Anatomy of speech production

This lecture: the sources of sound

Next lecture: video about the vocal folds

Following lecture: modifying that sound to make speech

- respiratory system
  - lungs, vocal tract (oral cavity, nasal cavity)
- speech organs: the sound source for voiced sounds
  - vocal folds

Airflow

- lungs
  - ingressive vs egressive sounds
- why is airflow needed?
  - as an energy source (refer back to lecture 2)
    - to drive the vocal folds
    - or for unvoiced sounds

The lungs

- primary function:
  - respiration – transfer oxygen into the blood and remove CO_2 from the blood
- secondary function
  - energy source for speech
    - air flow
    - air pressure
- typical lung capacity 2-5 litres

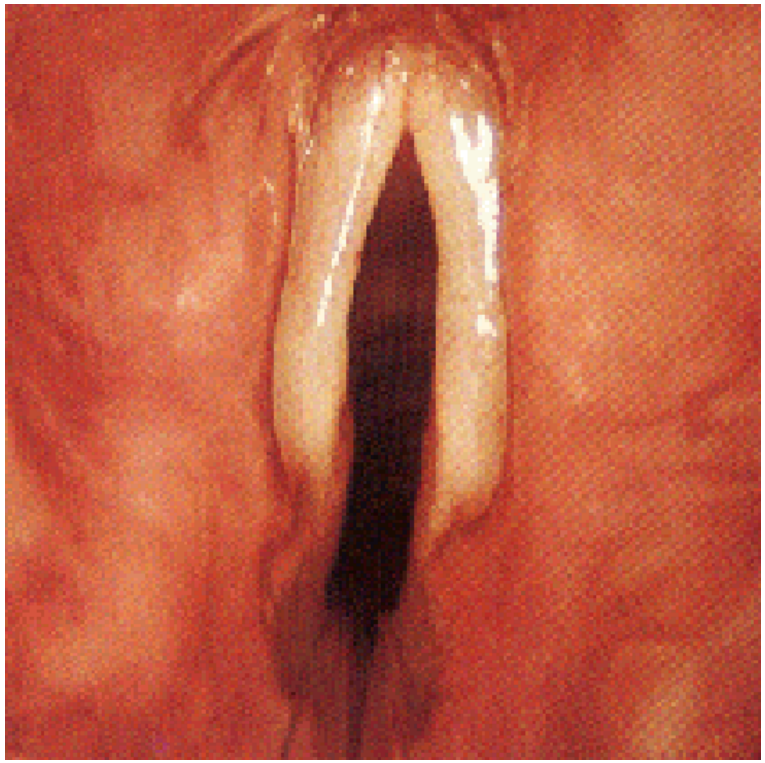
The vocal folds: airflow

- how do we get vibration from airflow?
- try “blowing a raspberry”
  - is there airflow?
  - can you do it without airflow?
  - what process is involved?

The vocal folds: vibration

- temporarily stop the airflow coming up from the lungs
- pressure builds up
- eventually, folds part and release pressure
  - a pulse of air travels up into the vocal tract
- fold close again due to muscular forces
- cycle repeats

The vocal folds: in action



[animation on web page]

The vocal folds: control

- how can we control
  - whether the folds vibrate at all
  - the frequency of vibration
- tension of the vocal folds
  - try blowing low and high frequency “raspberries”
  - what are you doing to control frequency?

Unvoiced sound sources

- fricative sounds
  - turbulence
- plosive sounds
  - sudden release of intra-oral air pressure
- we are not going to consider sounds with no air flow (e.g. clicks)

Turbulence

- steady smooth airflow
  - e.g. breathe out with relaxed vocal tract and relaxed vocal folds
  - no sound
- disturbed airflow
  - e.g. something placed in the airstream
  - compare to: leaning out of window of speeding car
    (*not* recommended for homework)

[animations on web page]

Turbulence

- in speech
  - airflow generated by lungs
  - what creates the turbulence?
- constrictions
  - e.g. tongue + upper incisors [th]
  - e.g. upper incisors and lower lip [f]

