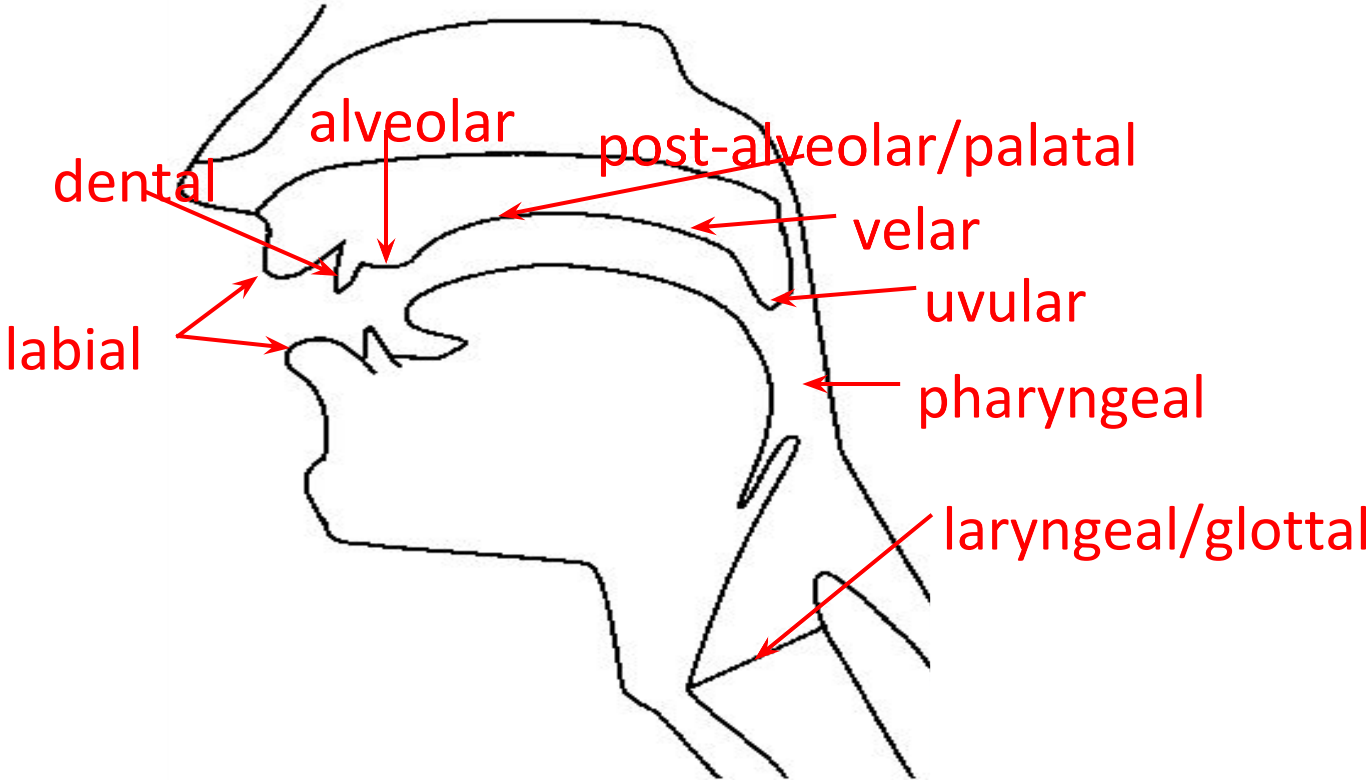
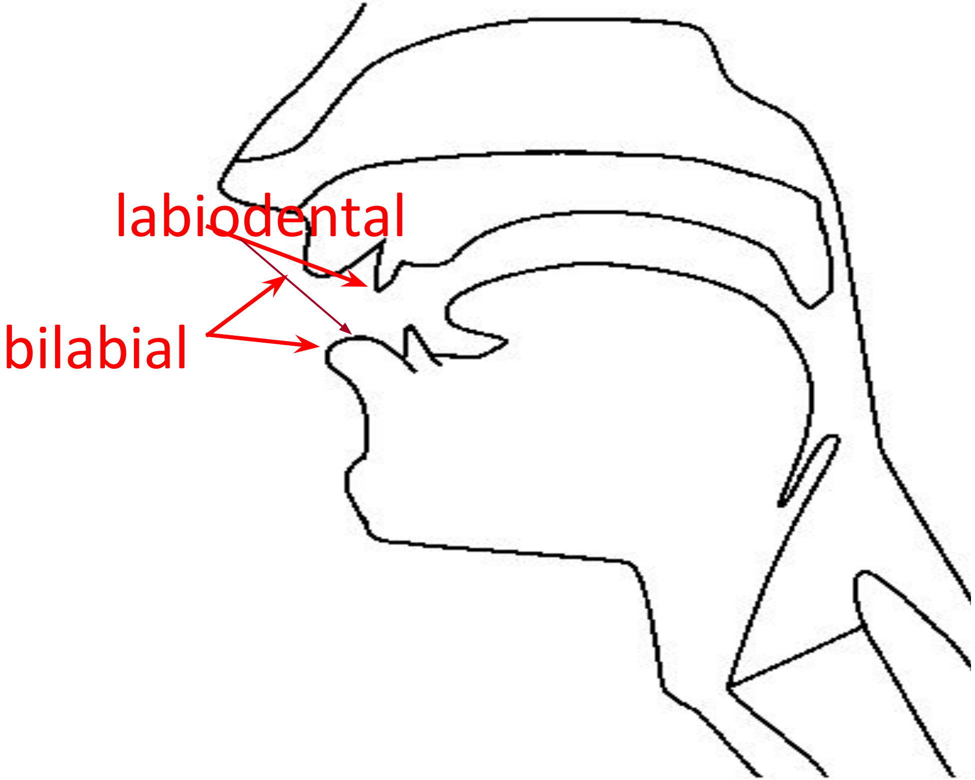


Figure thanks to Jennifer Venditti

Labial place



Bilabial:

