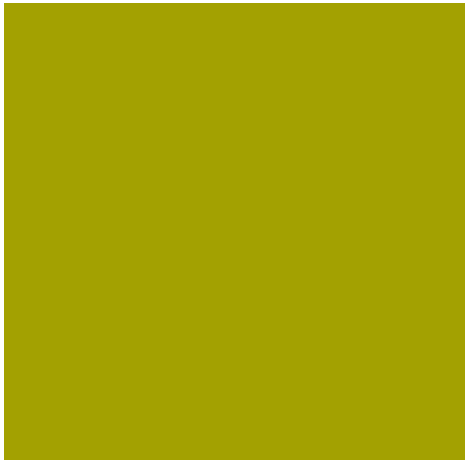
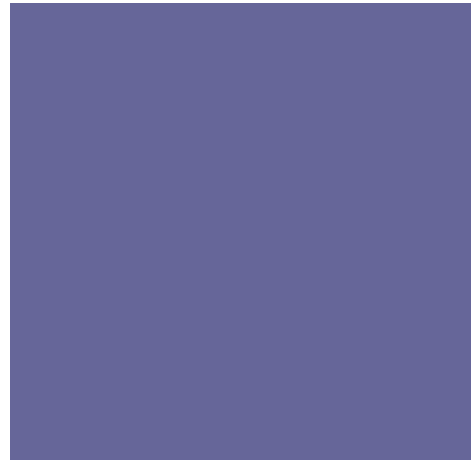
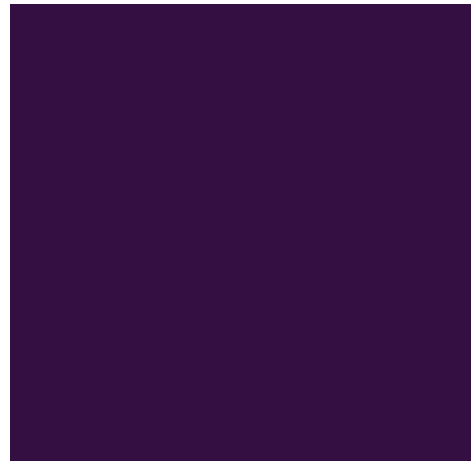
Recognition Architecture: Feature Extraction



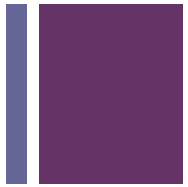




CS 136 Speech Recognition
January 21, 2020
Professor Meteer

Thanks to Dan Jurafsky for these slides

Phonetics



- Phonemes and the ARPAbet
  - An alphabet for transcribing American English phonetic sounds.
- Articulatory Phonetics
  - How speech sounds are made by articulators (moving organs) in mouth.
- Language resources and WFSTs

From speech to phonemes

- Phonemes are the minimal set of sounds to distinguish meaning
  - Pat – bat, tab – dab,
  - Fat – chat – that
  - Pack – pick – puck -- pike
- Uses the alphabet, but not isomorphic to spelling (especially in English)
- Standard used in speech recognition is the “ARPABET”
  - 46 total (17 vowels, 29 consonants) + 13 “extras
  - In practice there are many variations, but all are close
  - http://www.stanford.edu/class/cs224s/arpabet.html
  - NOTE: These are for English only—each language has its own set of phonemes

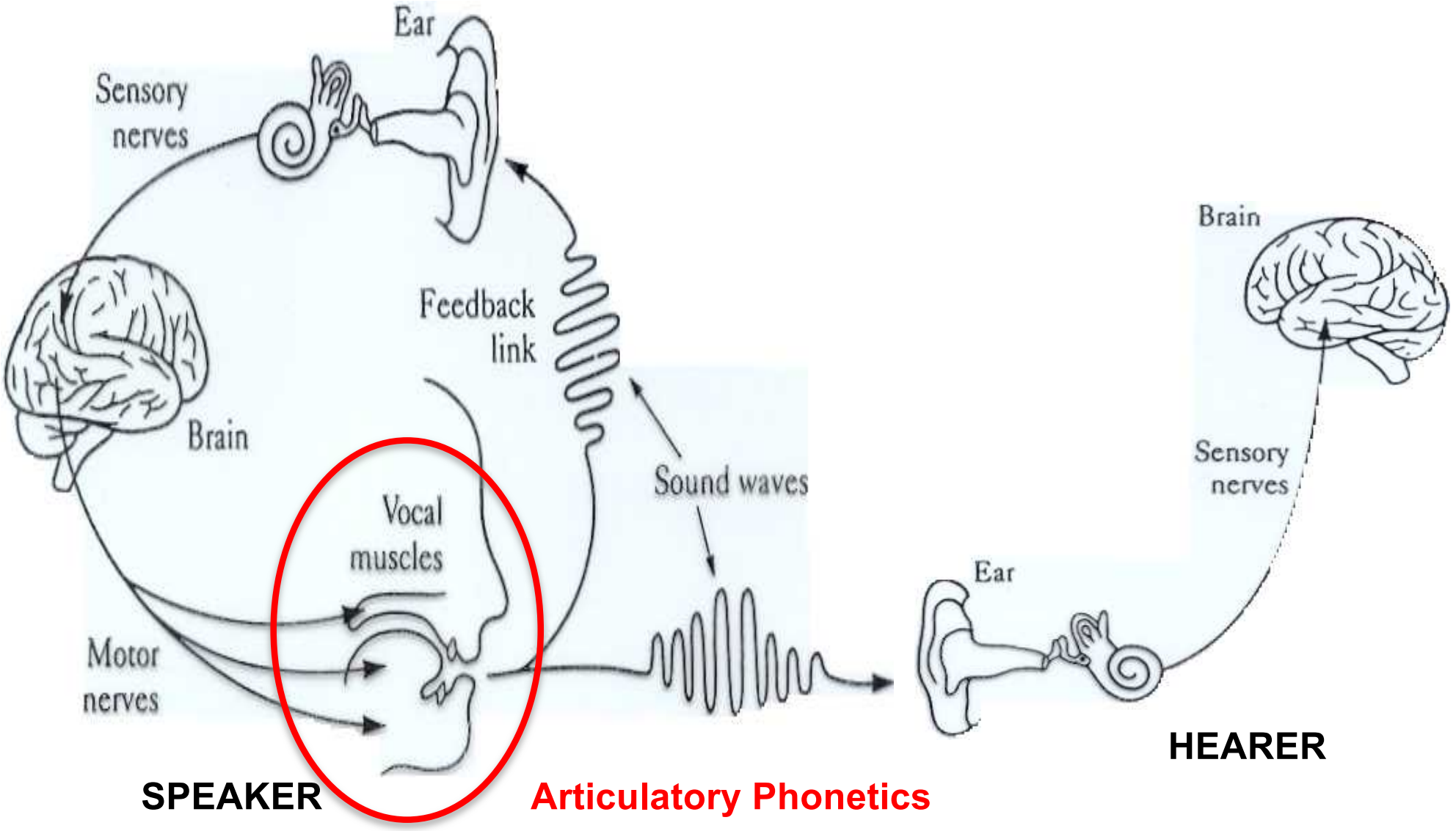
Thanks to Dan Jurafsky for these slides

ARPAbet Vowels

	b_d	ARPA		b_d	ARPA
1	bead	iy	9	bode	ow
2	bid	ih	10	booed	uw
3	bayed	ey	11	bud	ah
4	bed	eh	12	bird	er
5	bad	ae	13	bide	ay
6	bod(y)	aa	14	bowed	aw
7	bawd	ao	15	Boyd	oy
8	Budd(hist)	uh			

Note: Many speakers pronounce Buddhist with the vowel uw as in booed, So for them [uh] is instead the vowel in "put" or "book"

The Speech Chain (Denes and Pinson)

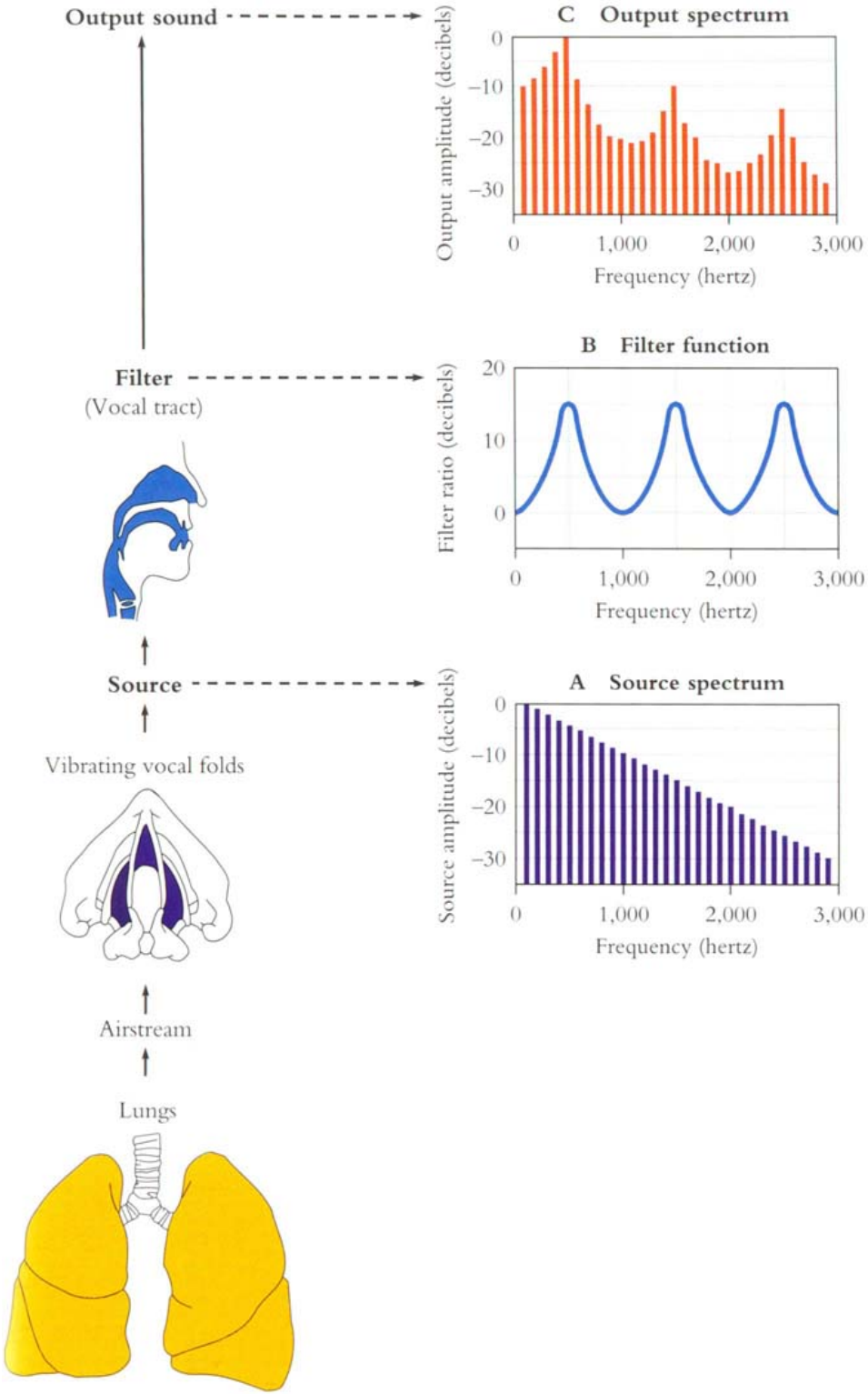


George Miller figure

Recognizing speech

Separating the filter from the source

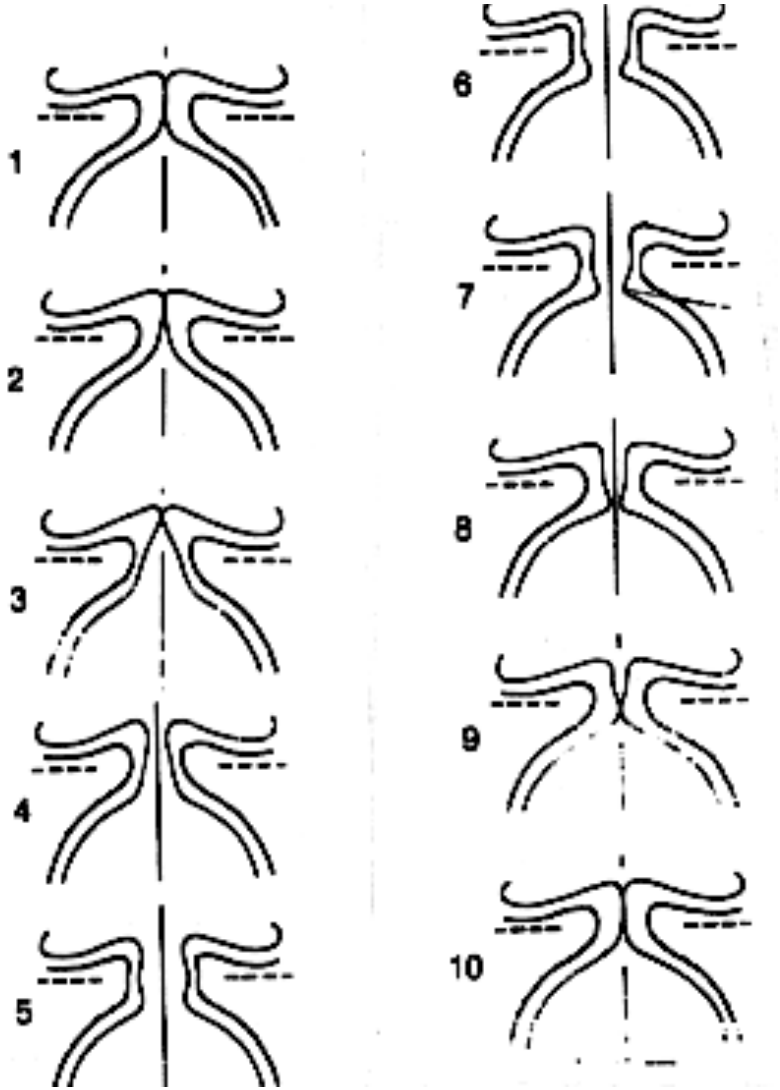
- Articulation and Resonance
  - Shape of vocal tract
- Phonation
  - Airstream sets vocal folds in motion. Vibration of vocal folds produces sounds.
- Respiration:
  - We (normally) speak while breathing out. Respiration provides airflow. “Pulmonic egressive airstream”



Phonation: Larynx and Vocal Folds

- The Larynx (voice box)
  - A structure made of cartilage and muscle
  - Located above the trachea (windpipe) and below the pharynx (throat)
  - Contains the vocal folds
  - (adjective for larynx: laryngeal)
- Vocal Folds (older term: vocal cords)
  - Two bands of muscle and tissue in the larynx
  - Can be set in motion to produce sound (voicing)

+ Voicing:



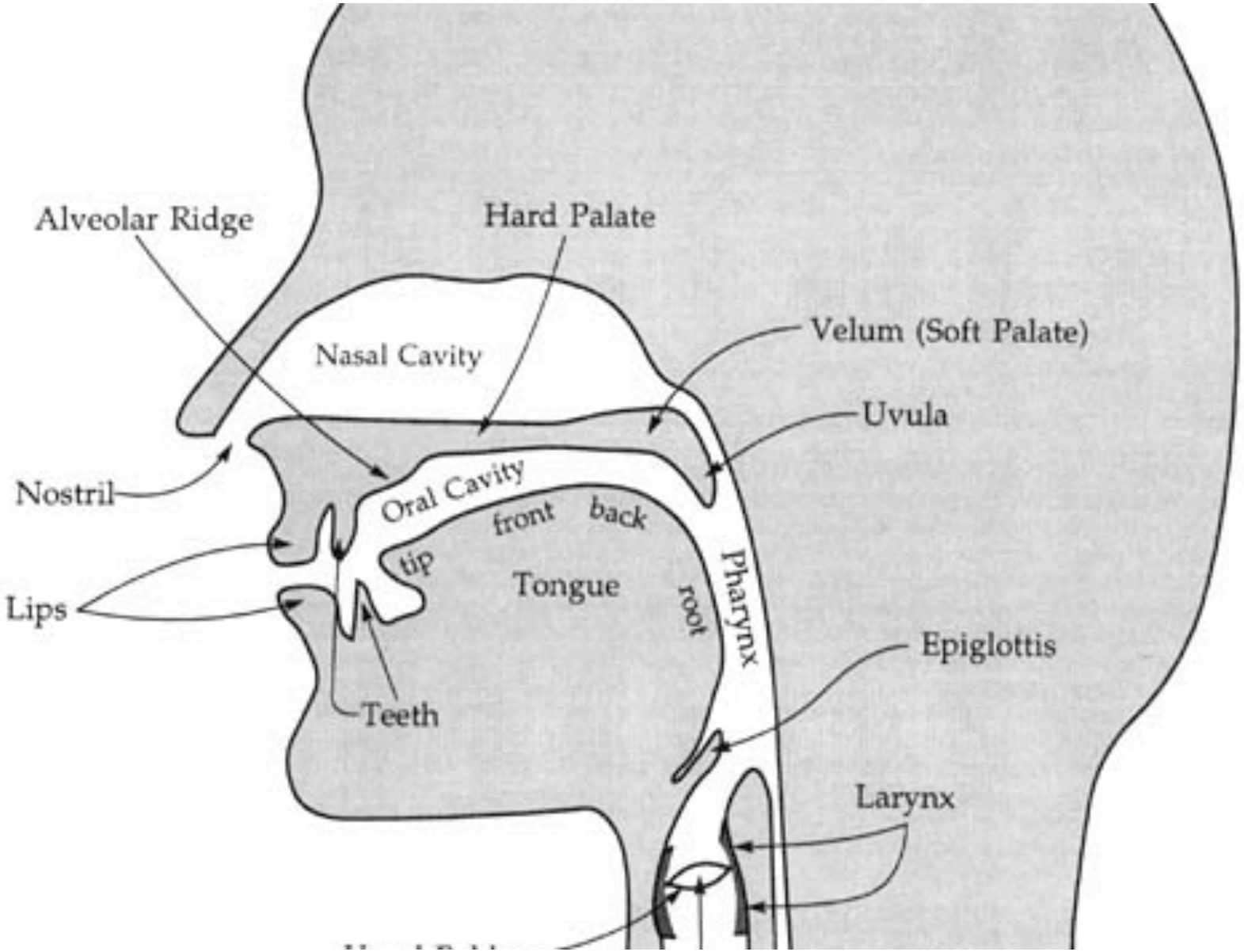
- Air comes up from lungs
- Forces its way through vocal folds, pushing open (2,3,4)
- This causes air pressure in glottis to fall, since:
  - when gas runs through constricted passage, its velocity increases (**Venturi tube effect**)
  - this increase in velocity results in a drop in pressure (**Bernoulli principle**)
- Because of drop in pressure, vocal cords snap together again (6-10)
- Single cycle: ~1/100 of a second.