Plosives

- temporary complete interruption of the airplow
  - no airflow out of the mouth or nose
  - air continues to flow from the lungs
  - pressure builds up
  - eventually pressure overcomes the “blockage” and air is suddenly released
- e.g. like a champagne cork popping

Homework

- do the “intro to the lab” practical (possibly spend extra time working in lab)
- vocal folds
  - make low and high frequency vibrations – feel what your larynx does as you change frequency
- turbulence: make as many different fricative sounds as you can
- make as many different plosive sounds as you can (you’ll need a vowel sound before and/or after them)

Sounds don’t have to be from your own language (or even from any language)

Anatomy of speech production, continued

Previous lectures: the source of sound, the vocal folds

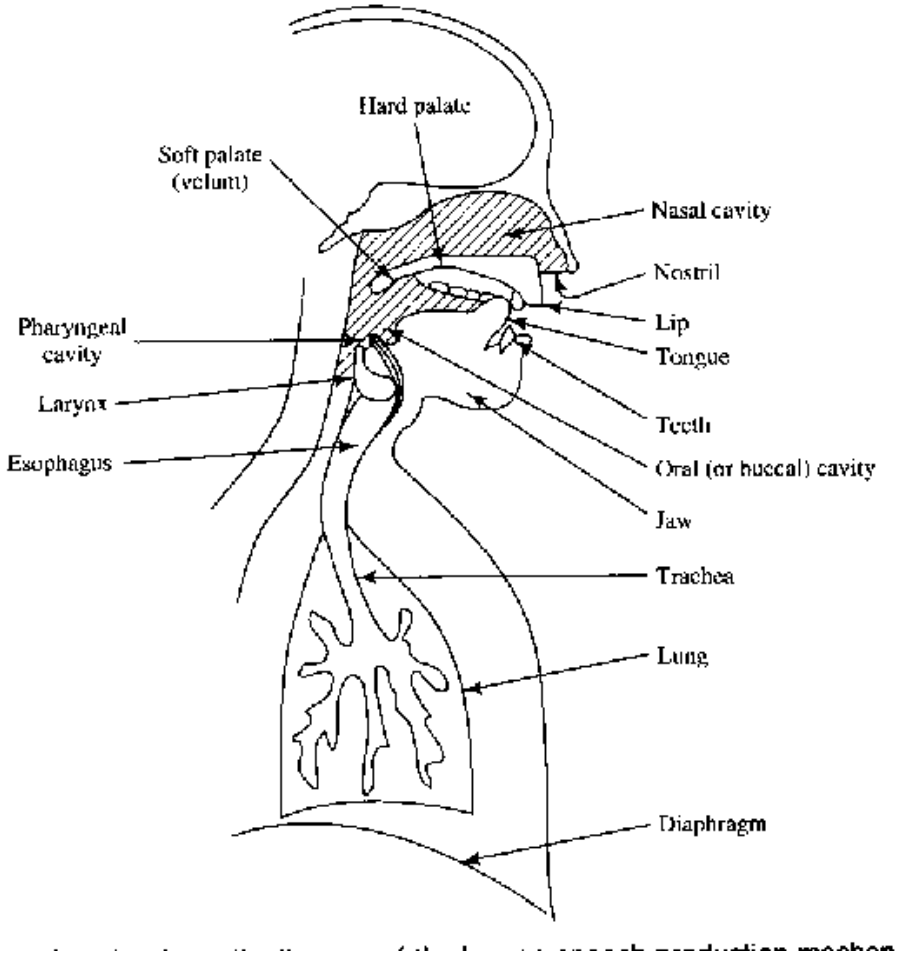
Starting in this lecture: modifying that sound to make speech

