

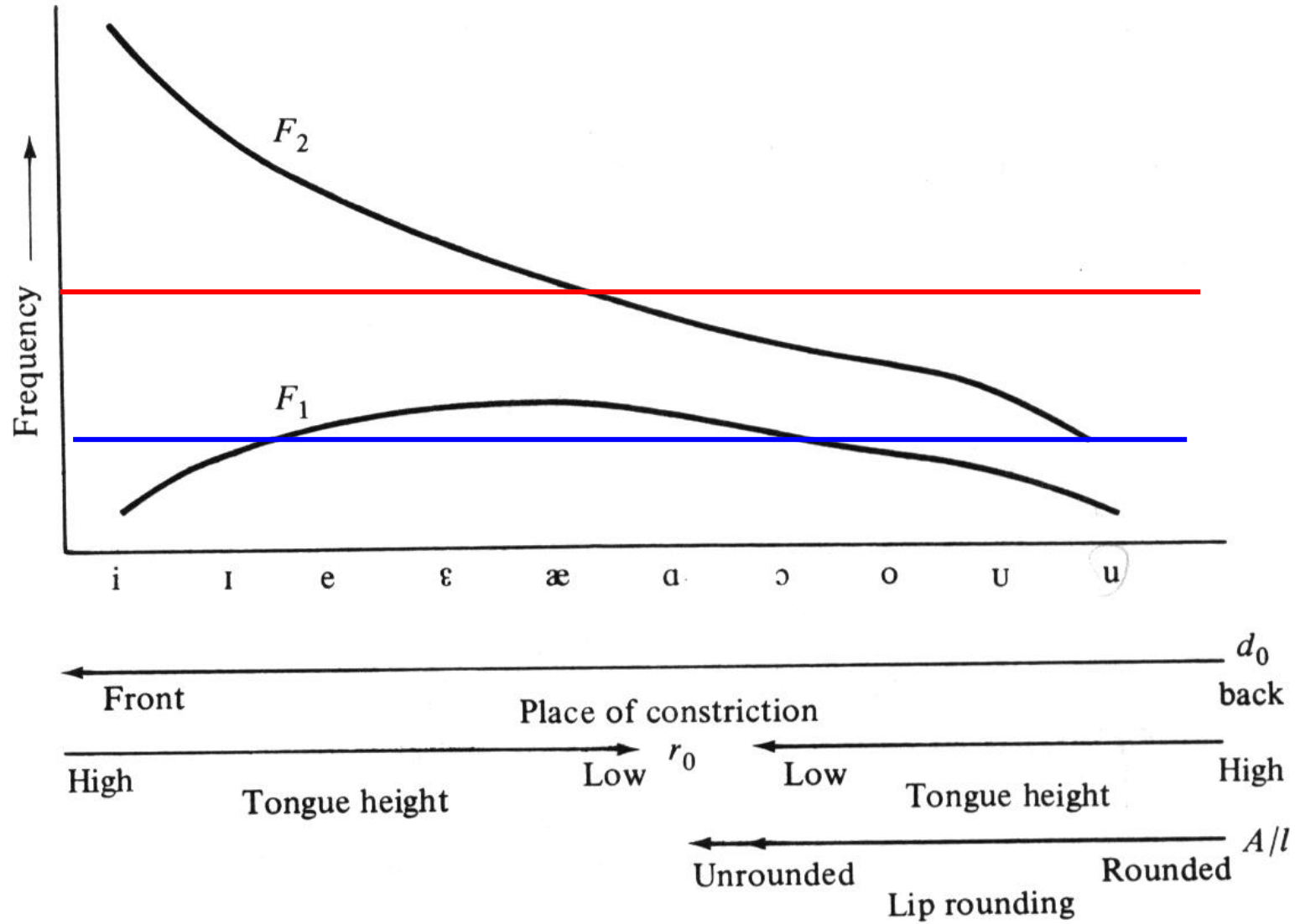
Lecture 04: Spoken Language Processing (2)



Instructor: Dr. Hossam Zawbaa

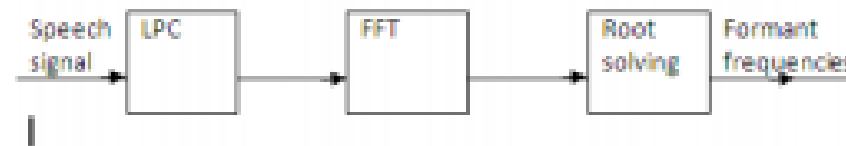
Formants

- **Formants are defined as the spectral peaks of sound spectrum of the speech.**
- In speech science and phonetics, formant frequencies refer to the acoustic resonance of the human vocal tract.
- They are often measured as amplitude peaks in the frequency spectrum of the sound wave.
- **Vowels largely distinguished by 2 characteristic pitches.**
- One of them (the higher of the two) **goes downward** throughout the series **iy ih eh ae aa ao ou u**
- The other **goes up** for the **first four vowels** and then **down** for the **next four**.
- These are called "formants" of the vowels, **lower is 1st formant, higher is 2nd formant.**



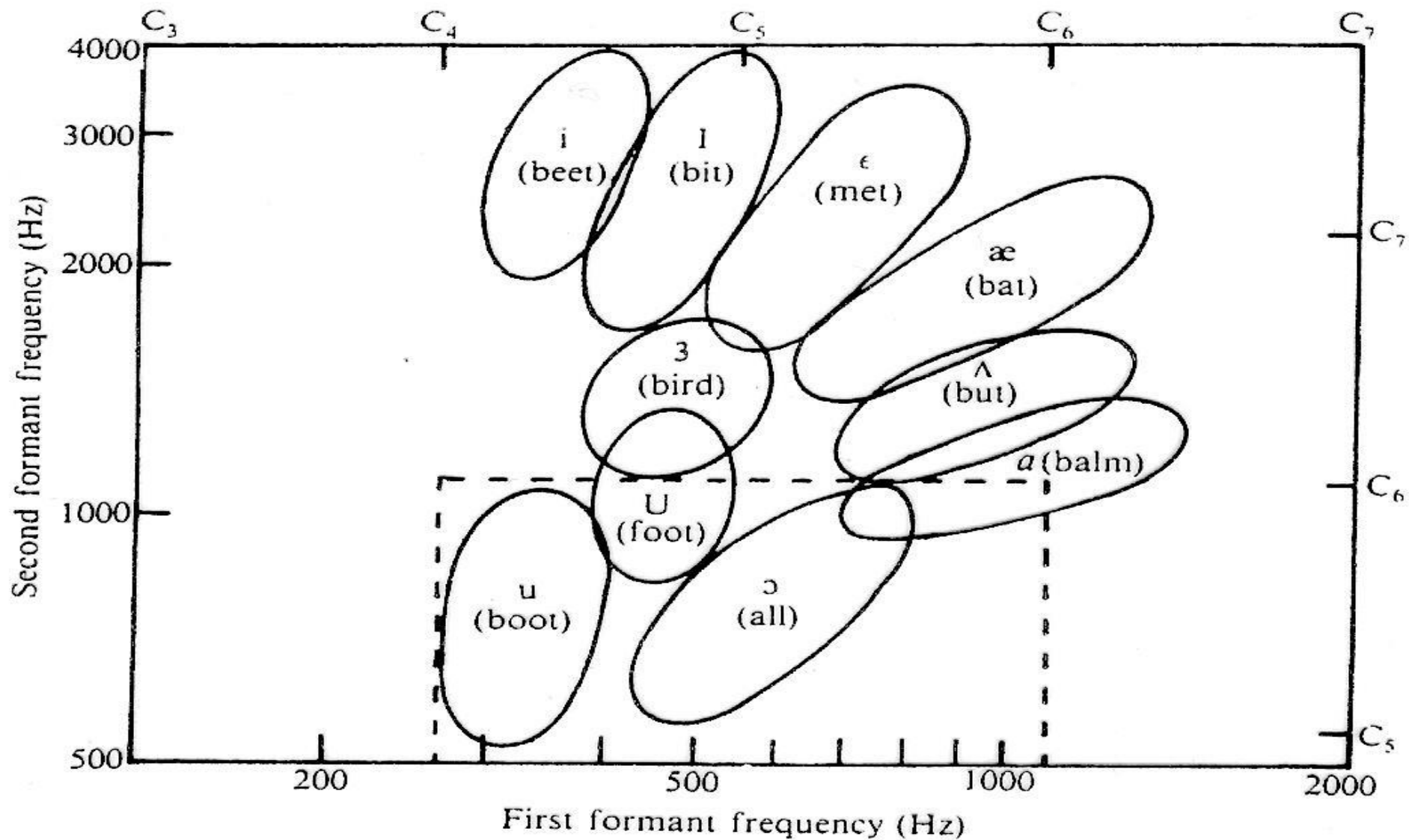
Formants

- We have considered the first 3 formants f_1 , f_2 , f_3 for analysis of emotions.
- For different vowels, the range of **f_1** lies between **270 to 730** Hz while the range of **f_2 and f_3** lie between **840 to 2290** and **1690 to 3010** Hz respectively.
- Formant frequencies are very important in the analysis of the emotional state of a person.
- Extraction of Formant Frequencies is done using Linear Predictive Coding (LPC) Based Formants Estimation Technique.



Block diagram of formant frequency detection using LPC.

FIRST AND SECOND FORMANTS OF VOWELS



Pitch

- **Pitch is fundamental frequency of speech signal.**
- The most widely considered areas of stress evaluation consider the characteristics of pitch.
- **Pitch is the mental sensation or perceptual correlated of F0**
- Relationship between pitch and F0 is not linear;
 - human pitch perception is most accurate between 100Hz and 1000Hz.
 - The frequencies of the vibrations must occur somewhere between 20 Hz and 20,000 Hz in order for humans to hear them through the air as “sound.”
- The average man’s speaking voice, for example, typically has a fundamental frequency between **85 Hz and 155 Hz.**
- A woman’s speech range is about **165 Hz to 255 Hz.**
- A child’s voice typically ranges from **250 Hz to 300 Hz** and higher.