Figure & text from John Coleman's web site

Voicelessness

- When vocal cords are open, air passes through unobstructed
- Voiceless sounds: p/t/k/s/f/sh/th/ch
- If the air moves very quickly, the turbulence causes a different kind of phonation: whisper

+ Articulators and resonance



From Mark Liberman's Web Site, from Language Files (7th ed)

Consonants and Vowels

- Consonants: phonetically, sounds with audible noise produced by a constriction

- Vowels: phonetically, sounds with no audible noise produced by a constriction

Text adapted from John Colema

Place of articulation

- Coronal (tip of the tongue)
  - Dental: th/dh
  - Alveolar: t/d/s/z/l
  - Post: sh/zh/y
- Dorsal (back of the tongue)
  - Velar: k/g/ng
- Lips
  - Bilabial: p//b/m
  - Labiodental: f/v
- Glottis:
  - Glotal stop, as in Cockney "bottle"

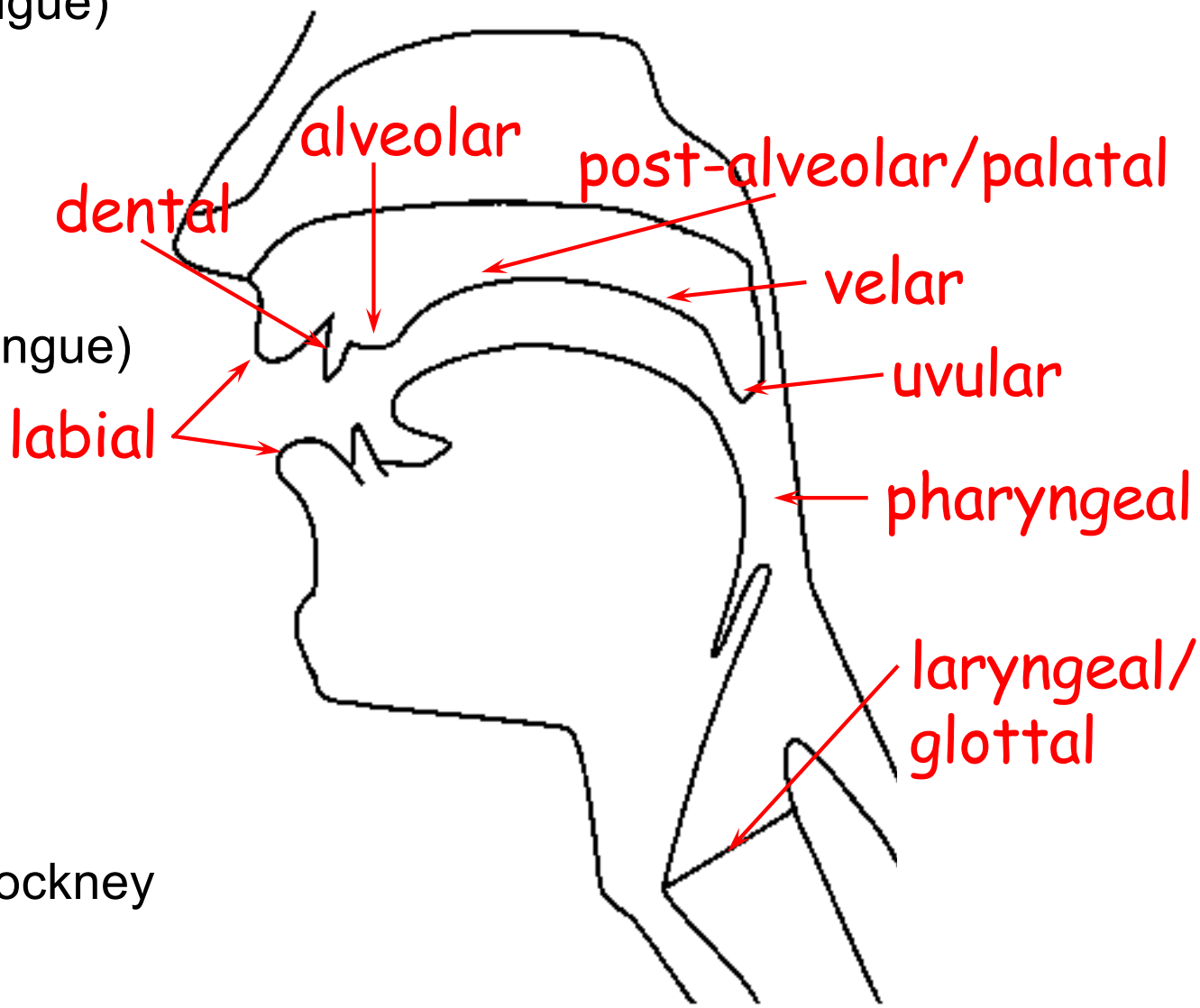


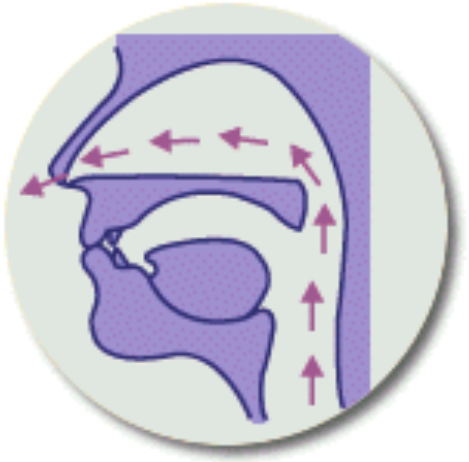
Figure thanks to Jennifer Venditti

Manner of Articulation

- Stop: complete closure of articulators, so no air escapes through mouth
  - Oral stop: palate is raised, no air escapes through nose. Air pressure builds up behind closure, explodes when released
    - p, t, k, b, d, g
  - Nasal stop: oral closure, but palate is lowered, air escapes through nose.
    - m, n, ng
- Fricative
  - Close approximation of two articulators, resulting in turbulent airflow between them
  - f, v, s, z, th, dh
- Affricate
- Approximant
- Tap or flap

Oral

Nasal



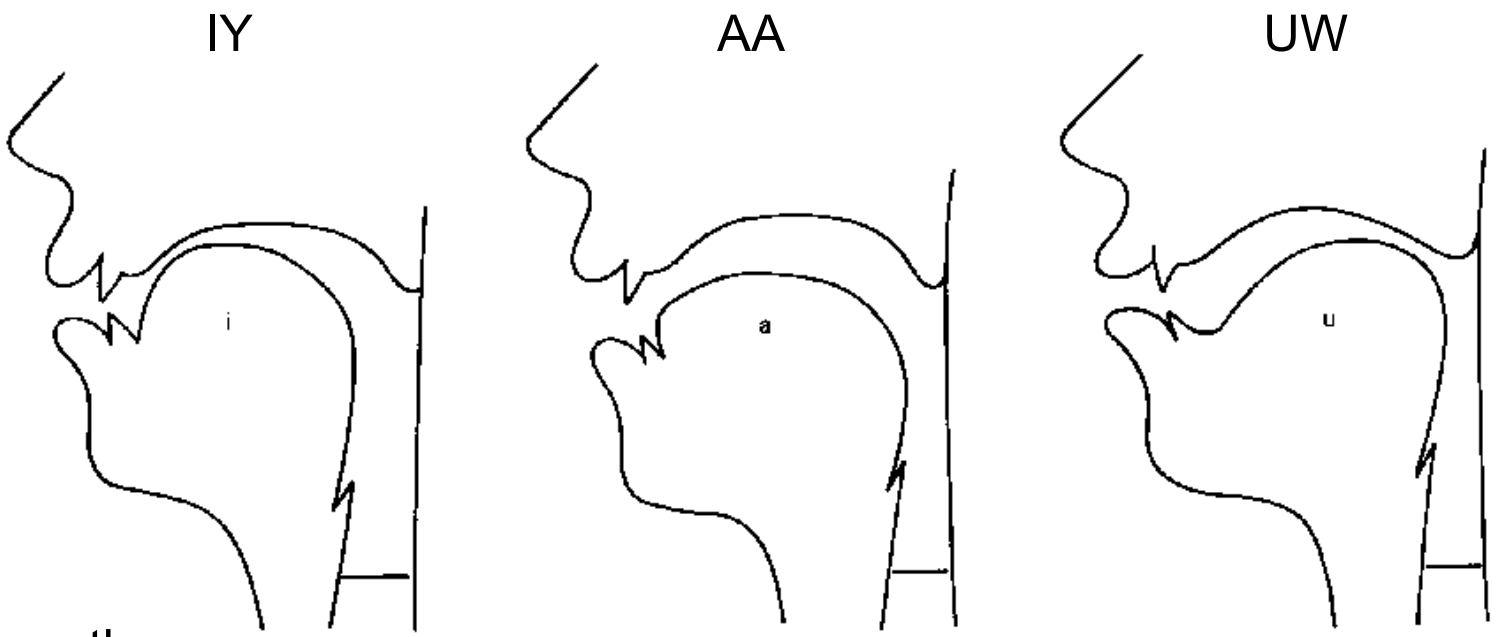
Articulatory parameters for English consonants (in ARPAbet)

MANNER OF ARTICULATION	PLACE OF ARTICULATION													
	bilabial		labio-dental		inter-dental		alveolar		palatal		velar		glottal	
stop	p	b					t	d			k	g	q	
fric.			f	v	th	dh	s	z	sh	zh			h	
affric.									ch	jh				
nasal		m						n				ng		
approx		w						l/r		y				
flap							dx							

VOICING: voiceless | voiced

Table from Jennifer Venditti

+ Vowels



Peaks are the Formants

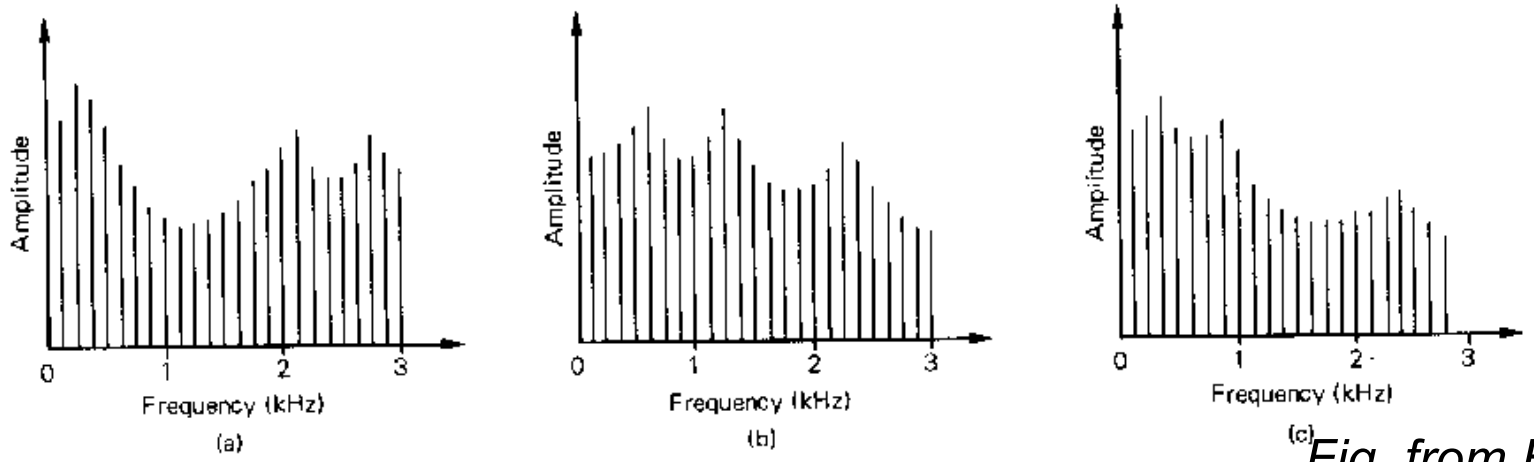
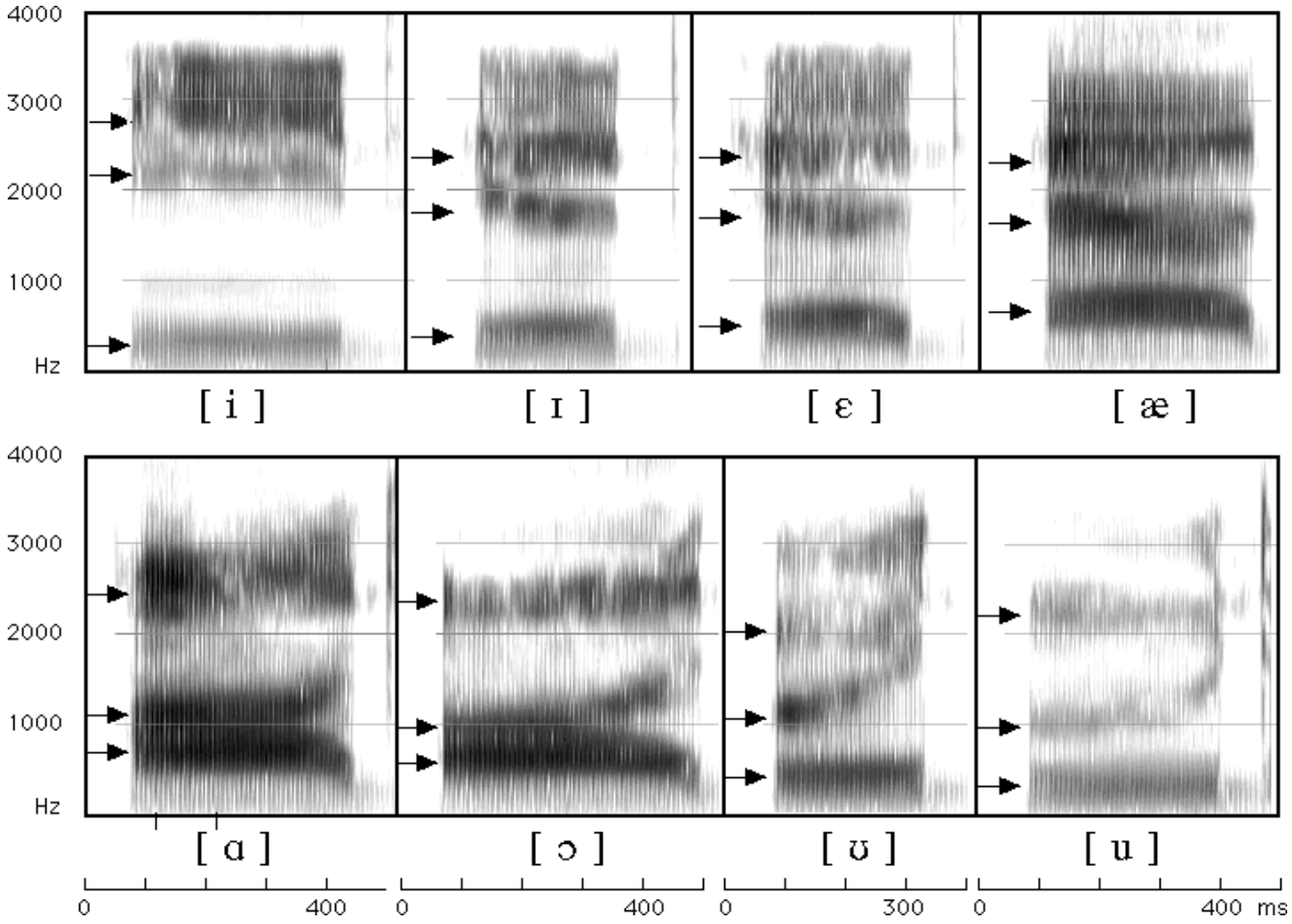


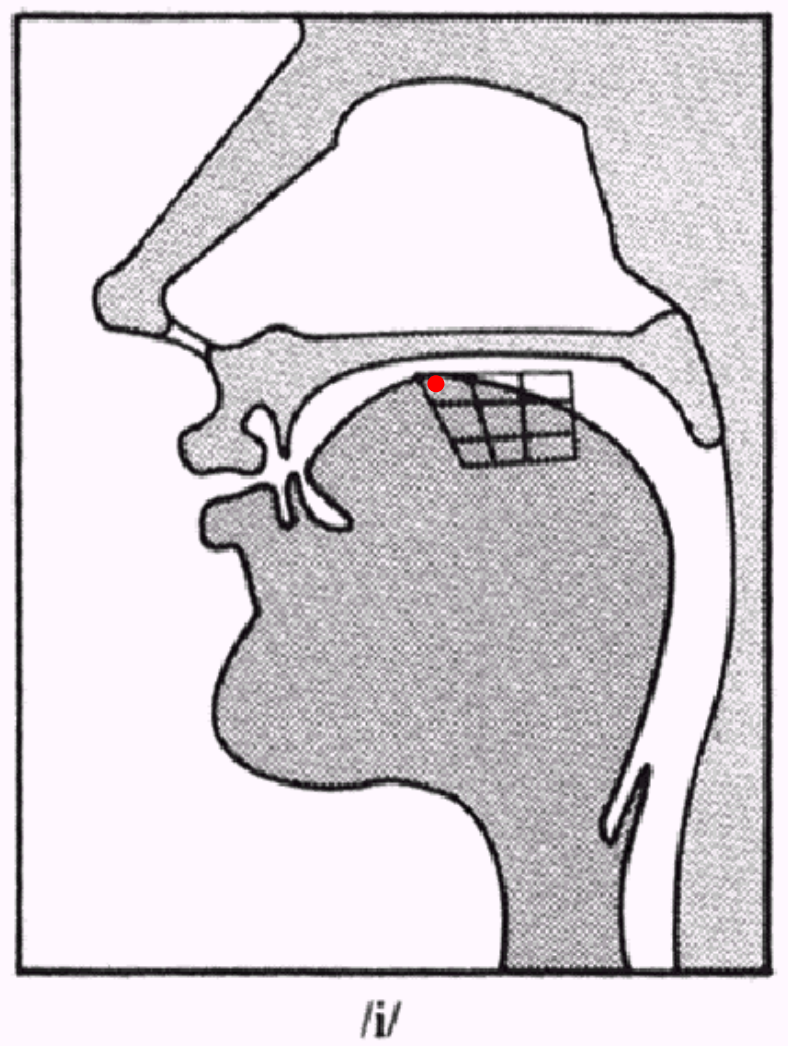
Fig. from Eric

Vowels

- Characterized by “formants”: Bands of energy
- Each vowel has 2 characteristic pitches
  - lower is 1st formant
  - higher is 2nd formant



+ [iy] vs. [uw]