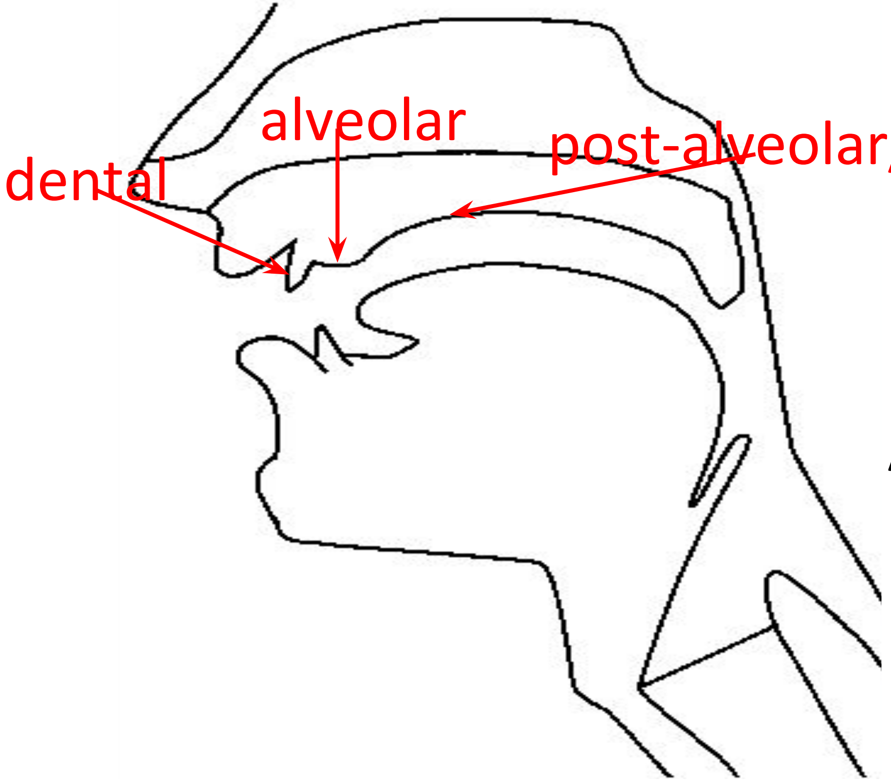
p, b, m

Labiodental:

f, v

Figure thanks to Jennifer Venditti

Coronal place



Dental:
  th/dh
Alveolar:
  t/d/s/z/l/n
Post:
  sh/zh/y

Figure thanks to Jennifer Venditti

Dorsal Place

Velar:

k/g/ng

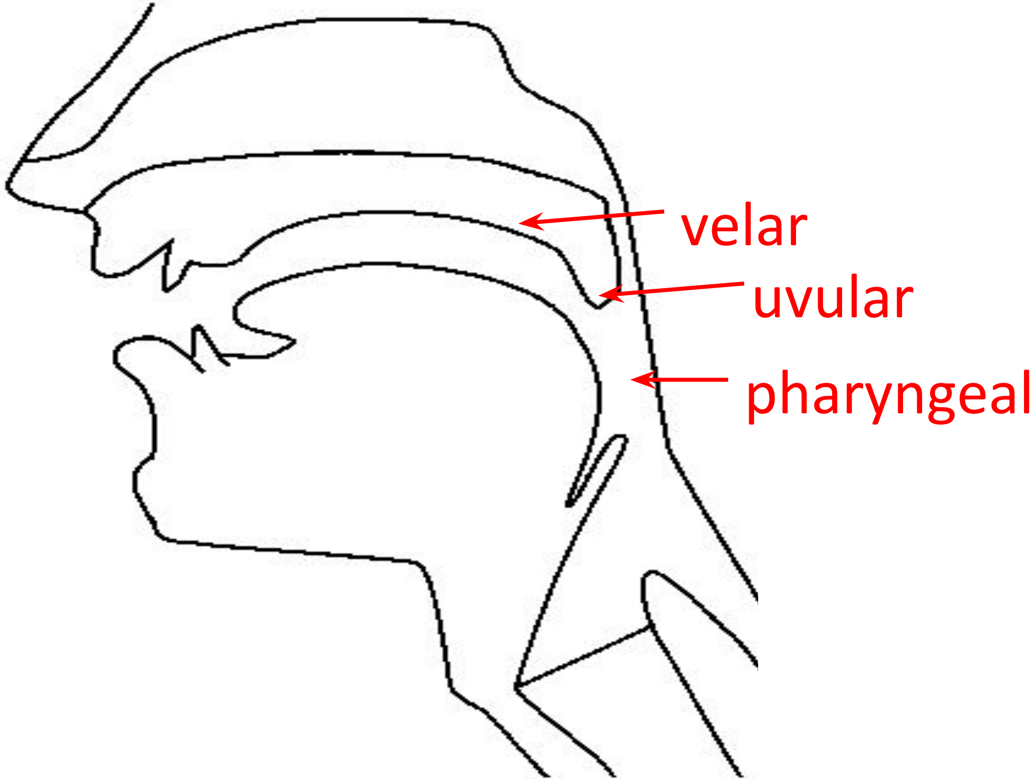


Figure thanks to Jennifer Venditti

Space of Phonemes

	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		я	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ		ꞎ	𝼆	𝼄				
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		𝼈						

- Standard international phonetic alphabet (IPA) chart of consonants

Manner

Manner of Articulation

- In addition to varying by place, sounds vary by manner

- Stop: complete closure of articulators, no air escapes via mouth
  - Oral stop: palate is raised (p, t, k, b, d, g)
  - Nasal stop: oral closure, but palate is lowered (m, n, ng)

- Fricatives: substantial closure, turbulent: (f, v, s, z)

- Approximants: slight closure, sonorant: (l, r, w)

- Vowels: no closure, sonorant: (i, e, a)

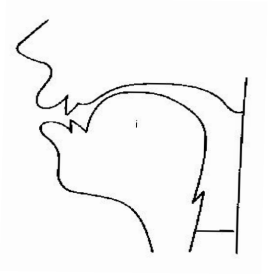
Space of Phonemes

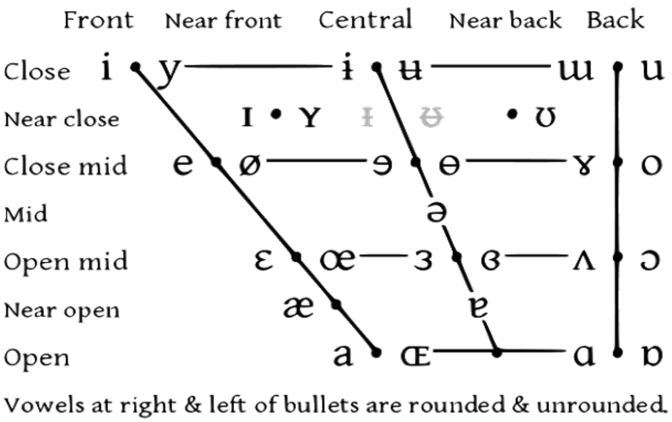
	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		я	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ		ꞎ	𝼆	𝼄				
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		𝼈						

