Speech Recognition and Synthesis

Dan Klein
UC Berkeley

Language Models

Noisy Channel Model: ASR

- We want to predict a sentence given acoustics:

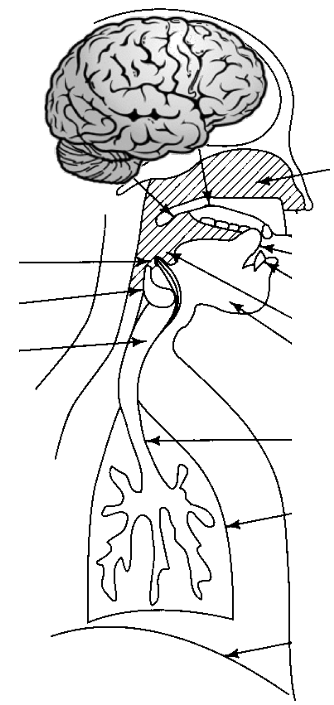
$$w^* = \arg\max_w P(w|a)$$

- The noisy-channel approach:

$$\begin{aligned} w^* &= \arg\max_w P(w|a) \\ &= \arg\max_w P(a|w)P(w)/P(a) \\ &\propto \arg\max_w P(a|w)P(w) \end{aligned}$$

Acoustic model: score fit between sounds and words

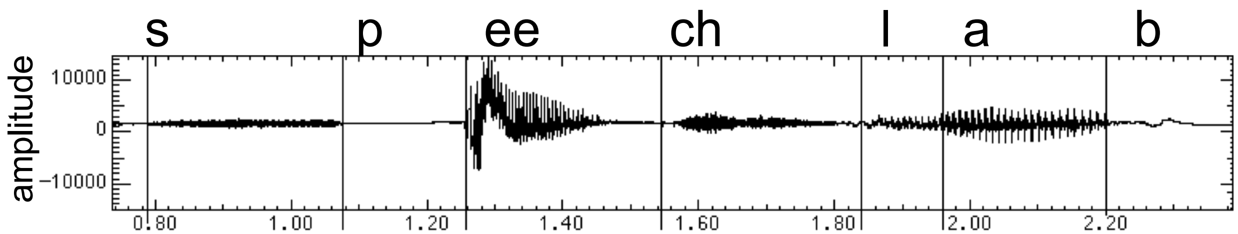
Language model: score plausibility of word sequences



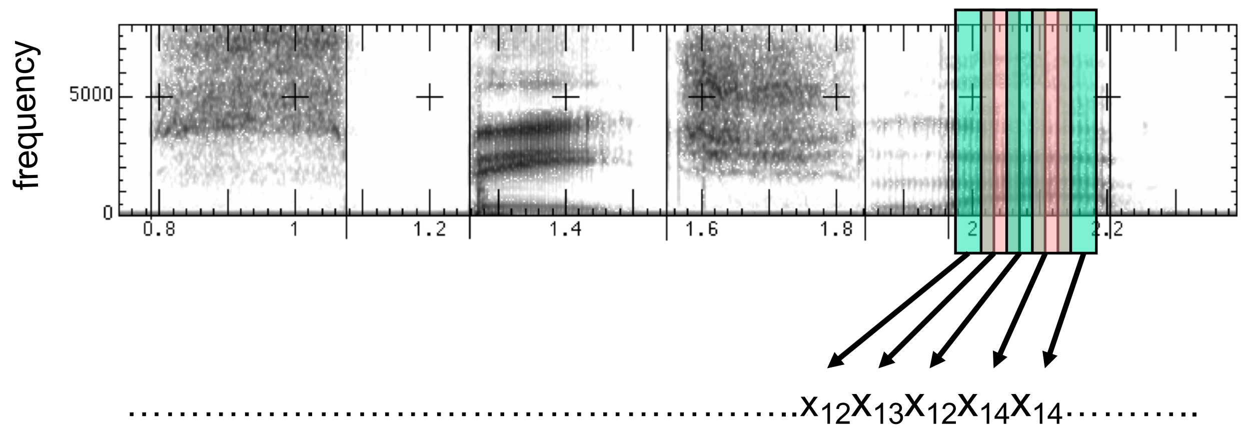
The Speech Signal

Speech in a Slide

- Frequency gives pitch; amplitude gives volume

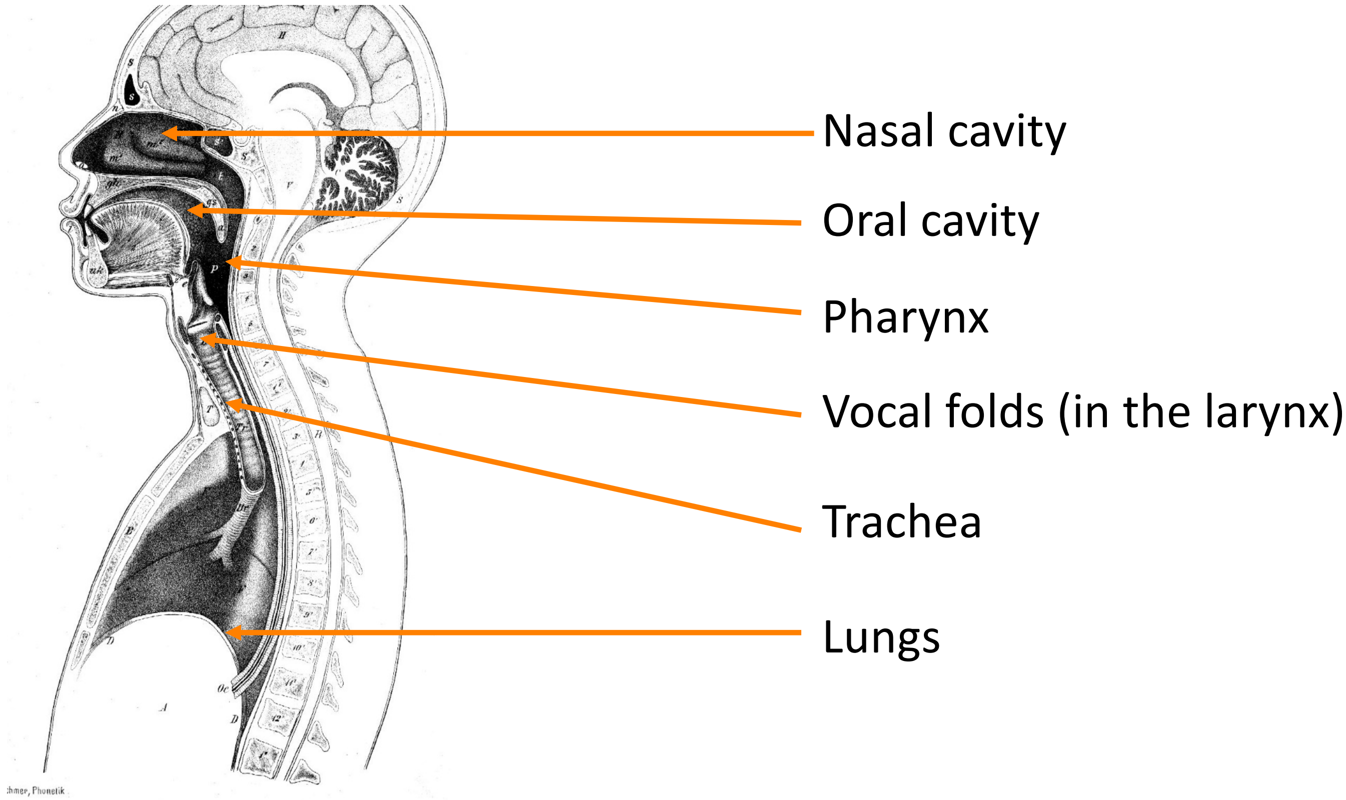


- Frequencies at each time slice processed into observation vectors



Articulation

Articulatory System



Sagittal section of the vocal tract (Techmer 1880)

Text from Ohala, Sept 2001, from Sharon Rose slide

Space of Phonemes

- Standard international phonetic alphabet (IPA) chart of consonants

	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	ʎ̝	ʟ̝				
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ		ɺ̢						

Articulation: Place

Places of Articulation

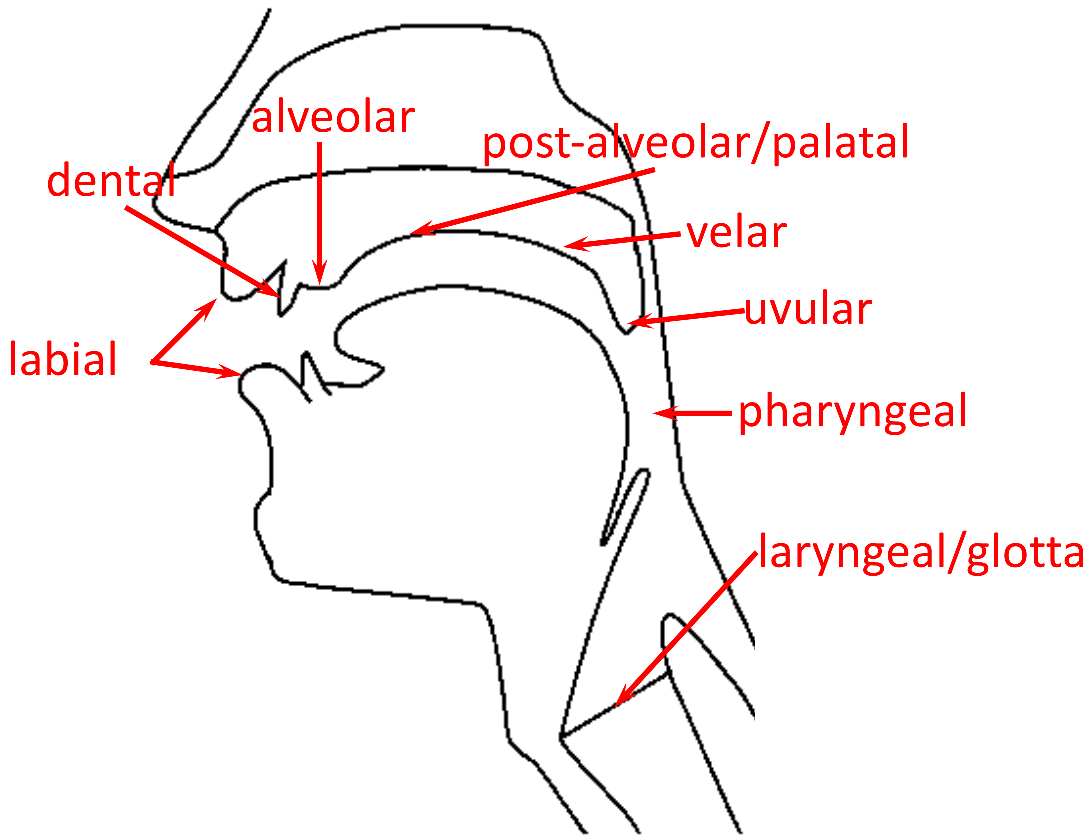
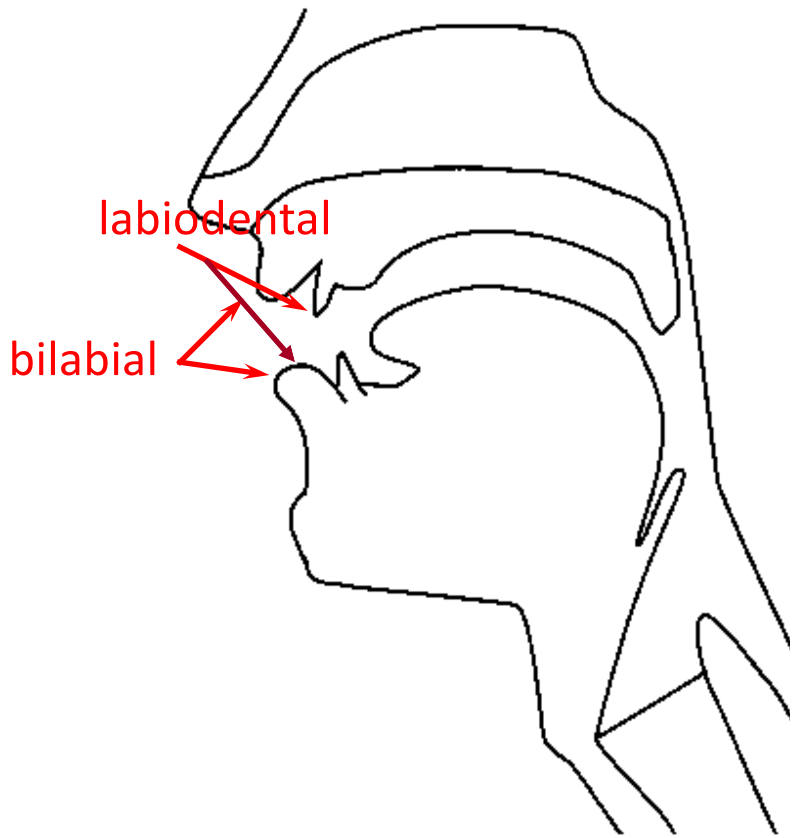


Figure thanks to Jennifer Venditti

Labial place



Bilabial:

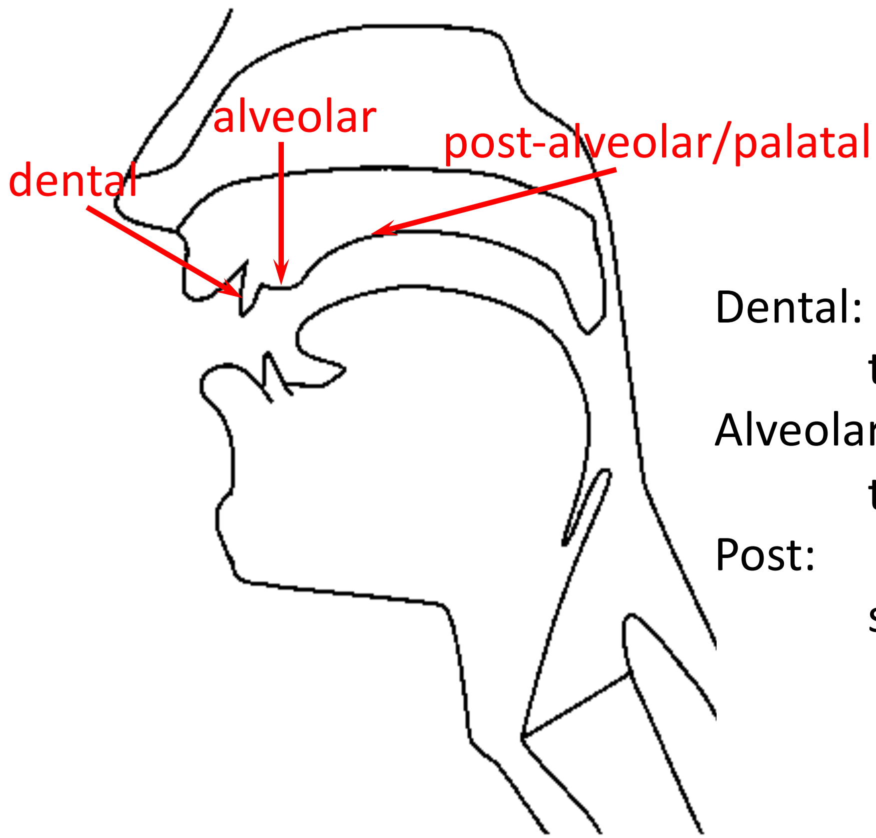
p, b, m

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

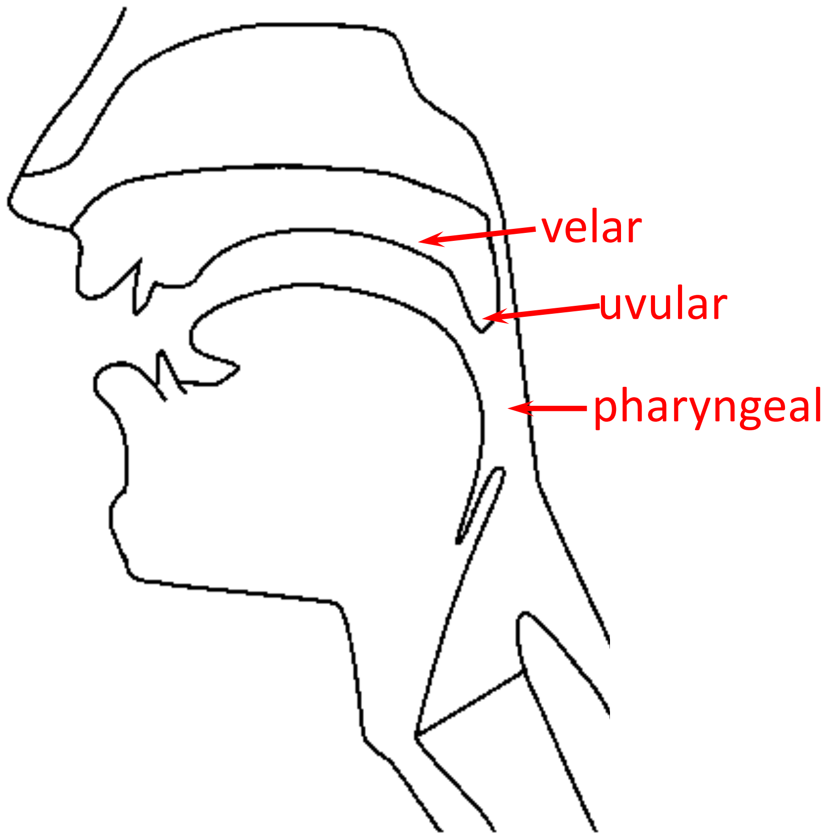


Figure thanks to Jennifer Venditti

Space of Phonemes

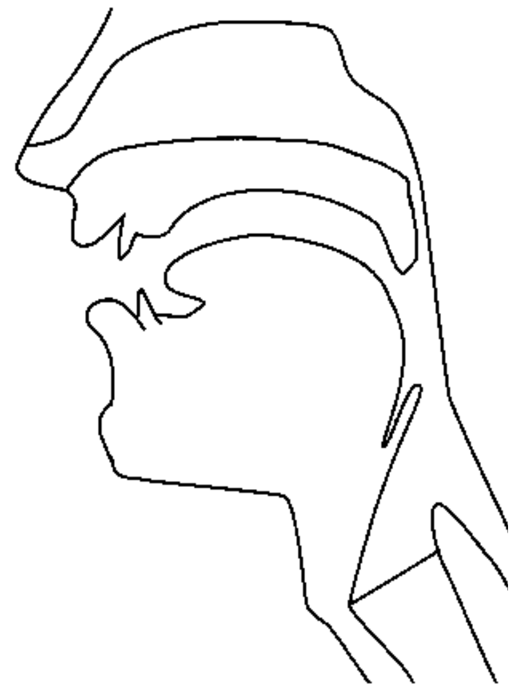
- Standard international phonetic alphabet (IPA) chart of consonants