Source-Filter Theory: Modeling Vowels

- Vocal Tract (VT) has an infinite number of resonances/formants
- Identification of vowel quality seems most dependent upon the location of F1, F2 & F3
- These observations are based on
 - Studies of vowel perception
 - Modeling efforts which suggest F4-F6 are relatively static

Problem

- VT length influences exact frequency location of formants
- Speakers vary in their vocal tract length
- men > women > children

Problem

/i/

	F_1	F_2	F_3
M	270	2290	3010
W	310	2790	3310
C	370	3200	3730

/i/

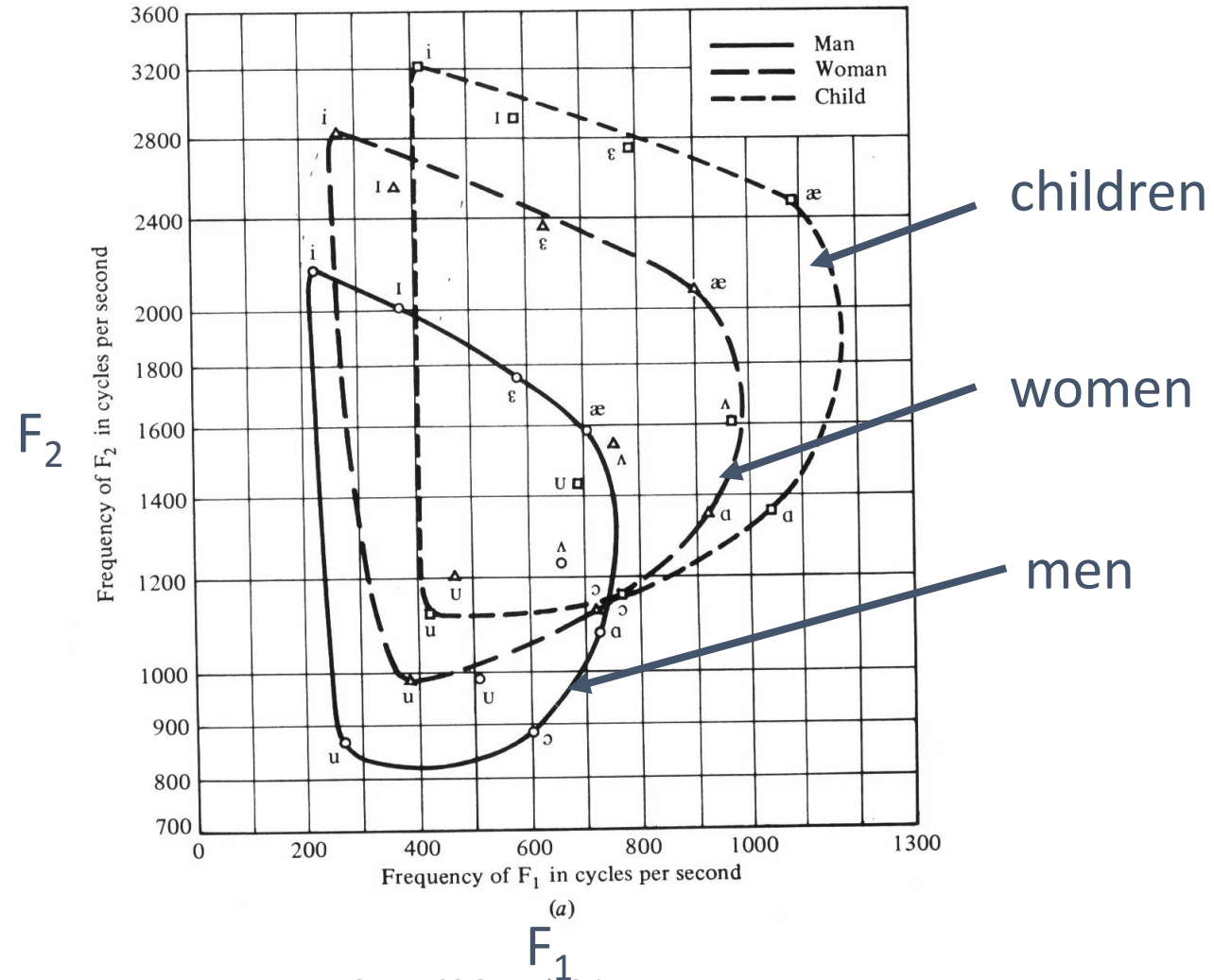
/u/

	F_1	F_2	F_3
M	300	870	2240
W	370	950	2670
C	430	1170	3260

/u/

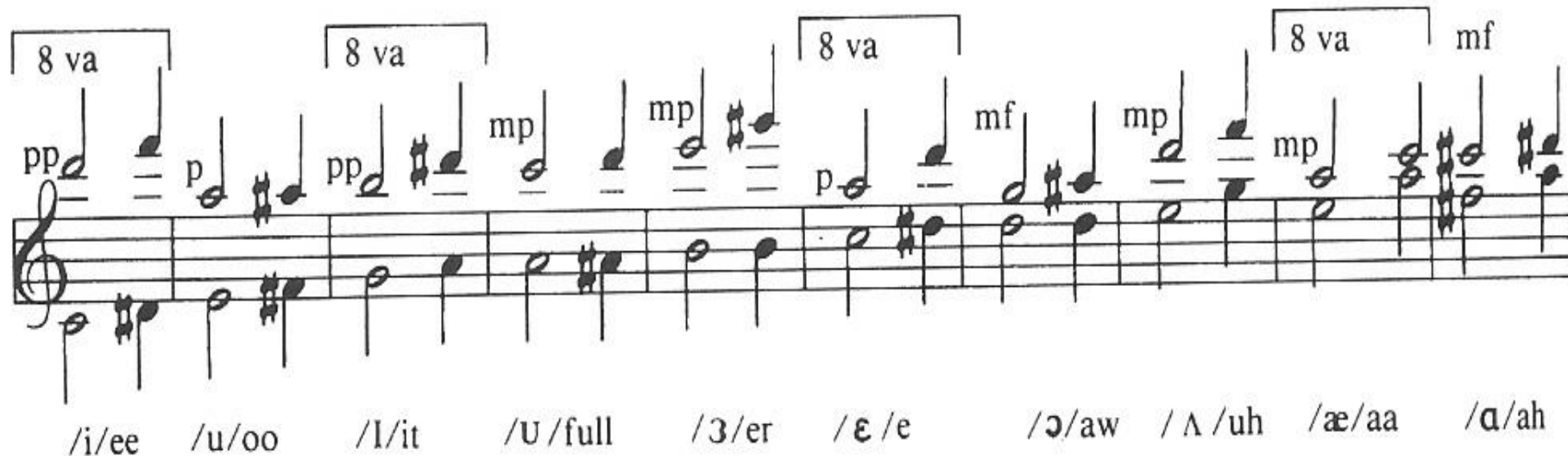
How do we know that a child, a man and a women all say /i/, when the acoustic values of formants are quite different?

A possible answer??



Formants and Pitch

- In both speech and singing, there is a division of labor between the vocal folds and vocal tract.
- **The vocal folds control the pitch, while the vocal tract determines the vowel sounds through formant and also articulates the consonants**

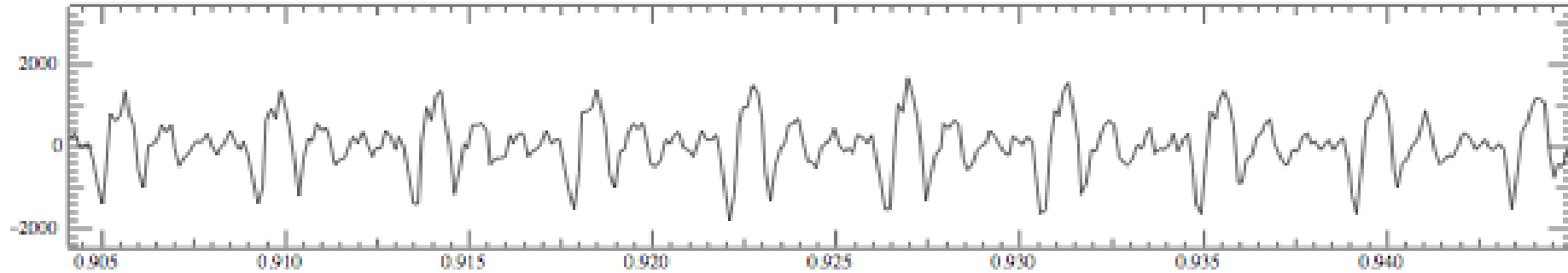


Typical formants of male and female speakers represented on a musical staff.