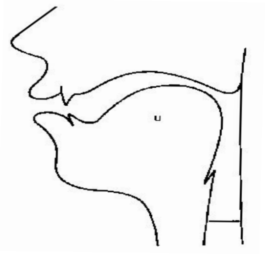
- Standard international phonetic alphabet (IPA) chart of consonants

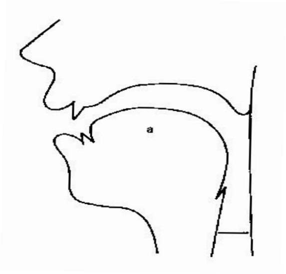
Vowels

Vowel Space

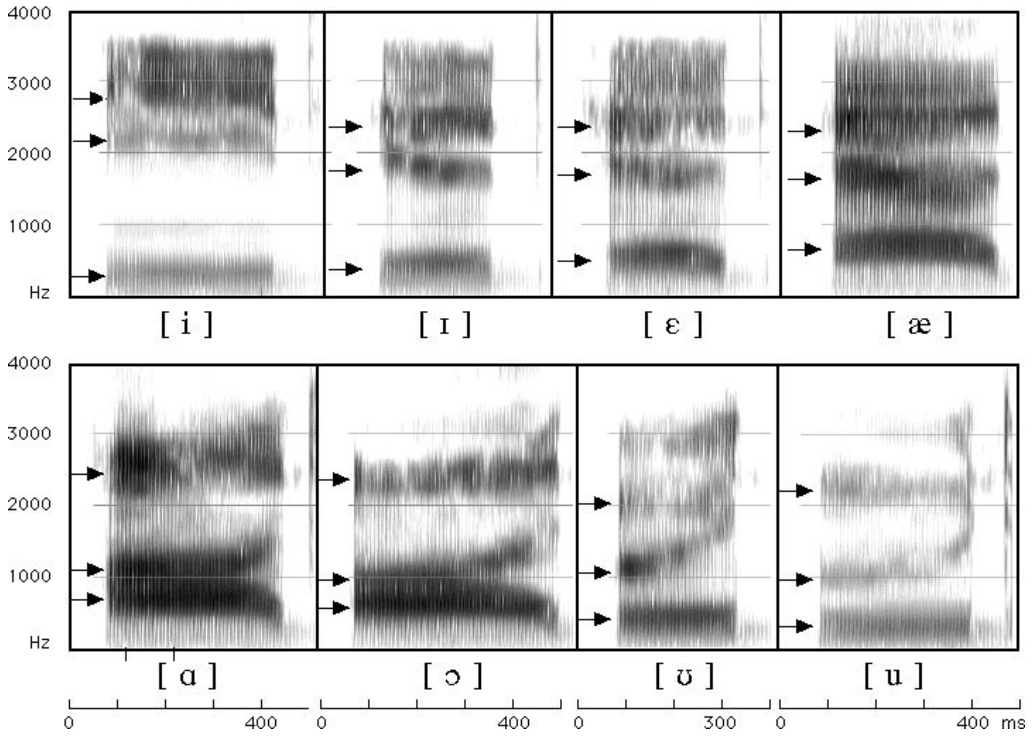



Front Near front Central Near back Back
Close i y ɨ ʉ ɯ u
Near close ɪ ʏ ʊ
Close mid e ø ɘ ɵ ɤ o
Mid ə
Open mid ɛ œ ɜ ɞ ʌ ɔ
Near open æ ɐ
Open a ɶ ɑ ɒ
Vowels at right & left of bullets are rounded & unrounded.

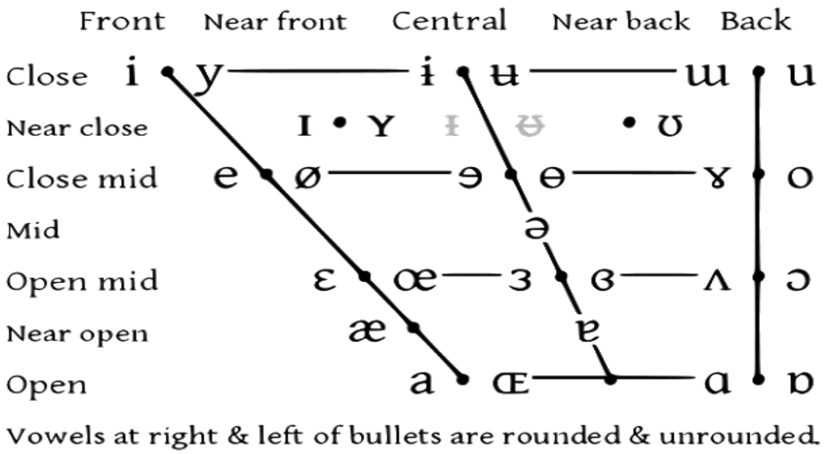


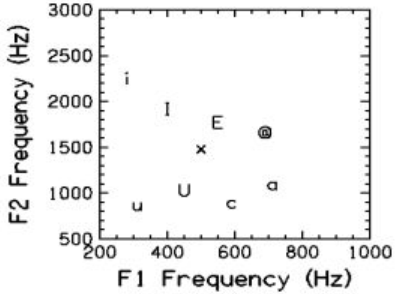


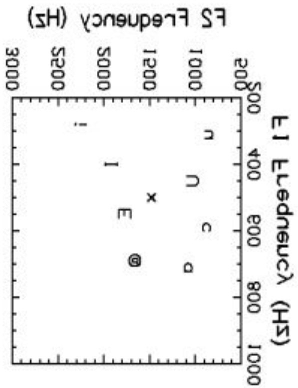
Seeing Formants: the Spectrogram



Vowel Space

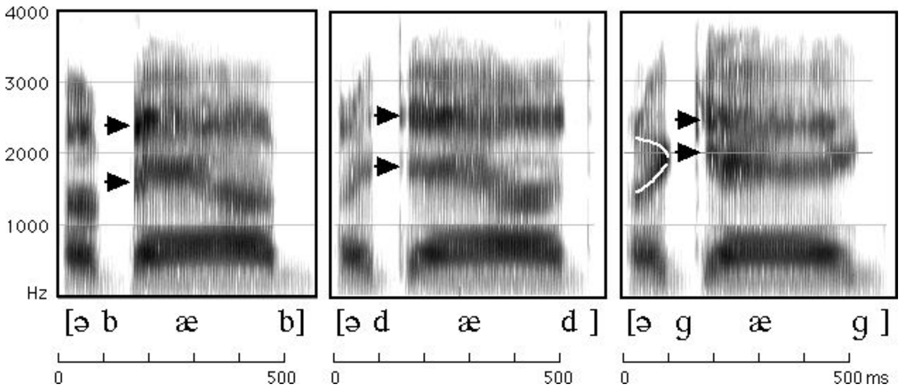

Front
Near front
Central
Near back
Back
Close
Near close
Close mid
Mid
Open mid
Near open
Open
i y ɨ ʉ ɯ u
ɪ ʏ ʊ
e ø ɘ ɵ ɤ o
ə
ɛ œ ɜ ɞ ʌ ɔ
æ ɐ
a ɶ ɑ ɒ
Vowels at right & left of bullets are rounded & unrounded.