	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						

Articulation: 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)



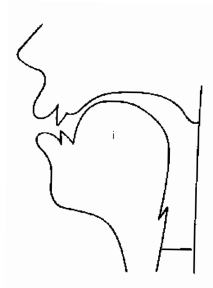
Space of Phonemes

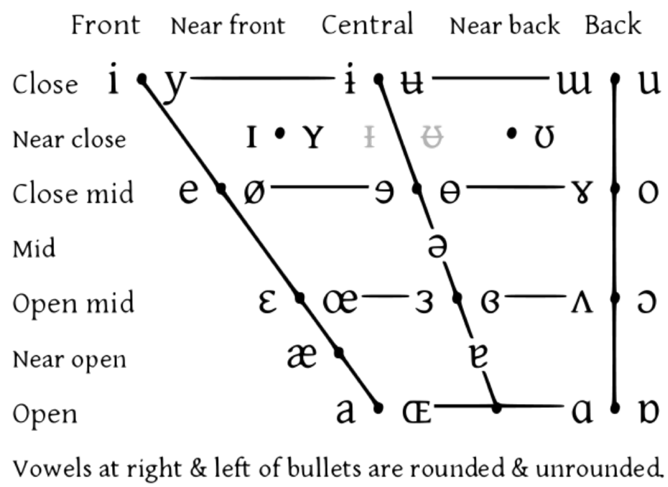
- Standard international phonetic alphabet (IPA) chart of consonants

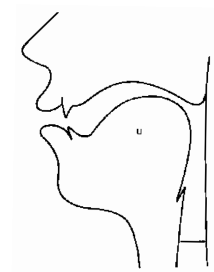
	LABIAL		CORONAL				DORSAL			RADICAL		LARYNGEAL
	Bilabial	Labio-dental	Dental	Alveolar	Palato-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b	ȹ ȸ		t d		ʈ ɖ	c ɟ	k g	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant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

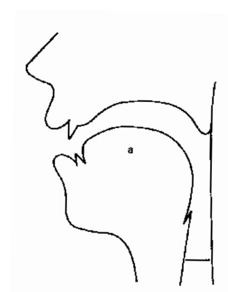
Articulation: Vowels

Vowel Space



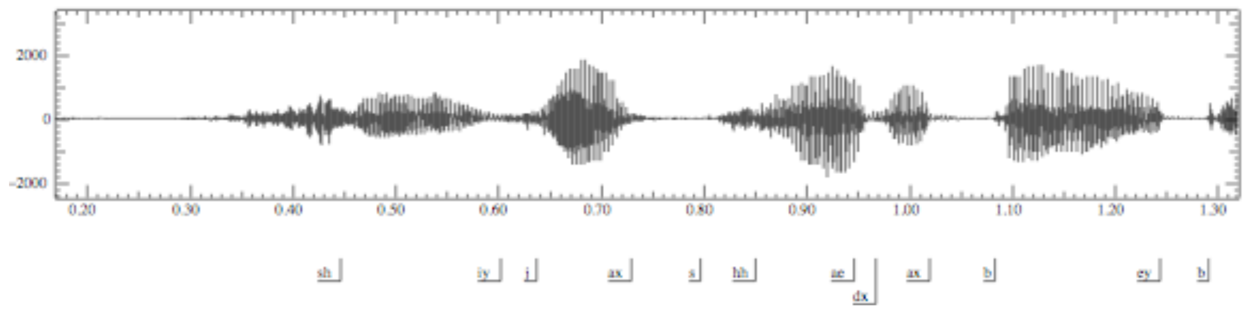

Front
Near front
Central
Near back
Back
Close i • y — ɨ • ʉ — ɯ • u
Near close ɪ • ʏ ɪ̈ ʊ̈ • ʊ
Close mid e • ø — ɘ • ɵ — ɤ • o
Mid ə
Open mid ɛ • œ — ɜ • ɞ — ʌ • ɔ
Near open æ ɐ
Open a • ɶ — ɑ • ɒ
Vowels at right & left of bullets are rounded & unrounded.





Acoustics

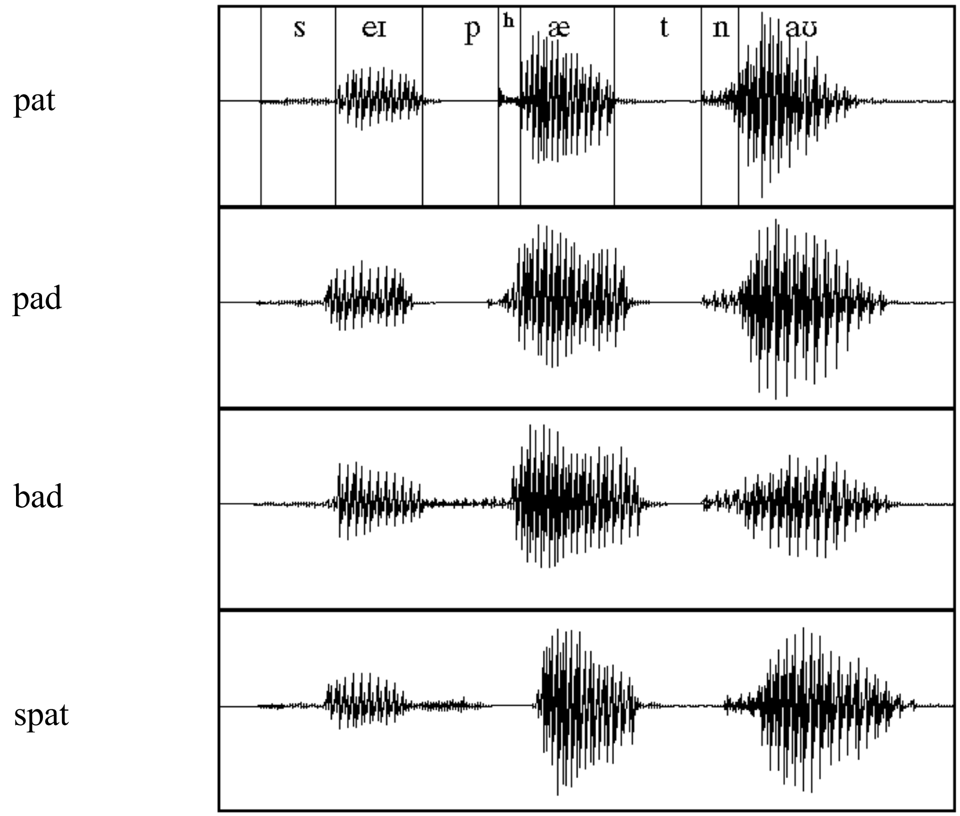
“She just had a baby”



What can we learn from a wavefile?

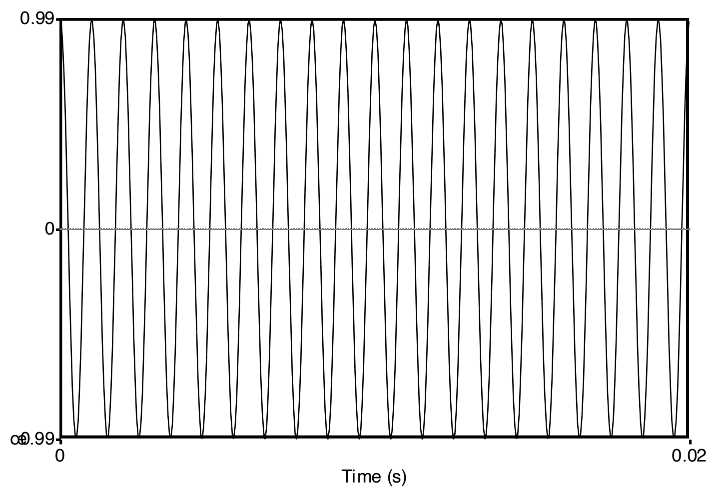
- No gaps between words (!)
- Vowels are voiced, long, loud
- Length in time = length in space in waveform picture
- Voicing: regular peaks in amplitude
- When stops closed: no peaks, silence
- Peaks = voicing: .46 to .58 (vowel [iy], from second .65 to .74 (vowel [ax]) and so on
- Silence of stop closure (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

Time-Domain Information



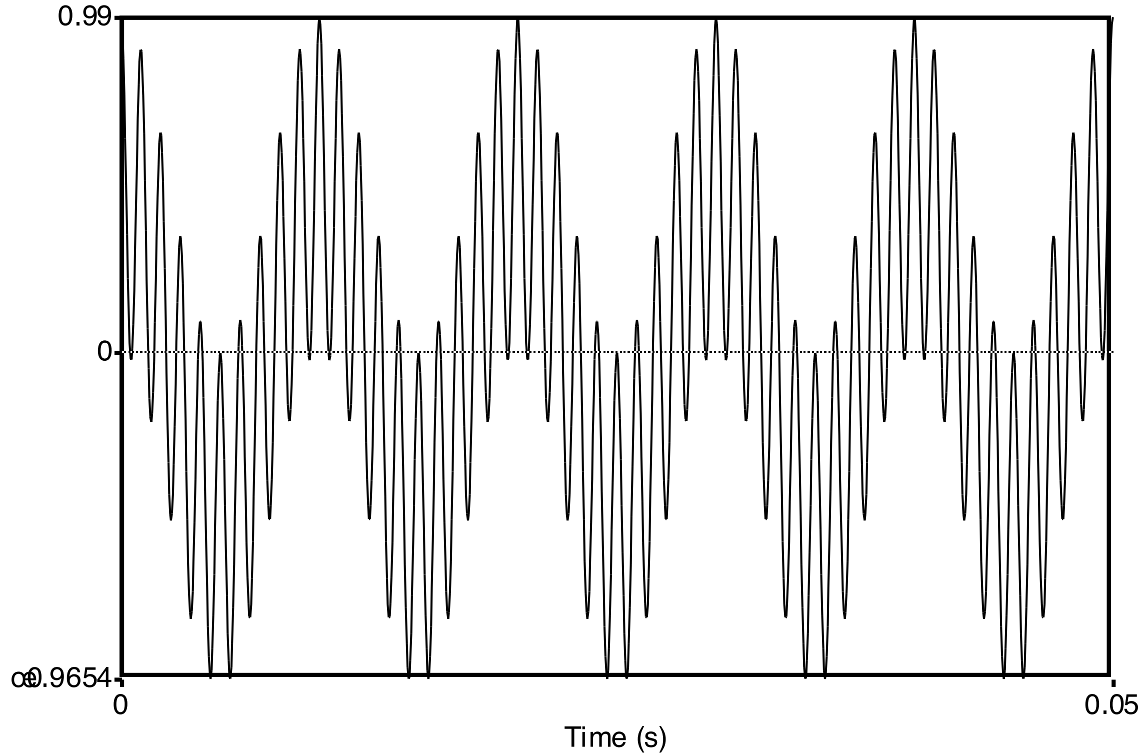
Example from Ladefoged

Simple Periodic Waves of Sound



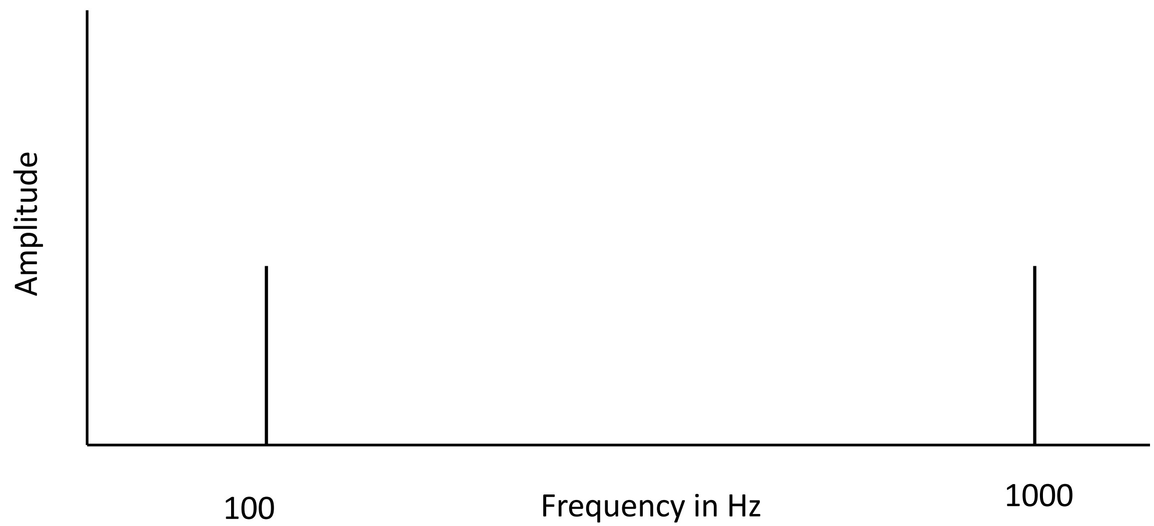
- Y axis: Amplitude = amount of air pressure at that point in time
  - Zero is normal air pressure, negative is rarefaction
- X axis: Time
- Frequency = number of cycles per second
- 20 cycles in .02 seconds = 1000 cycles/second = 1000 Hz

Complex Waves: 100Hz+1000Hz

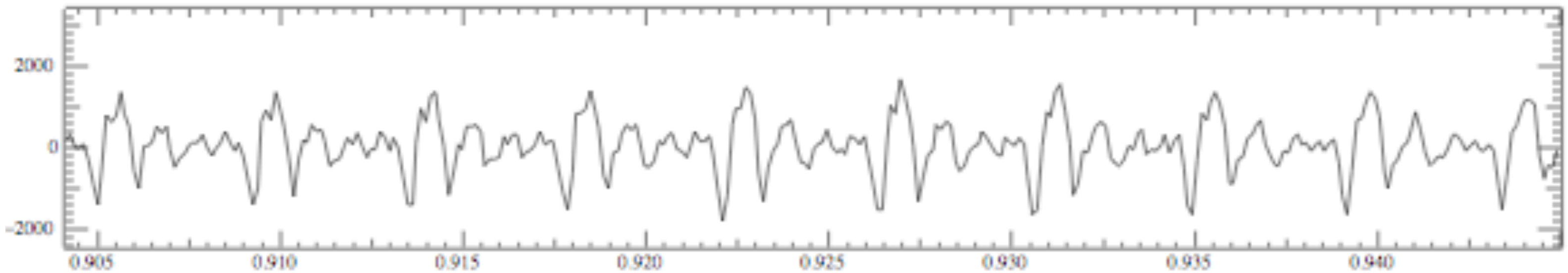


Spectrum

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

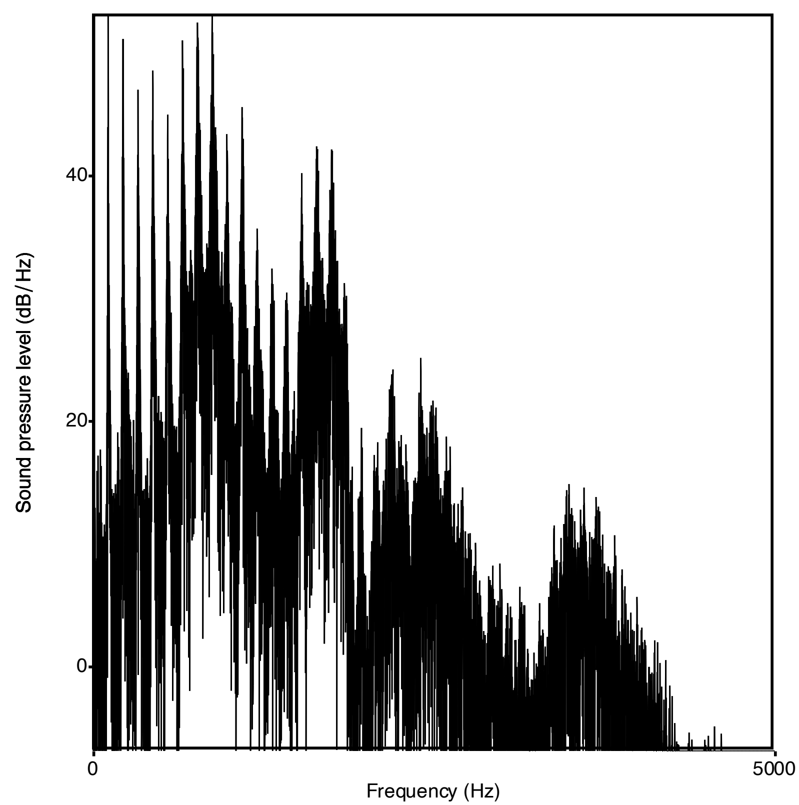


Part of [ae] waveform from “had”