Formant frequencies of basic sung vowels

Formant frequency (Hz)		/i/	/I/	/e/	/æ/	/ɑ/	/ɔ/	/U/	/u/	/ʌ/
		(ee)	(i)	(e)	(aa)	(ah)	(aw)	(ù)	(oo)	(u)
F_1	M	300	375	530	620	700	610	400	350	500
	W	400	475	550	600	700	625	425	400	550
F_2	M	1950	1810	1500	1490	1200	1000	720	640	1200
	W	2250	2100	1750	1650	1300	1240	900	800	1300
F_3	M	2750	2500	2500	2250	2600	2600	2500	2550	2675
	W	3300	3450	3250	3000	3250	3250	3375	3250	3250

Part of [ae] waveform from “had”

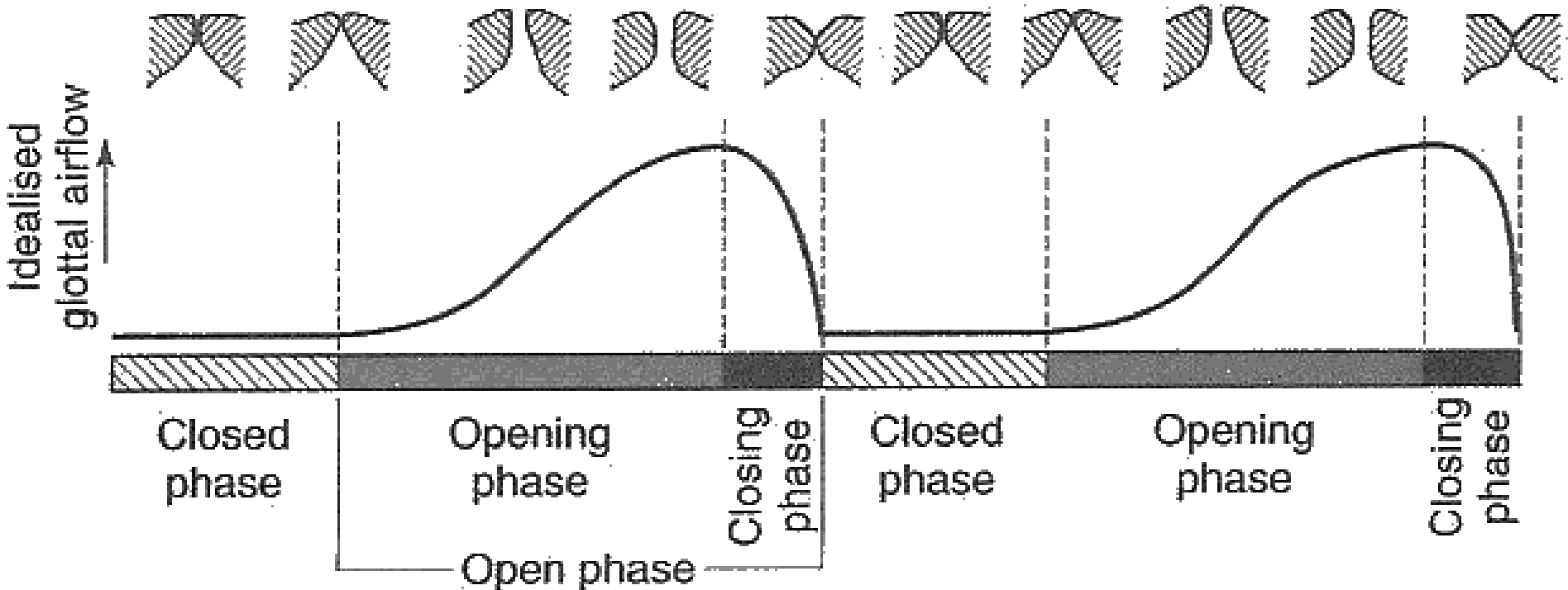


- Note complex wave repeating nine times in figure
- Plus smaller waves which repeats 4 times for every large pattern
- **Large wave has frequency of 250 Hz (9 times in 0.036 seconds)**

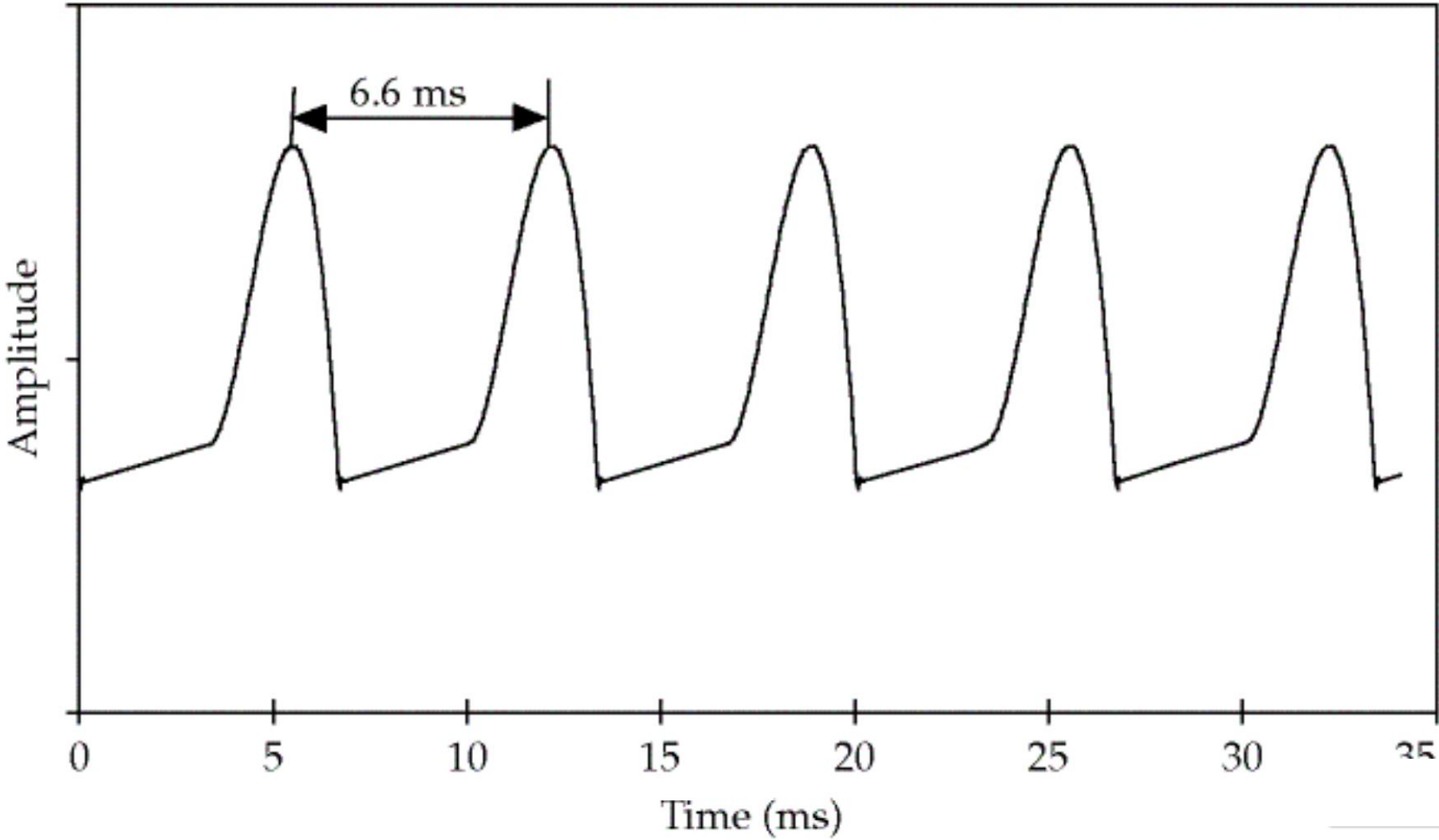
Different vowels have different formants

- Every time the vocal folds open and close, pulse of air from the lungs is sharp tap on air in vocal tract.
- Setting air in vocal cavity vibrating, producing different harmonics
- Harmonic is a component frequency of an oscillation or wave.