F2 Frequency (Hz)
F1 Frequency (Hz)
3000
2500
2000
1500
1000
500
200
400
600
800
1000
i
I
E
@
x
u
U
c
a



Spectrograms

Pronunciation is Context Dependent



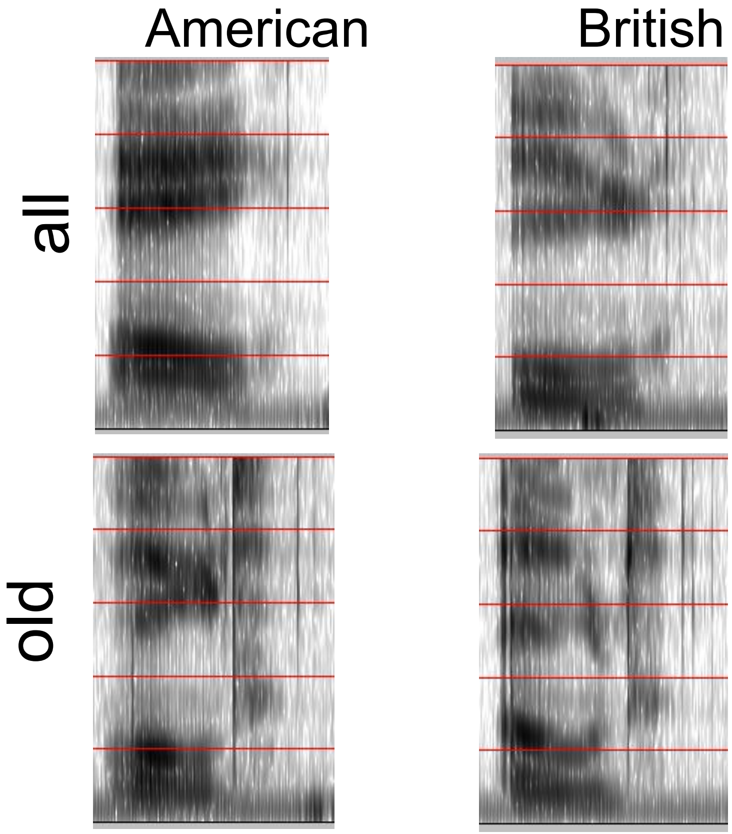
- [bab]: closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab"
- [dad]: first formant increases, but F2 and F3 slight fall
- [gag]: F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials

From Ladefoged “A Course in Phonetics”

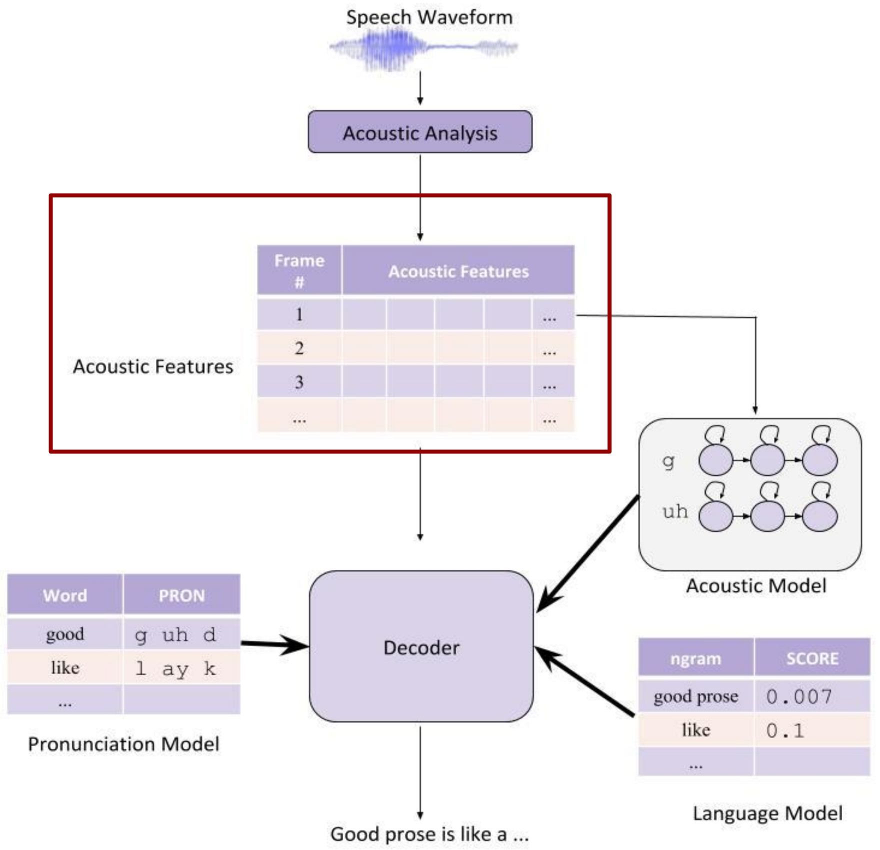
Dialect Issues

- Speech varies from dialect to dialect (examples are American vs. British English)
  - Syntactic ("I could" vs. "I could do")
  - Lexical ("elevator" vs. "lift")
  - Phonological
  - Phonetic

- Mismatch between training and testing dialects can cause a large increase in error rate