- Note complex wave repeating nine times in figure
- Smaller waves which repeat 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

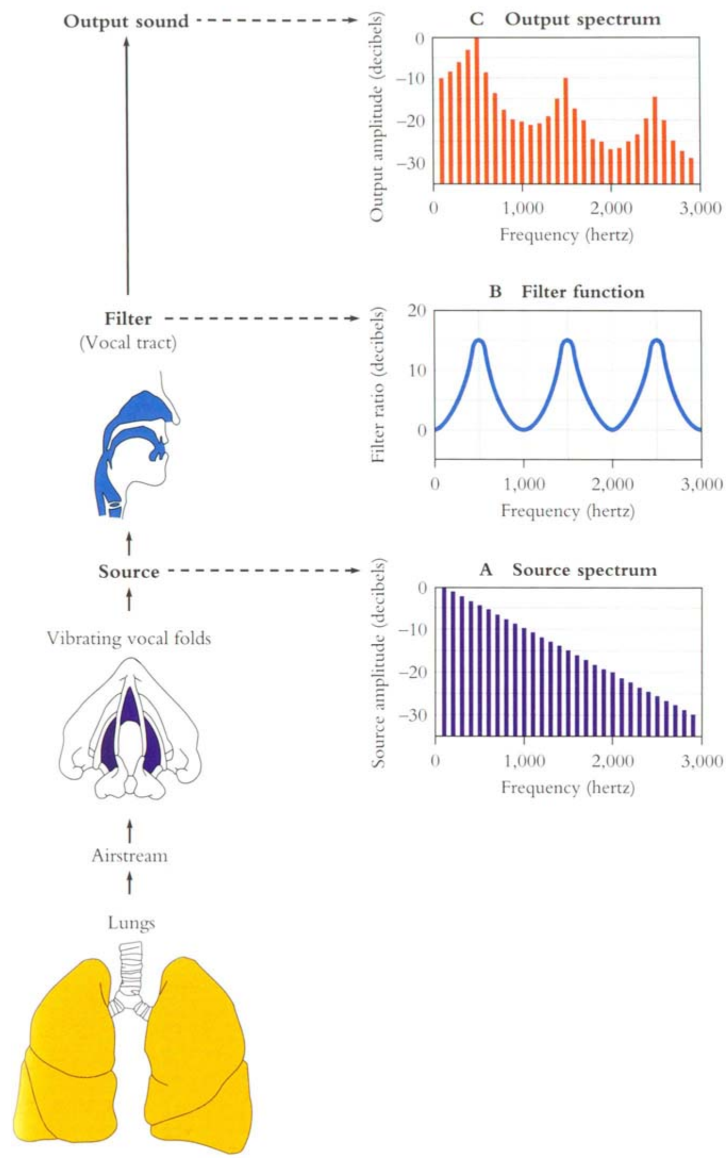
Spectrum of an Actual Soundwave



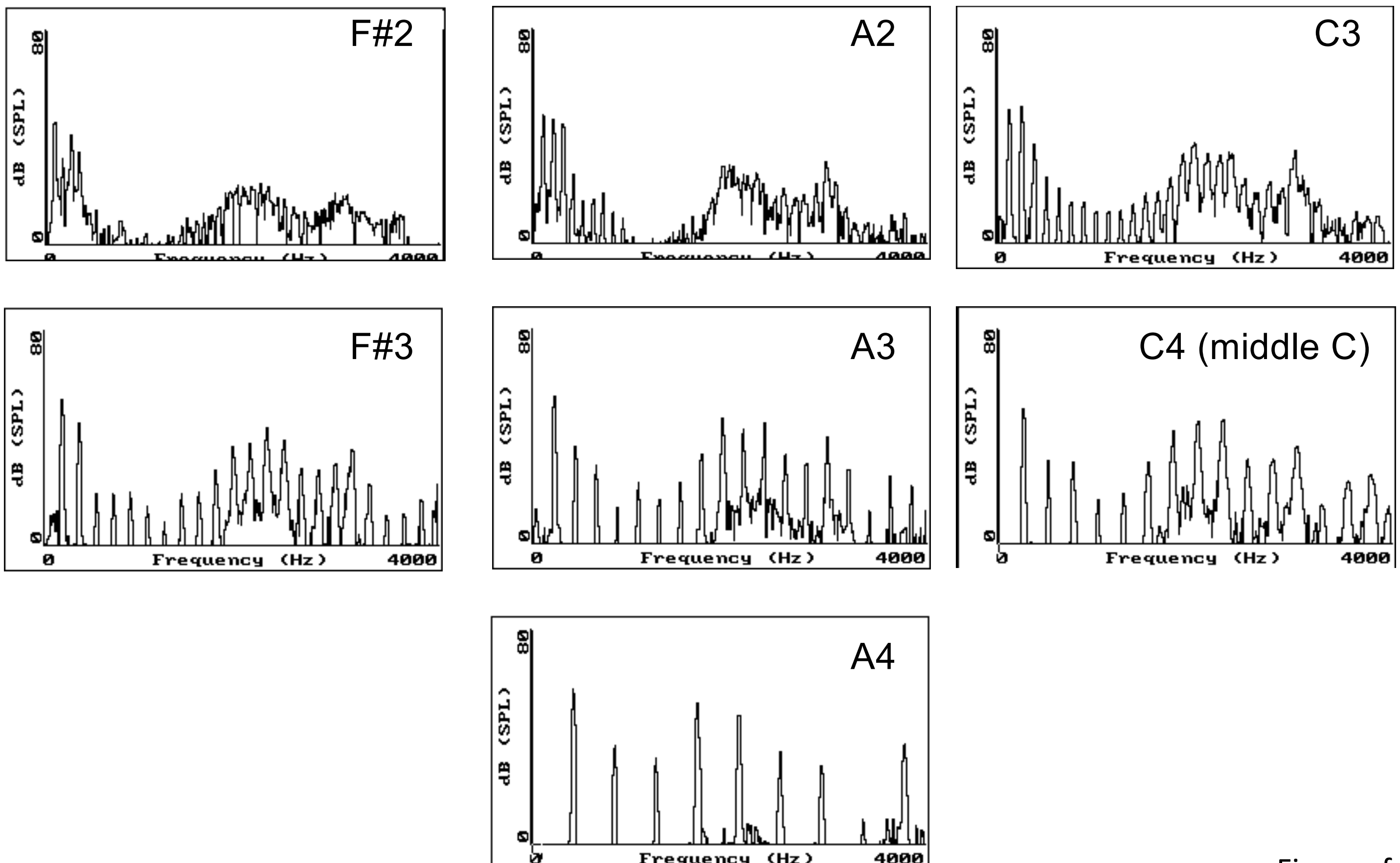
Source / Channel

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



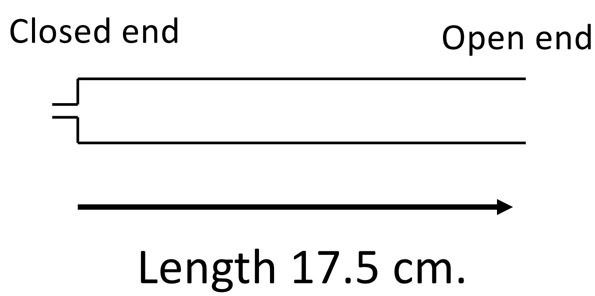
Vowel [i] at increasing pitches



Figures from Ratree Wayland

Resonances of the Vocal Tract

- The human vocal tract as an open tube:



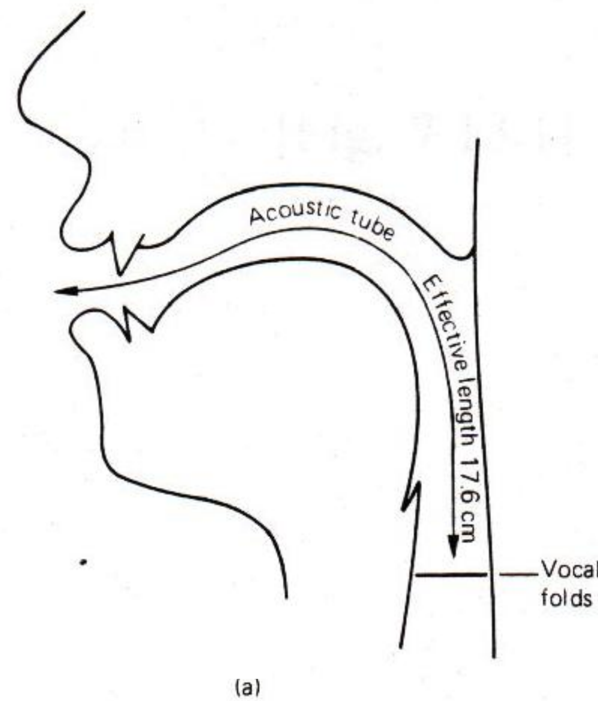
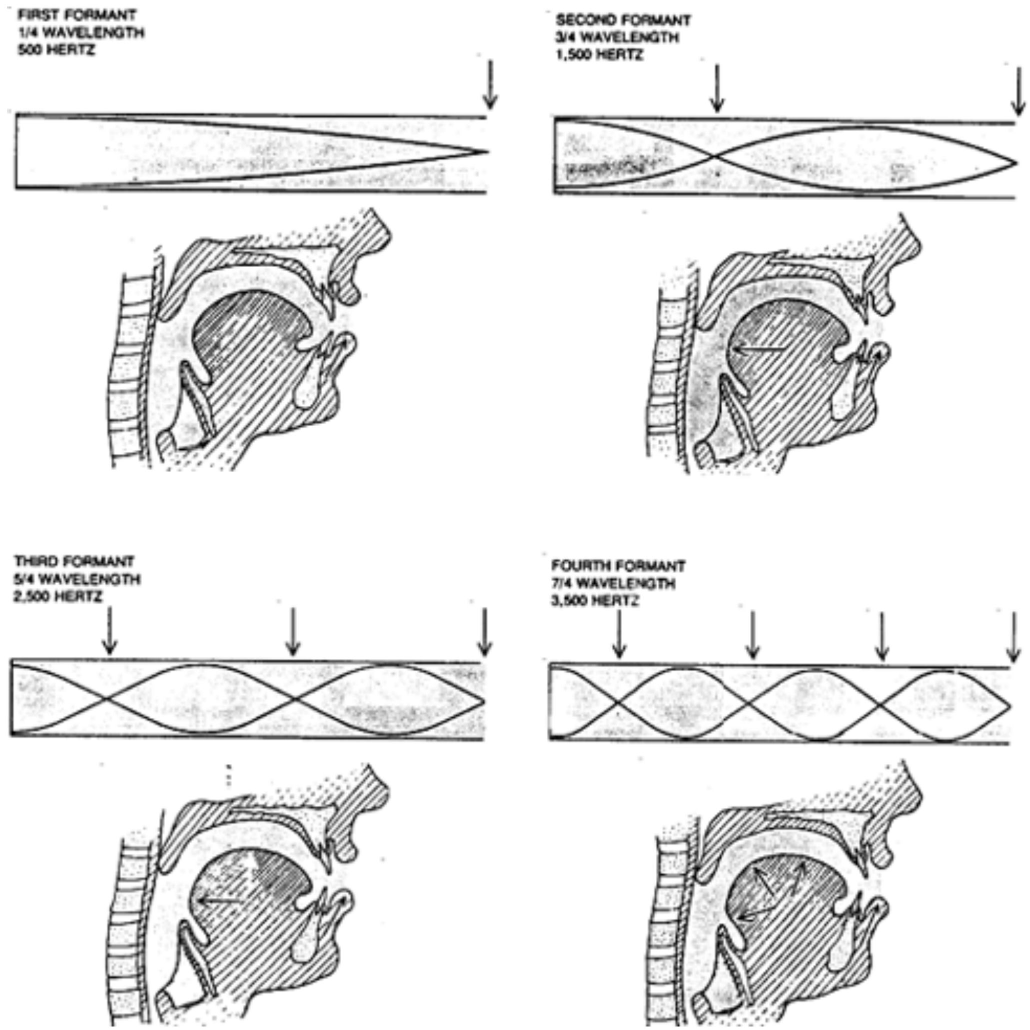


Figure from W. Barry

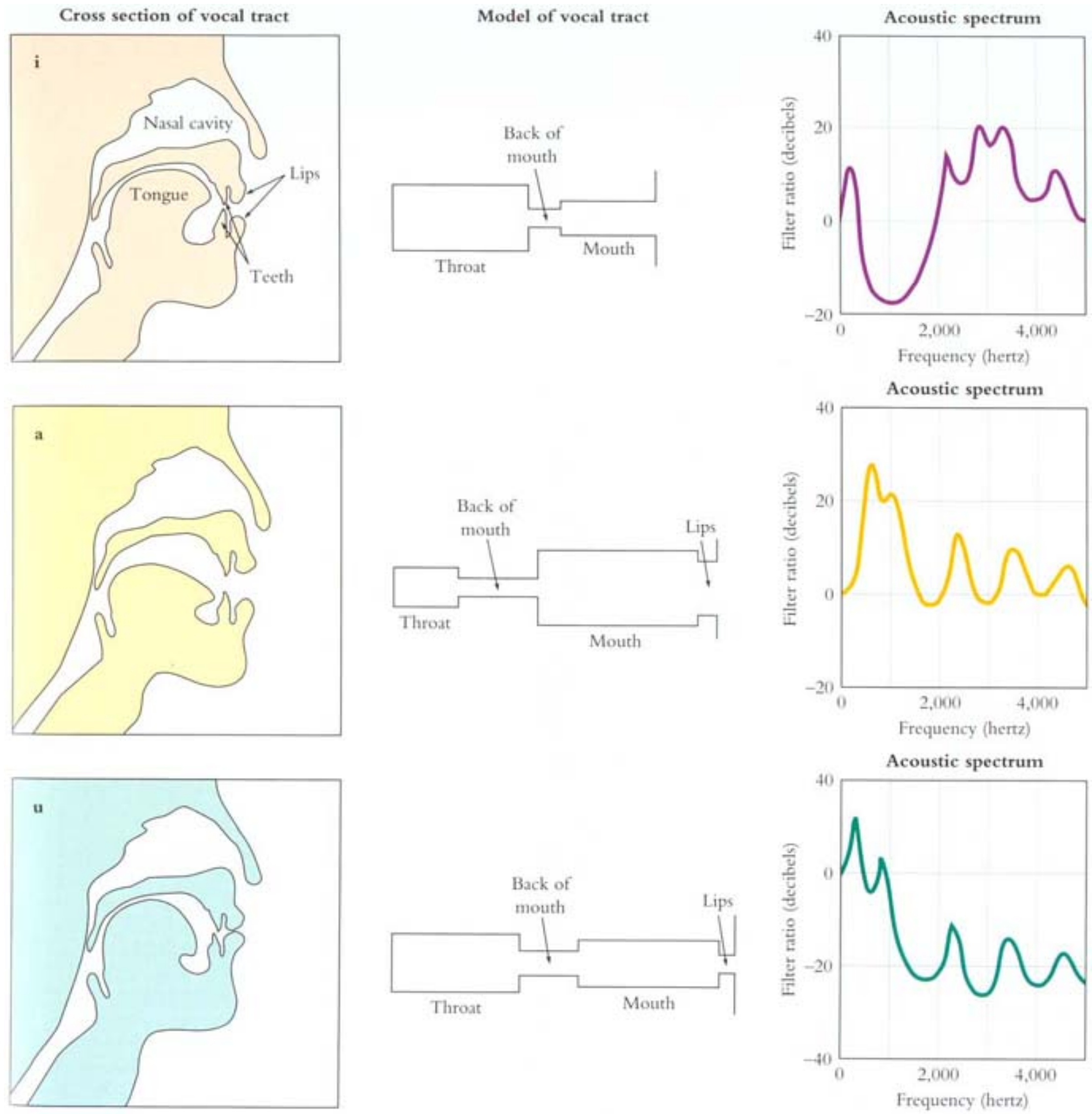
- Air in a tube of a given length will tend to vibrate at resonance frequency of tube.
- Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end.