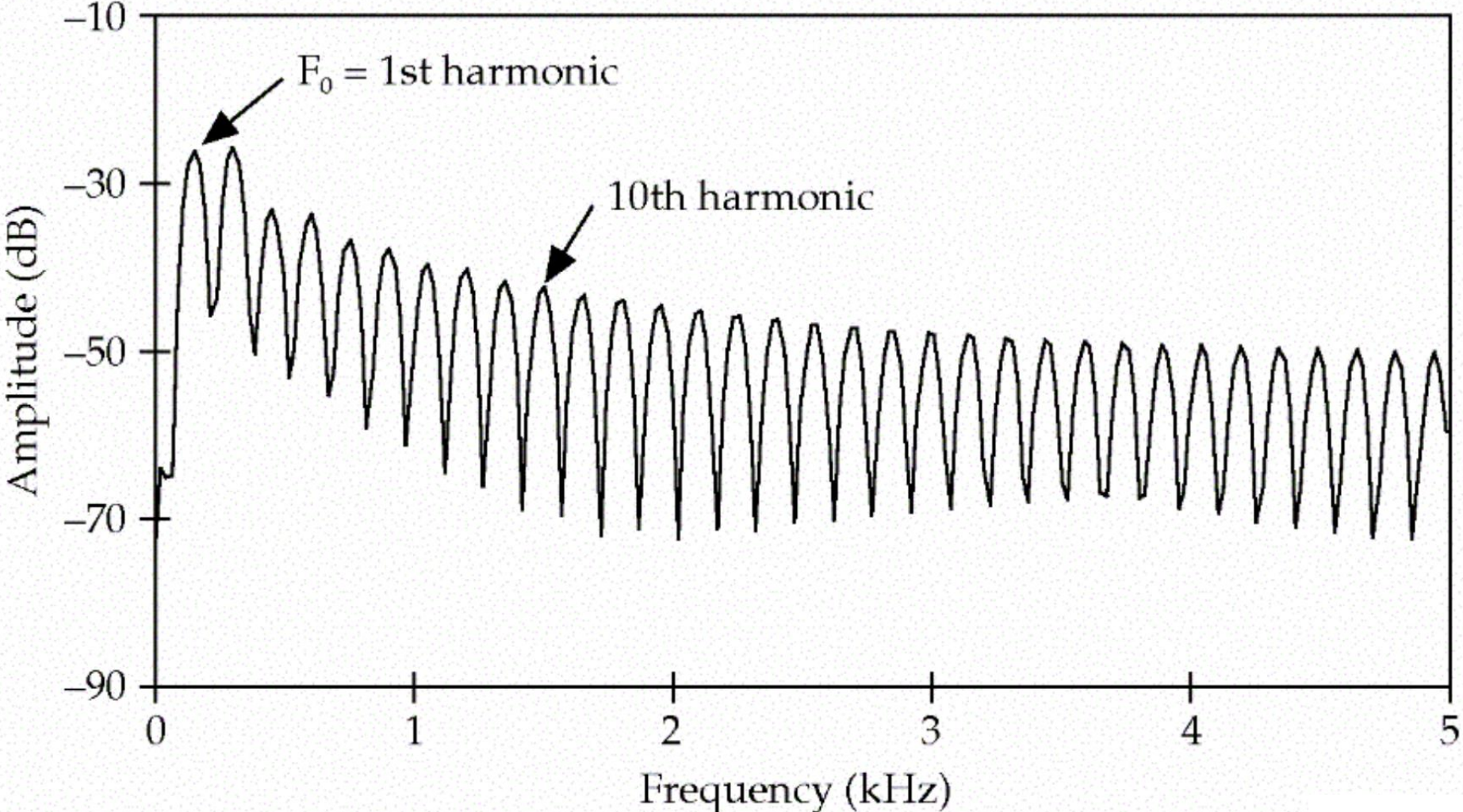
Vocal Fold Cycles



The vocal source at 150 Hz



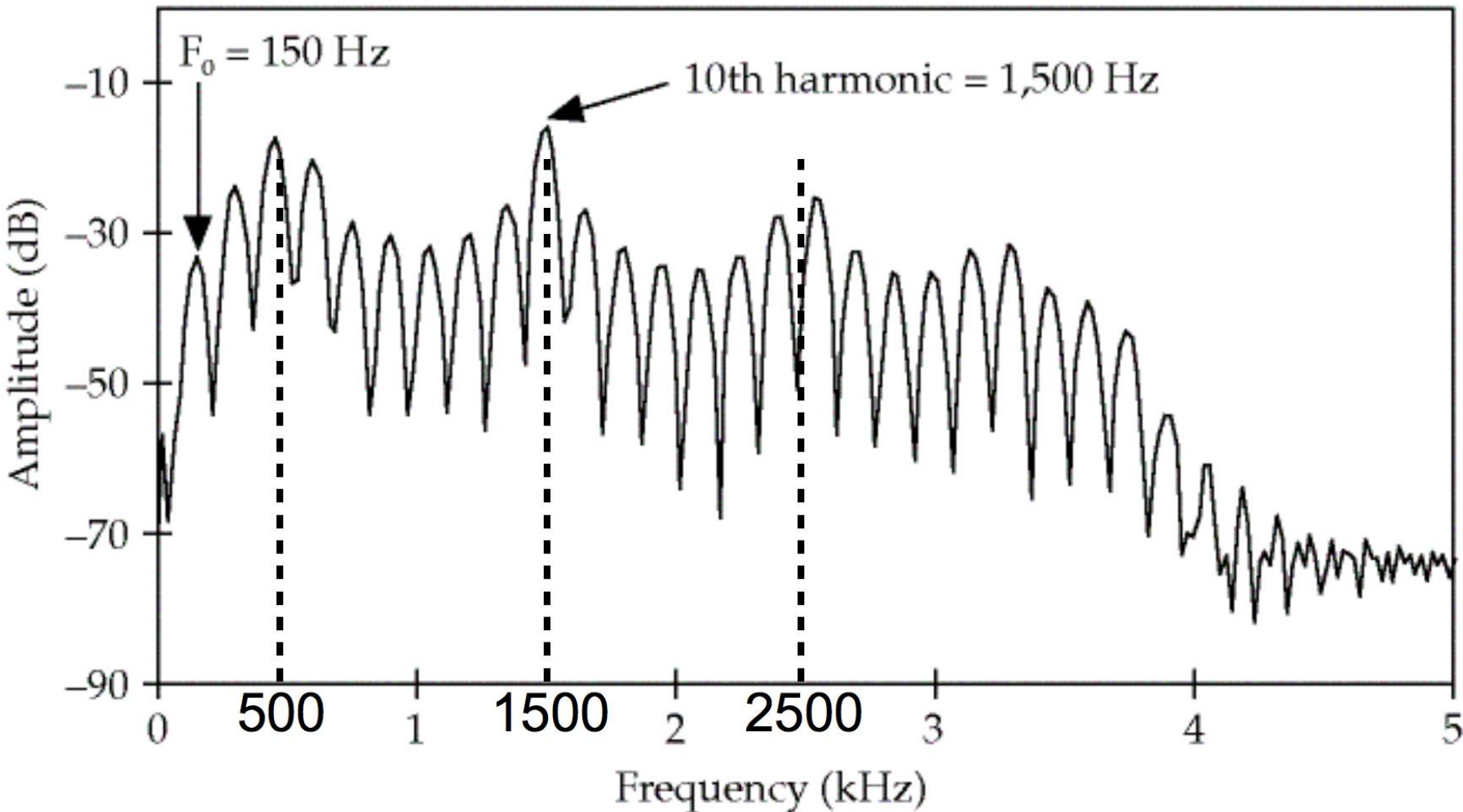
The harmonics



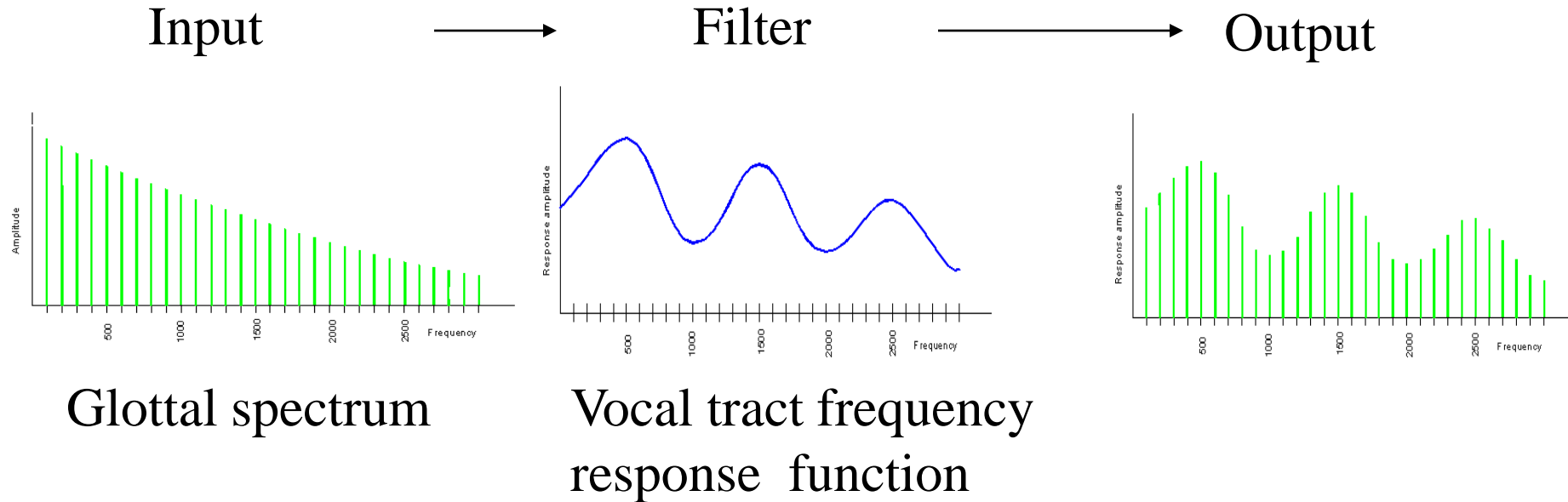
Source filter model of vowels

- Any body of air will vibrate in a way that depends on its size and shape.
- Vocal tract as "amplifier"; amplifies certain harmonics
- Formants are result of different shapes of vocal tract.