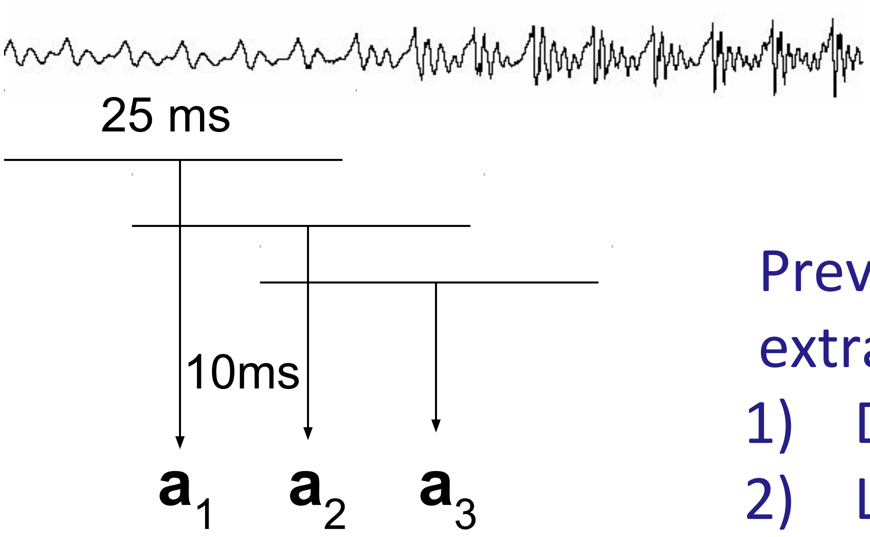
Acoustic Analysis



Slide by Preethi Jyothi

Frame Extraction

- A frame (25 ms wide) extracted every 10 ms



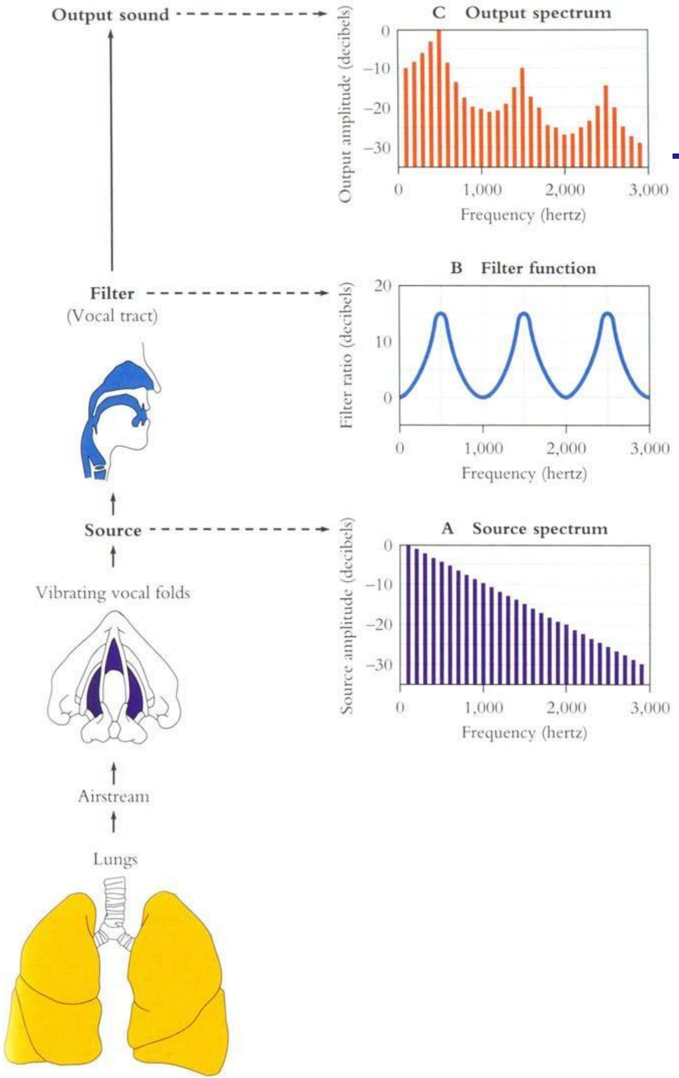
Preview of feature extraction for each frame:

1) DFT (Spectrum)
2) Log (Calibrate)
3) another DFT (!!??)

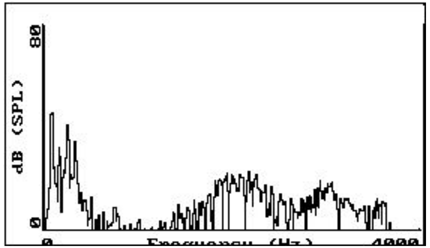
Figure: Simon Arnfield

Why these Peaks?

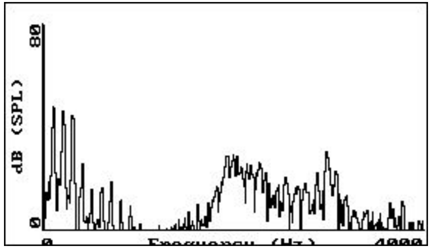
- Articulation process:
  - The vocal cord vibrations create harmonics
  - The mouth is an amplifier
  - Depending on shape of mouth, some harmonics are amplified more than others