From Sundberg

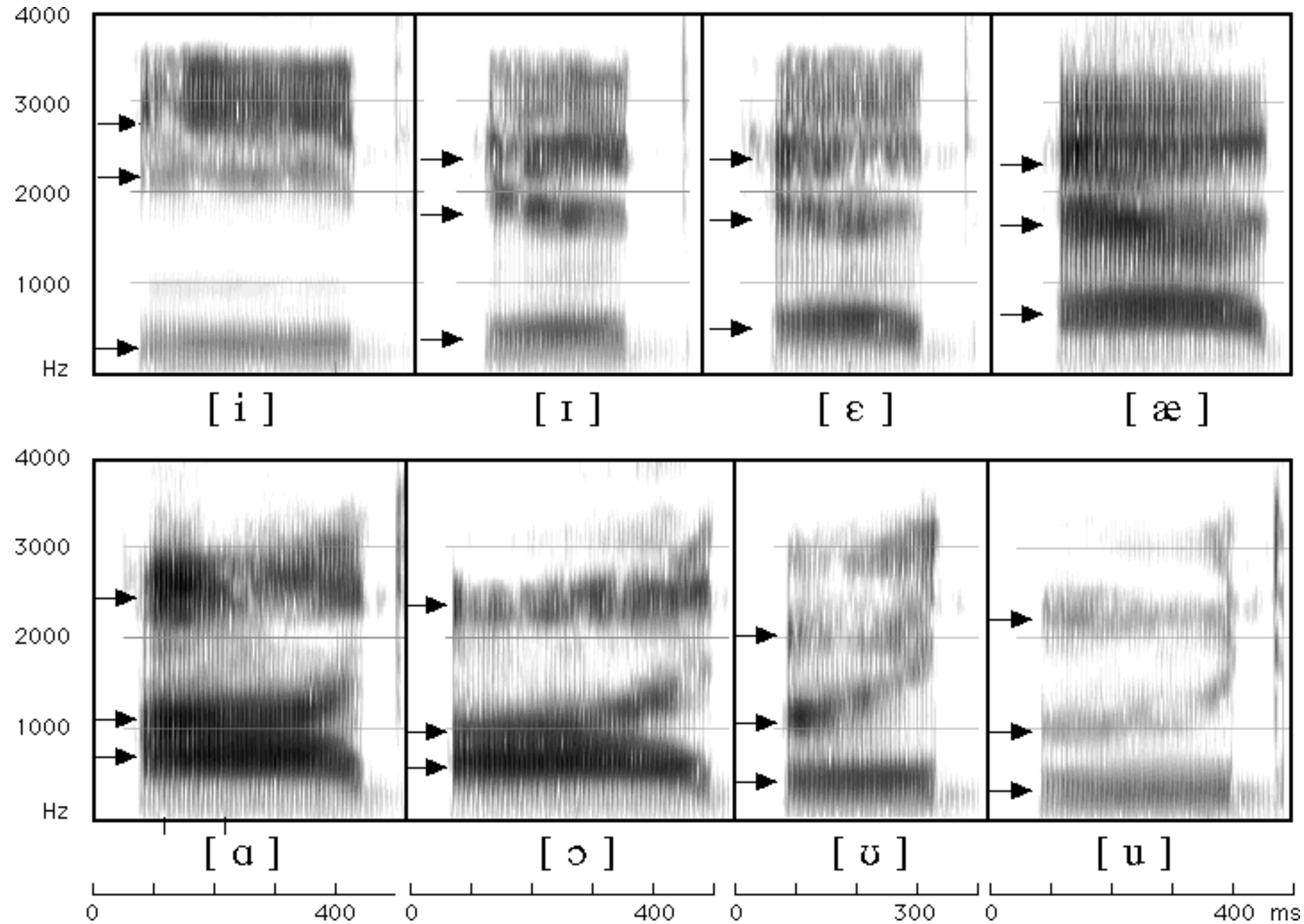
Computing the 3 Formants of Schwa

- Let the length of the tube be L
  - $F_1 = c/\lambda_1 = c/(4L) = 35{,}000/4*17.5 = 500\text{Hz}$
  - $F_2 = c/\lambda_2 = c/(4/3L) = 3c/4L = 3*35{,}000/4*17.5 = 1500\text{Hz}$
  - $F_3 = c/\lambda_3 = c/(4/5L) = 5c/4L = 5*35{,}000/4*17.5 = 2500\text{Hz}$
- So we expect a neutral vowel to have 3 resonances at 500, 1500, and 2500 Hz
- These vowel resonances are called formants

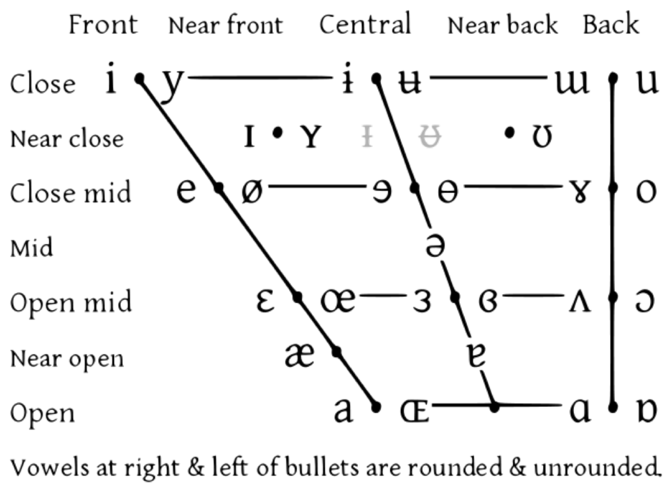


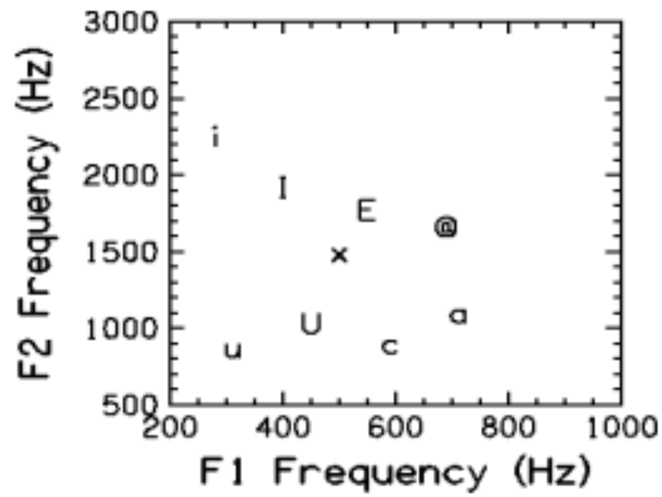
From
Mark
Liberman

Seeing Formants: the Spectrogram



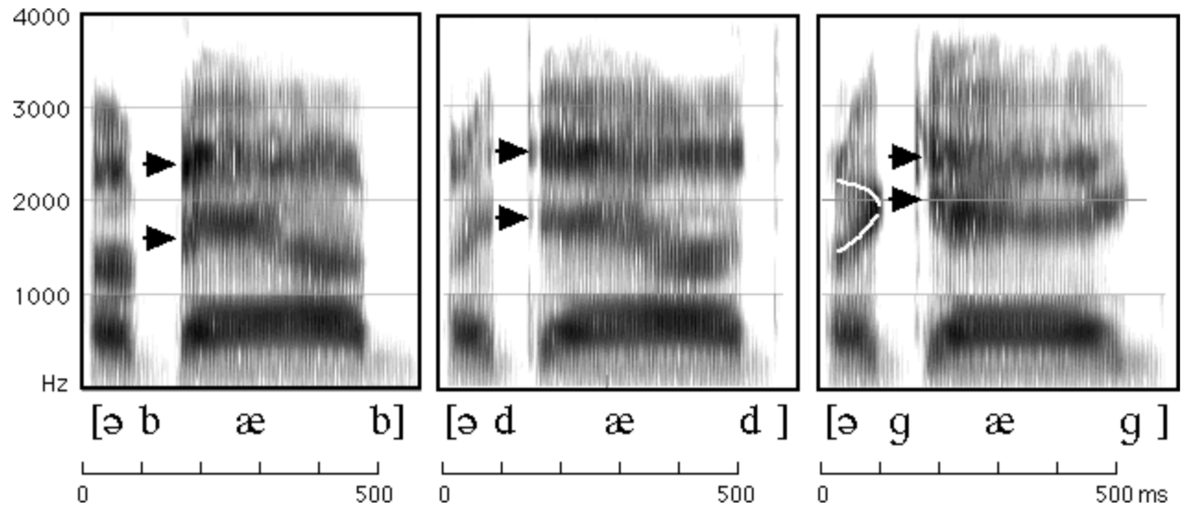
Vowel Space


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 ᵻ ᵿ 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.


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

Spectrograms

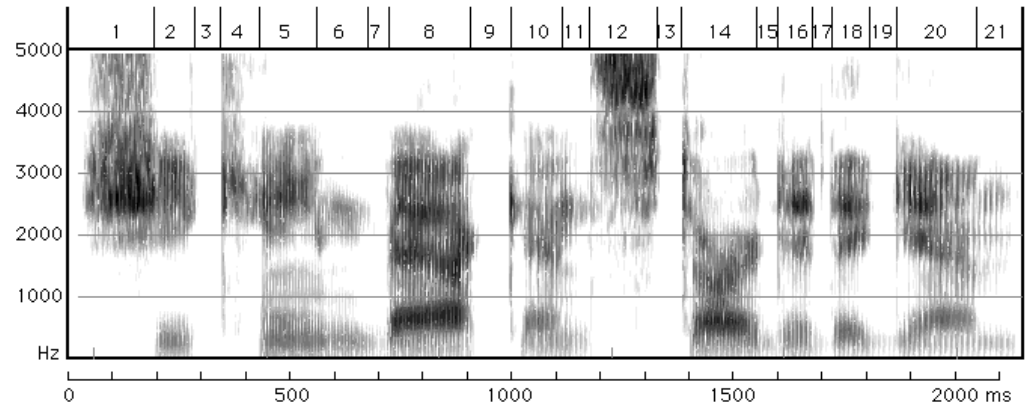
How to Read Spectrograms



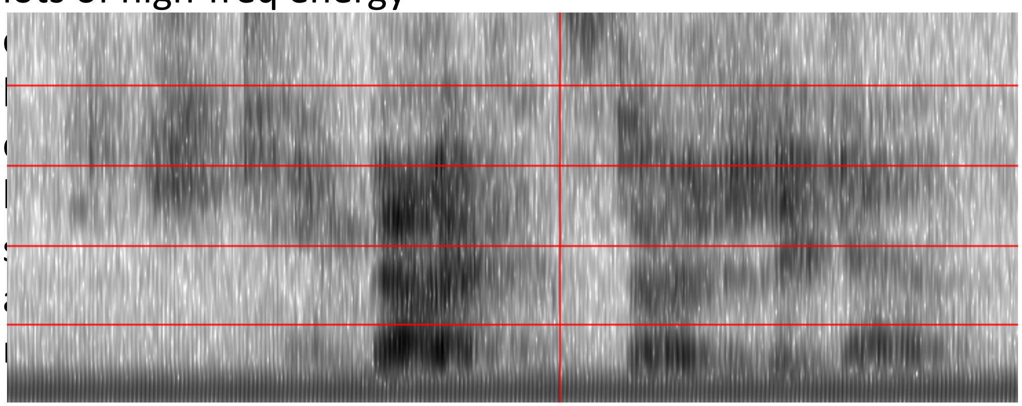
- [bab]: closure of lips lowers all formants: so rapid increase in all formants at beginning of "bab"
- [dad]: first formant increases, but F2 and F3 slight fall
- [gag]: F2 and F3 come together: this is a characteristic of velars. Formant transitions take longer in velars than in alveolars or labials

From Ladefoged "A Course in Phonetics"

“She came back and started again”



1. lots of high-freq energy
3.
4.
5.
6.
7.
8.
9.



From Ladefoged “A Course in Phonetics”

Speech Recognition

Speech Recognition Architecture

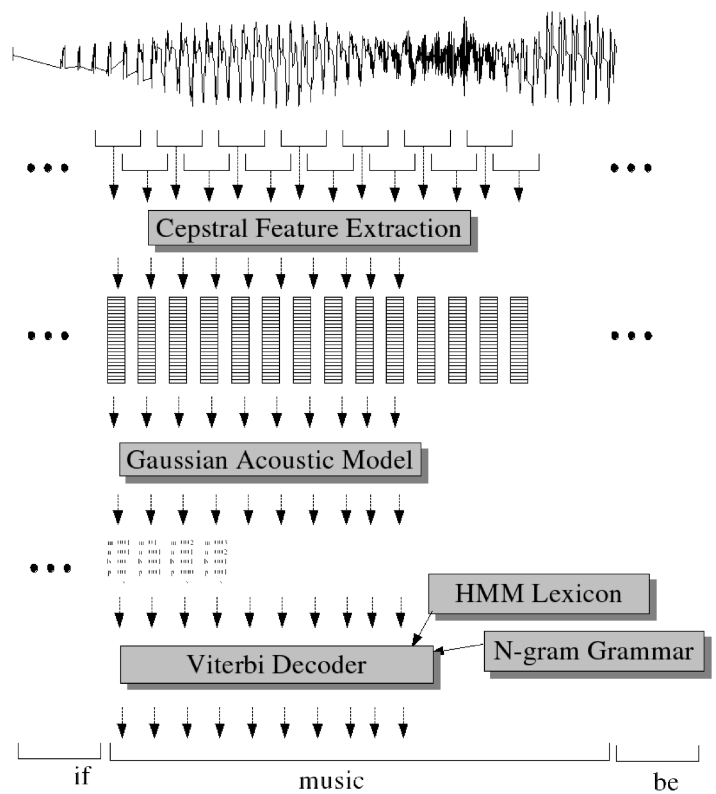


Figure: J & M

Feature Extraction

Digitizing Speech

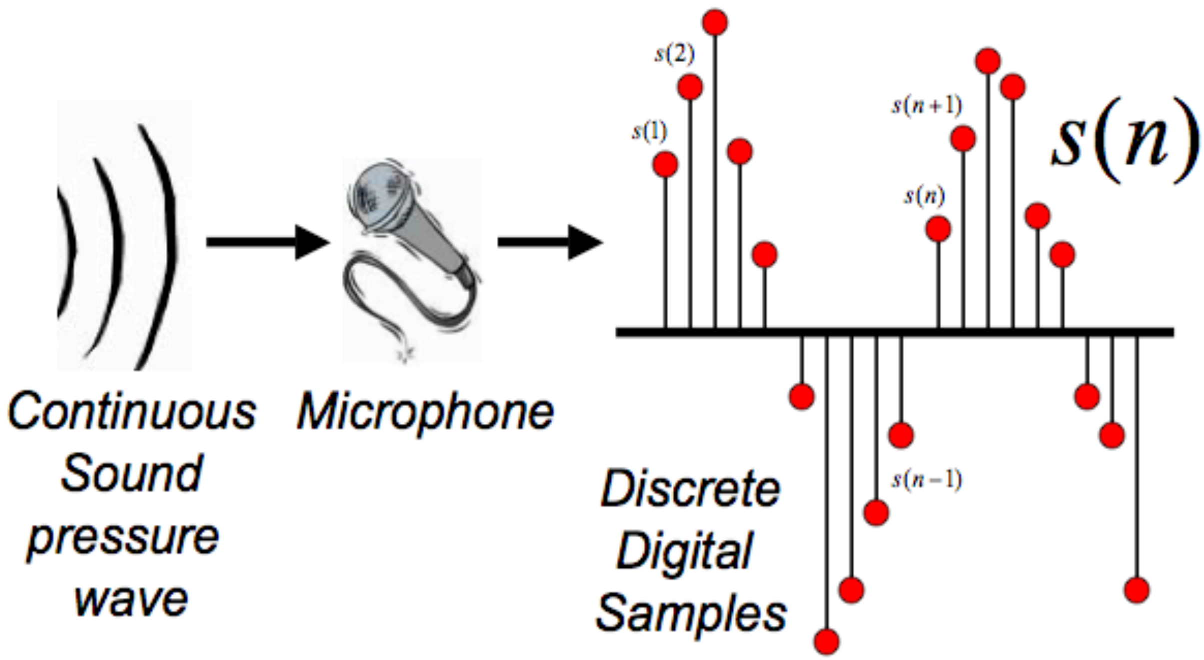


Figure: Bryan Pellom

Frame Extraction

- A 25 ms wide frame is extracted every 10 ms

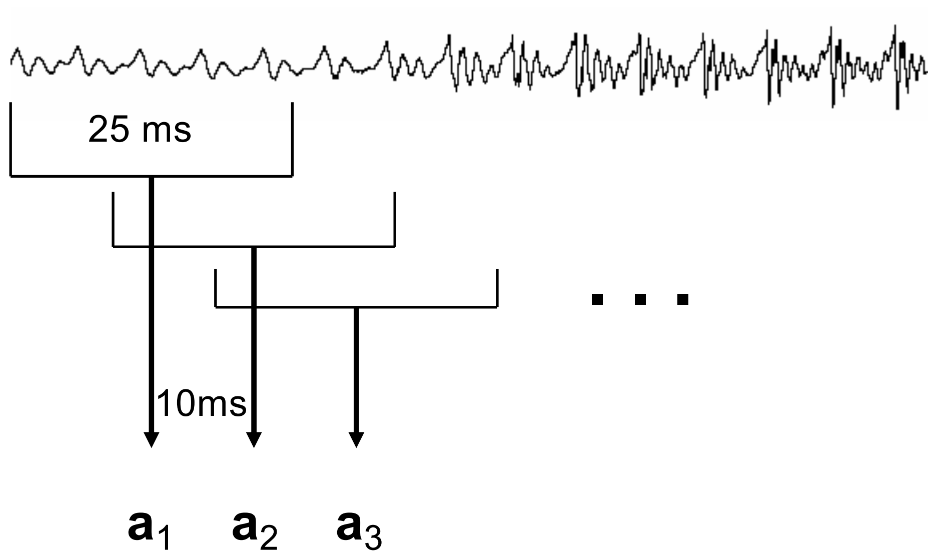
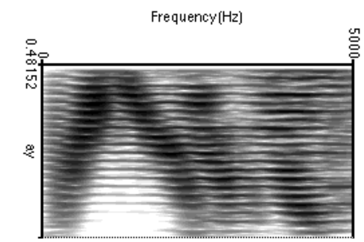
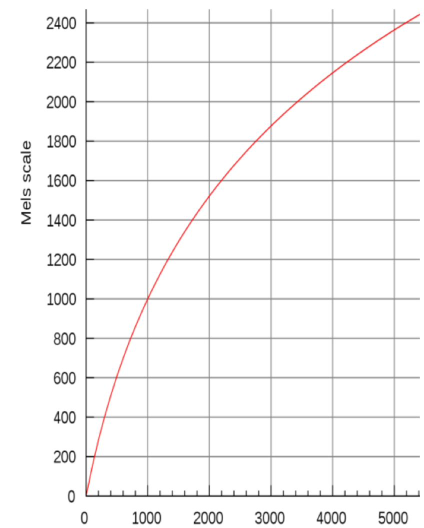