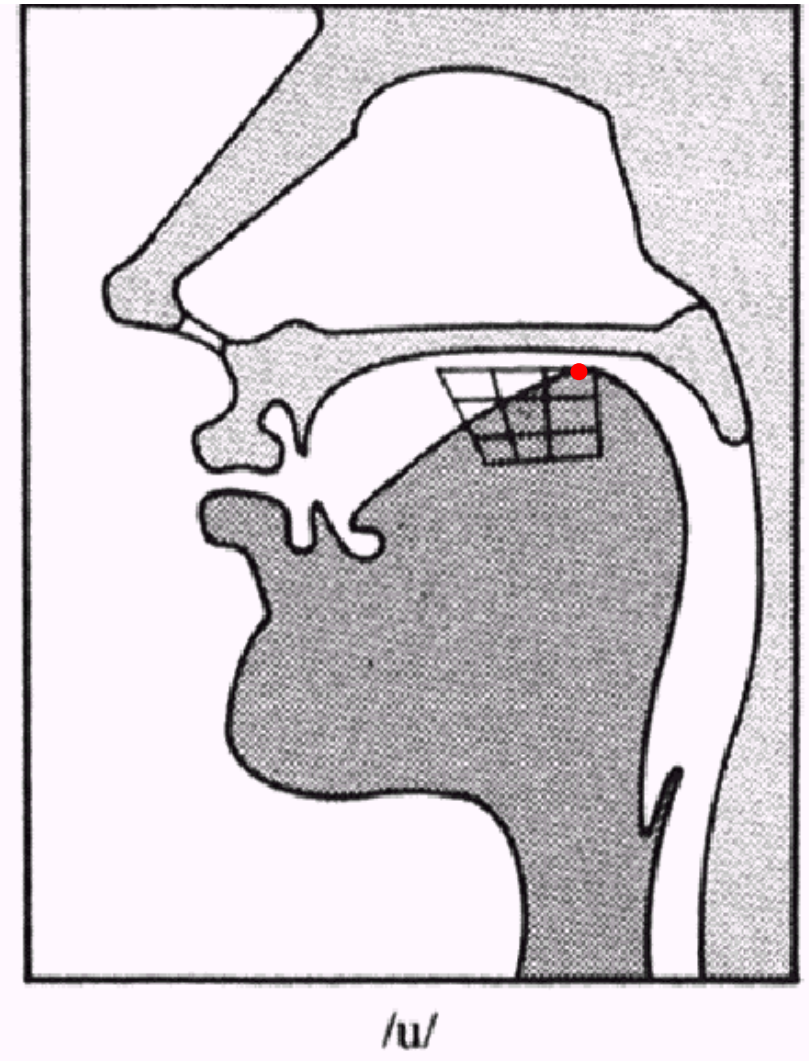


Figure from Jennifer Venditti, from a lecture given by Rochelle Newma

American English Vowel Space

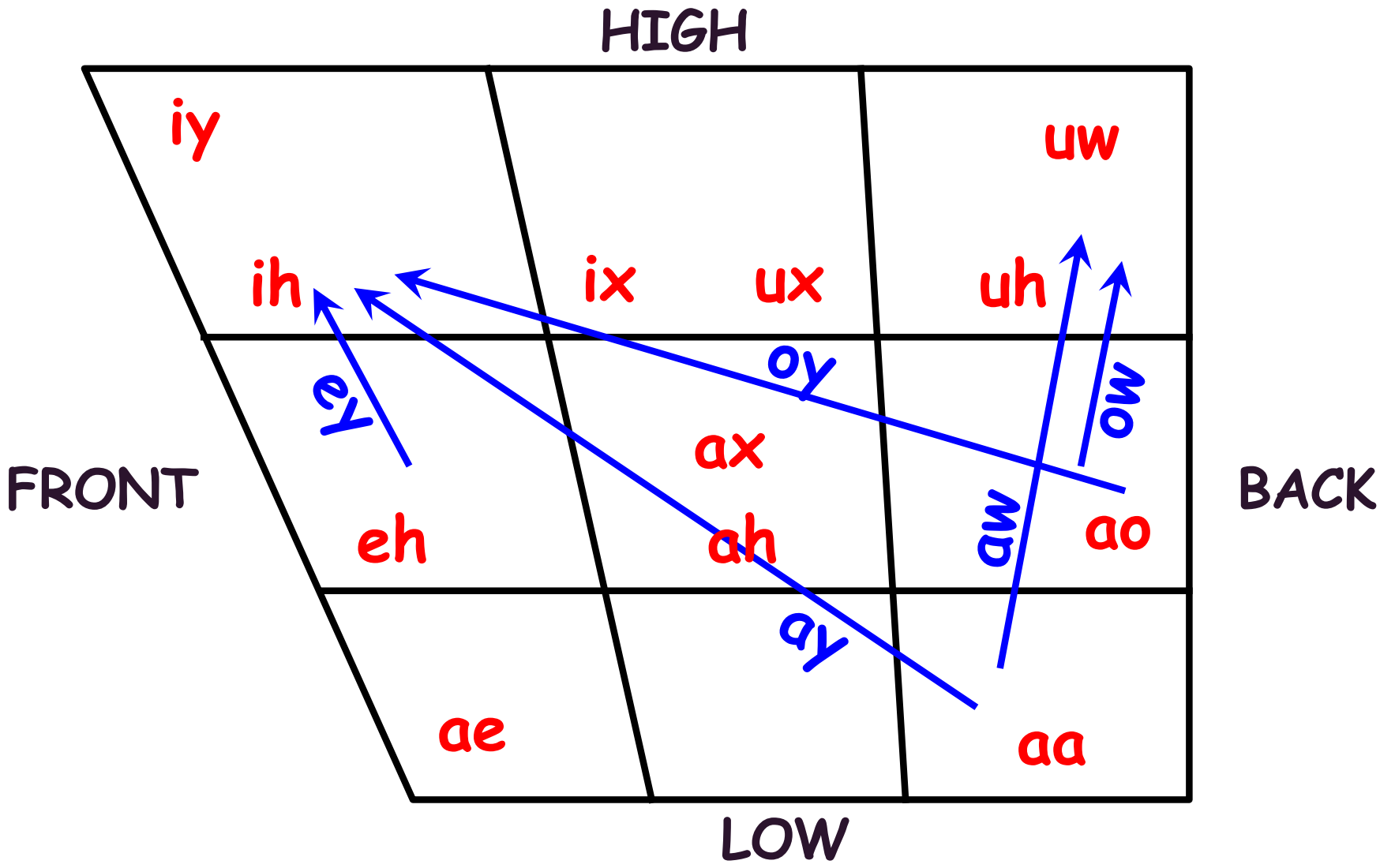
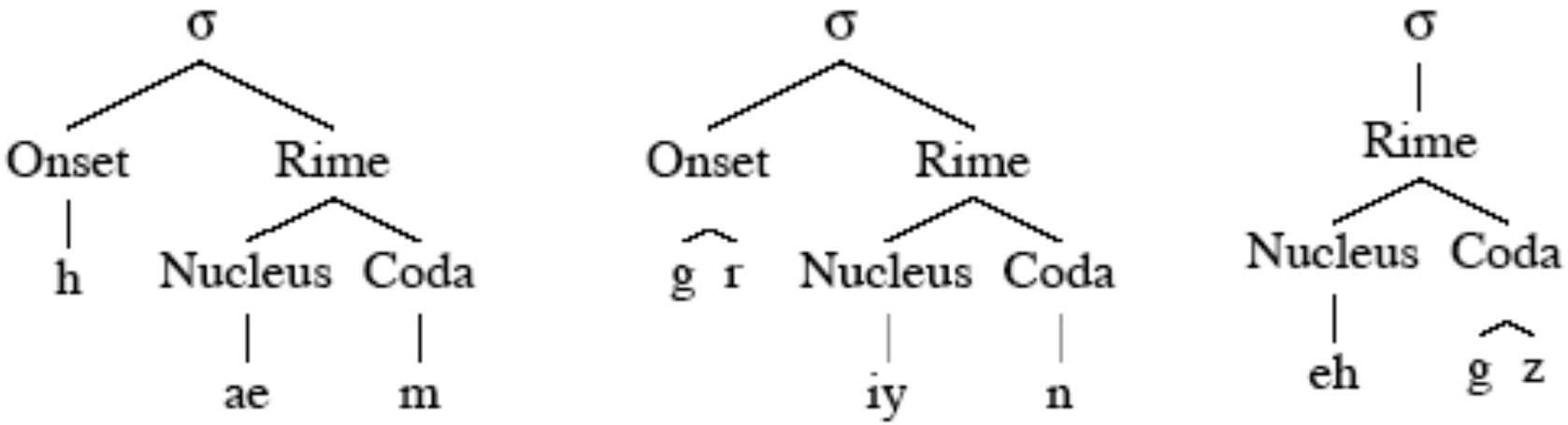


Figure from Jennifer Venditti

More phonetic structure

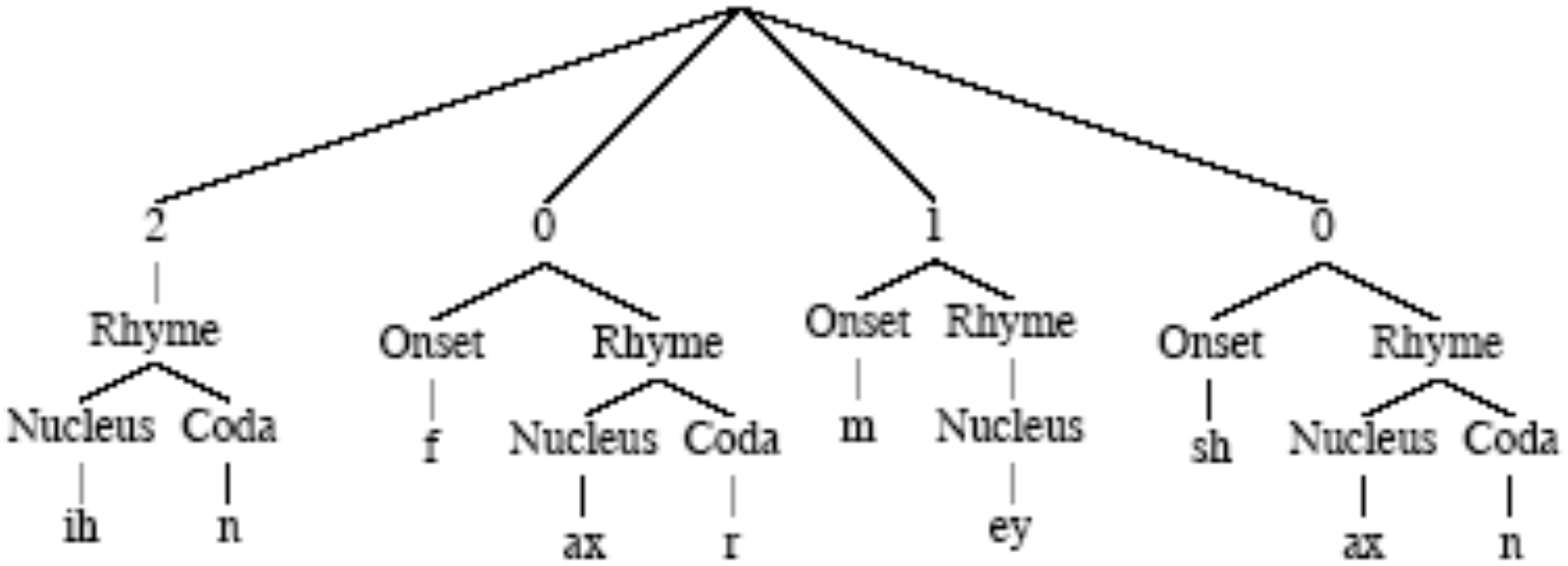
- Syllables
  - Composed of vowels and consonants. Not well defined. Something like a “vowel nucleus with some of its surrounding consonants”.



More phonetic structure

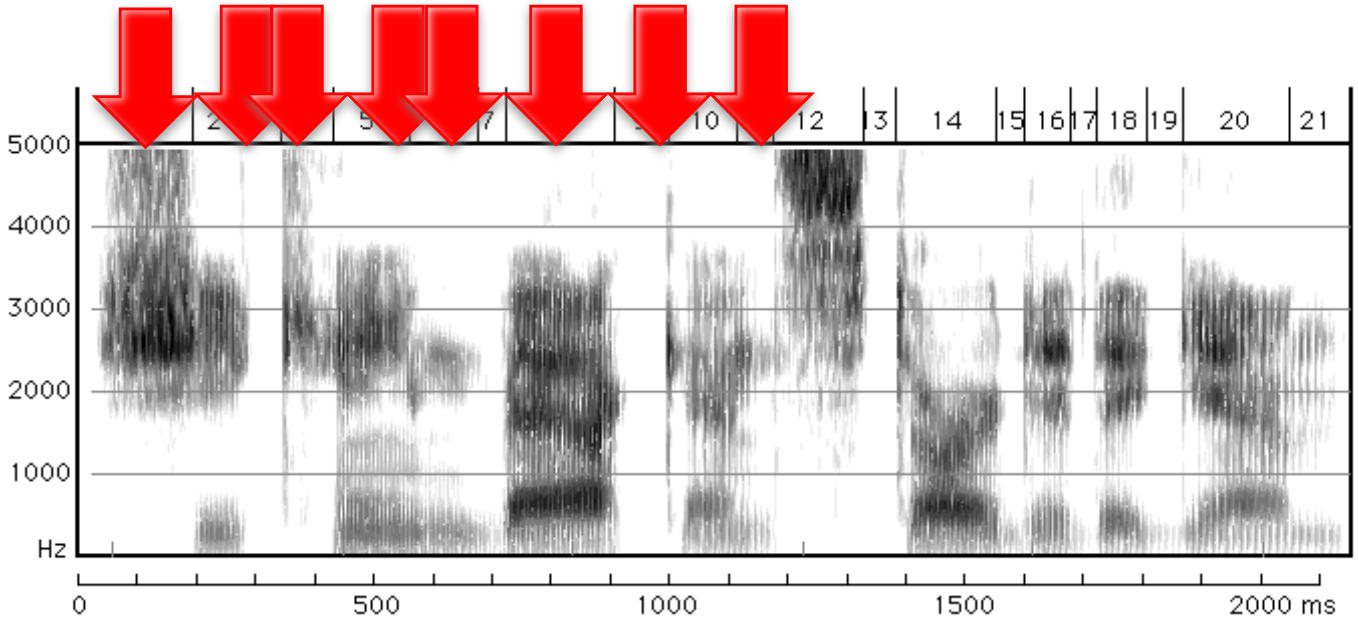
- Stress
  - Some syllables have more energy than others
  - Stressed syllables versus unstressed syllables
  - (an) ‘INsult vs. (to) in’SULT
  - (an) ‘OBject vs. (to) ob’JECT
- Simple model: every multi-syllabic word has one syllable with:
  - **“primary stress”**
    - We can represent by using the number “1” on the vowel (and an implicit unmarking on the other vowels)
    - “table”: t ey1 b ax l
    - “machine: m ax sh iy1 n
  - Also possible: “secondary stress”, marked with a “2”
    - ih-2 n f axr m ey-1 sh ax n
  - Third category: **reduced**: schwa:
    - ax

Multi syllable words



She came back and started again

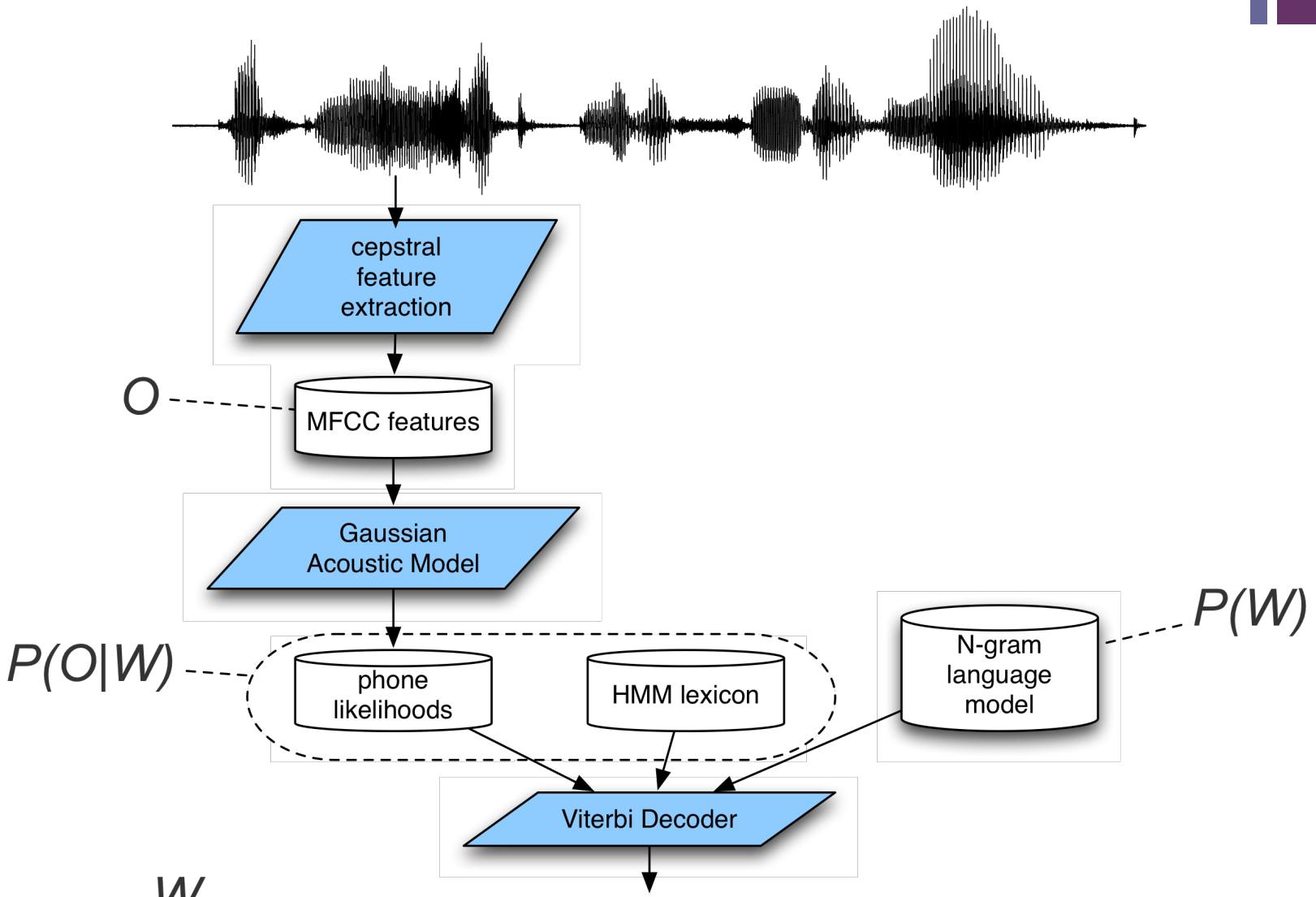
SH–IY-K-EY-M-B-AE-K-AX-N-D-S-T-AA-R-T-DX-IX-D-AX-G-EH-N



1. SH- lots of high-freq energy
3. closure for K in came
4. burst of aspiration for K
5. EY vowel;faint 1100 Hz formant is nasalization
6. bilabial nasal
8. ae; note upward transitions after bilabial stop at beginning
9. note F2 and F3 coming together for "K"
10. D is lost between N and S

From Ladefoged "A Course in Phonetics"

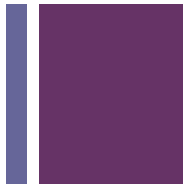
Speech Recognition Architecture



$$\hat{W} = \arg\max_{W \in L} P(O \mid W) P(W)$$

Thanks to Dan Jurafsky for these slides

ASR components

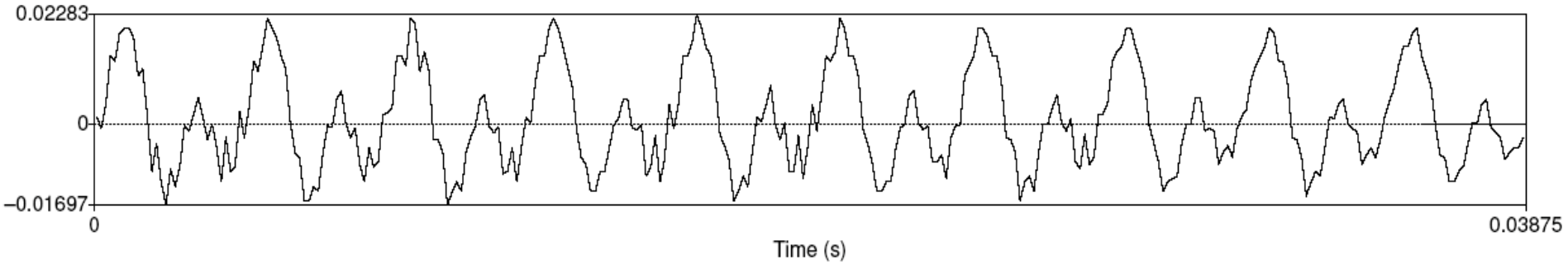


- Feature Extraction, MFCCs, start of AM
- HMMs, Forward, Viterbi,
- Baum-Welch (Forward-Backward)
- Acoustic Modeling and GMMs
- N-grams and Language Modeling
- Search and Advanced Decoding
- Dealing with Variation