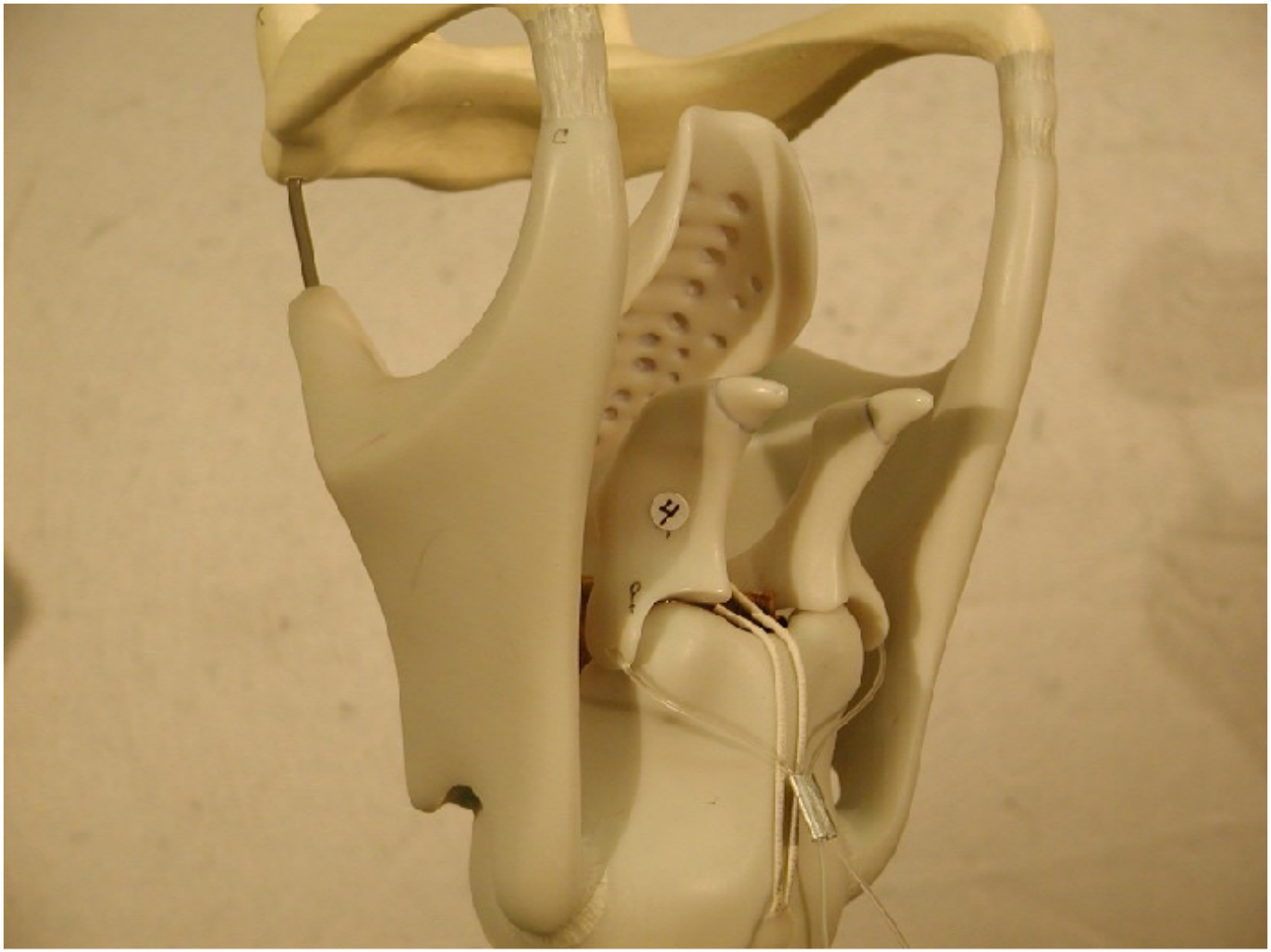
Today: anatomy of the articulators

- velum
- tongue
- jaw
- lips

The vocal tract

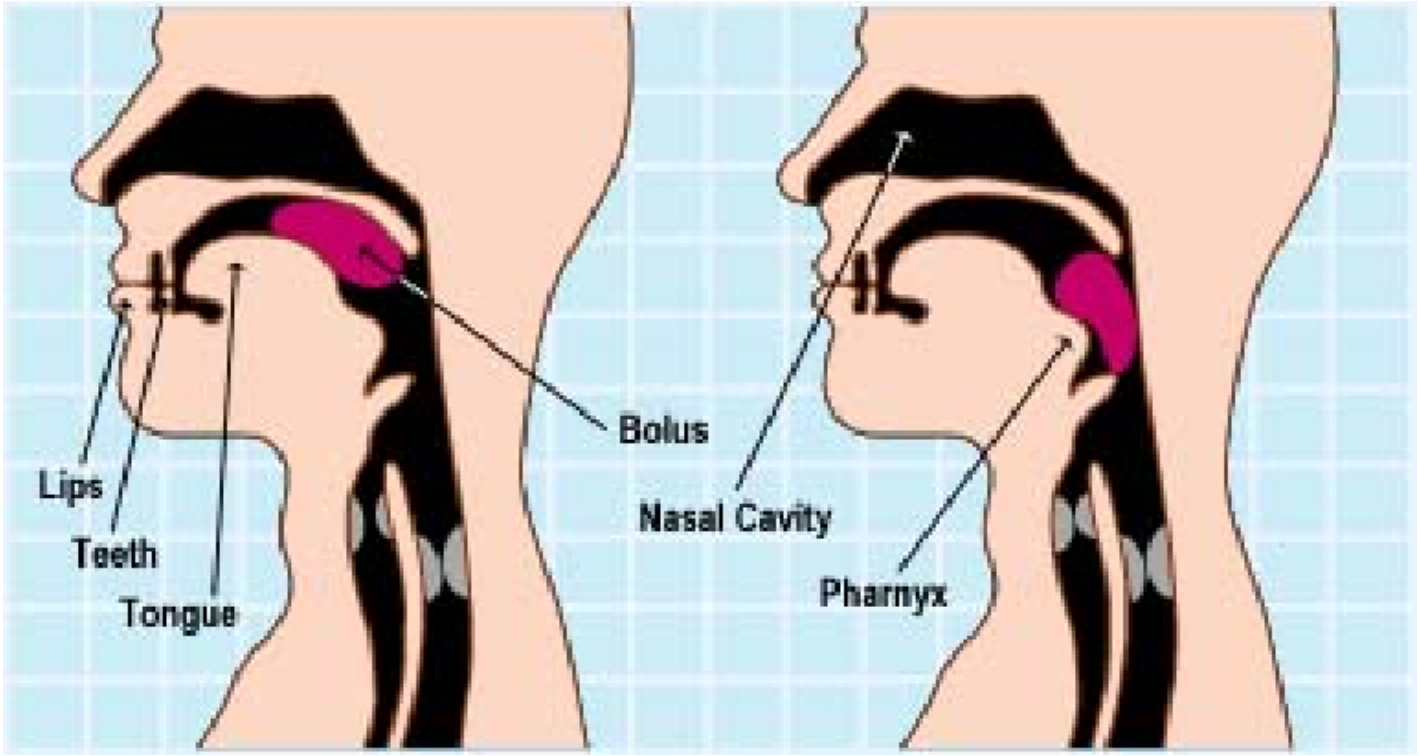
Speech production - general anatomy



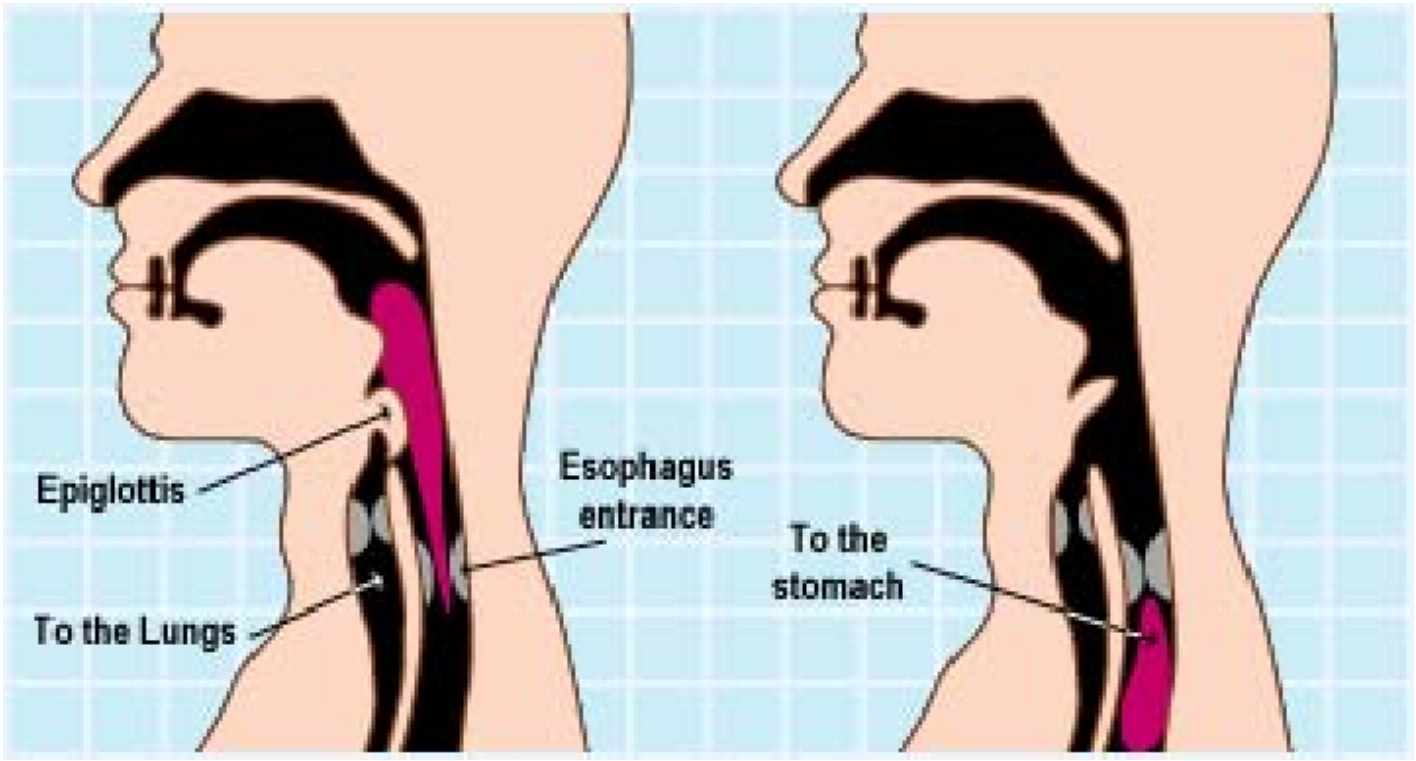


5:4

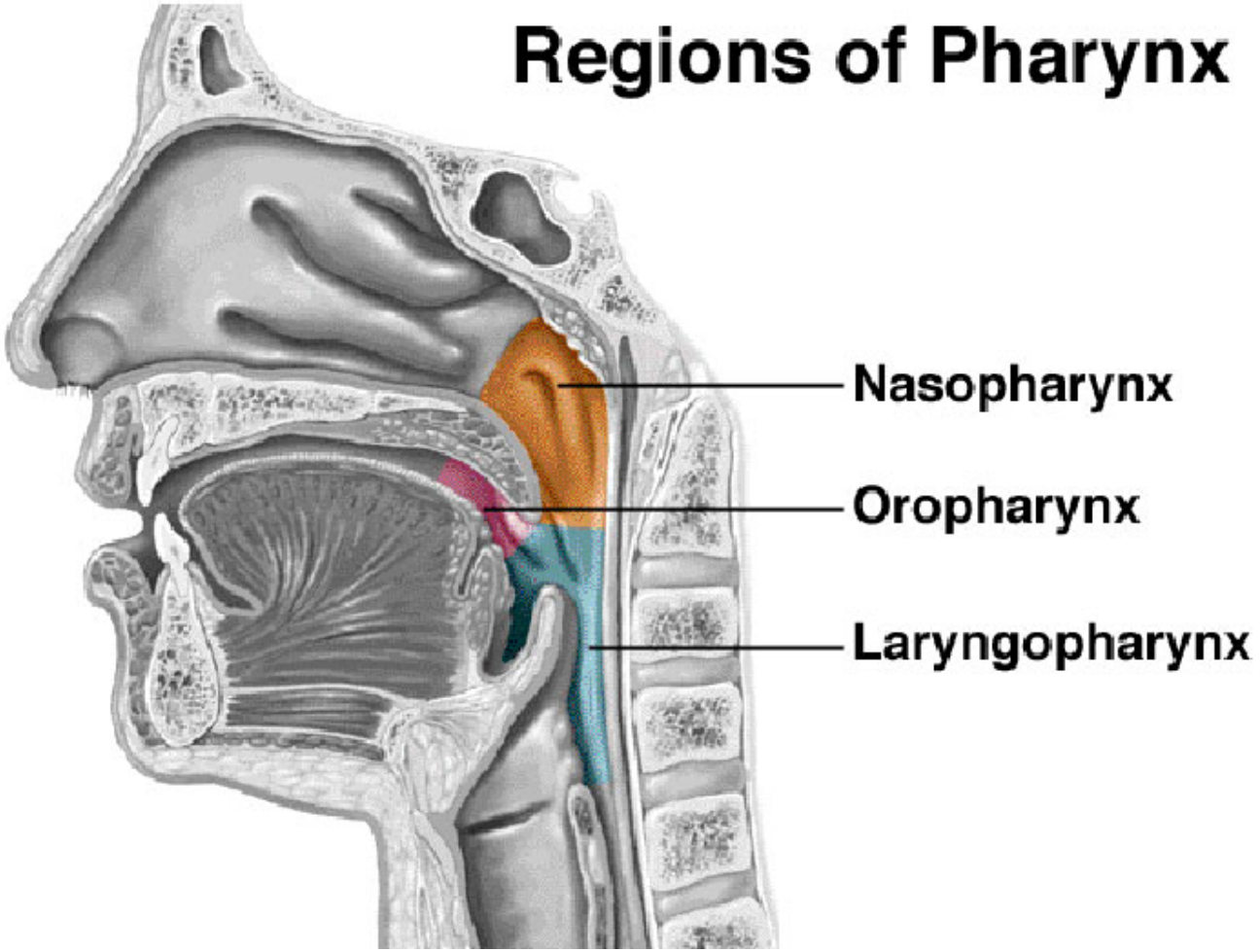
Swallowing - 1



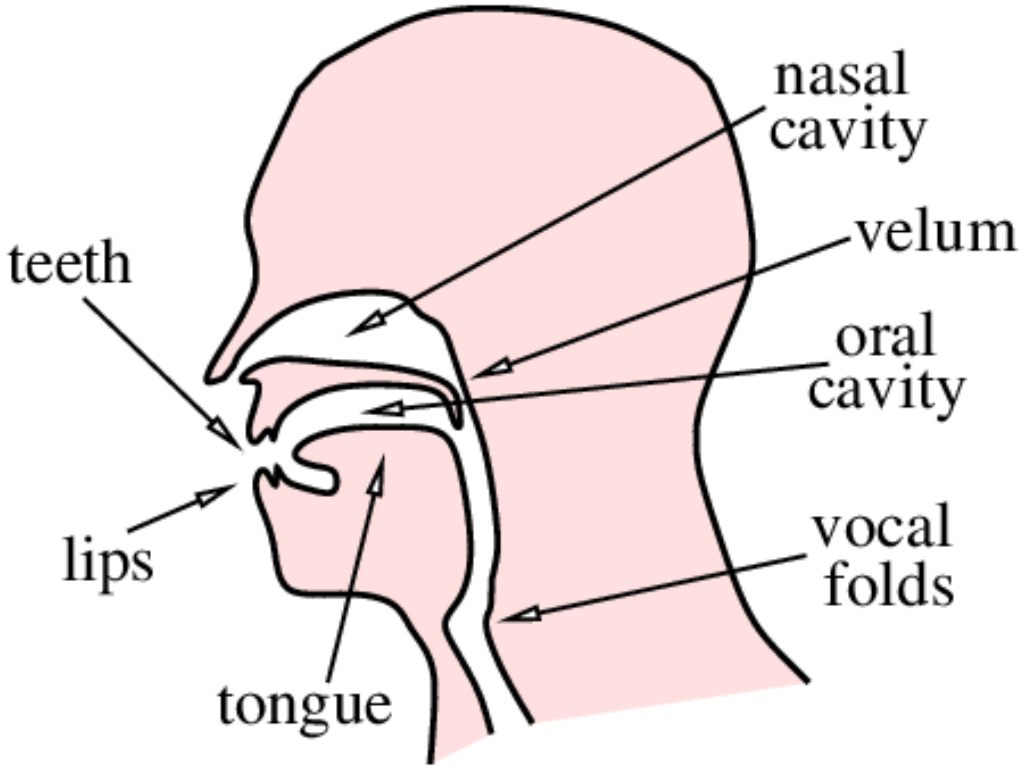
Swallowing - 2



Regions of Pharynx



The oral and nasal cavities



The articulators: velum

- need to connect/disconnect nasal cavity from oral cavity
- velum resting position is down
  - nasal cavity connected
  - can breathe in/out through mouth or nose