Figure: Simon Arnfield

Mel Freq. Cepstral Coefficients

- Do FFT to get spectral information
  - Like the spectrogram we saw earlier
- Apply Mel scaling
  - Models human ear; more sensitivity in lower freqs
  - Approx linear below 1kHz, log above, equal samples above and below 1kHz
- Plus discrete cosine transform





[Graph: Wikipedia]

Final Feature Vector

- 39 (real) features per 10 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)

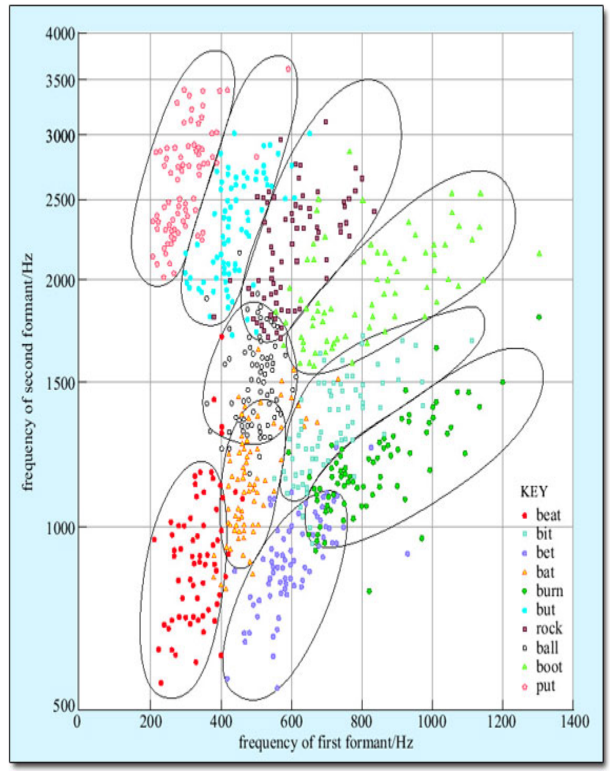
- So each frame is represented by a 39D vector

Emission Model

HMMs for Continuous Observations

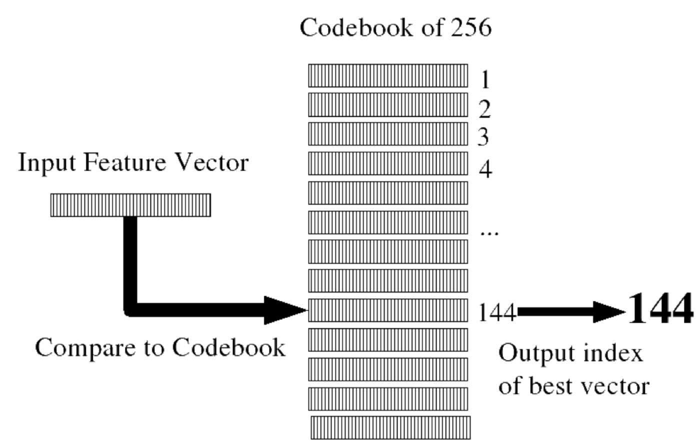
- Solution 1: discretization
- Solution 2: continuous emission models
  - Gaussians
  - Multivariate Gaussians
  - Mixtures of multivariate Gaussians
- Solution 3: neural classifiers

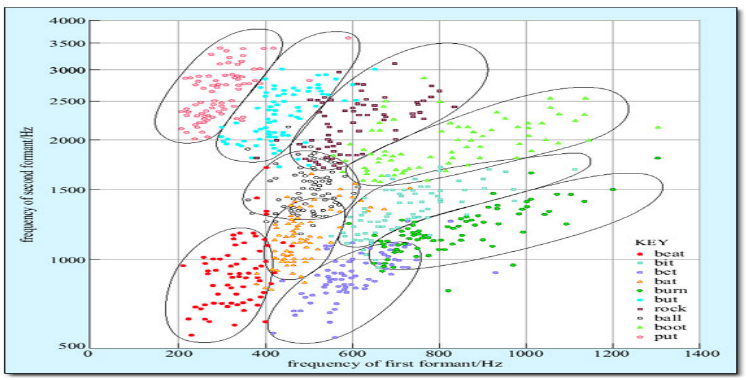
- A state is progressively
  - Context independent subphone (~3 per phone)
  - Context dependent phone (triphones)
  - State tying of CD phone



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, and for understanding neural methods

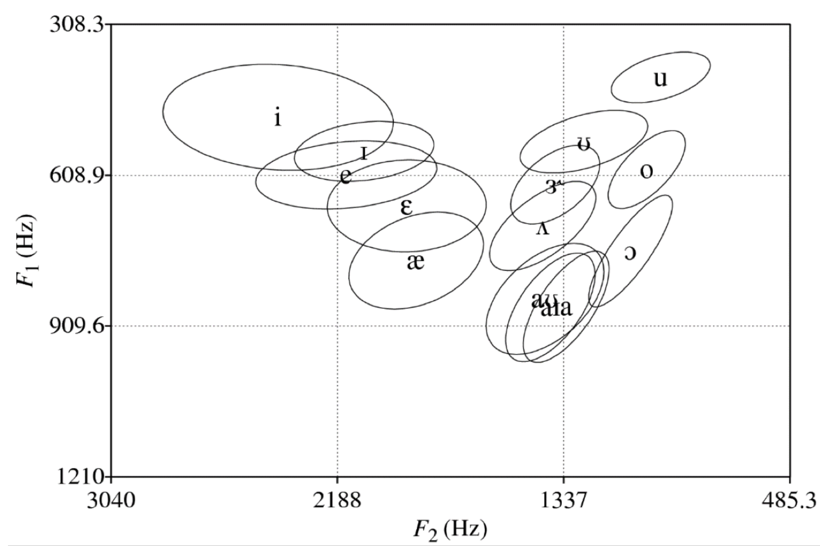




Gaussian Emissions

- VQ is insufficient for top-quality ASR
  - Hard to cover high-dimensional space with codebook
  - Moves ambiguity from the model to the preprocessing

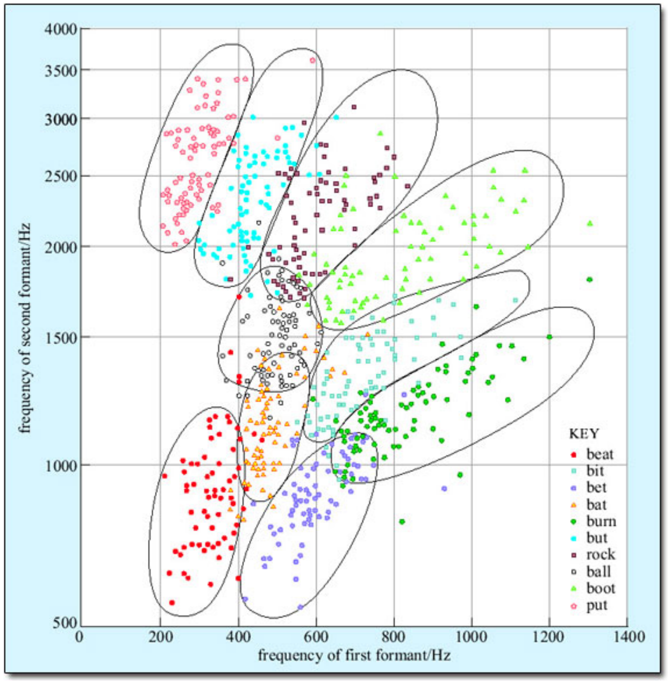
- Instead: assume the possible values of the observation vectors are normally distributed.
  - Represent the observation likelihood function as a Gaussian?



From bartus.org/akustyk

But we're not there yet

- Single Gaussians may do a bad job of modeling a complex distribution in any dimension
- Even worse for diagonal covariances
- Classic solution: mixtures of Gaussians
- Modern solution: NN-based acoustic models map feature vectors to (sub)states

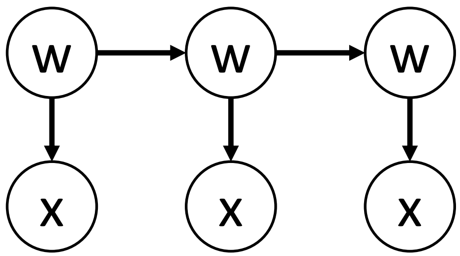


From openlearn.open.ac.uk

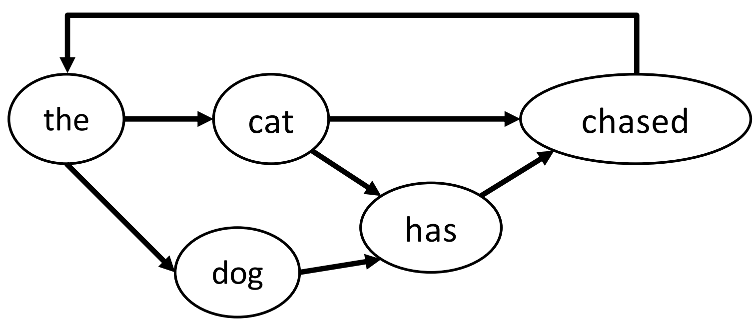
HMM / State Model

State Transition Diagrams

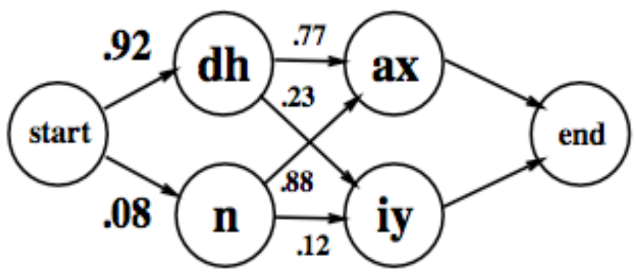
- Bayes Net: HMM as a Graphical Model



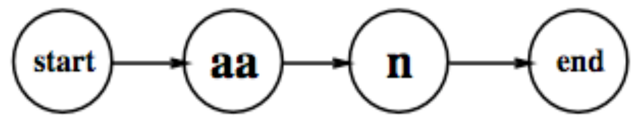
- State Transition Diagram: Markov Model as a Weighted FSA



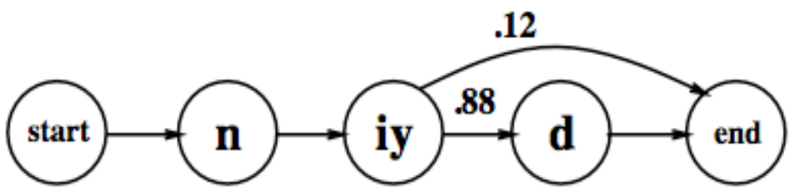
ASR Lexicon



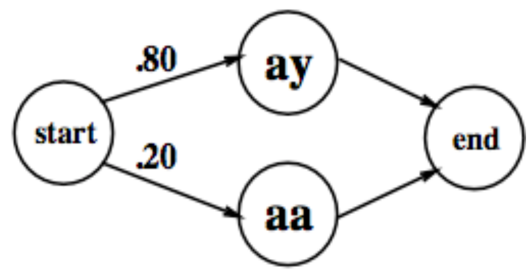
Word model for "the"



Word model for "on"



Word model for "need"



Word model for "I"