Acoustic Phonetics

- Waves, sound waves, and spectra
  - (**Informally!** We'll see it with more math when we do feature extraction)
- Speech waveforms
- F0, pitch, intensity
- Spectra
  - Spectrograms
  - Formants
  - Reading spectrograms
- Resources: dictionaries and phonetically-labeled corpora

Speech sound waves

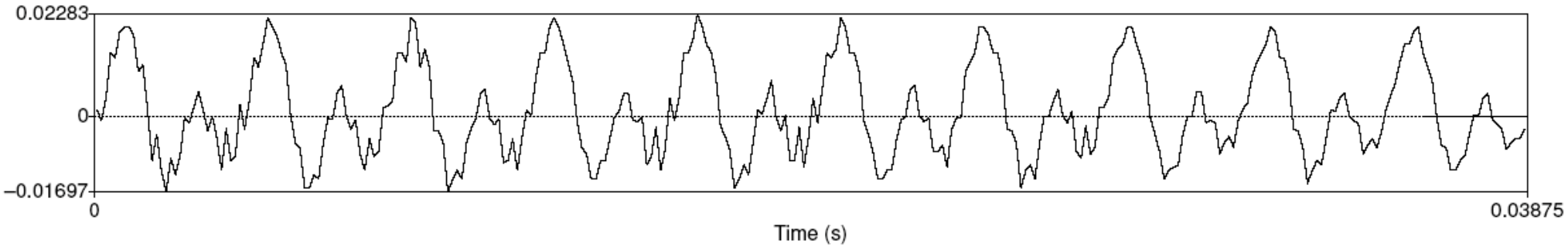


- Positive is compression
- Zero is normal air pressure,
- negative is rarefaction

- X axis: time: .03875 seconds

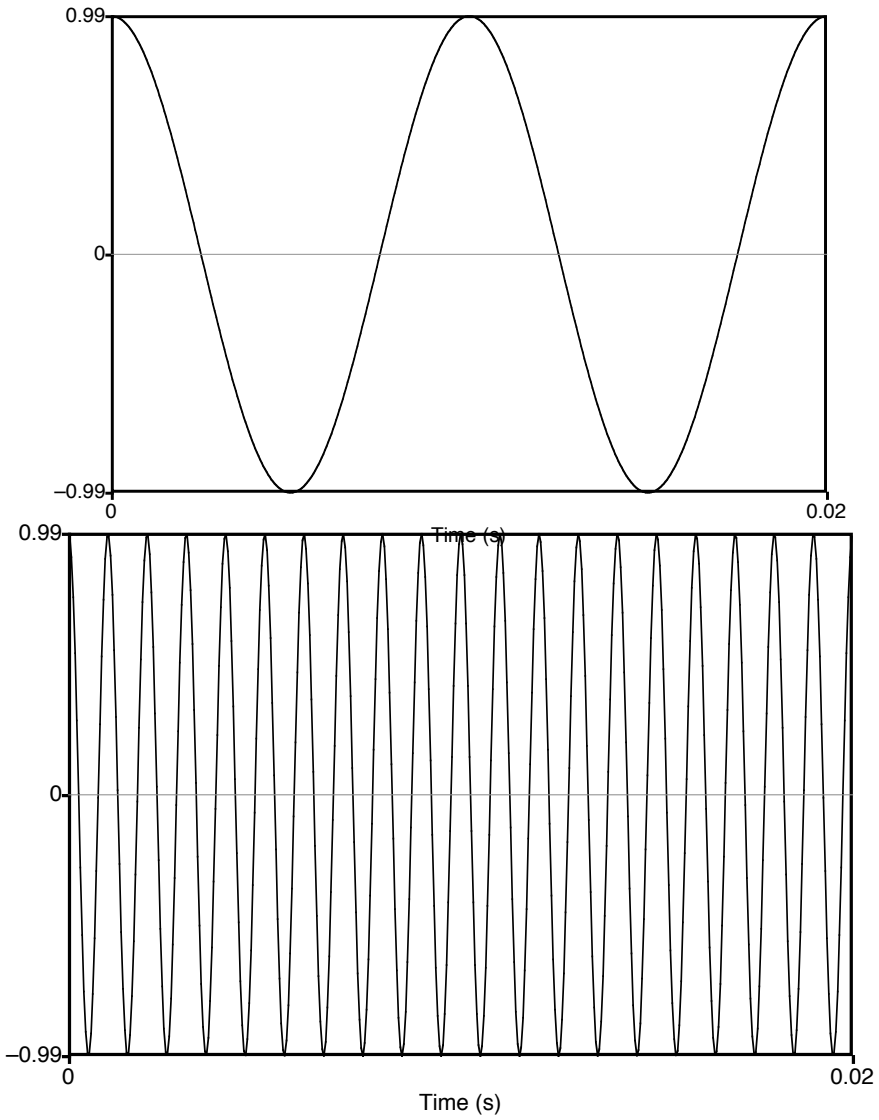
Fundamental frequency

- Waveform of the vowel [iy]



- Frequency: repetitions/second of a wave
- Above vowel has 10 reps in .03875 secs
- So freq is 10/.03875 = 258 Hz
- This is speed that vocal folds move, hence voicing
- Each peak corresponds to an opening of the vocal folds
- The frequency of the complex wave is called the fundamental frequency of the wave or F0

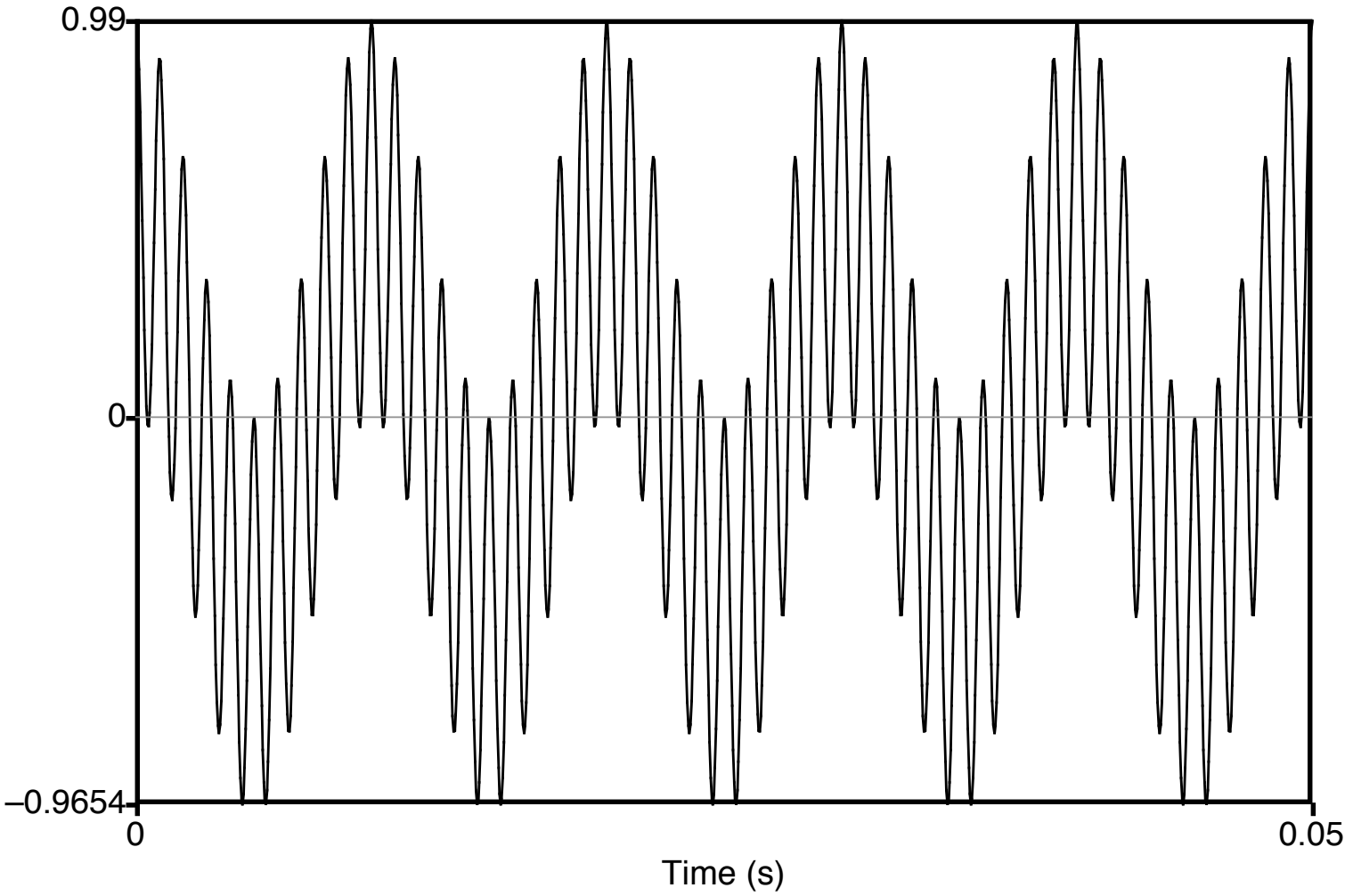
+ Waves have different frequencies



100 Hz

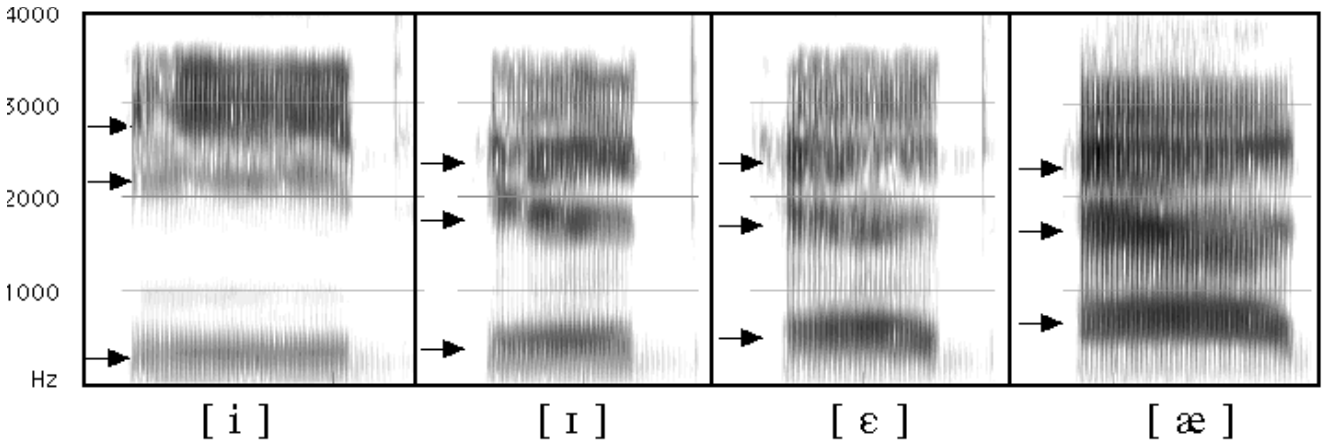
1000 Hz

Complex waves: Adding a 100 Hz and 1000 Hz wave together



+ Spectral characteristics of vowels

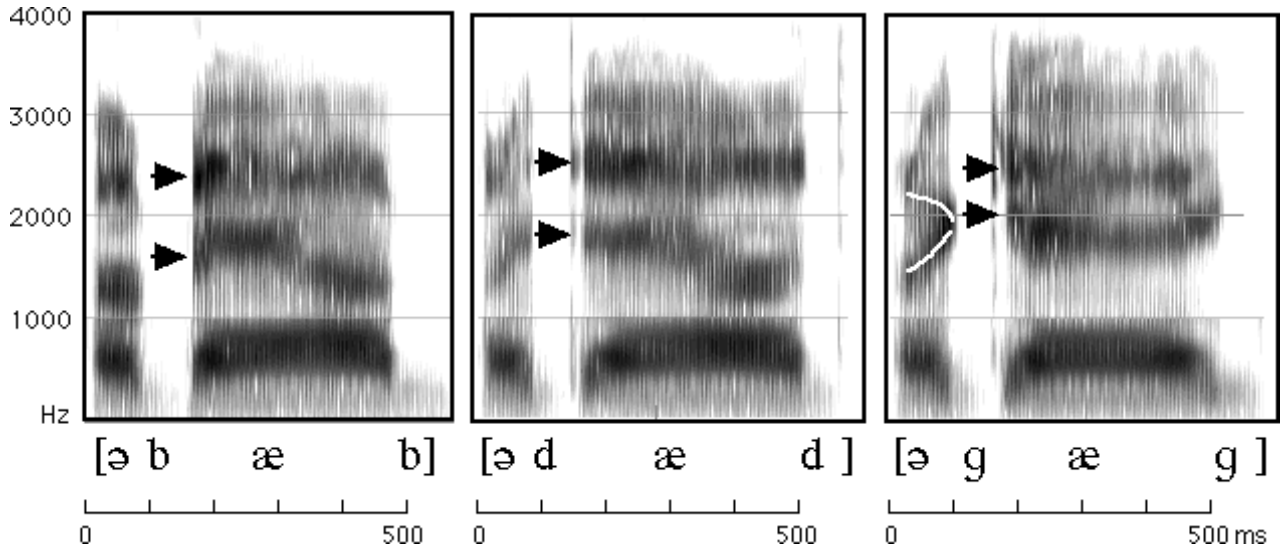
- Any body of air will vibrate in a way that depends on its size and shape of its container
  - Air in vocal tract is set in vibration by action of vocal cords.
    - Every time the vocal cords open and close, pulse of air from the lungs, acting like sharp taps on air in vocal tract
    - Setting resonating cavities into vibration so produce a number of different frequencies.
  - Vocal tract as "amplifier"; amplifies different frequencies
- Formants are result of different shapes of vocal tract.



Again: why is a speech sound wave composed of these peaks?

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

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 Phor

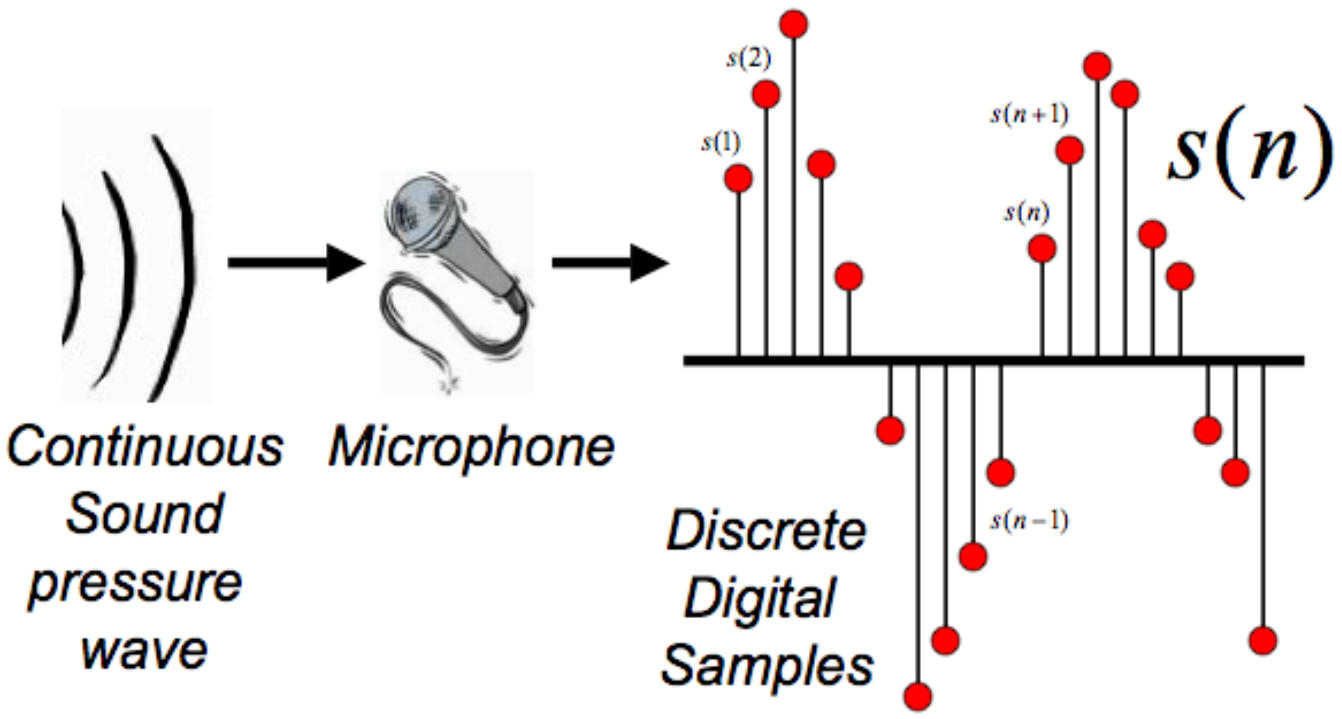
Front End Processing

- To go from a continuous analog signal to a tractable number of values that represent the features **most important to distinguishing speech sounds**
- Multiple signal processing algorithms in a sequence to foreground important distinctions and background unimportant ones
- Ending up with a structure that is usable in a HMM model

Thanks to Dan Jurafsky for these slides

Digitizing Speech

- Analog-to-digital conversion
- Or A-D conversion.
- Two steps
  - Sampling
  - Quantization



Thanks to Bryan Pellom for this slide!