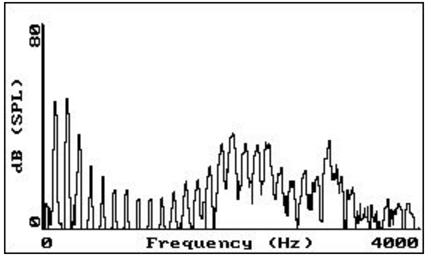
Vowel [i] at increasing pitches



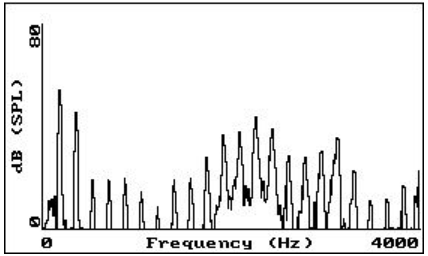
F#2



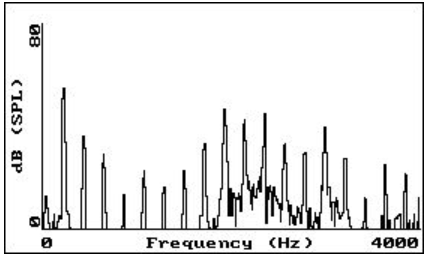
A2



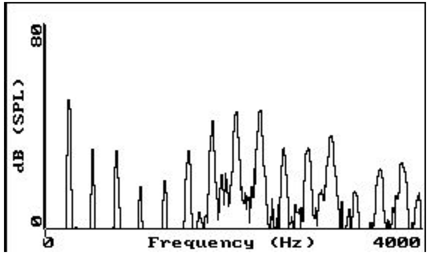
C3



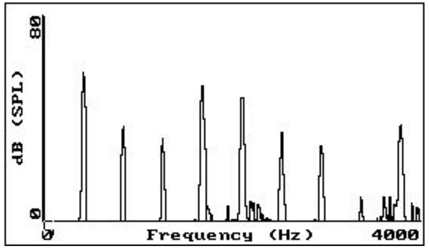
F#3



A3



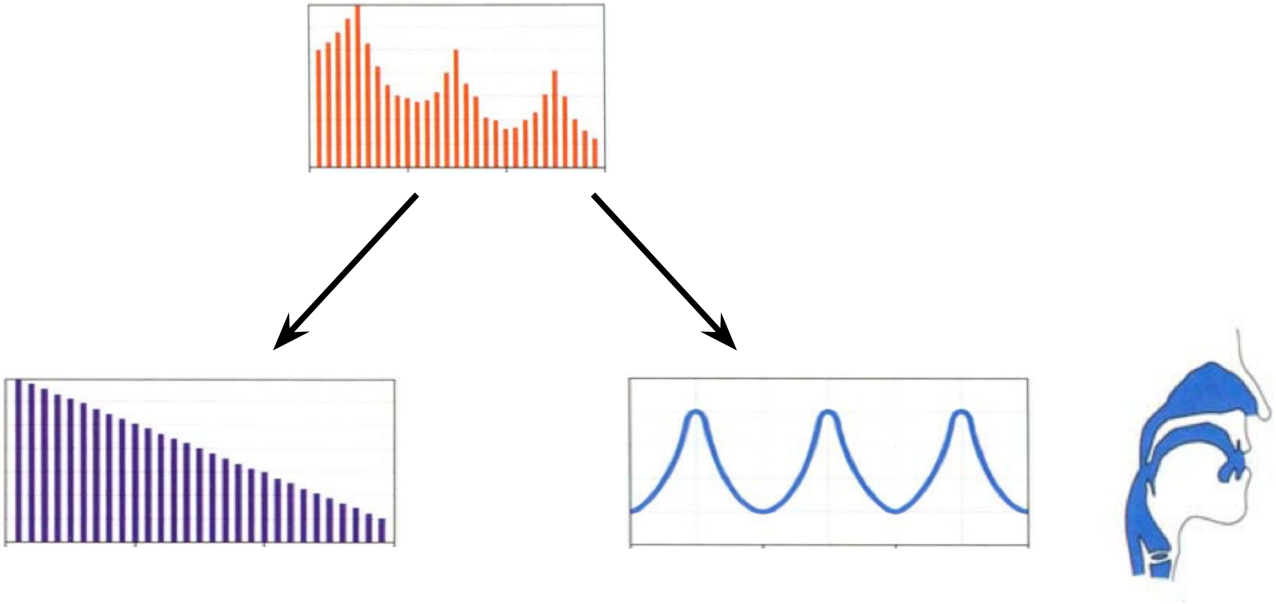
C4



A4

Figures from Ratree Wayland

Deconvolution / The Cepstrum

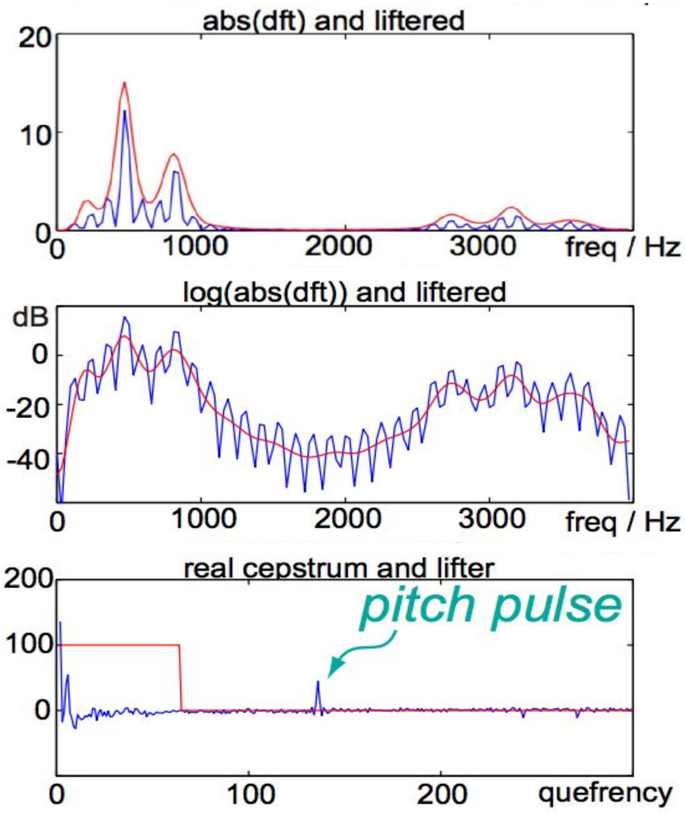


Deconvolution / The Cepstrum

$$s = e \circ f$$

$$\log(s) = \log(e) + \log(f)$$

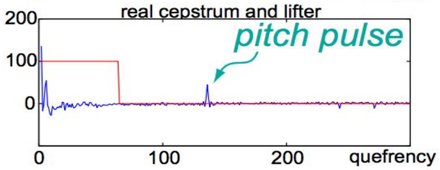
$$\mathrm{IDFT}(\log(s))$$

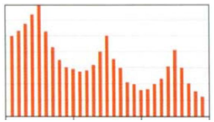


Graphs from Dan Ellis

Final Feature Vector

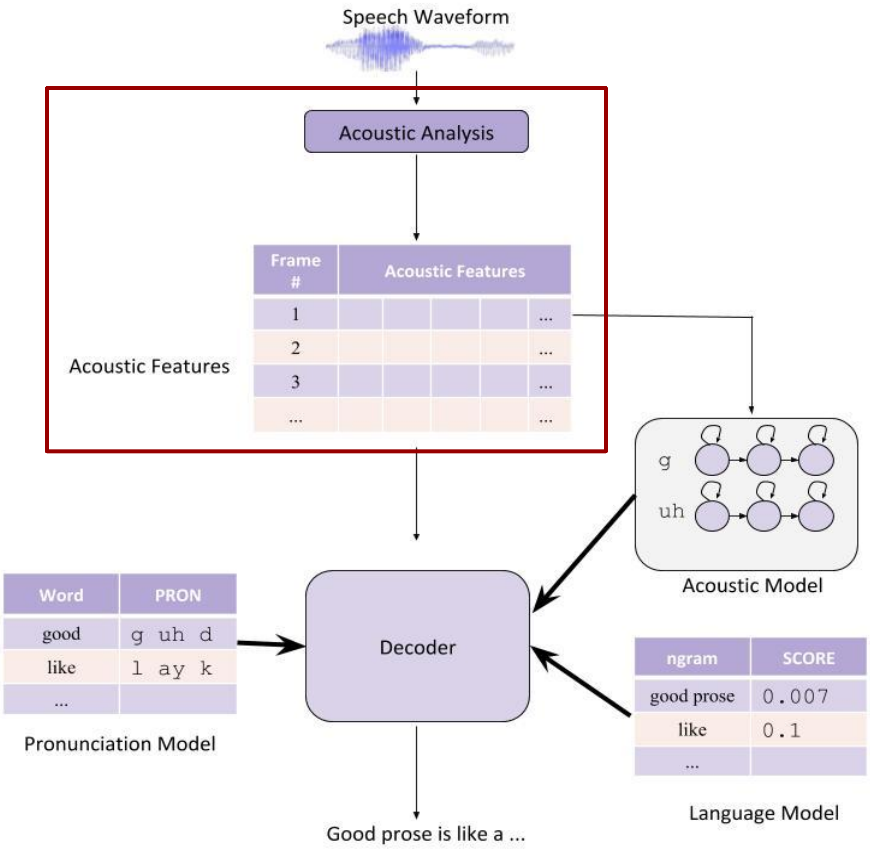
- 39 (real) features per 25 ms frame:
  - 12 MFCC features
  - 12 delta MFCC features
  - 12 delta-delta MFCC features
  - 1 (log) frame energy
  - 1 delta (log) frame energy
  - 1 delta-delta (log frame energy)