Figure: J & M

Lexical State Structure

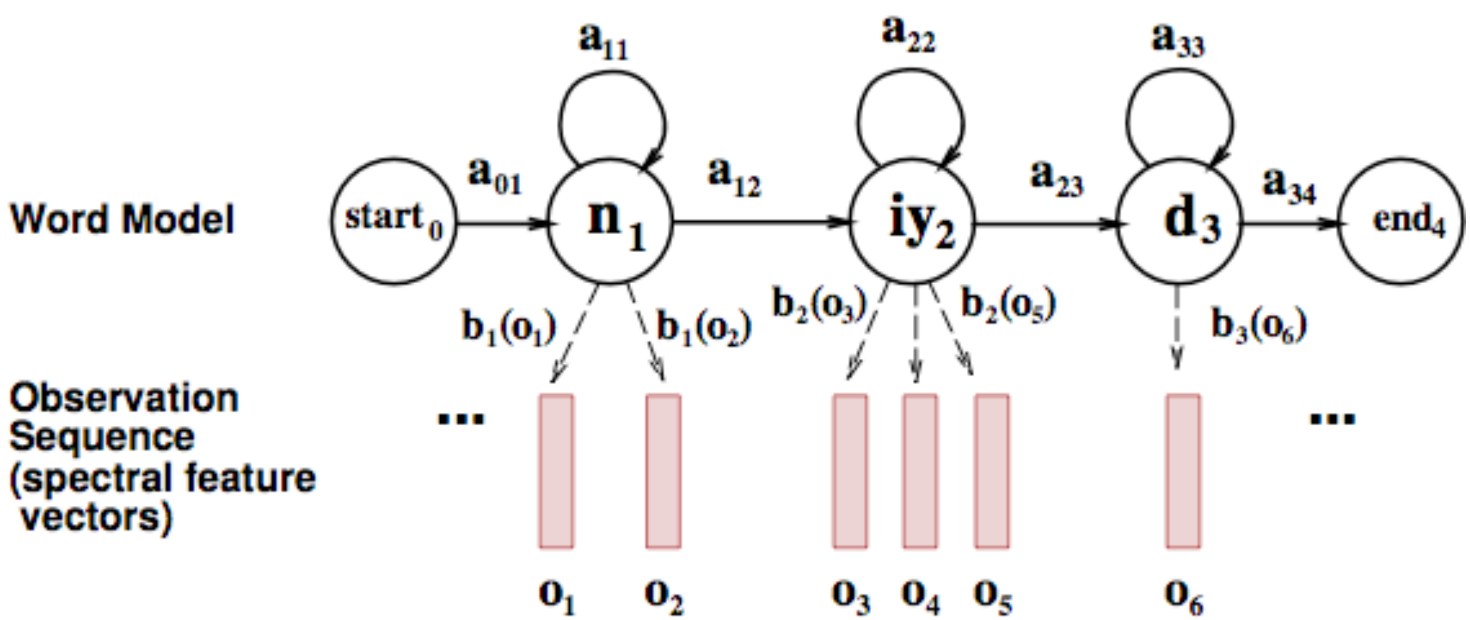


Figure: J & M

Adding an LM

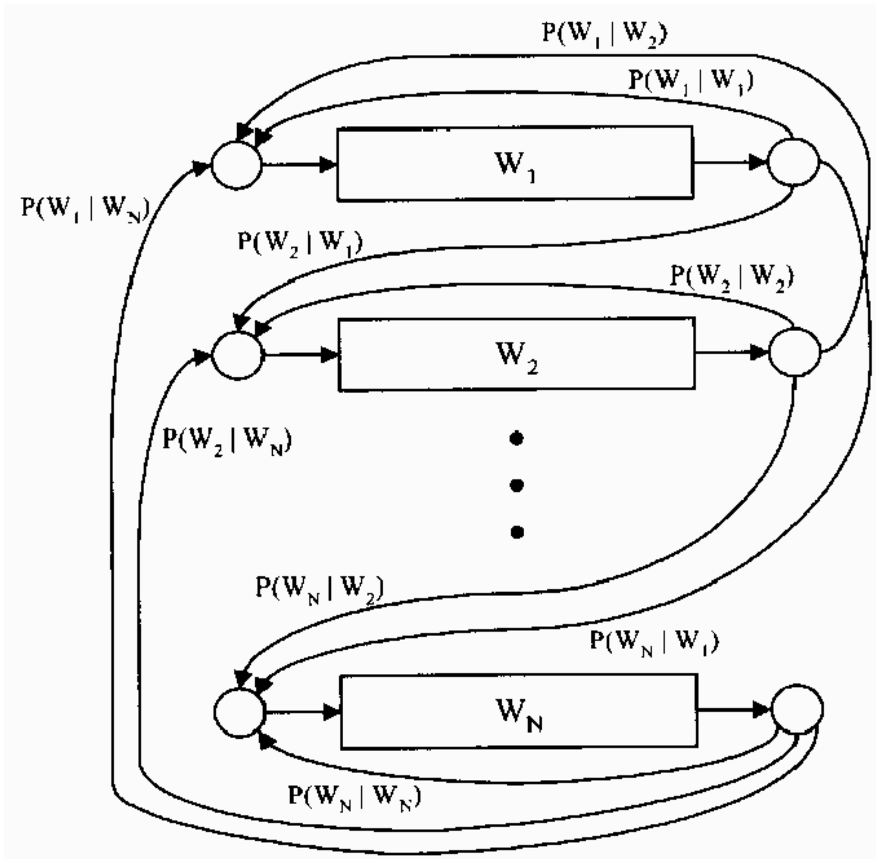


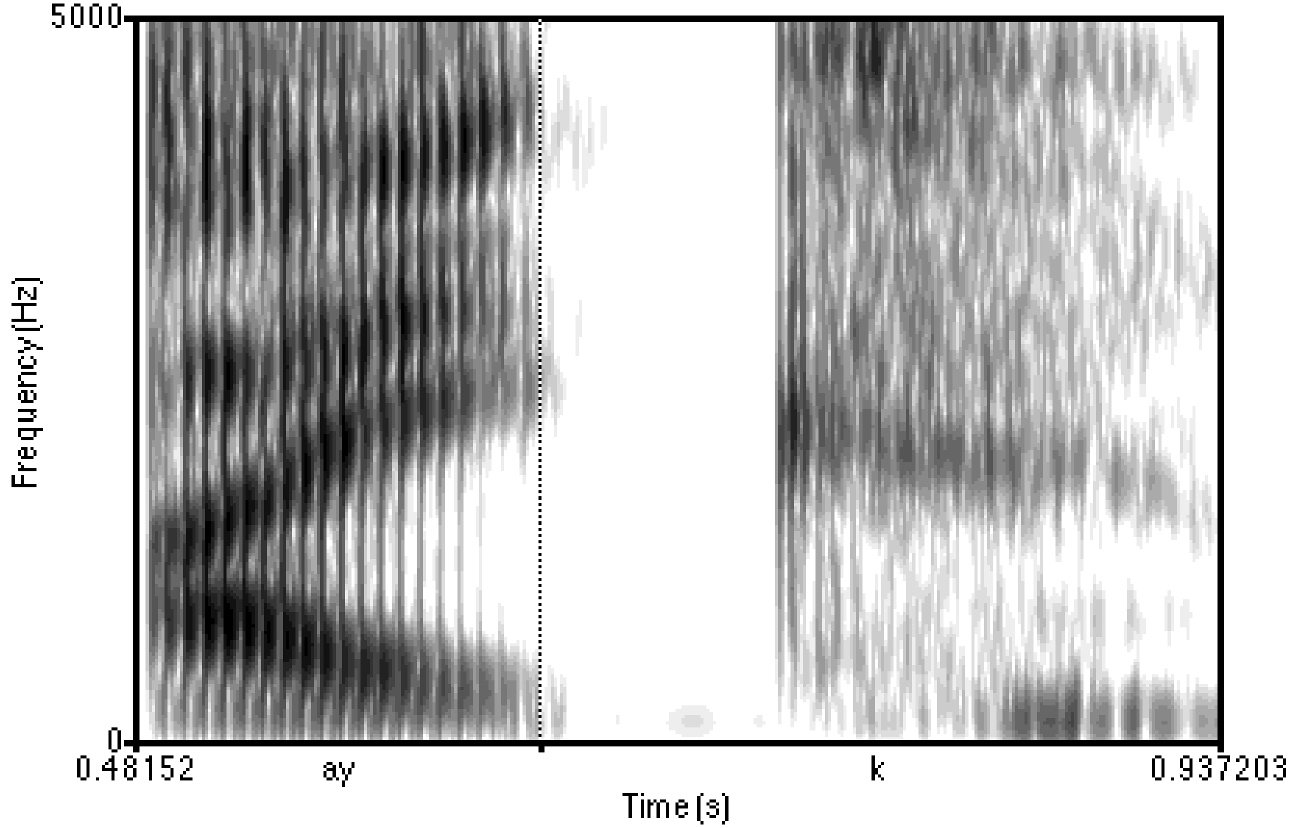
Figure from Huang et al page 618

State Space

- State space must include
  - Current word (|V| on order of 50K+)
  - Index within current word (|L| on order of 5)
  - E.g. (lec[t]ure) (though not in orthography!)
- Acoustic probabilities only depend on (contextual) phone type
  - E.g. P(x|lec[t]ure) = P(x|t)
- From a state sequence, can read a word sequence

State Refinement

Phones Aren't Homogeneous



Subphones

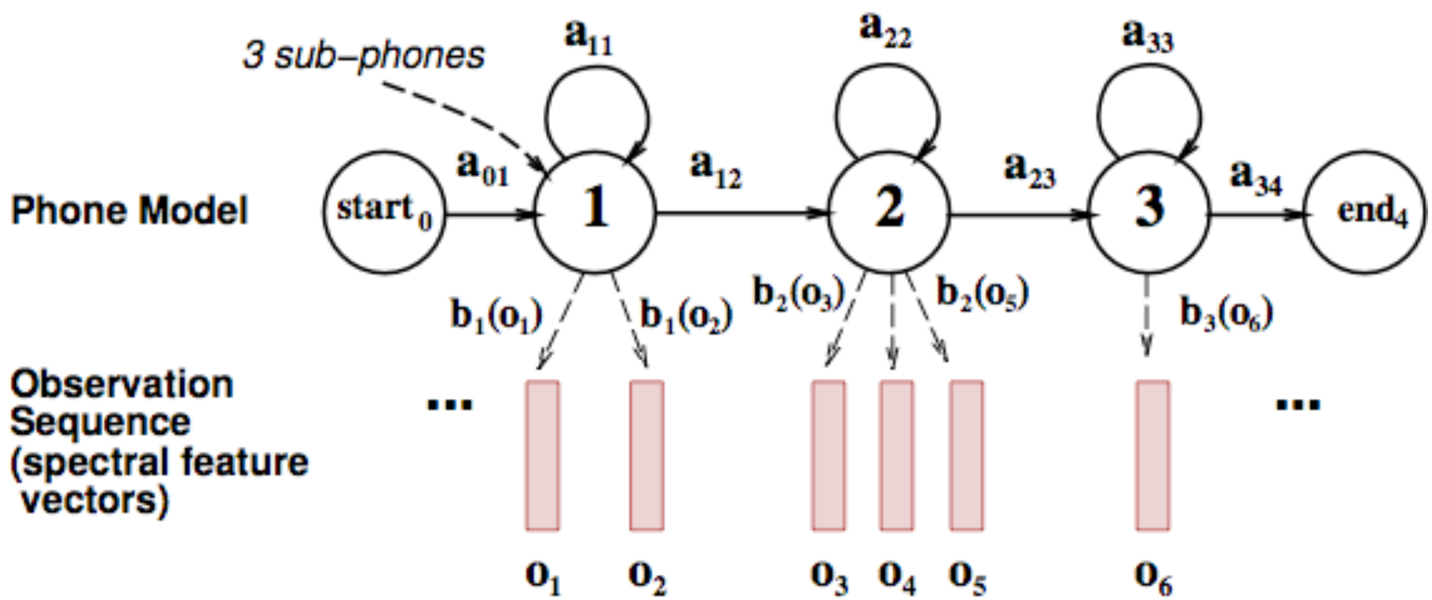


Figure: J & M

A Word with Subphones

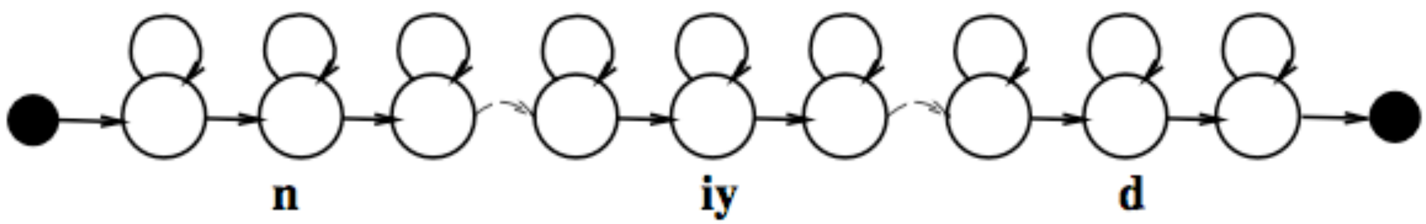
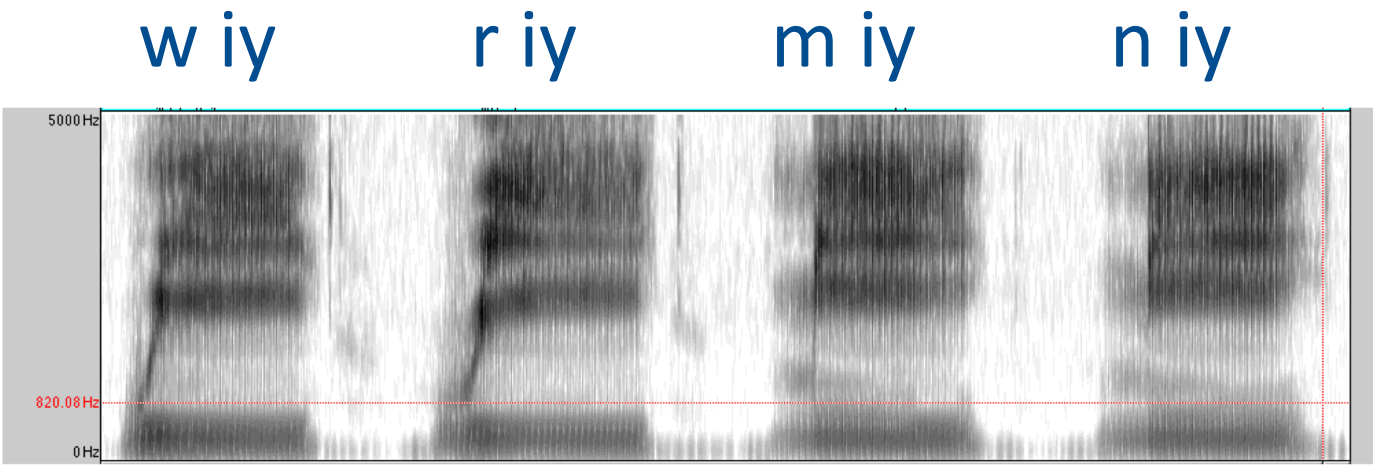


Figure: J & M

Modeling phonetic context



“Need” with triphone models

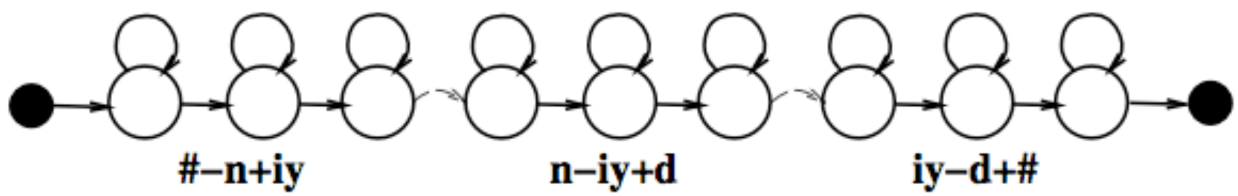


Figure: J & M

Lots of Triphones

- Possible triphones: 50x50x50=125,000
- How many triphone types actually occur?
- 20K word WSJ Task (from Bryan Pellom)
  - Word internal models: need 14,300 triphones
  - Cross word models: need 54,400 triphones
- Need to generalize models, tie triphones

State Tying / Clustering

- [Young, Odell, Woodland 1994]
- How do we decide which triphones to cluster together?
- Use phonetic features (or `broad phonetic classes')
  - Stop
  - Nasal
  - Fricative
  - Sibilant
  - Vowel
  - lateral

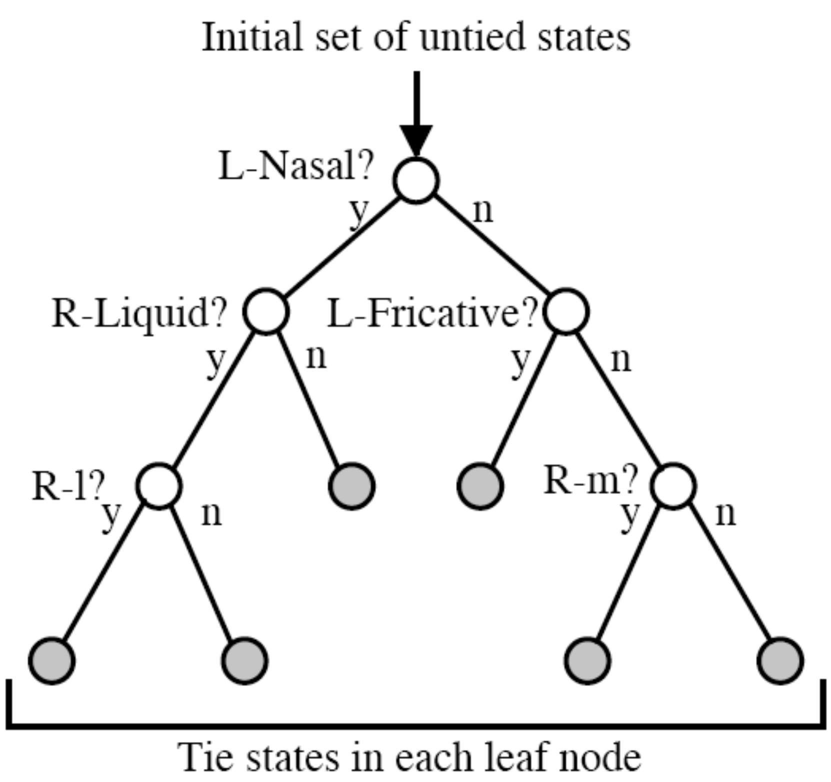


Figure: J & M

State Space

- Full state space

  (LM context, lexicon index, subphone)

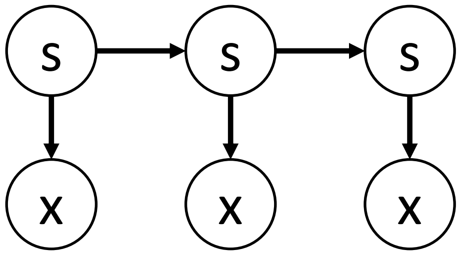
- Details:
  - LM context is the past n-1 words
  - Lexicon index is a phone position within a word (or a trie of the lexicon)
  - Subphone is begin, middle, or end
  - E.g. (after the, lec[t-mid]ure)

- Acoustic model depends on clustered phone context
  - But this doesn’t grow the state space

Learning Acoustic Models

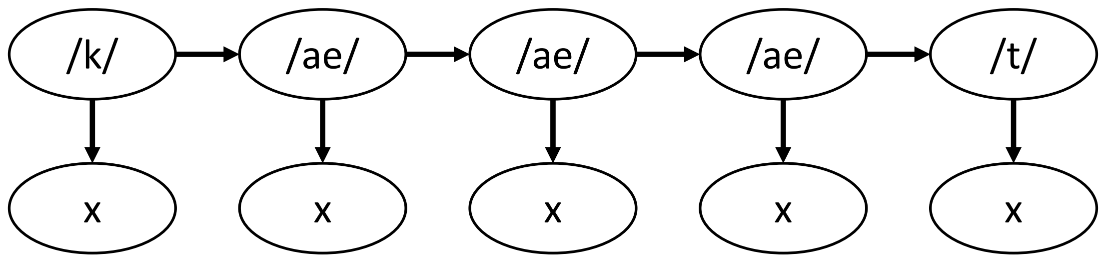
What Needs to be Learned?

- Emissions: P(x | phone class)
  - X is MFCC-valued
  - In neural methods, actually have P(phone | window around x) and then coerce those scores into P(x | phone)
- Transitions: P(state | prev state)
  - If between words, this is P(word | history)
  - If inside words, this is P(advance | phone class)
  - (Really a hierarchical model)



Estimation from Aligned Data

- What if each time step were labeled with its (context-dependent sub) phone?