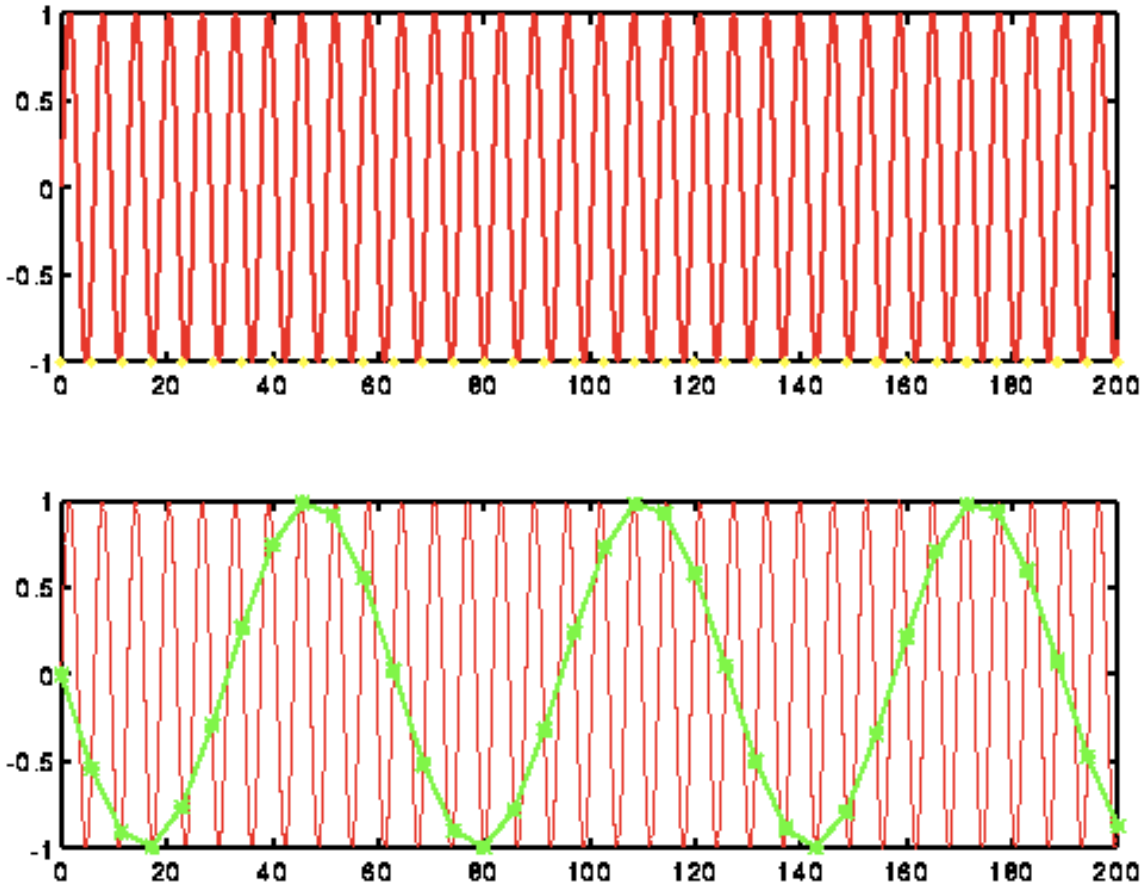
Sampling

- Measuring amplitude of a signal at time *t*
- The sample rate needs to have at least two samples for each cycle
  - One for the positive, and one for the negative half of each cycle
  - More than two samples per cycle is ok
  - Less than two samples will cause frequencies to be missed
- So the maximum frequency that can be measured is one that is half the sampling rate.
- The maximum frequency for a given sampling rate called Nyquist frequency

Sampling

If measure at green dots, will see a lower frequency wave and miss the correct higher frequency one!

Original signal in red:

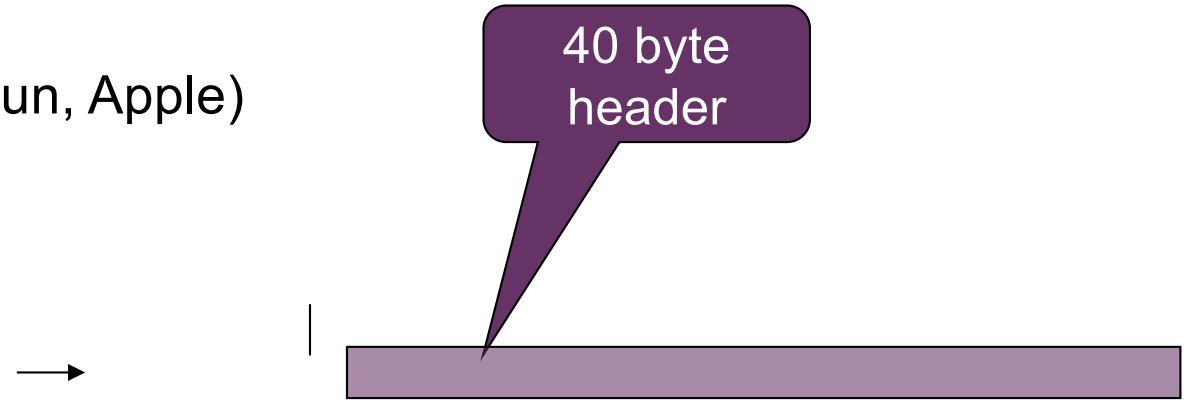


Sampling

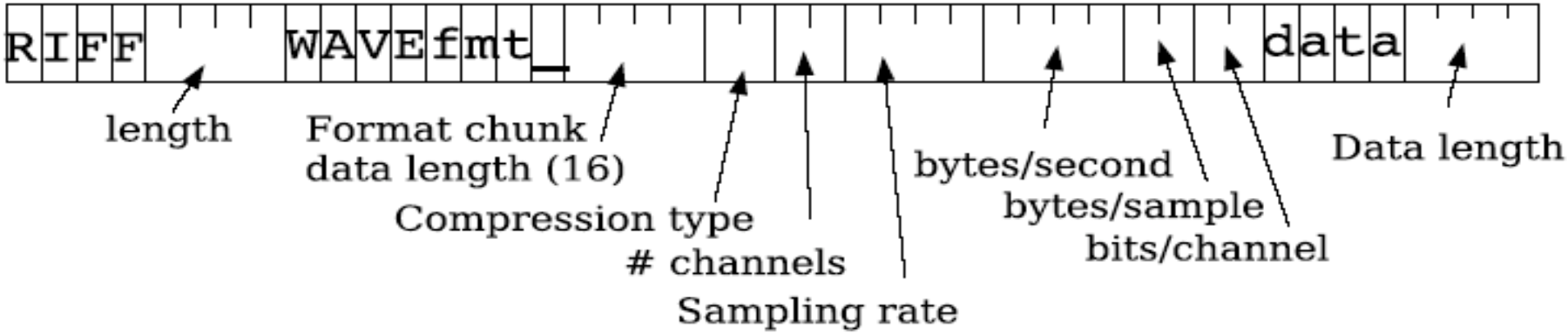
- In practice we use the following sample rates
  - 16,000 Hz (samples/sec), for microphones, "wideband"
  - 8,000 Hz (samples/sec) Telephone
- Why?
  - Need at least 2 samples per cycle
  - Max measurable frequency is half the sampling rate
  - Human speech < 10KHz, so need max 20K
  - Telephone is filtered at 4K, so 8K is enough.

Quantization

- Quantization
  - Representing real value of each amplitude as integer
  - 8-bit (-128 to 127) or 16-bit (-32768 to 32767)
- Formats:
  - 16 bit PCM
  - 8 bit mu-law; log compression
- Byte order
  - LSB (Intel) vs. MSB (Sun, Apple)
- Headers:
  - Raw (no header)
  - Microsoft wav
  - Sun .au

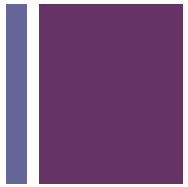


WAV format



(Open a .wav form in a text editor and you will see this)

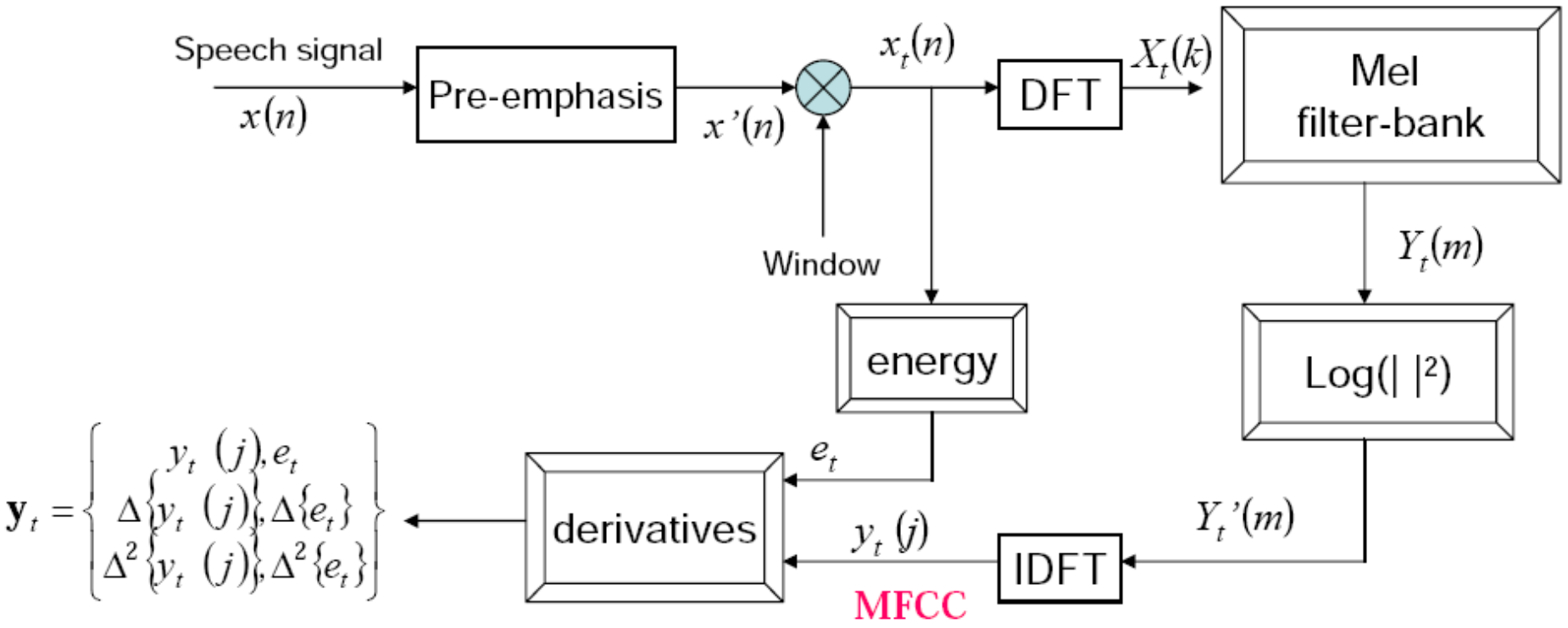
Manipulating audio



- Nice sound manipulation tool: sox.
  - change sampling rate
  - convert speech formats
  - Check out where in Kaldi sox is used

MFCC

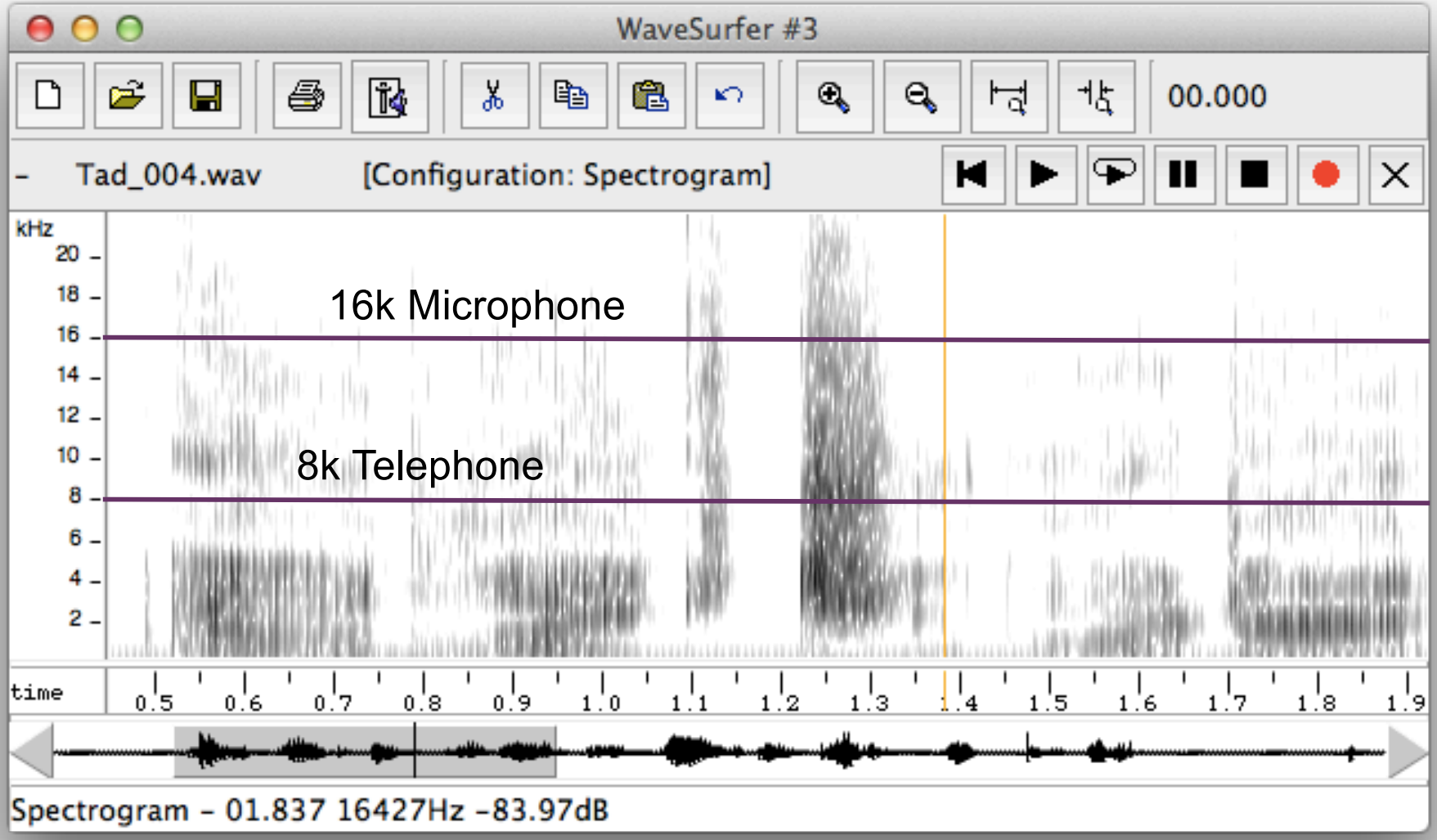
- Mel-Frequency Cepstral Coefficient (MFCC)
  - Most widely used spectral representation in ASR



Pre-Emphasis

- Pre-emphasis: boosting the energy in the high frequencies
- Q: Why do this?
- A: The spectrum for voiced segments has more energy at lower frequencies than higher frequencies.
  - This is called **spectral tilt**
  - Spectral tilt is caused by the nature of the glottal pulse
- Boosting high-frequency energy gives more info to Acoustic Model
  - Improves phone recognition performance

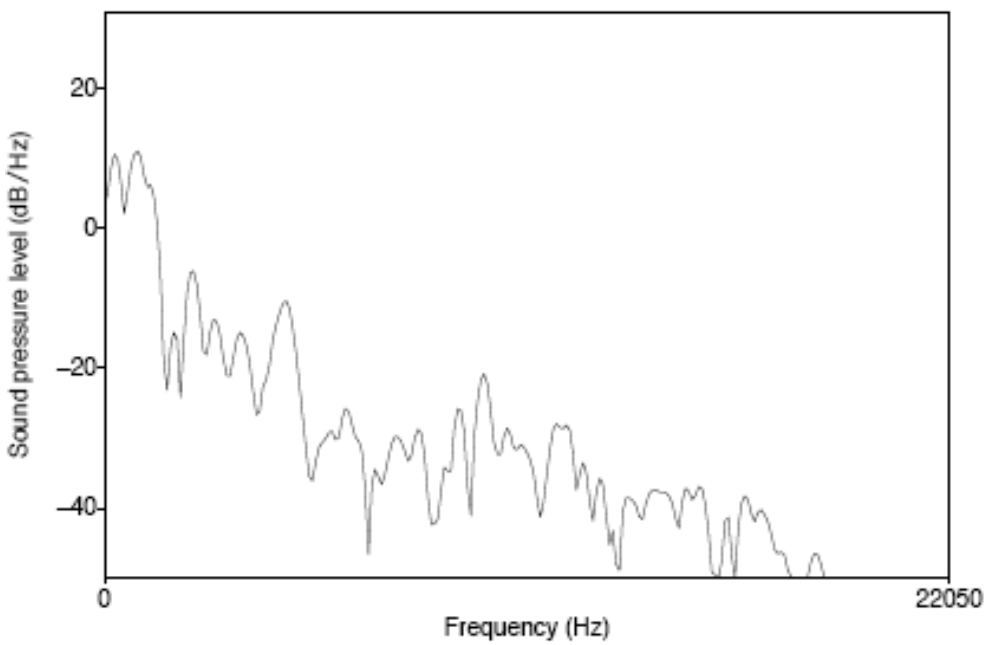
More energy at lower frequencies than higher frequencies

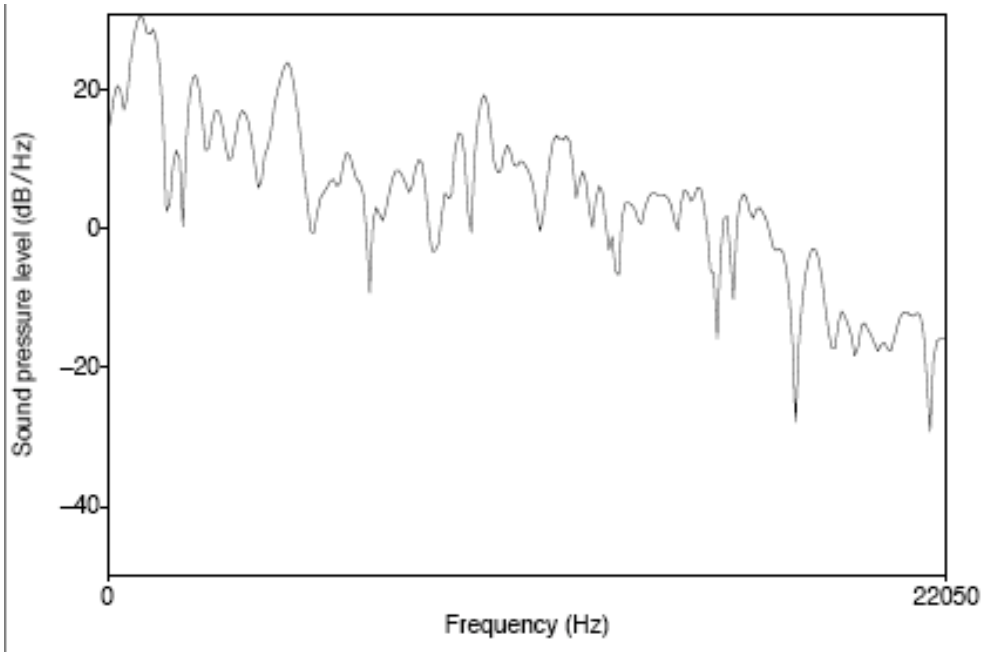


Thanks to Dan Jurafsky for these slides

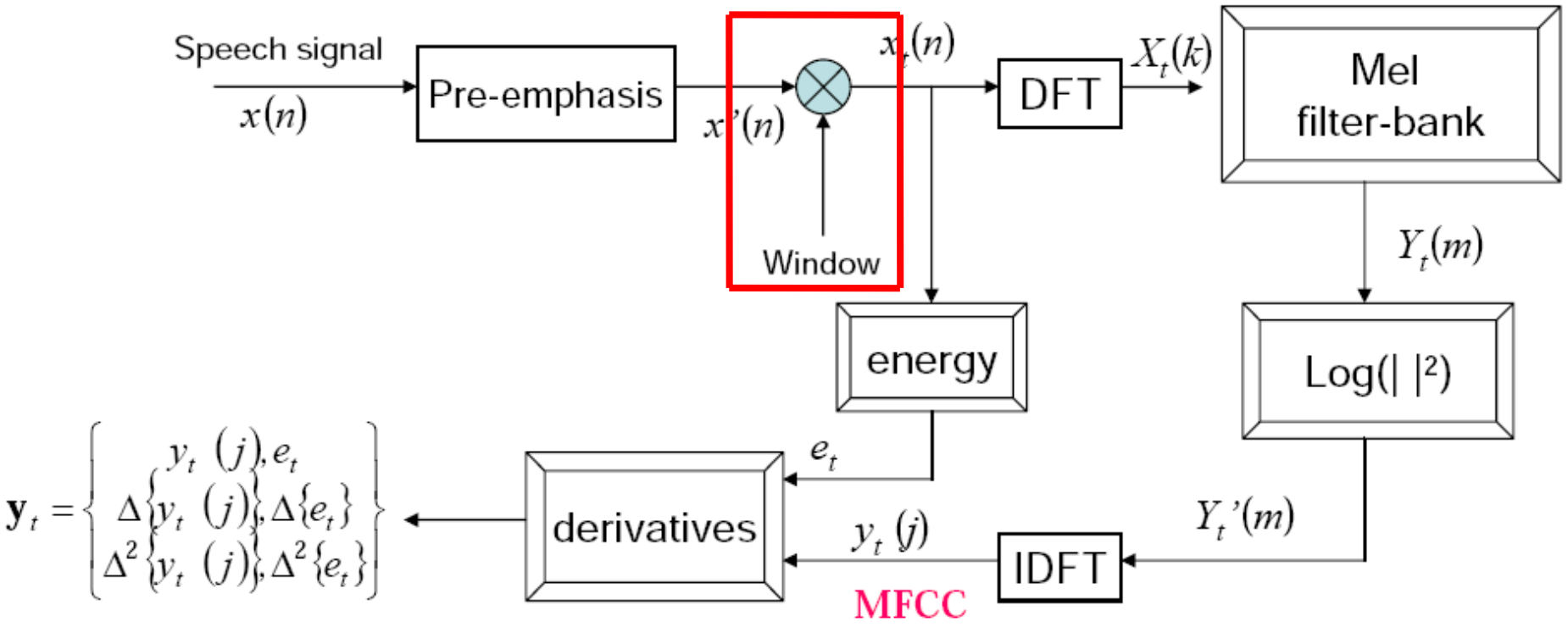
Example of pre-emphasis

- Before and after pre-emphasis
  - Spectral slice from the vowel [aa]