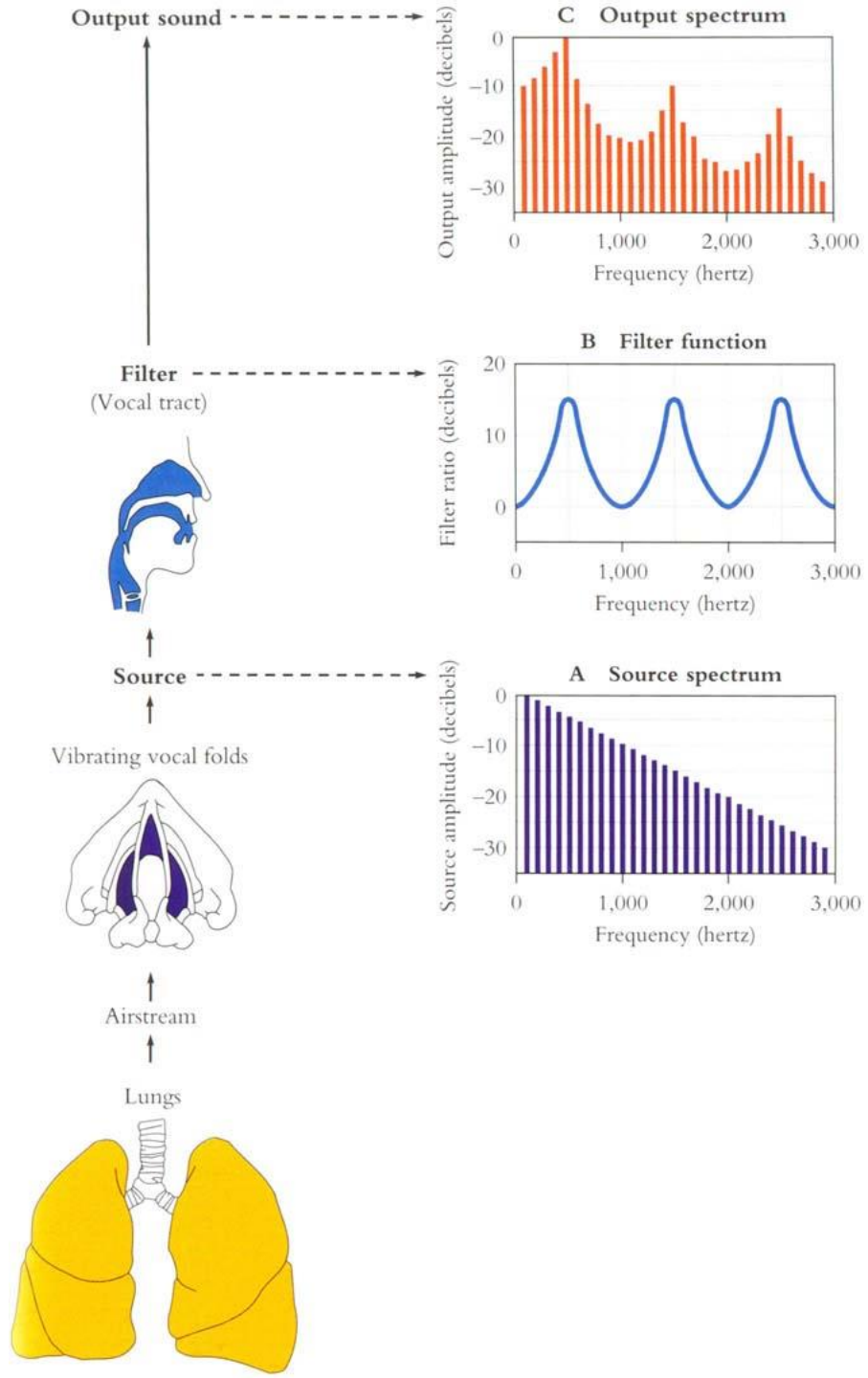
The oral cavity amplifies some harmonics



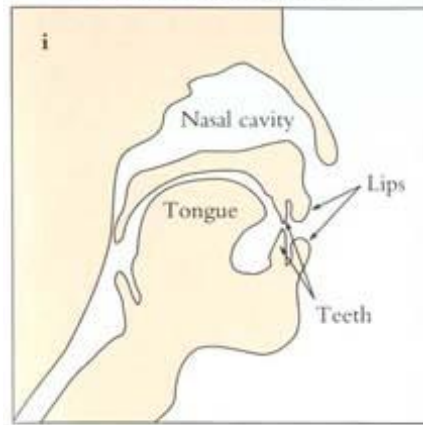
Source-filter model of speech production



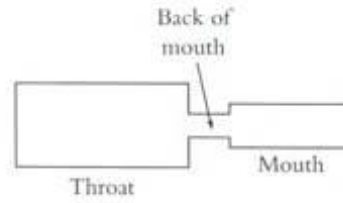
Source and filter are independent, so:
Different vowels can have same pitch
The same vowel can have different pitch



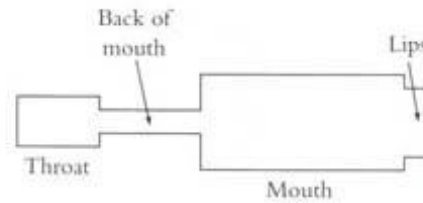
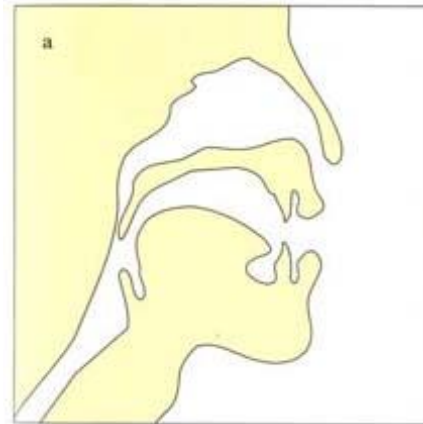
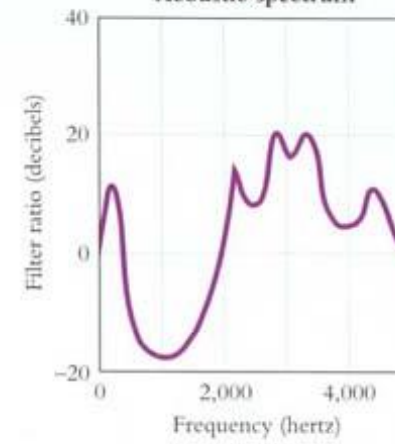
Cross section of vocal tract



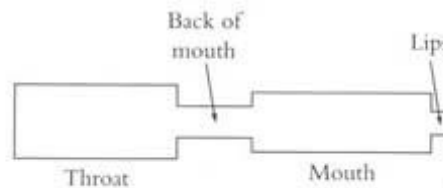
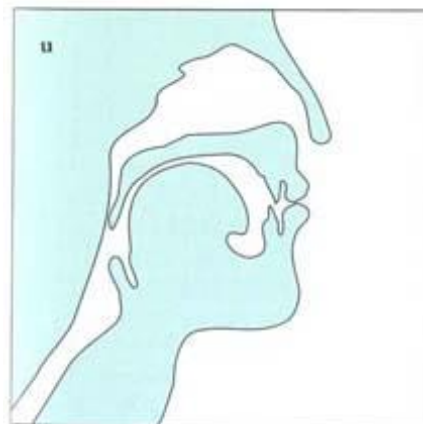
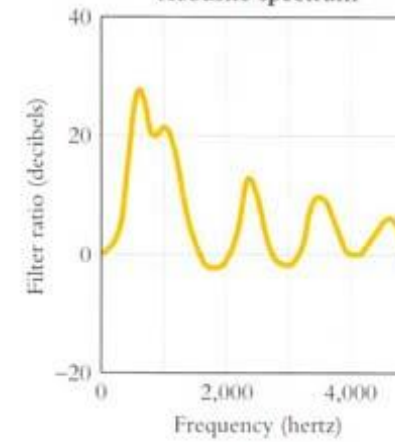
Model of vocal tract



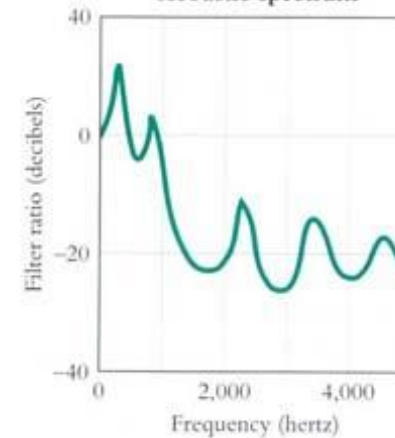
Acoustic spectrum



Acoustic spectrum



Acoustic spectrum



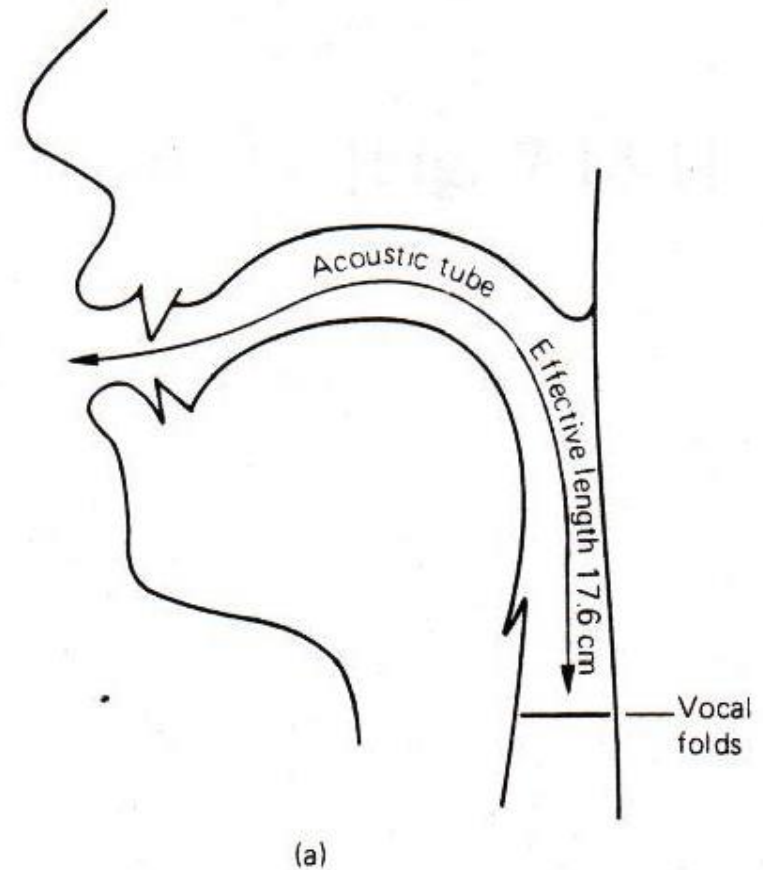
From
Mark
Liberman's
Web site

Resonances of the vocal tract

- The human vocal tract as an open tube



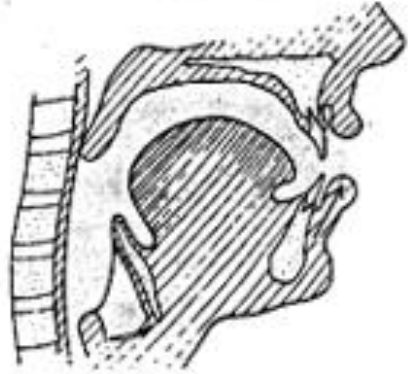
- Air in a tube of a given length will tend to vibrate at resonance frequency of tube.