- Can estimate P(x|/ae/) as empirical mean and (co-)variance of x's with label /ae/, or mixture, etc/

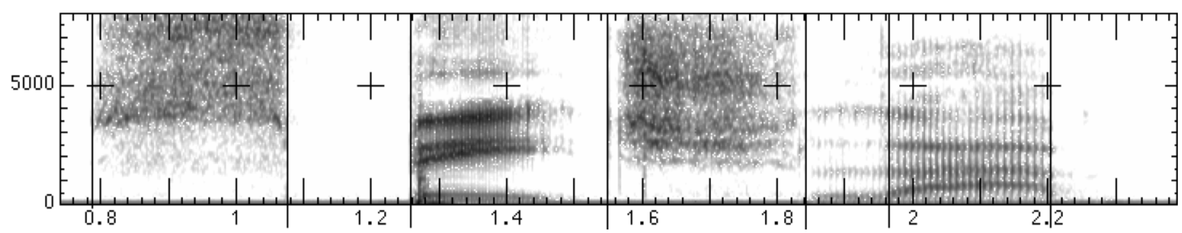
- Problem: Don't know alignment at the frame and phone level

Forced Alignment

- What if the acoustic model P(x|phone) were known (or approximately known)?
  - … and also the correct sequences of words / phones
- Can predict the best alignment of frames to phones

“speech lab”

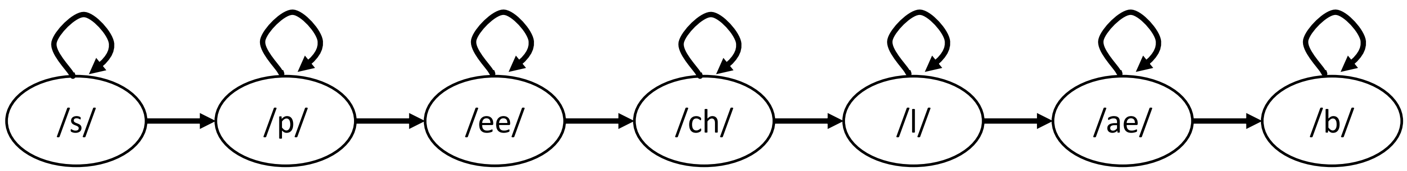
sssssssppppeeeeeeetshshshshllllaeaeaebbbbb



- Called “forced alignment”

Forced Alignment

- Create a new state space that forces the hidden variables to transition through phones in the (known) order



- Still have uncertainty about durations: this key uncertainty persists in neural models (and in some ways is worse now)

- In this HMM, all the parameters are known
  - Transitions determined by known utterance
  - Emissions assumed to be known
  - Minor detail: self-loop probabilities

- Just run Viterbi (or approximations) to get the best alignment

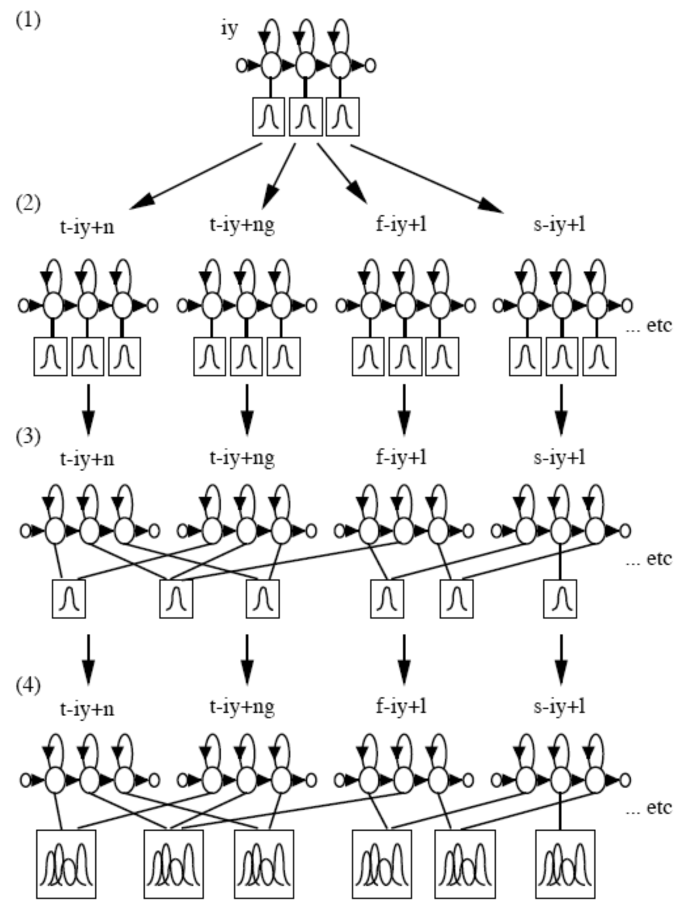
EM for Alignment

- Input: acoustic sequences with word-level transcriptions
- We don’t know either the emission model or the frame alignments
- Expectation Maximization
  - Alternating optimization
  - Impute completions for unlabeled variables (here, the states at each time step)
  - Re-estimate model parameters (here, Gaussian means, variances, mixture ids)
  - Repeat
  - One of the earliest uses of EM for structured problems

Staged Training and State Tying

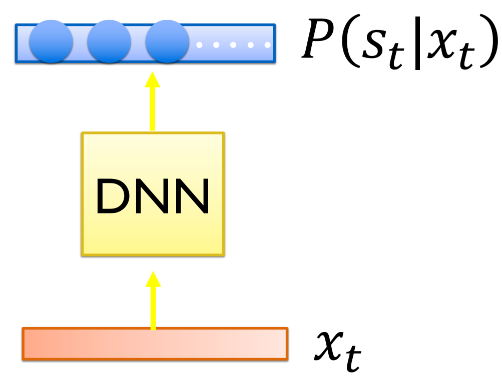
- Creating CD phones:
  - Start with monophone, do EM training
  - Clone Gaussians into triphones
  - Build decision tree and cluster Gaussians
  - Clone and train mixtures (GMMs)

- General idea:
  - Introduce complexity gradually
  - Interleave constraint with flexibility



Neural Acoustic Models

- Given an input x, map to s; this score coerced into generative P(x|s) via Bayes rule (liberally ignoring terms)
  - One major advantage of the neural net is that you can look at many x's at once to capture dynamics (important!)



[Diagram from Hung-yi Li]

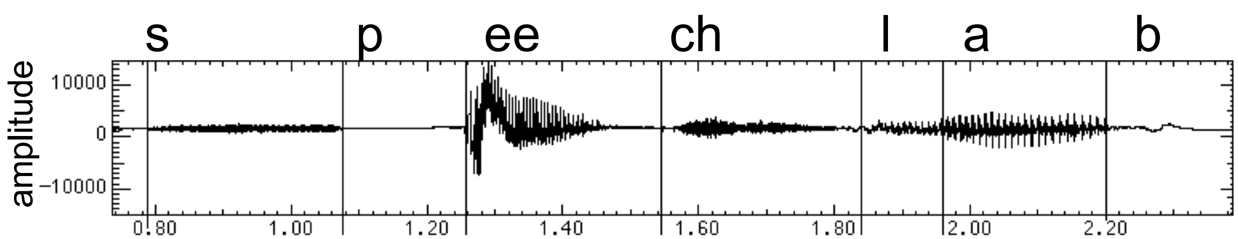
Speech Recognition and Synthesis

Berkeley
N L P

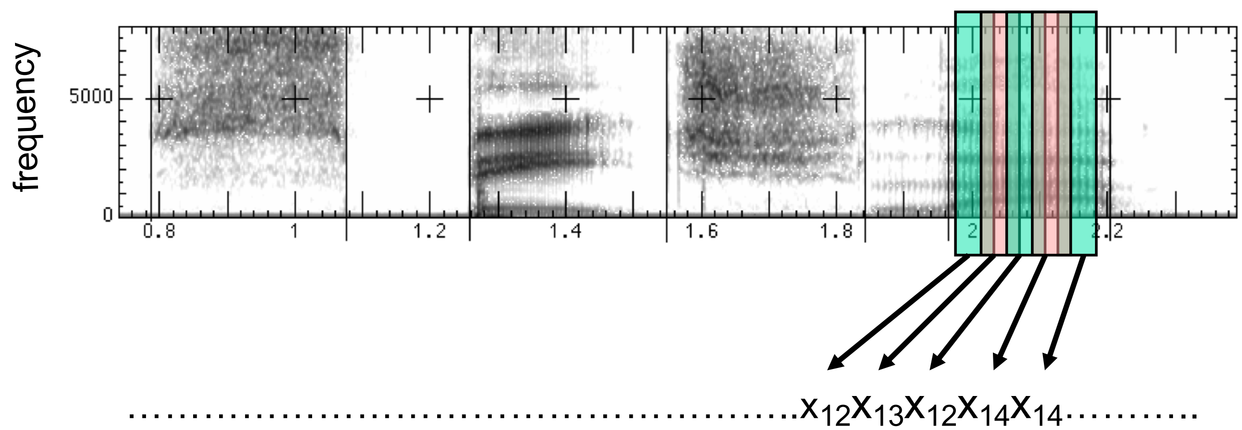
Dan Klein
UC Berkeley

Speech in a Slide

- Frequency gives pitch; amplitude gives volume

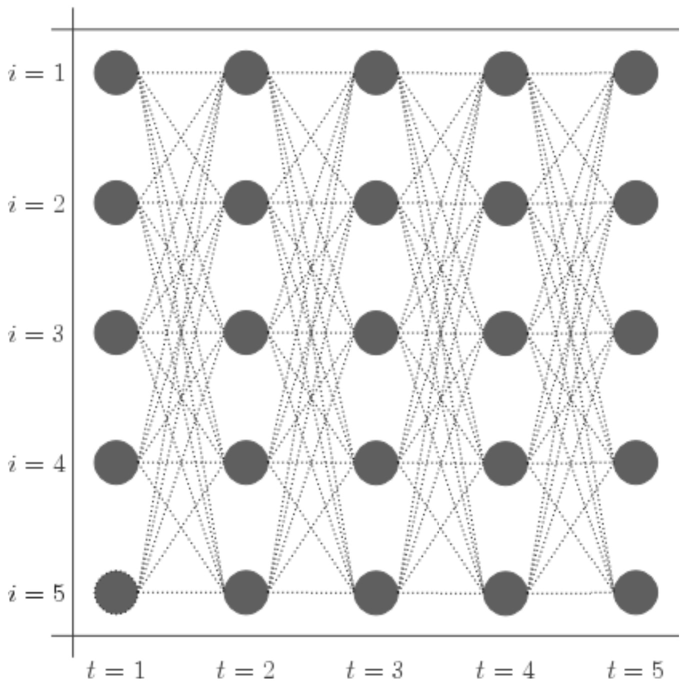


- Frequencies at each time slice processed into observation vectors



Decoding

State Trellis



$$\phi_t(s_{t-1}, s_t) = P(x_t|s_t)P(s_t|s_{t-1})$$

$$\begin{aligned} P(x, s) &= \prod_i P(x_i|s_i)P(s_i|s_{i-1}) \\ &= \prod_i \phi_t(s_{i-1}, s_i) \end{aligned}$$

Figure: Enrique Benimeli

Beam Search

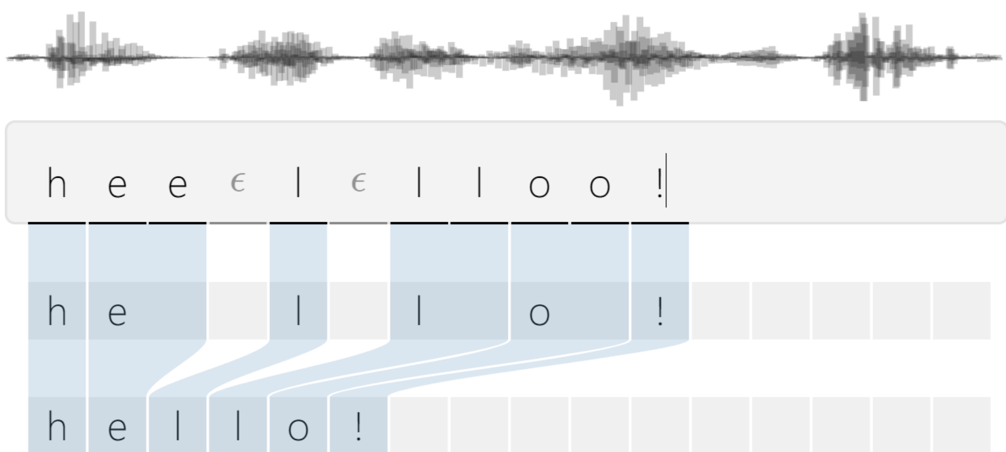
- Lattice is not regular in structure! Dynamic vs static decoding
- At each time step
  - Start: Beam (collection) v_t of hypotheses s at time t
  - For each s in v_t
    - Compute all extensions s' at time t+1
    - Score s' from s
    - Put s' in v_{t+1} replacing existing s' if better
  - Advance to t+1
- Beams are priority queues of fixed size* k (e.g. 30) and retain only the top k hypotheses

Dynamic vs Static Decoding

- Dynamic decoding
  - Build transitions on the fly based on model / grammar / etc
  - Very flexible, allows heterogeneous contexts easily (eg complex LMs)
- Static decoding
  - Compile entire subphone/vocabulary/LM into a huge weighted FST and use FST optimization methods (eg pushing, merging)
  - Much more common at scale, better eng and speed properties

Direct Neural Decoders

- Lots of work in decoders that skip explicit / discrete alignment
  - Decode to phone, or character, or word
  - Handle alignments softly (eg attention) or discretely (eg CTC)



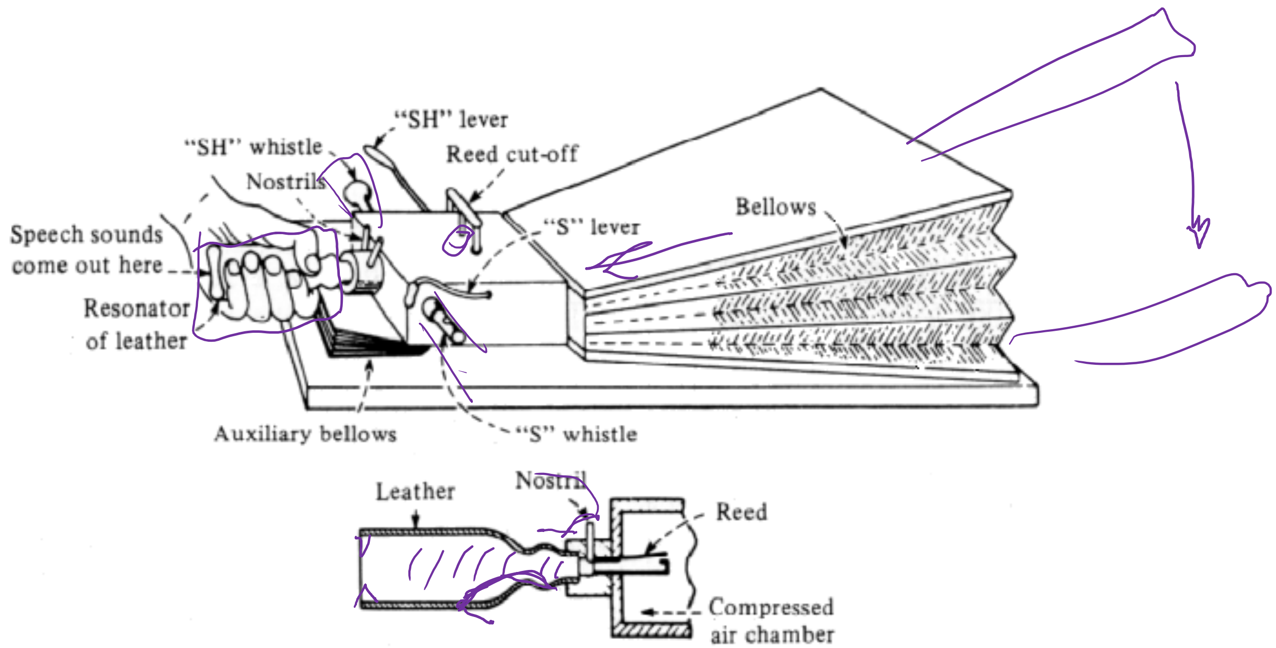
[CTC: Graves 06; Diagram from https://distill.pub/2017/ctc/]

Speech Synthesis

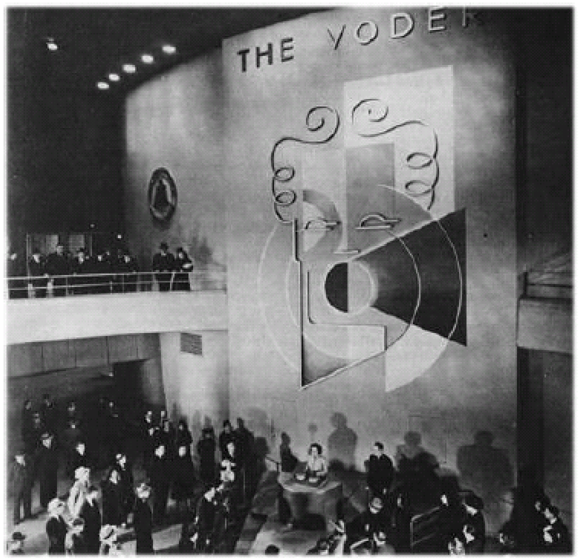
[Many slides from Dan Jurafsky]