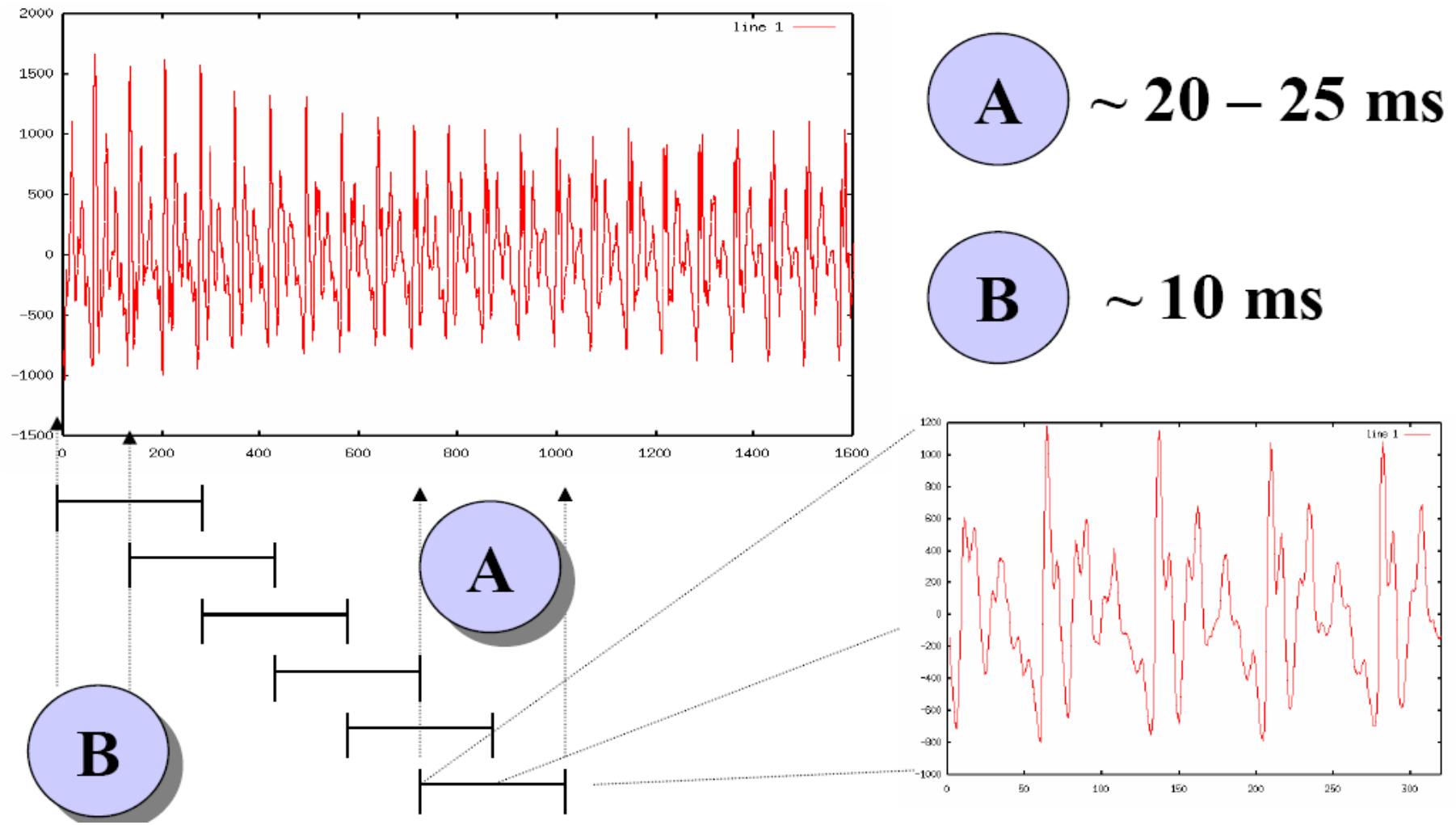




+ MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
MFCC
energy
e_t
derivatives
$\mathbf{y}_t = \left\{ \begin{matrix} y_t(j), e_t \\ \Delta\{y_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t(j)\}, \Delta^2\{e_t\} \end{matrix} \right\}$

Windowing: “Observations” are successive overlapping frames



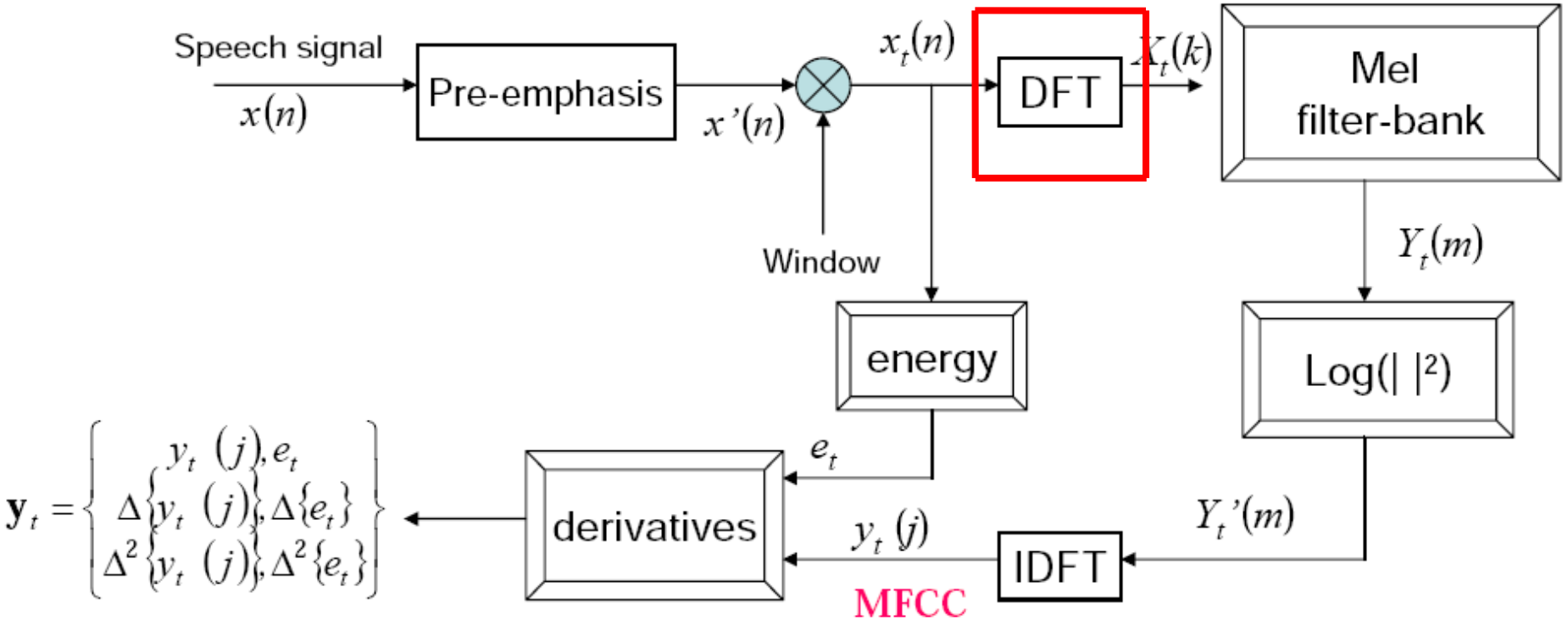
Slide from Bryan Pellom

Windowing

- Why divide speech signal into successive overlapping frames?
  - Speech is not a stationary signal; we want information about a small enough region that the spectral information is a useful cue.

- Frames
  - Frame size: typically, 10-25ms
  - Frame shift: the length of time between successive frames, typically, 5-10ms

MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
$x_t(n)$
DFT
$X_t(k)$
Mel filter-bank
$Y_t(m)$
Window
energy
Log(| |²)
e_t
$Y_t'(m)$
$\mathbf{y}_t = \begin{Bmatrix} y_t(j), e_t \\ \Delta\{y_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t(j)\}, \Delta^2\{e_t\} \end{Bmatrix}$
derivatives
$y_t(j)$
IDFT
MFCC

Discrete Fourier Transform

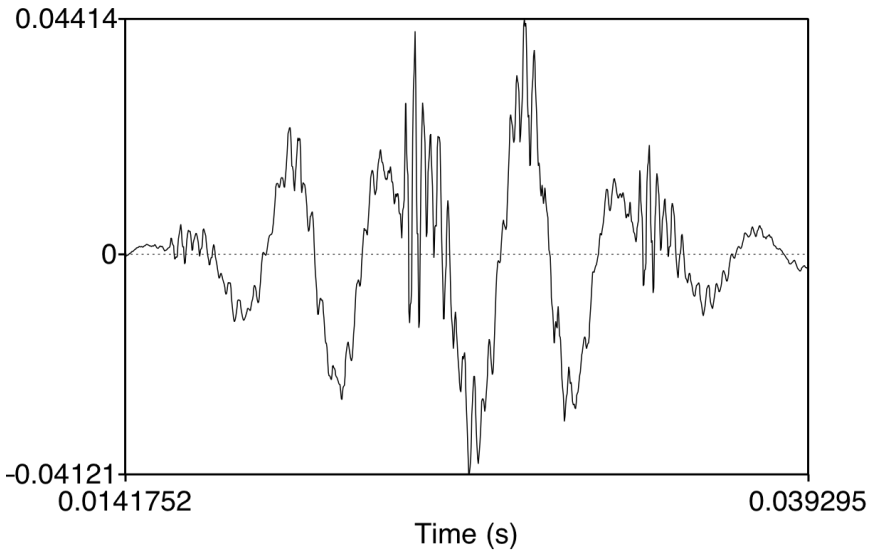
- Input:
  - Windowed signal x[n]…x[m]
- Output:
  - For each of N discrete frequency bands
  - A complex number X[k] representing magnitude and phase of that frequency component in the original signal
- Discrete Fourier Transform (DFT)

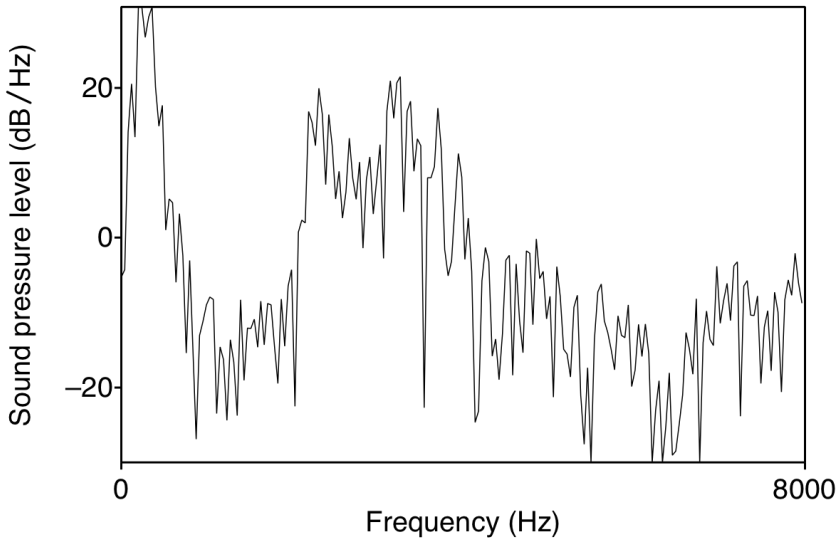
$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\frac{\pi}{N}kn}$$

- Standard algorithm for computing DFT:
  - Fast Fourier Transform (FFT) with complexity N*log(N)
  - In general, choose N=512 or 1024

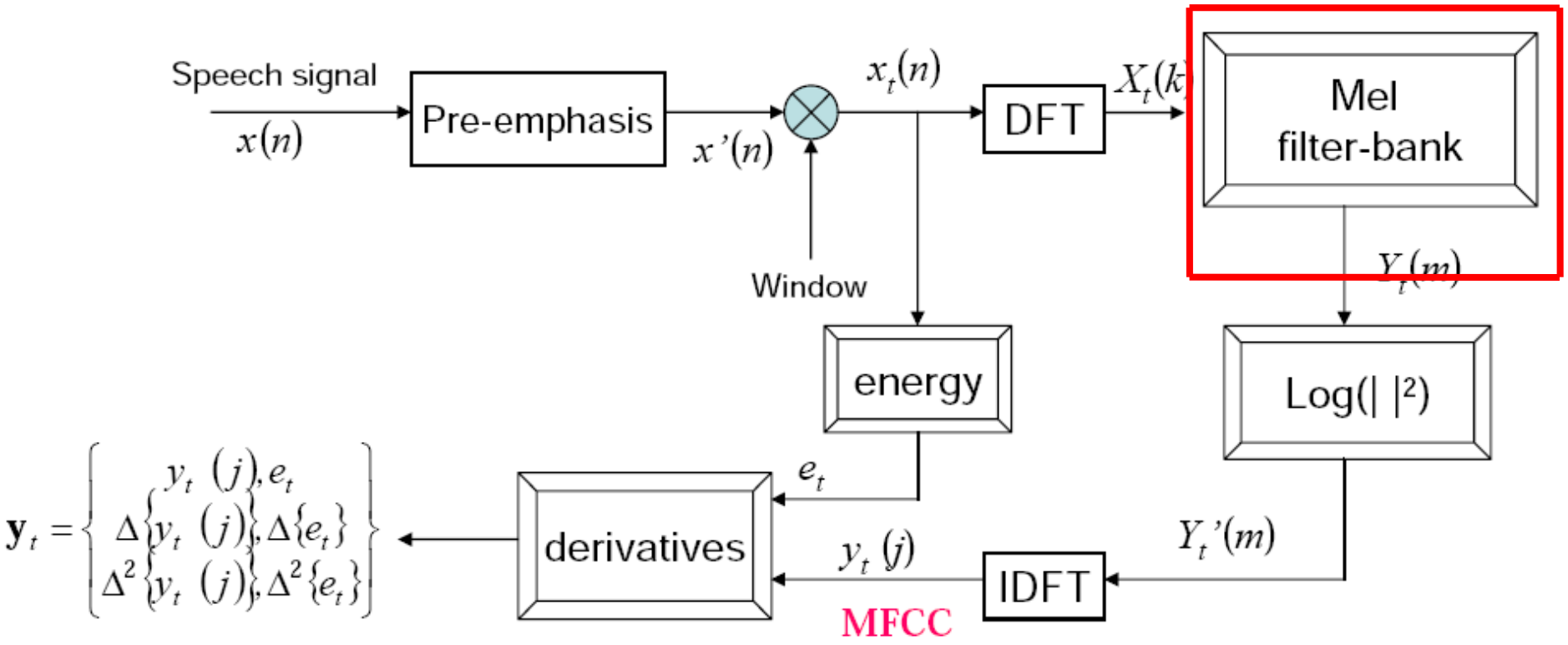
Discrete Fourier Transform computing a spectrum

- A 25 ms Hamming-windowed signal from [iy]
  - And its spectrum as computed by DFT (plus other smoothing)





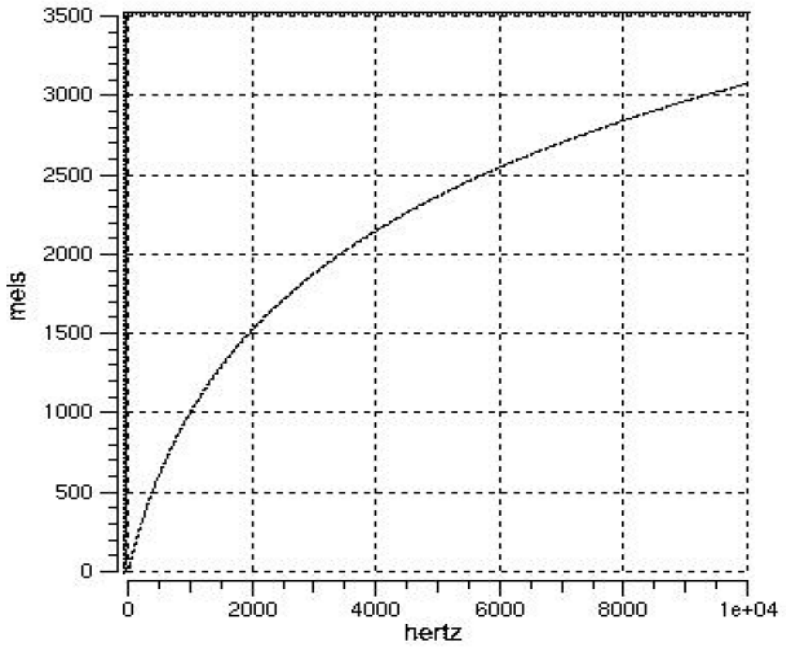
+ MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
energy
e_t
derivatives
$\mathbf{y}_t = \left\{ \begin{matrix} y_t(j), e_t \\ \Delta\{y_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t(j)\}, \Delta^2\{e_t\} \end{matrix} \right\}$
MFCC

Mel-scale

- Human hearing is not equally sensitive to all frequency bands
- Less sensitive at higher frequencies, roughly > 1000 Hz

I.e. human perception of frequency is non-linear:



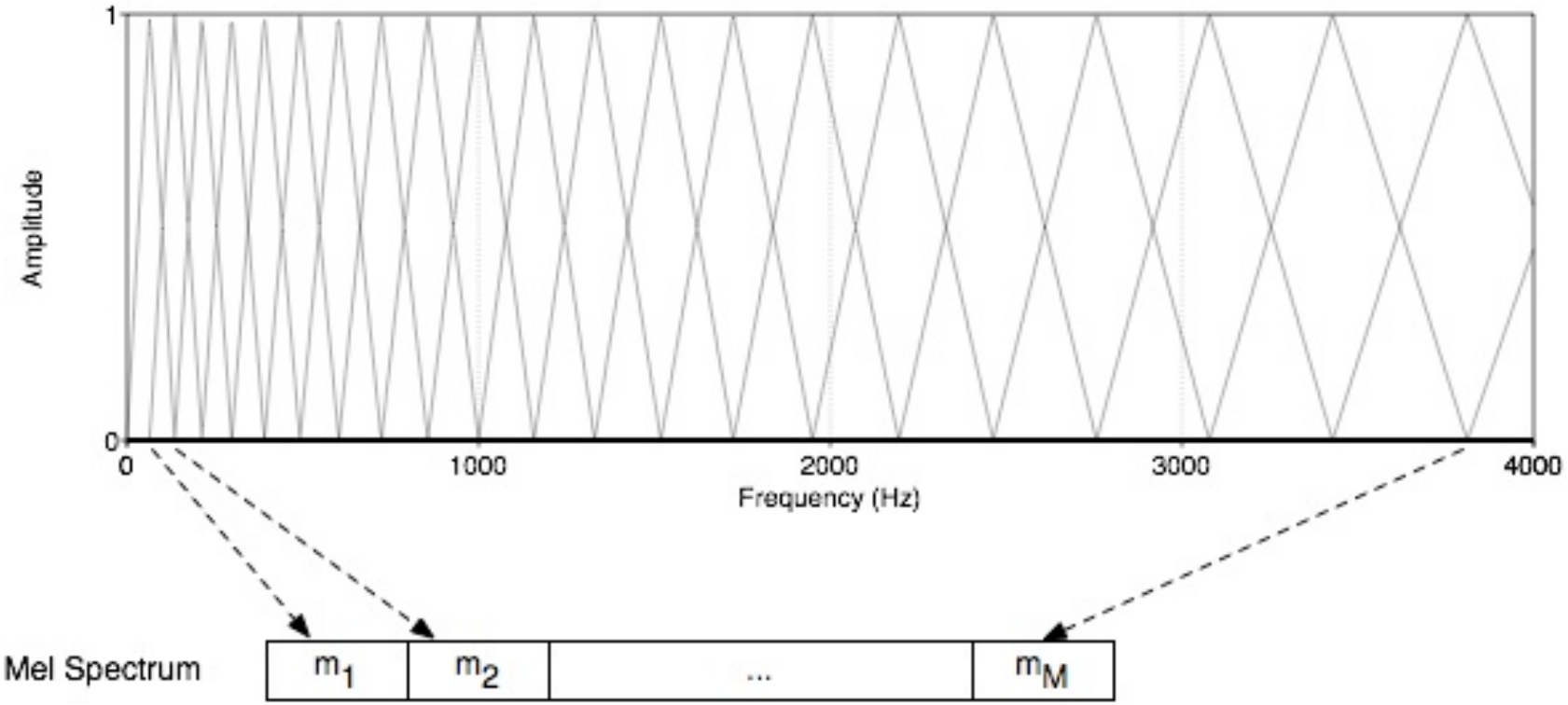
Mel-scale

- A mel is a unit of pitch
  - Definition:
    - Pairs of sounds perceptually equidistant in pitch
    - Are separated by an equal number of mels:
- Mel-scale is approximately linear below 1 kHz and logarithmic above 1 kHz
- Definition:

$$Mel(f) = 2595 \log_{10}(1 + \frac{f}{700})$$

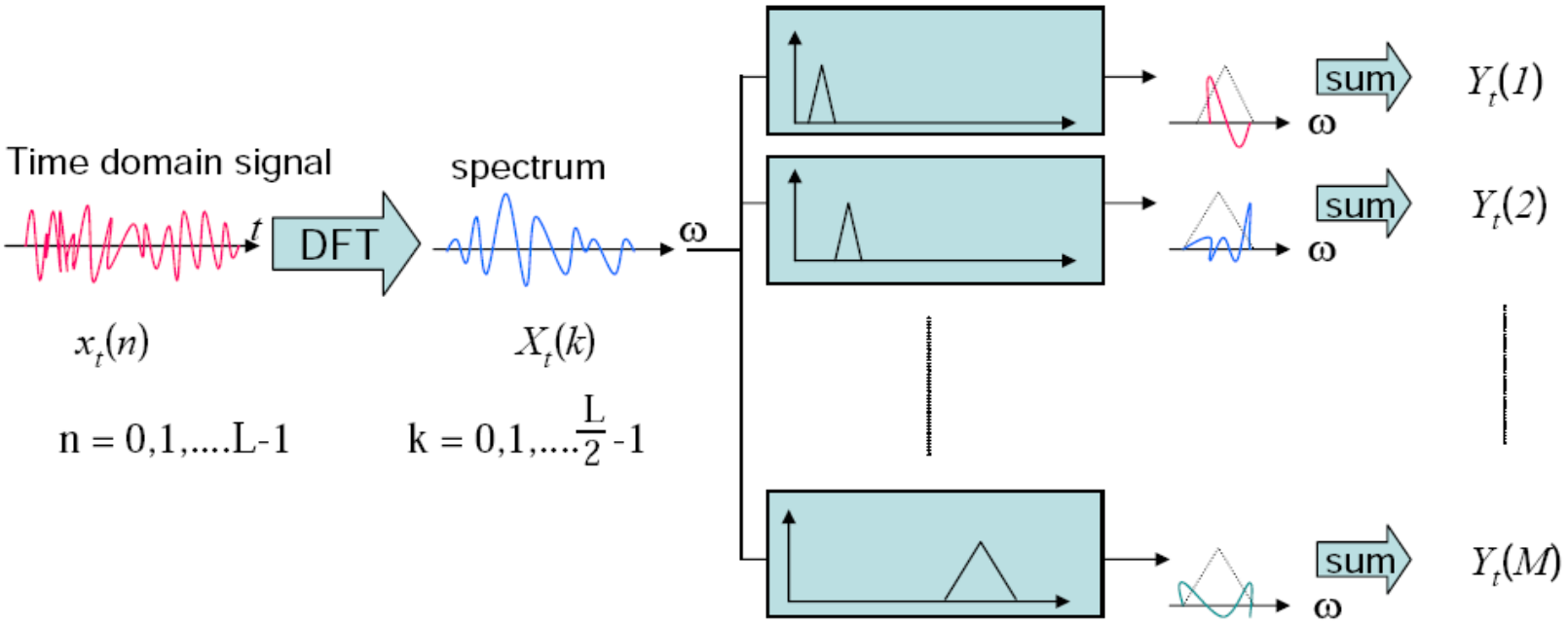
Mel Filter Bank Processing

- Mel Filter bank
  - Uniformly spaced before 1 kHz
  - logarithmic scale after 1 kHz

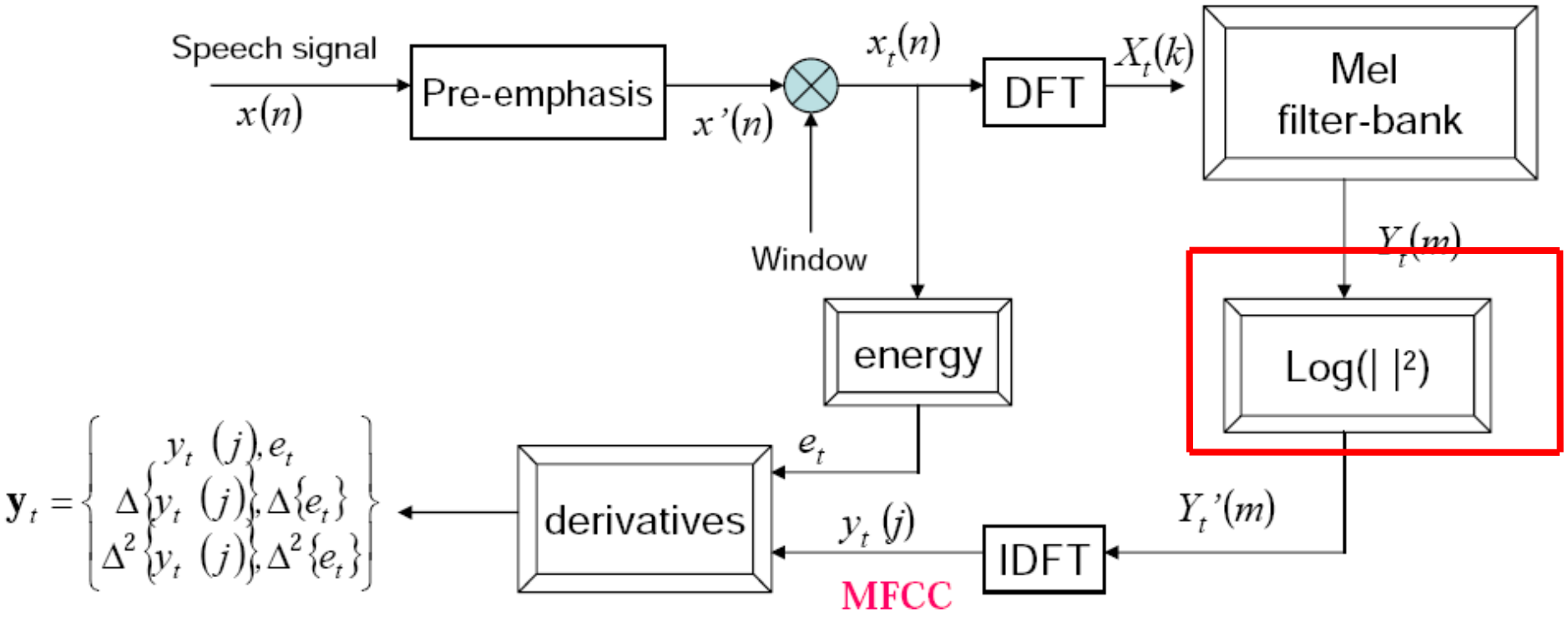


Mel-filter Bank Processing

- Apply the bank of filters according Mel scale to the spectrum
- Each filter output is the sum of its filtered spectral components

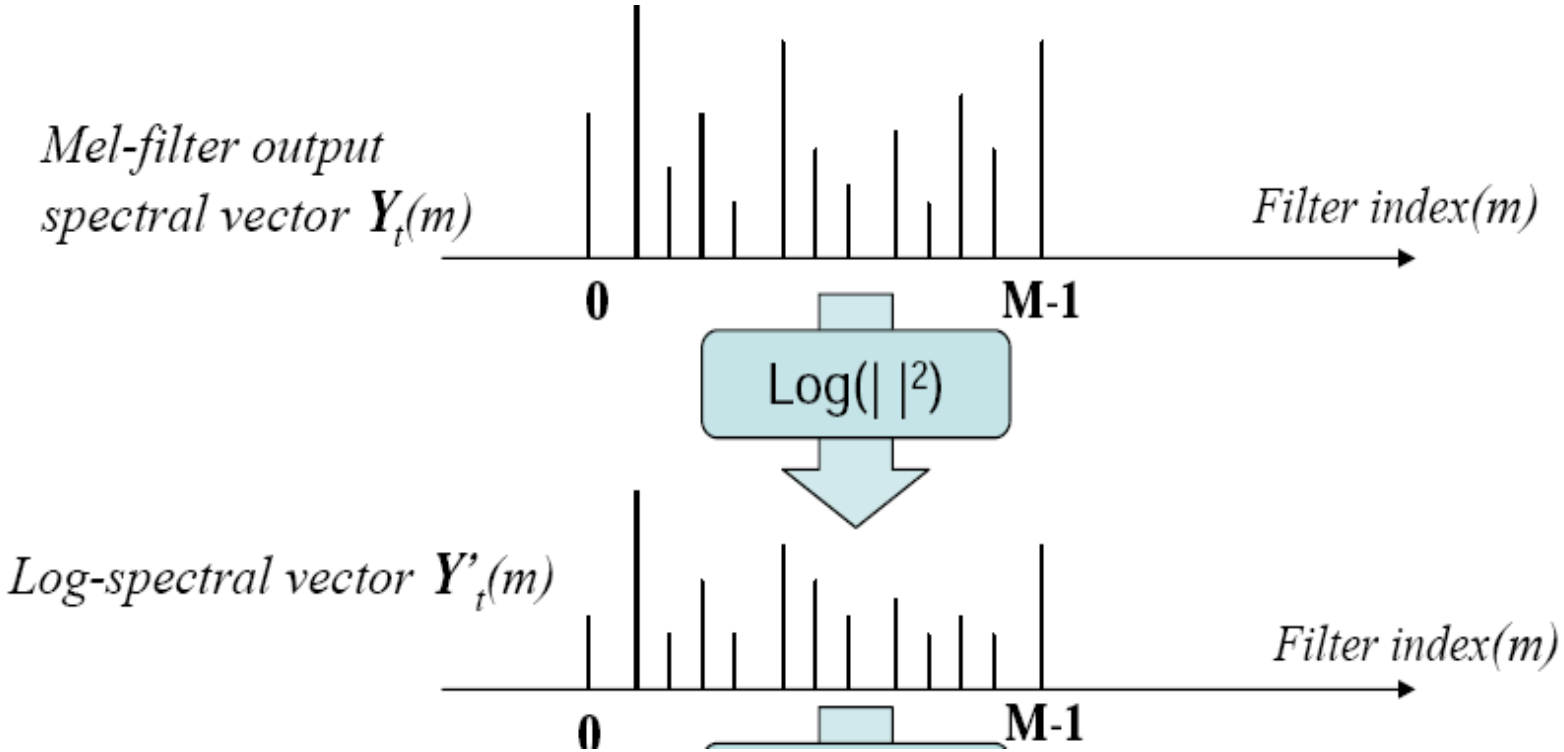


MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
MFCC
energy
e_t
derivatives
$\mathbf{y}_t = \begin{Bmatrix} y_t(j), e_t \\ \Delta\{y_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t(j)\}, \Delta^2\{e_t\} \end{Bmatrix}$

Log energy computation

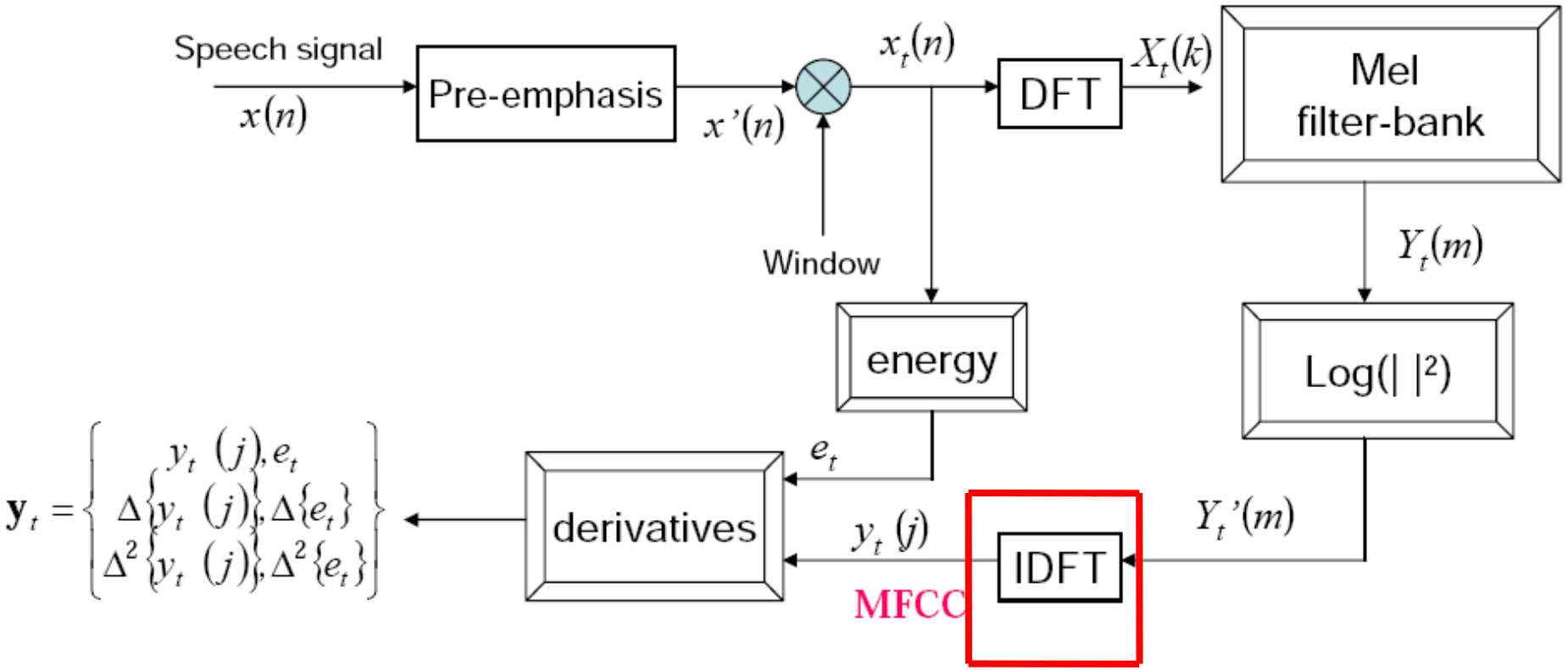
- Compute the logarithm of the square magnitude of the output of Mel-filter bank



Log energy computation

- Why log energy?
- Logarithm compresses dynamic range of values
  - Human response to signal level is logarithmic
  - Humans are less sensitive to slight differences in amplitude at high amplitudes than low amplitudes
- Makes frequency estimates less sensitive to slight variations in input (power variation due to speaker's mouth moving closer to mike)

MFCC


Speech signal
$x(n)$
Pre-emphasis
$x'(n)$
Window
$x_t(n)$
DFT
$X_t(k)$
Mel
filter-bank
$Y_t(m)$
Log(| |²)
$Y_t'(m)$
IDFT
$y_t(j)$
MFCC
energy
e_t
derivatives
$\mathbf{y}_t = \left\{ \begin{matrix} y_t(j), e_t \\ \Delta\{y_t(j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t(j)\}, \Delta^2\{e_t\} \end{matrix} \right\}$

The Cepstrum

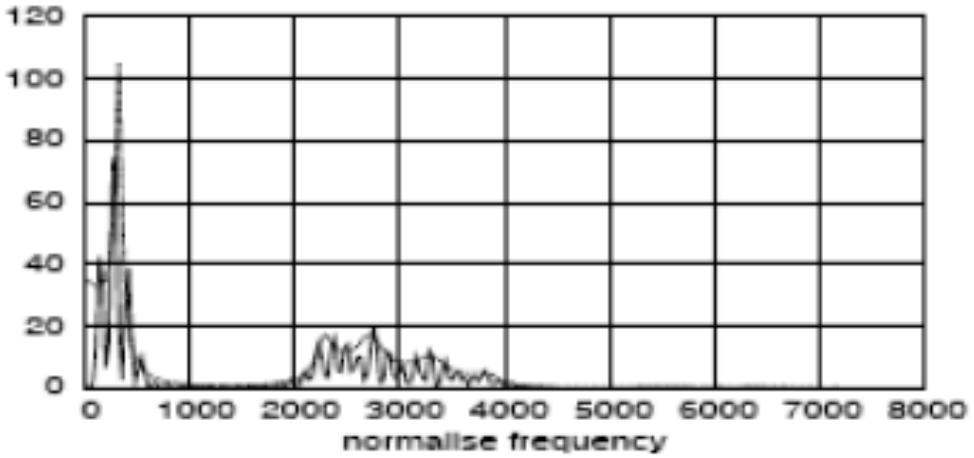
- One way to think about this
  - Separating the **source** and **filter**
  - Speech waveform is created by
    - A glottal source waveform
    - Passes through a vocal tract which because of its shape has a particular filtering characteristic
- Articulatory facts:
  - The vocal cord vibrations create harmonics
  - The mouth is an amplifier
  - Depending on shape of oral cavity, some harmonics are amplified more than others