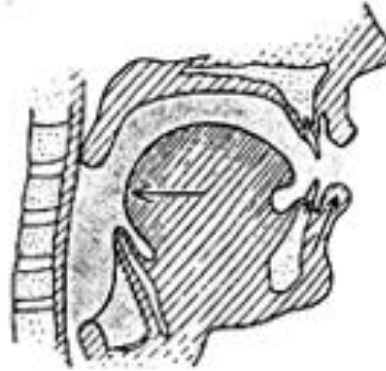
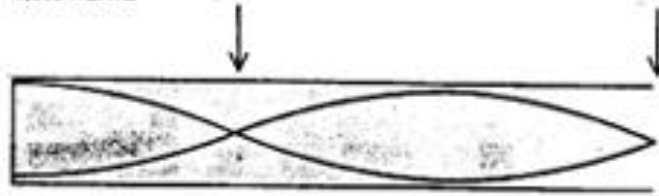
Resonances of the vocal tract

- Vocal tract is cylindrical tube open at one end
- Standing waves form in tubes
- Waves will resonate if their wavelength corresponds to dimensions of tube
- Constraint: Pressure differential should be maximal at (closed) glottal end and minimal at (open) lip end.

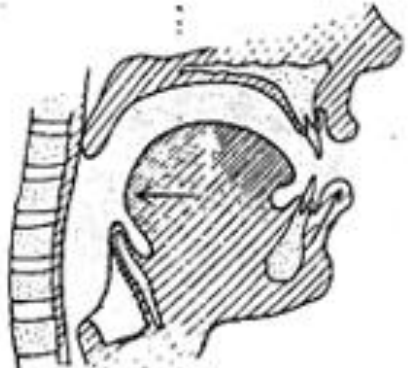
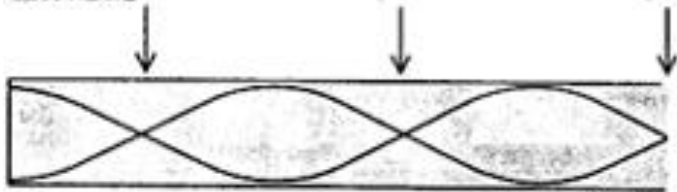
FIRST FORMANT
1/4 WAVELENGTH
500 HERTZ



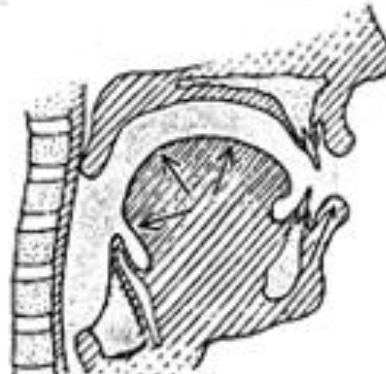
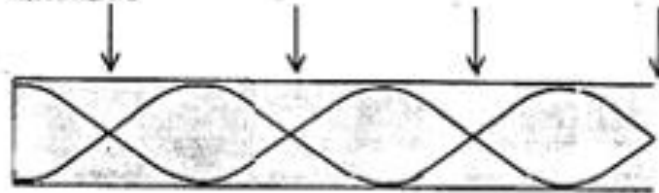
SECOND FORMANT
3/4 WAVELENGTH
1,500 HERTZ



THIRD FORMANT
5/4 WAVELENGTH
2,500 HERTZ



FOURTH FORMANT
7/4 WAVELENGTH
3,500 HERTZ



Three aspects of prosody

- **Prominence**: some syllables/words are more prominent (important) than others
- **Structure/boundaries**: sentences have prosodic structure
 - Some words group naturally together
 - Others have a noticeable break or disjuncture between them
- **Tune**: the intonation melody of an utterance (part of speech).

Placement of Pitch Accents

Stress vs. accent




- *Stress* is a structural property of a word
 - it marks a potential location for an accent to occur, if there is one.
- *Accent* is a property of a word in context
 - it is a way to mark intonation prominence in order to ‘highlight’ important words in the discourse.

(x)				(x)				(accented syll)
x				x				stressed syll
x			x	x				full vowels
x	x	x		x	x	x	x	syllables
vi	ta	mins		Ca	li	for	nia	

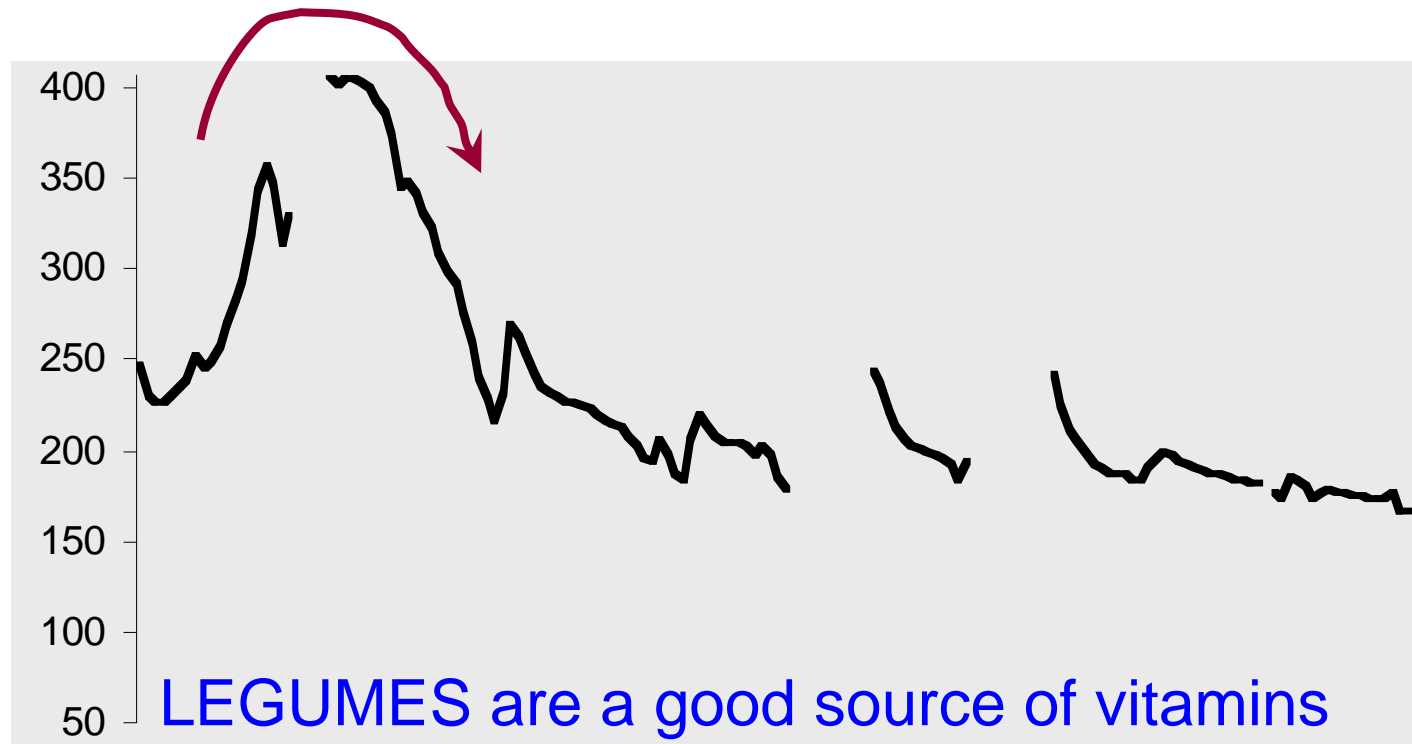
Stress vs. accent (2)

- The speaker decides to make the word **vitamin** more prominent by accenting it.
- Lexical stress tell us that this prominence will appear on the first syllable, hence **VI**itamin.
- So prosodic prominence is a function of
 - lexicon
 - context
- I'm a little **surPRISED** to hear it **CHAR**acterized as **up**BEAT

Which word receives an accent?