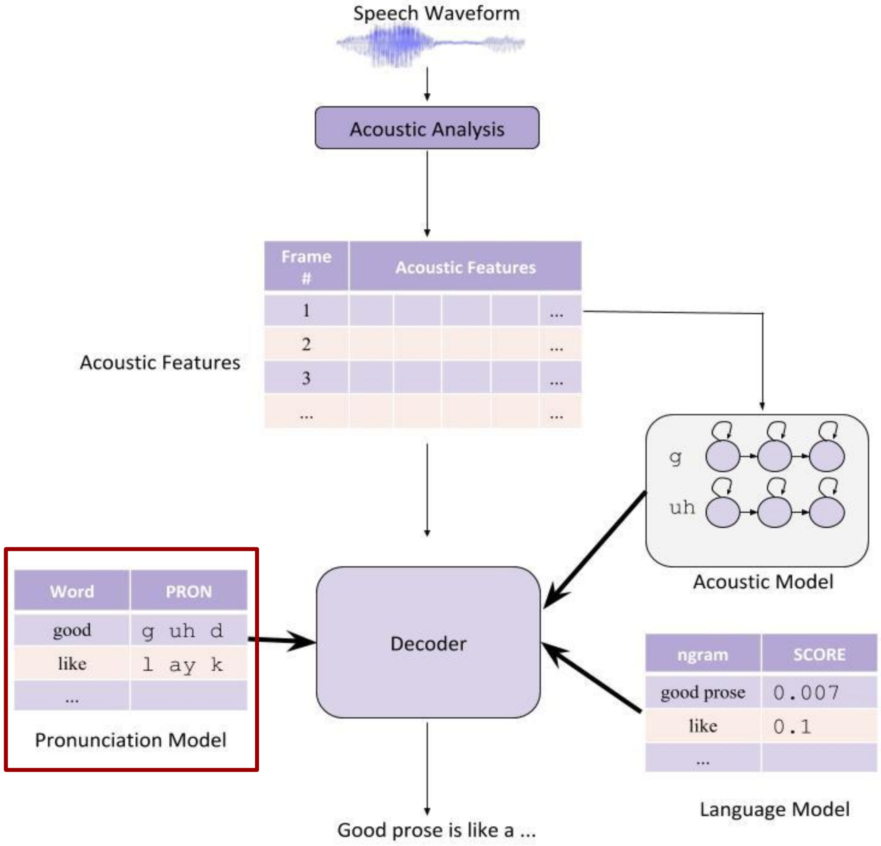
- So each frame is represented by a 39D vector