Early TTS

- Von Kempelen, 1791



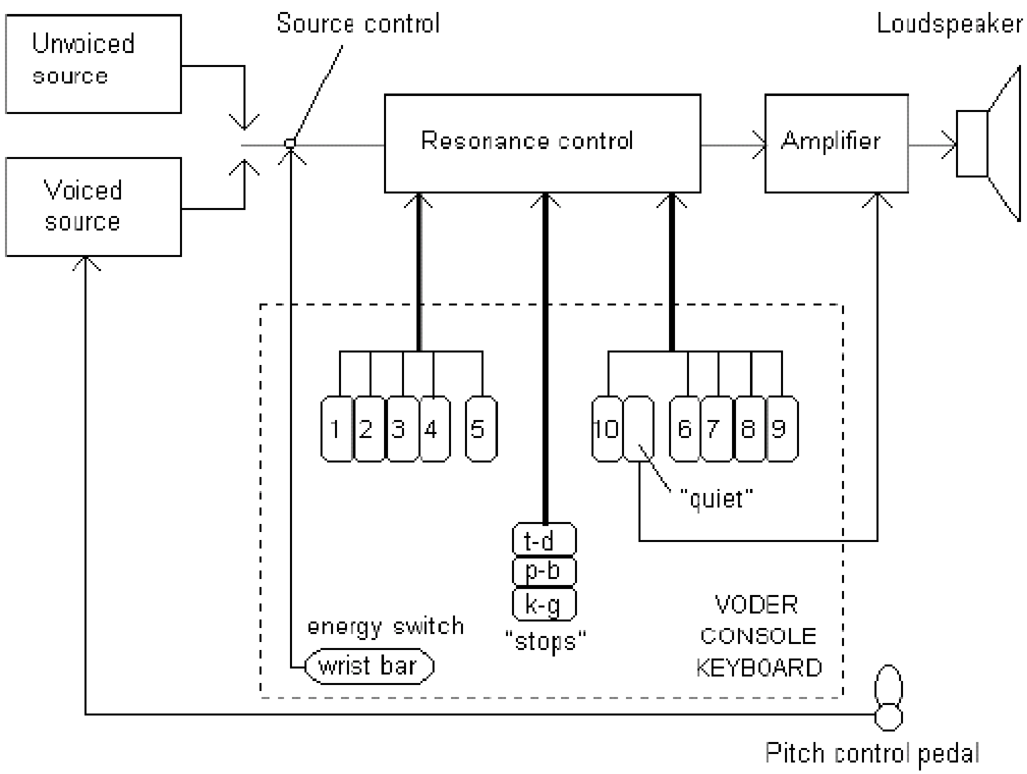
The Voder



Developed by Homer Dudley at Bell Telephone Laboratories, 1939

Voder Architecture

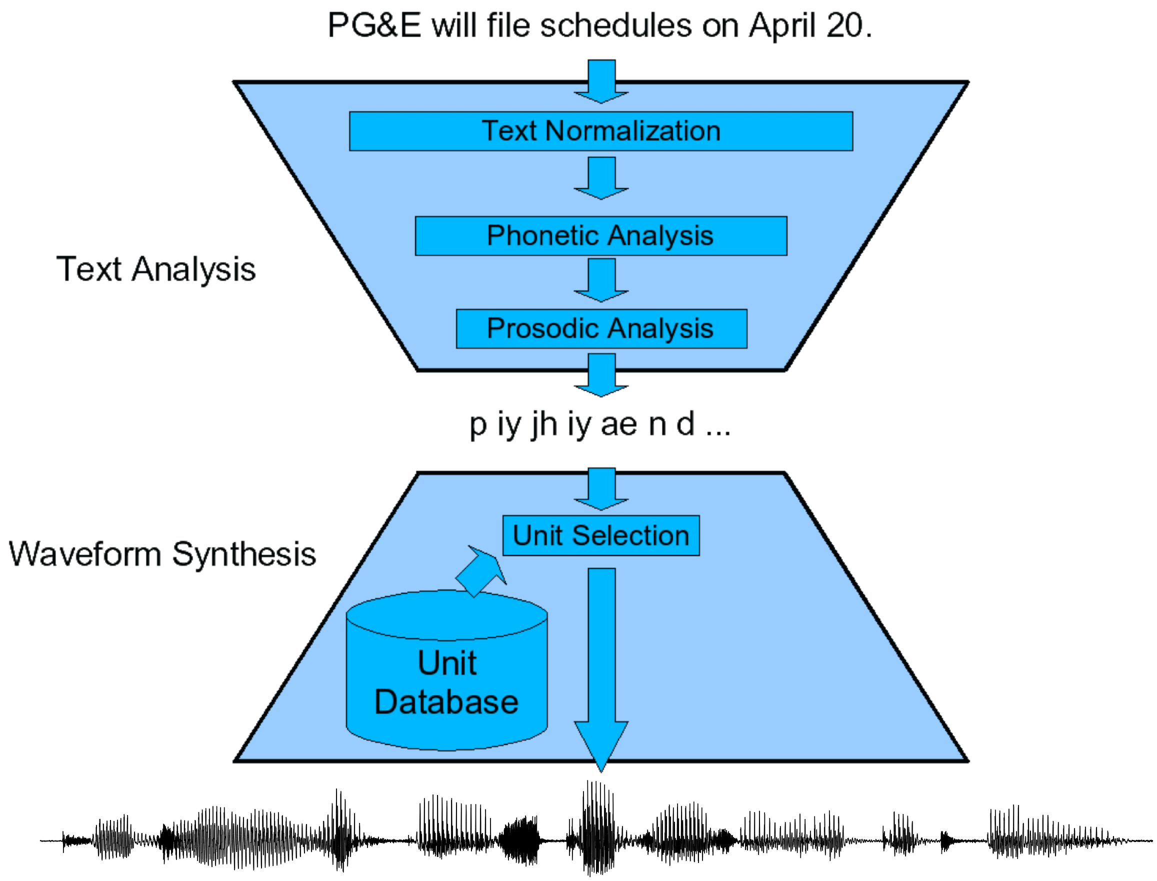
- An early hardware solution that already captured the flow of parametric synthesizers



Modern TTS

- 1960's first full TTS: Umeda et al (1968)
- 1970's
  - Joe Olive 1977 concatenation of linear-prediction diphones
  - Speak and Spell
- 1980's
  - 1979 MIT MITalk (Allen, Hunnicut, Klatt)
- 1990's – 2000's
  - Diphone synthesis
  - Unit selection synthesis
- Recent
  - Parametric synthesis returns!

TTS Architecture

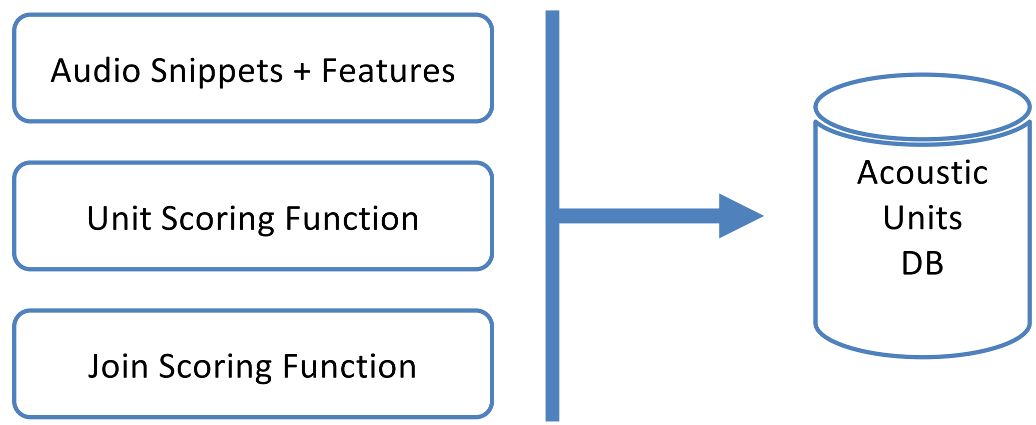


Typical Data for Classic TTS

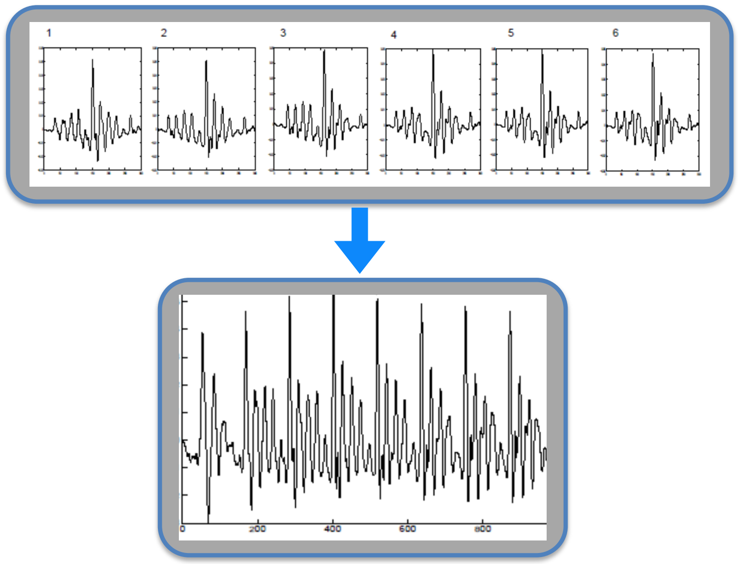
- Professional voice actor
- Carefully selected material
- High-quality recordings
  - 10-100 hours @ 44kHz
  - High signal-to-noise ratio
  - Consistent audio levels
  - No vocal issues (creaky voice)
  - Anechoic-like environment
- Usually lots of post-processing (alignments, pronunciations, …)

Concatenative Synthesis

- Commercially dominant (diphones, unit-selection, etc)

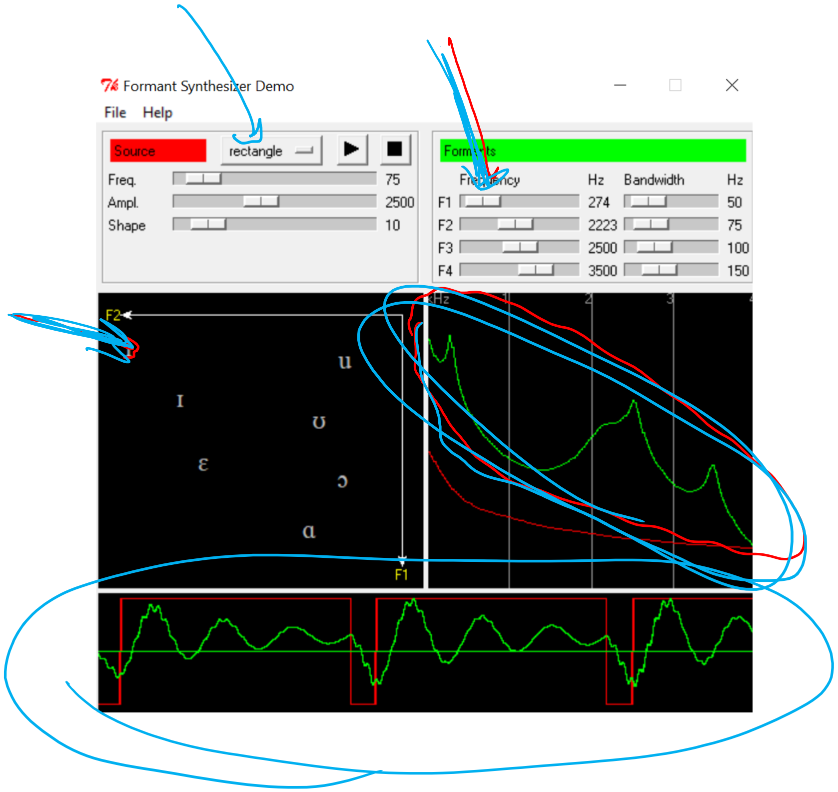


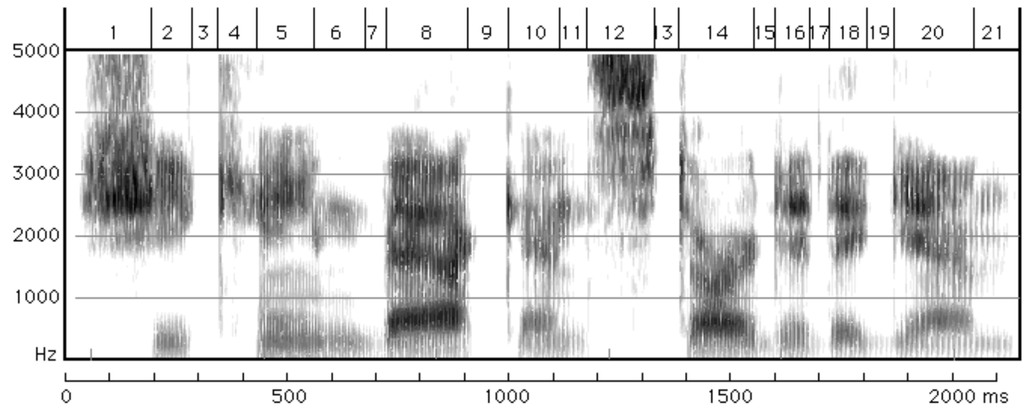
PSOLA



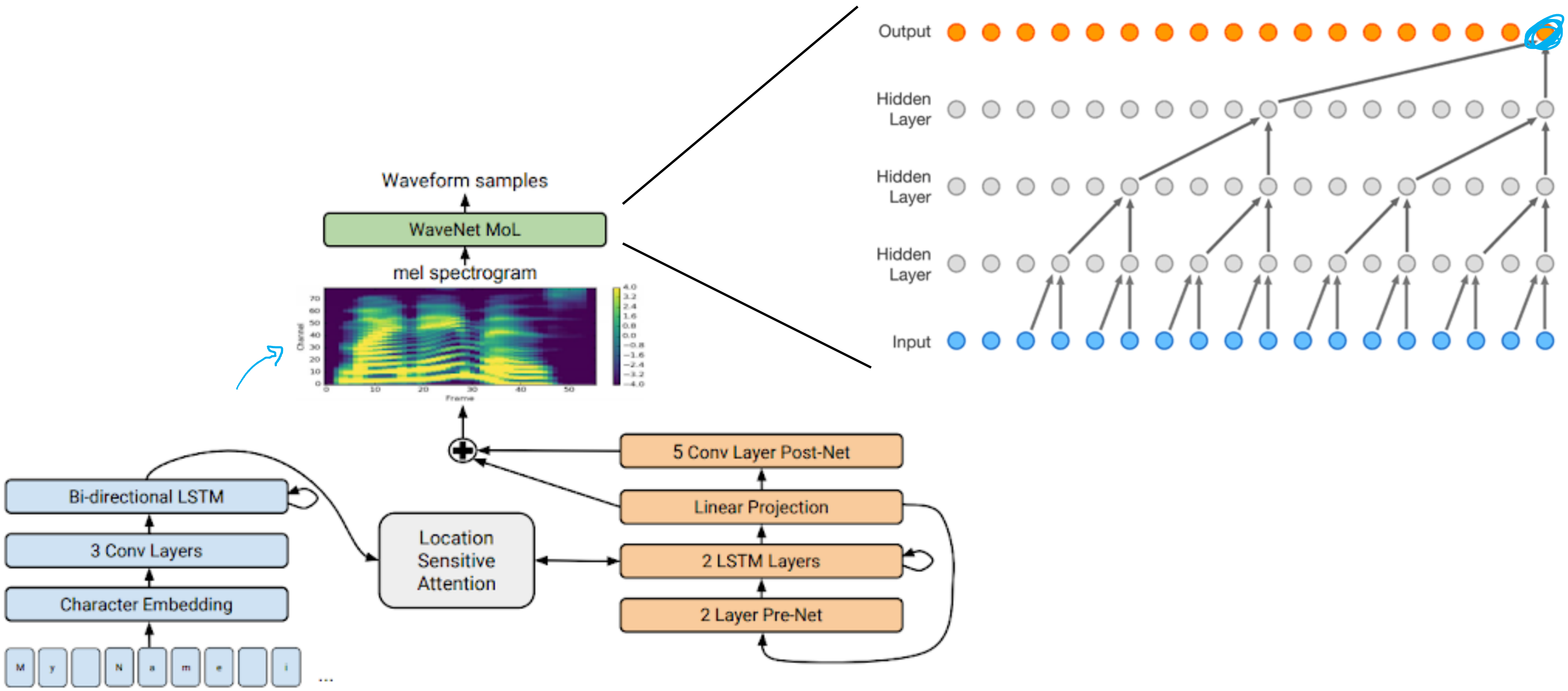
Time-domain Pitch-Synchronous Overlap and Add (TD-PSOLA)

Formant Synthesis





Direct-to-Wave Synthesis



https://ai.googleblog.com/2017/12/tacotron-2-generating-human-like-speech.html

Examples

Concatenative

Parametric

WaveNet

[https://www.deepmind.com/blog/wavenet-a-generative-model-for-raw-audio]