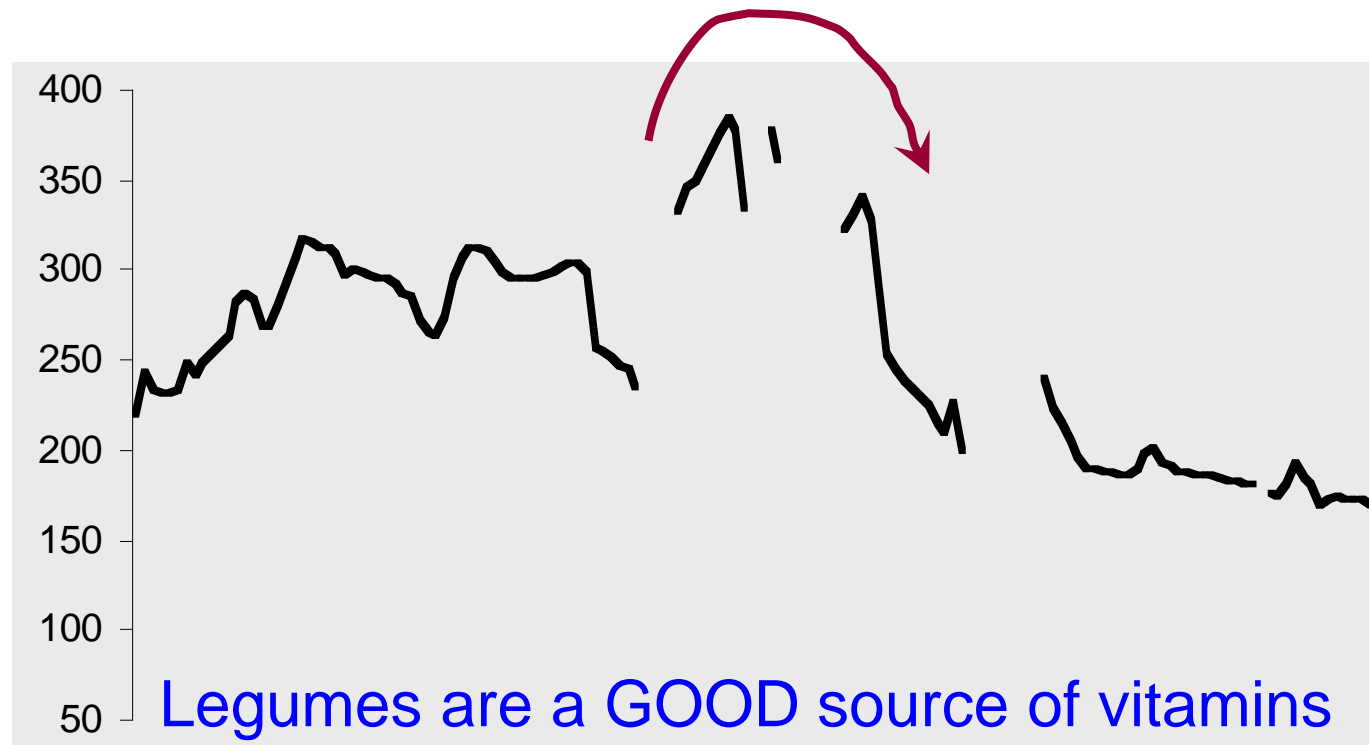
- It depends on the context.
 - **The 'new' information in the answer to a question is often accented**
 - while the 'old' information is usually not.
- Q1: What types of foods are a good source of vitamins?
- A1: LEGUMES are a good source of vitamins. 
- Q2: Are legumes a source of vitamins?
- A2: Legumes are a GOOD source of vitamins. 
- Q3: I've heard that legumes are healthy, but what are they a good source of ?
- A3: Legumes are a good source of VITAMINS. 

Same 'tune', different alignment



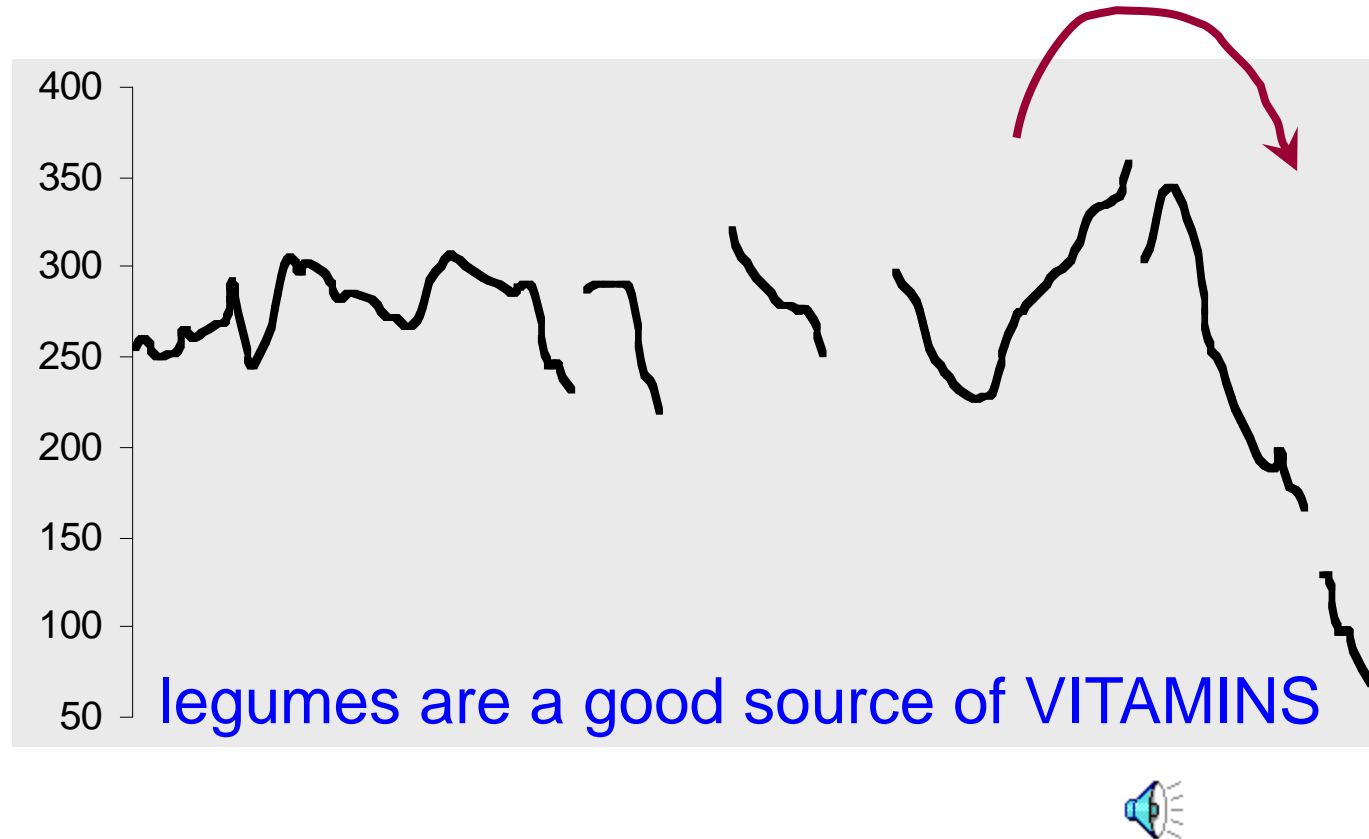
The main **rise-fall** accent (= “I assert this”) shifts locations.

Same 'tune', different alignment



The main **rise-fall** accent (= “I assert this”) shifts locations.

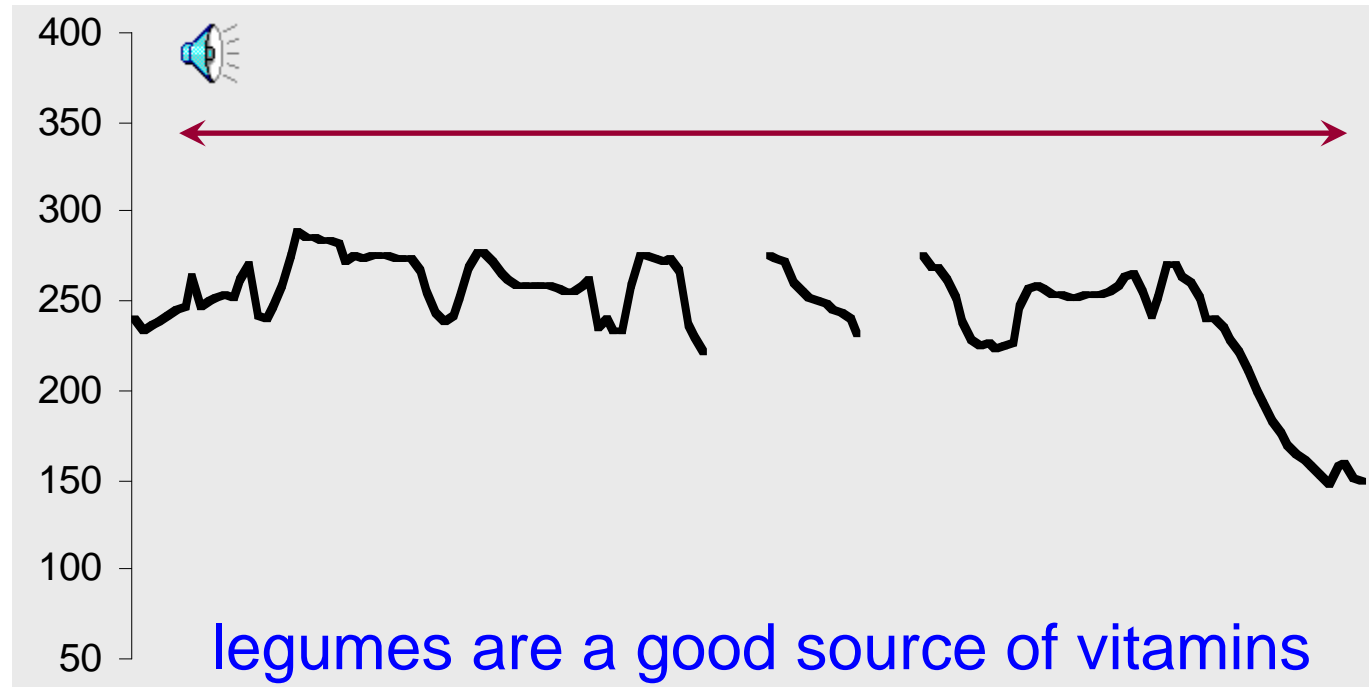
Same 'tune', different alignment



The main **rise-fall** accent (= “I assert this”) shifts locations.