Acoustic Analysis



Slide by Preethi Jyothi

Phonetic Analysis

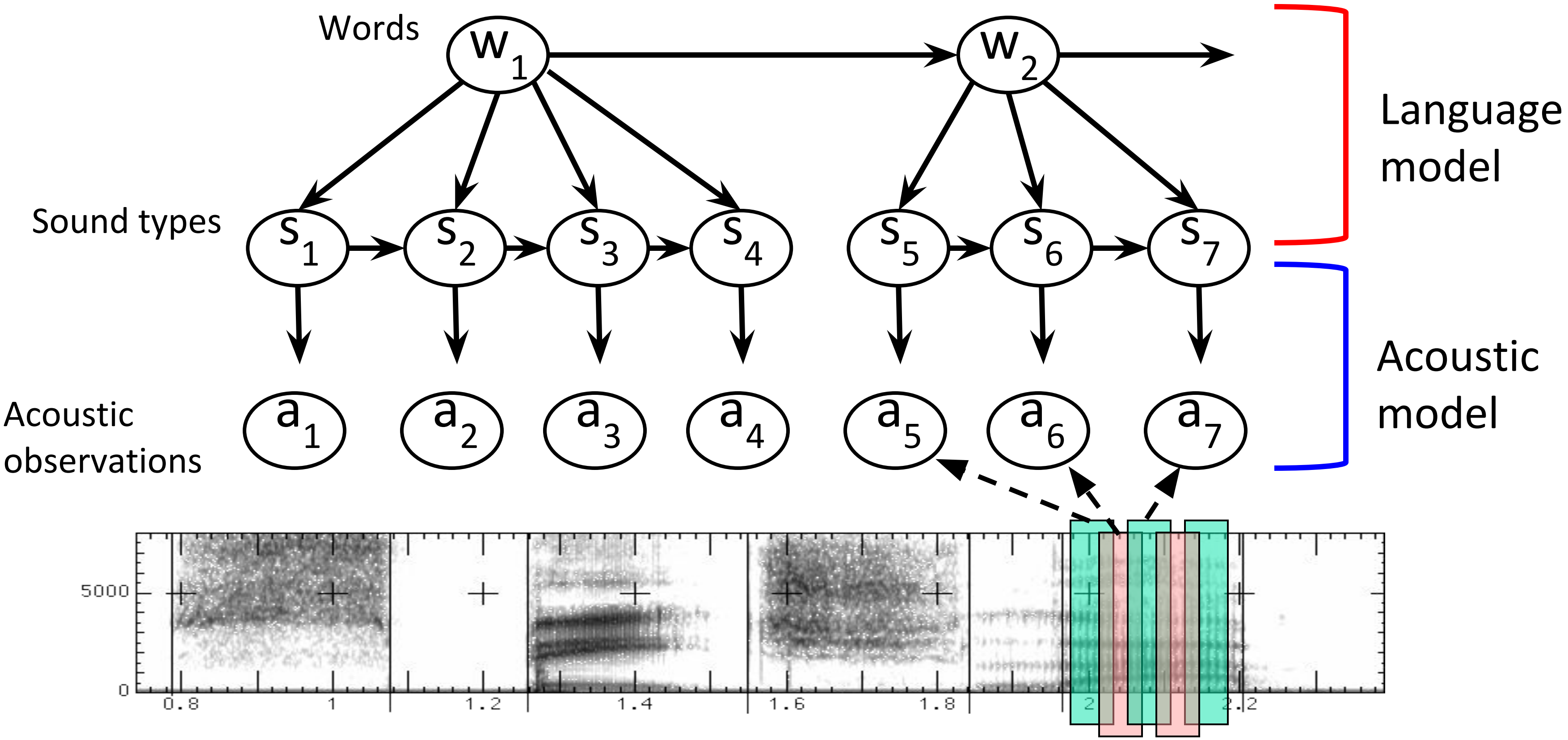


Slide by Preethi Jyothi

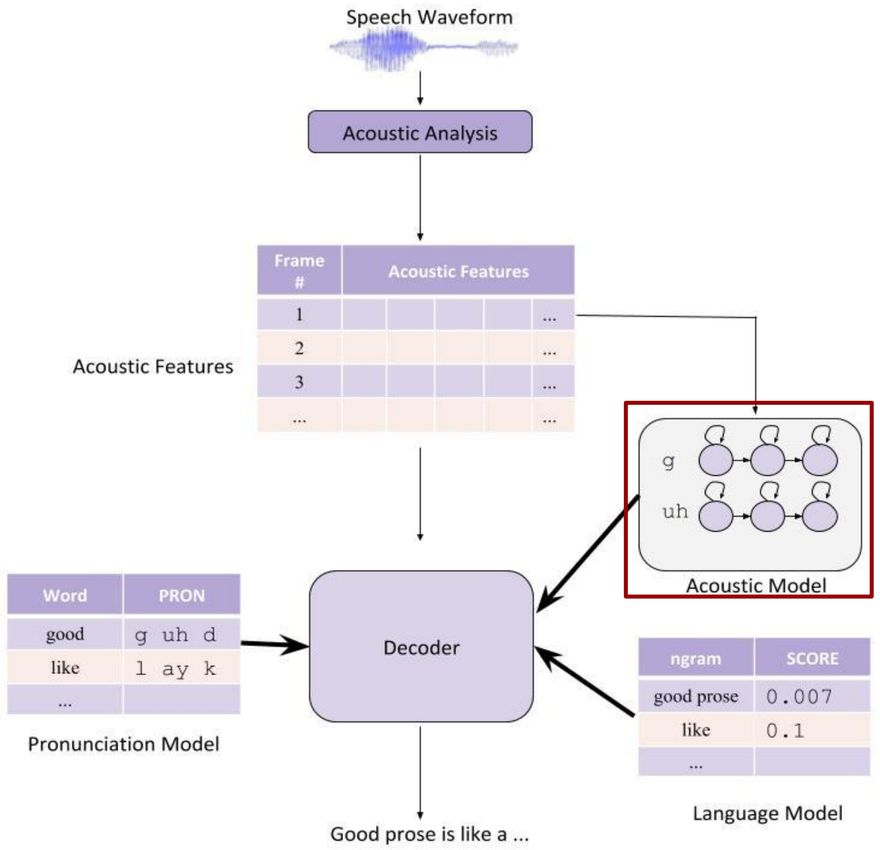
CMU Pronunciation Dict

ABBREVIATE	AH B R IY V IY EY T
ABBREVIATED	AH B R IY V IY EY T AH D
ABBREVIATED(2)	AH B R IY V IY EY T IH D
ABBREVIATES	AH B R IY V IY EY T S
ABBREVIATING	AH B R IY V IY EY T IH NG
ABBREVIATION	AH B R IY V IY EY SH AH N
ABBREVIATIONS	AH B R IY V IY EY SH AH N Z
ABBRUZZESE	AA B R UW T S EY Z IY
ABBS	AE B Z
ABBY	AE B IY
ABCO	AE B K OW
ABCOTEK	AE B K OW T EH K
ABDALLA	AE B D AE L AH
ABDALLAH	AE B D AE L AH

Speech Model



Acoustic Modeling



Slide by Preethi Jyothi

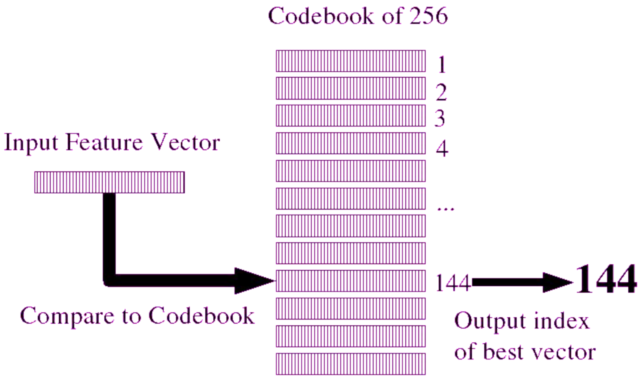
Vector Quantization

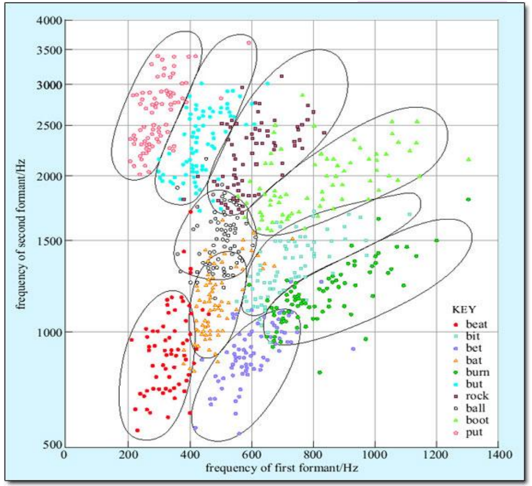
- Idea: discretization
  - Map MFCC vectors onto discrete symbols
  - Compute probabilities just by counting

- This is called vector quantization or VQ

- Not used for ASR any more

- But: useful to consider as a starting point





Next class: HMMs for Continuous Observations

- Feature vectors are real-valued
- Solution 1: discretization
- Solution 2: continuous emissions
  - Gaussians
  - Multivariate Gaussians
  - Mixtures of multivariate Gaussians