We care about the filter not the source

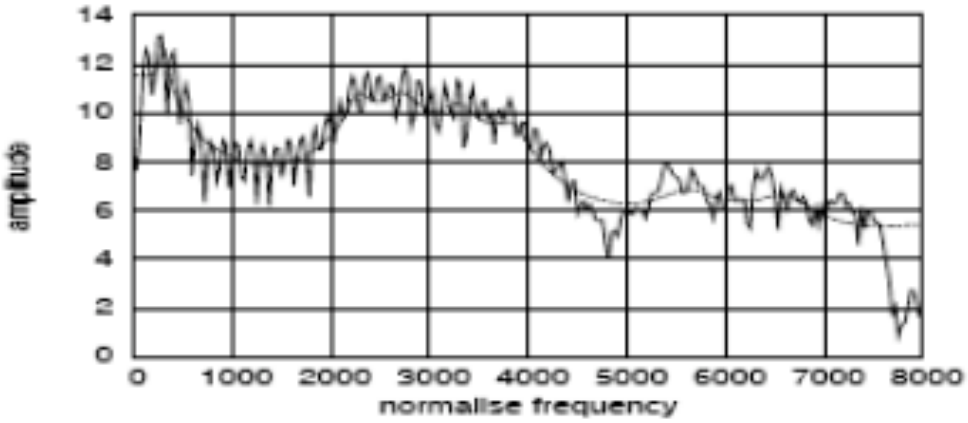
- Most characteristics of the source
  - F0
  - Details of glottal pulse
  - Don't matter for phone detection
- What we care about is the **filter**
  - The exact position of the articulators in the oral tract
- So we want a way to separate these
  - And use only the filter function

The Cepstrum

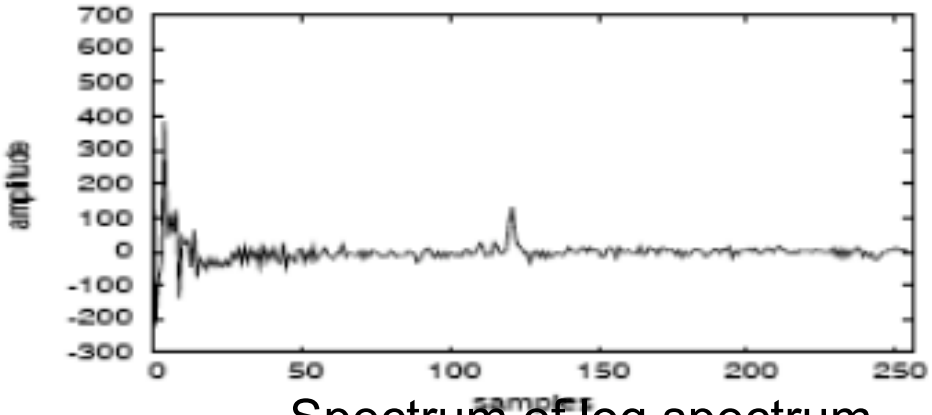
- The spectrum of the log of the spectrum



Spectrum



Log spectrum



Spectrum of log spectrum

Mel Frequency cepstrum

- The cepstrum requires Fourier analysis
  - from frequency space back to time
- So we actually apply inverse DFT

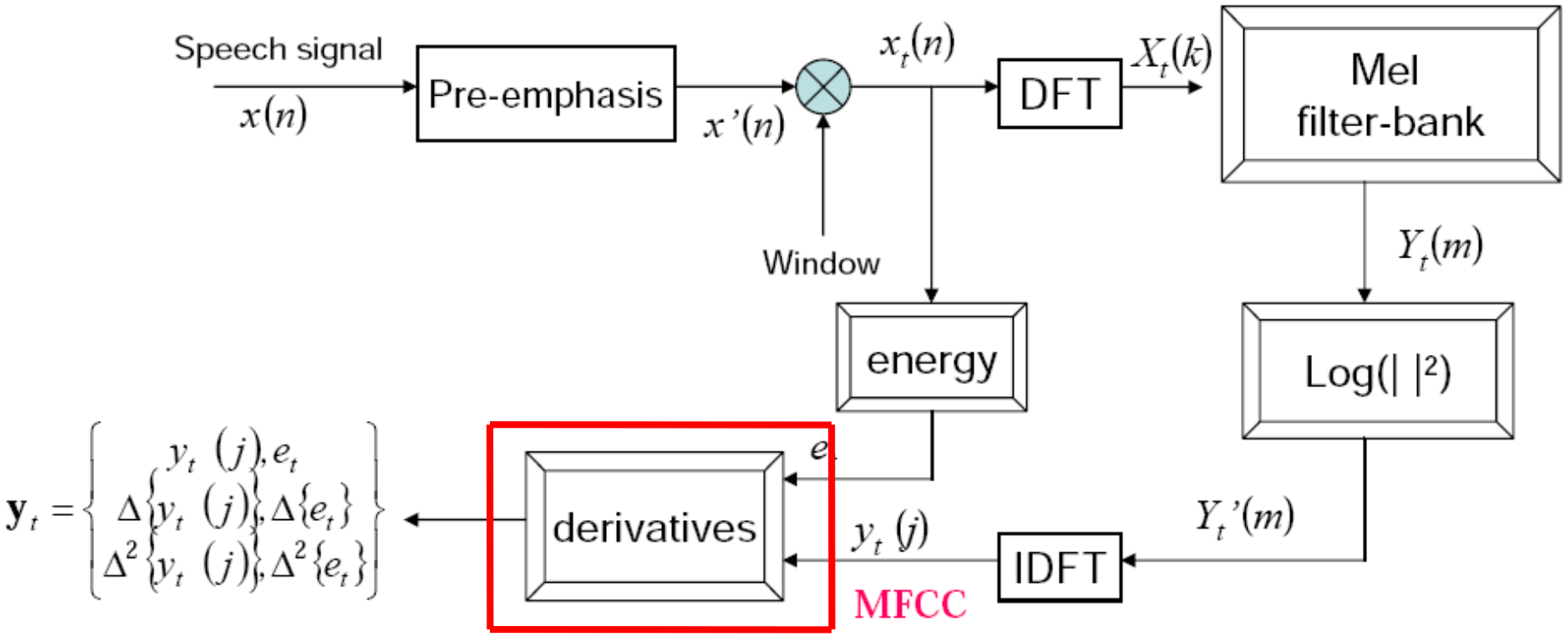
$$y_t[k] = \sum_{m=1}^{M} \log(|Y_t(m)|) \cos(k(m-0.5)\frac{\pi}{M}),\ \text{k=0,...,J}$$

- Details for signal processing gurus: Since the log power spectrum is real and symmetric, inverse DFT reduces to a Discrete Cosine Transform (DCT)

Another advantage of the Cepstrum

- DCT produces highly **uncorrelated** features
- We'll see when we get to acoustic modelling that these will be much easier to model than the spectrum
  - Simply modelled by linear combinations of Gaussian density functions with diagonal covariance matrices
- In general we'll just use the first 12 cepstral coefficients (we don't want the later ones which have the F0 spike)

MFCC


Speech signal
x(n)
Pre-emphasis
x'(n)
Window
x_t(n)
DFT
X_t(k)
Mel filter-bank
Y_t(m)
Log(| |²)
Y_t'(m)
IDFT
y_t (j)
MFCC
energy
e_t
derivatives
y_t (j), e_t
Δ{y_t (j)}, Δ{e_t}
Δ²{y_t (j)}, Δ²{e_t}
y_t = {...}

Dynamic Cepstral Coefficient

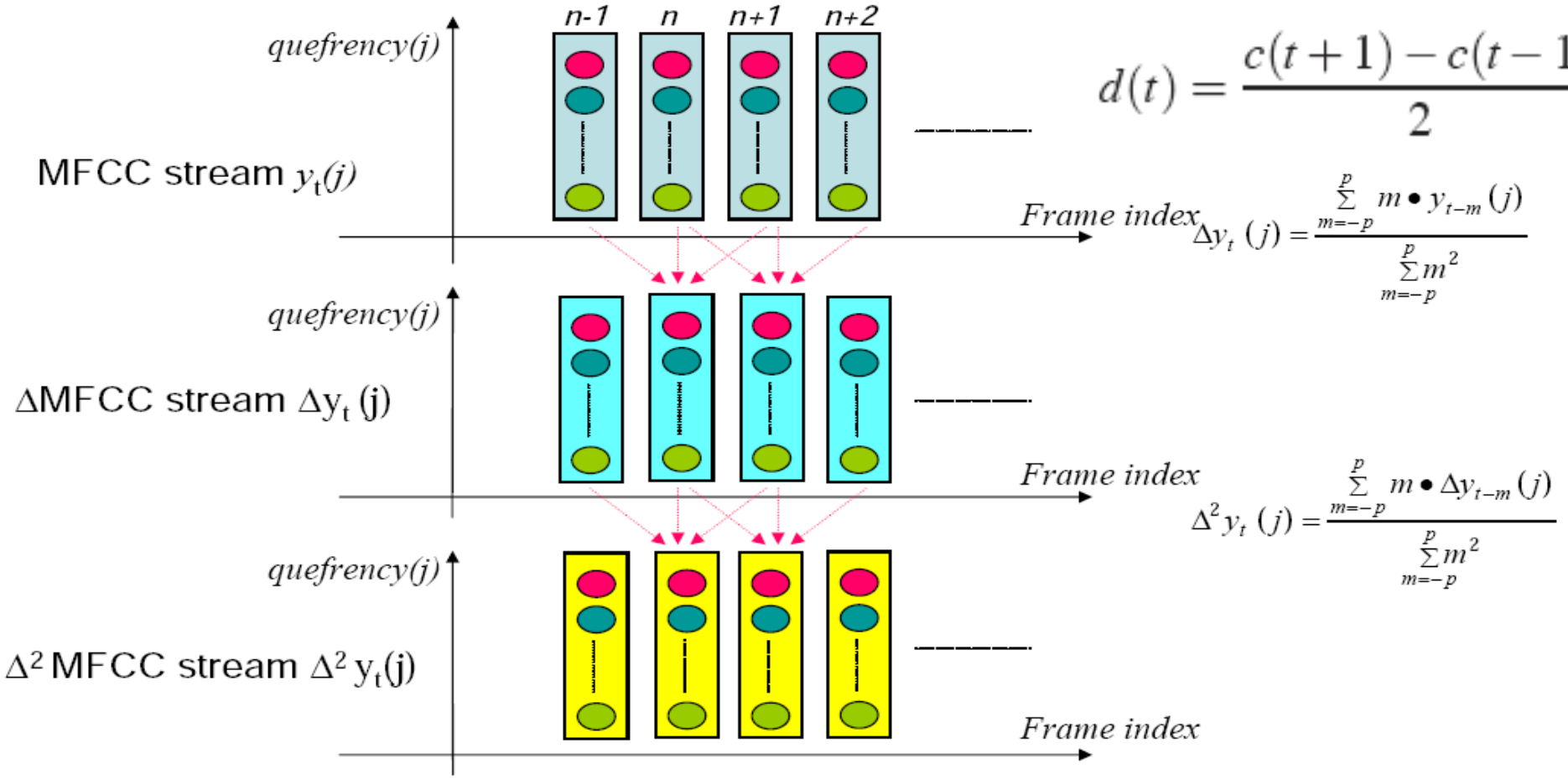
- The cepstral coefficients do not capture energy
  - So we add an energy feature

$$Energy = \sum_{t=t_1}^{t_2} x^2[t]$$

- Also, we know that speech signal is not constant (slope of formants, change from stop burst to release).
- So we want to add the changes in features (the slopes).
  - We call these delta features
  - We also add double-delta acceleration features

Delta and double-delta

- Derivative: in order to obtain temporal information



$$d(t) = \frac{c(t+1) - c(t-1)}{2}$$

$$\Delta y_t(j) = \frac{\sum_{m=-p}^{p} m \bullet y_{t-m}(j)}{\sum_{m=-p}^{p} m^2}$$

$$\Delta^2 y_t(j) = \frac{\sum_{m=-p}^{p} m \bullet \Delta y_{t-m}(j)}{\sum_{m=-p}^{p} m^2}$$

Summary: Typical MFCC features

- Window size: 25ms
- Window shift: 10ms
- Pre-emphasis coefficient: 0.97
- MFCC:
  - 12 MFCC (mel frequency cepstral coefficients)
  - 1 energy feature
  - 12 delta MFCC features
  - 12 double-delta MFCC features
  - 1 delta energy feature
  - 1 double-delta energy feature
- Total 39-dimensional features

Why is MFCC so popular?

- Efficient to compute
- Incorporates a perceptual Mel frequency scale
- Separates the source and filter
- IDFT(DCT) decorrelates the features
  - Improves diagonal assumption in HMM modeling
- Alternative
  - PLP
- Another look at this
  - http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

Problem: how to apply HMM model to continuous observations?

- We have assumed that the output alphabet V has a finite number of symbols
- But spectral feature vectors are real-valued!
- How to deal with real-valued features?
  - Decoding: Given o_t, how to compute $P(o_t|q)$
  - Learning: How to modify EM to deal with real-valued features

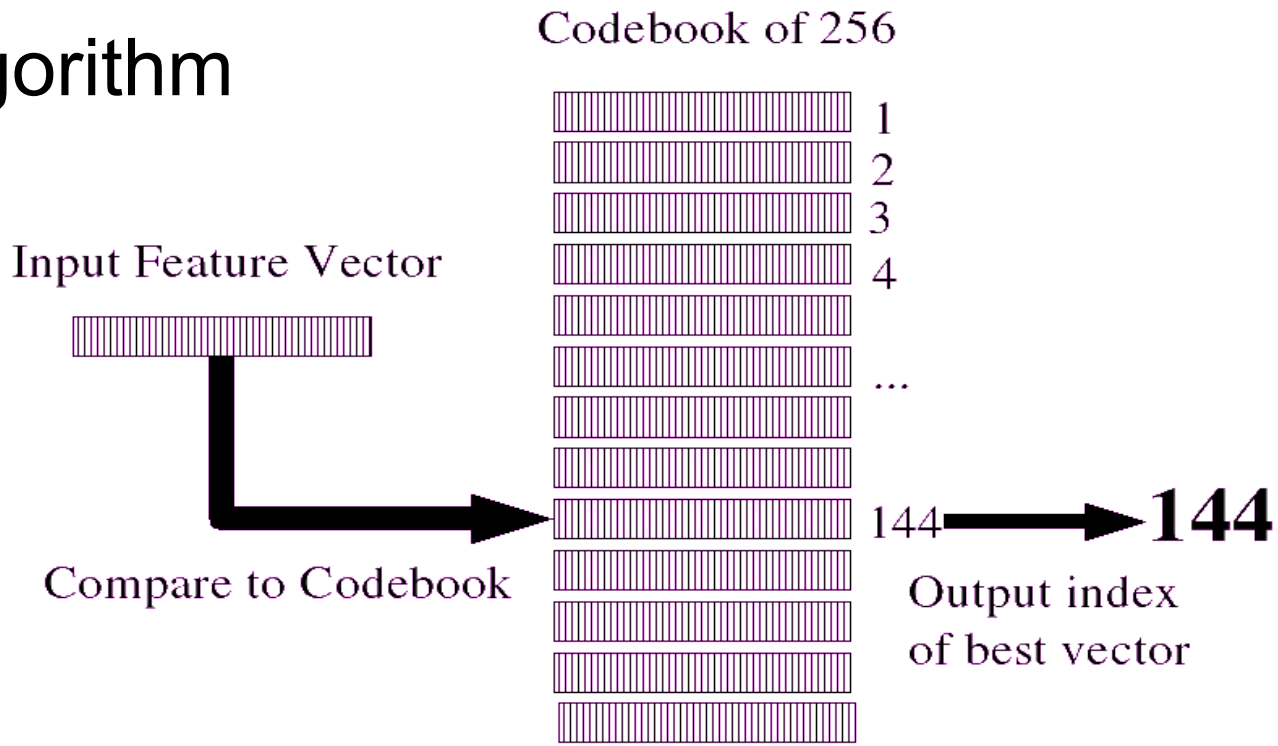
Vector Quantization

- Create a training set of feature vectors
- Cluster them into a small number of classes
- Represent each class by a discrete symbol
- For each class v_k, we can compute the probability that it is generated by a given HMM state using Baum-Welch

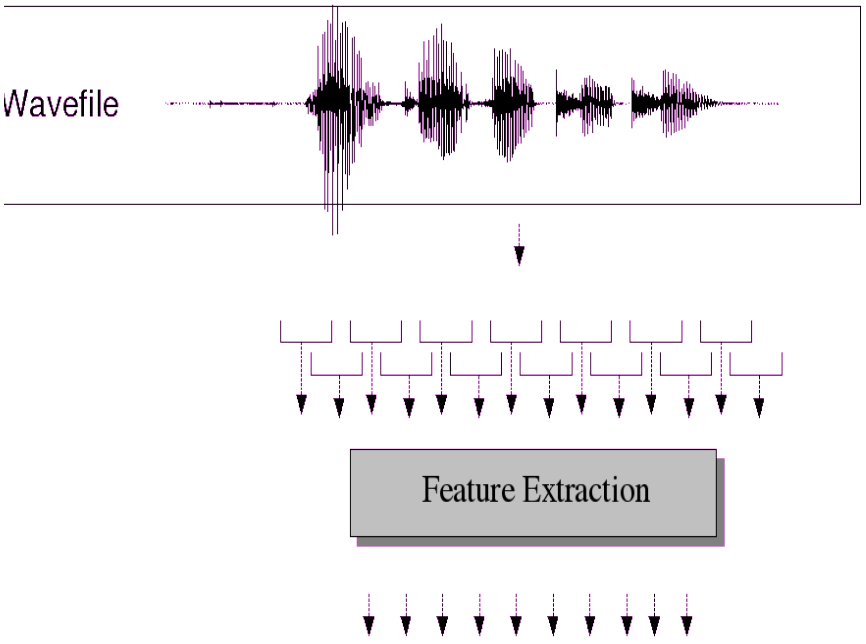
VQ requirements

- A distance metric or distortion metric
  - Specifies how similar two vectors are
  - Used:
    - to build clusters
    - To find prototype vector for cluster
    - And to compare incoming vector to prototypes
- A clustering algorithm
  - K-means, etc.

Why don't we do this now?



Front End


Wavefile
Feature Extraction
Observations
Feature
Vectors
State Sequence
Start
s_0
s_1
s_2
ih_0
ih_1
ih_2
k_0
k_1
k_2
s_0
s_1
s_2
End

Embedded Training