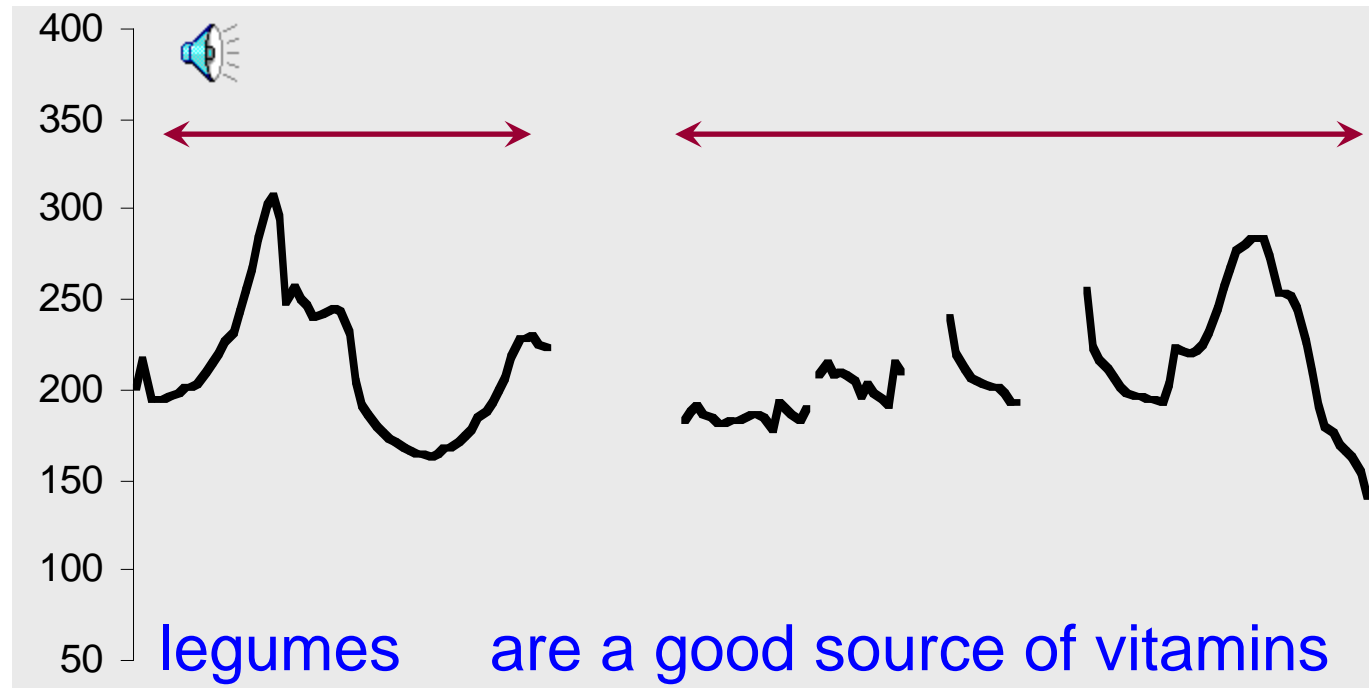
Intonation phrasing/boundaries

A single intonation phrase



Broad focus statement consisting of one intonation phrase (that is, one intonation tune spans the whole unit).

Multiple phrases



Part of speech can be ‘chunked’ up into smaller phrases in order to signal the importance of information in each unit.

Phrasing can disambiguate

- **Global ambiguity:**

The old men and women stayed home.

The old men % and women % stayed home.

Sally saw % the man with the binoculars.

Sally saw the man % with the binoculars.

John doesn't drink because he's unhappy.

John doesn't drink % because he's unhappy.

Phrasing sometimes helps disambiguate

- **Temporary ambiguity:**

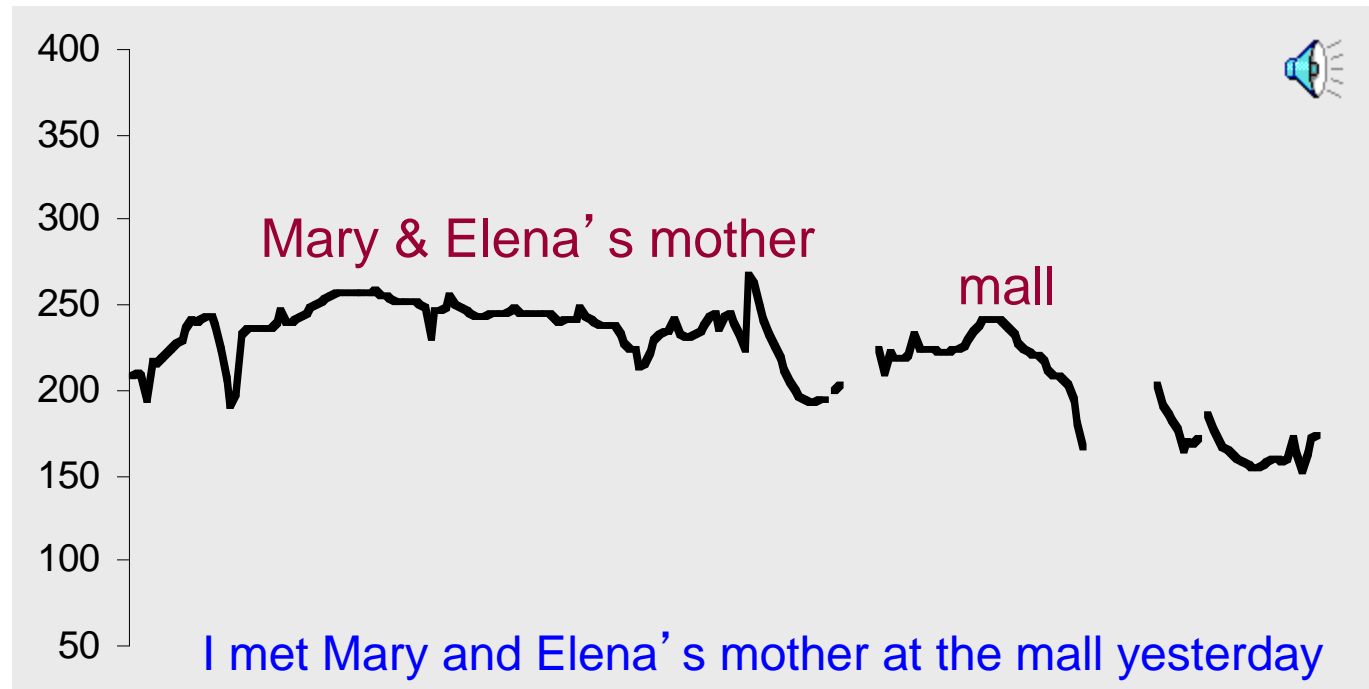
When Madonna sings the song is a hit.

When Madonna sings % the song is a hit.

When Madonna sings the song % it's a hit.

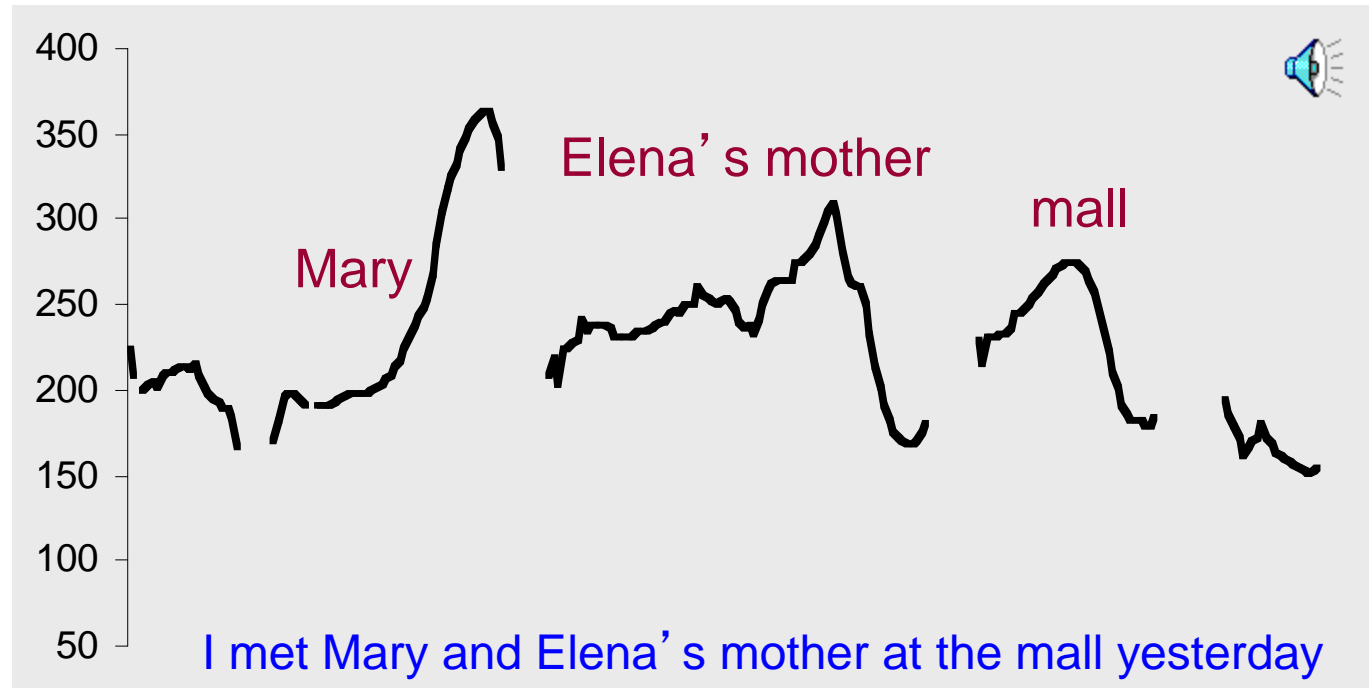
[from Speer & Kjelgaard (1992)]

Phrasing sometimes helps disambiguate



One intonation phrase with relatively flat overall pitch range.

Phrasing sometimes helps disambiguate



Separate phrases, with expanded pitch movements.

Using Intonation in Spoken Language Processing

- 1) **Prominence/Accent**: Tells us about focus of utterance
- 2) **Tune**: whether utterance is question/statement, important for affect extraction
- 3) **Boundaries**: can help parsing