- for most speech
  - velum moves up to close off nasal cavity
- for nasal sounds
  - velum remains down

The articulators: tongue

- most important articulator
- capable of complex movements
- can change shape
  - can raise different portions of the tongue
  - can alter cross-section of tongue
  - can raise different portions of the tongue
  - can make or approximate contact with the palate
    - contact only at edges: grooving, e.g. /s/, /z/

        * contact only in centre, open at sides: laterals, e.g. /l/
        * complete closure: oral stops, e.g. /t/, /d/

- controls shape of oral cavity
    - up/down, front/back movements
    - affects shape and size of cavities
    - cavities resonate = formant frequencies

We will be looking at exactly how these resonances can be modelled and predicted in a few lectures time

The articulators: jaw

- often forgotten, but an important articulator!
- controls size of oral cavity
- controls position of lower incisors
  - e.g. constriction required to produce /f/, /v/

The articulators: lips

- shape of lips affects
  - length of vocal tract
    - rounded lips protrude – vocal tract lengthens
    - longer vocal tract = lower formants

As we will see, for some vowels the tongue effectively divides the vocal tract into two cavities, each with it's own resonance.

In that case, lip rounding enlarges the front cavity and lowers it's resonant frequency.

One step closer to a model

- how does all this anatomy help?
  - vocal tract / oral and nasal cavities
  - can model them as tubes
  - tubes resonate
- we will see how to predict the resonant frequency of a tube
- and of connected tubes

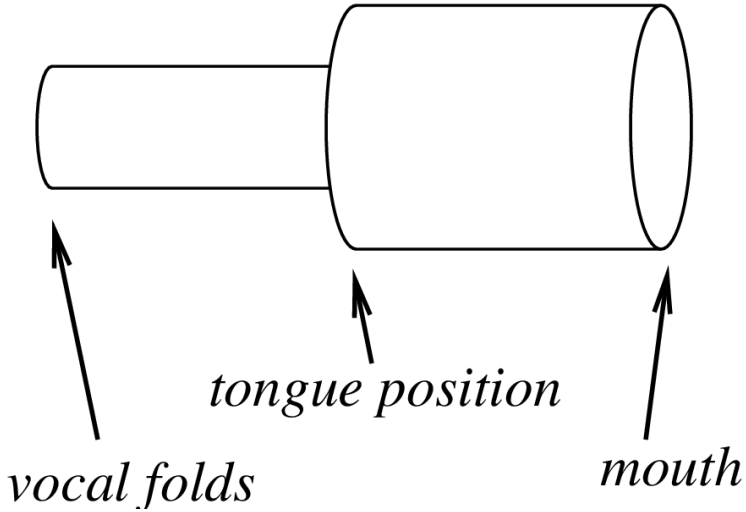
Vocal tract as two tubes

- tongue creates a constriction
  - part of tongue is raised towards palate
- position and size of constriction can move

  front $\longleftrightarrow$ back

  high $\longleftrightarrow$ low
- vocal tract can be modelled as two connected tubes

Two tube model (preview)



- tract has
  - back cavity
  - front cavity

Measuring articulation

Before going on with tube models – an interesting interlude:

How can we really know exactly what the articulators are doing?

- tongue position
- tongue–palate contact
- velum position
- vocal fold activity
- airflow

Measuring tongue contact and vocal fold activity

- Electropaletography (EPG)
- Laryngograph (with demo?)

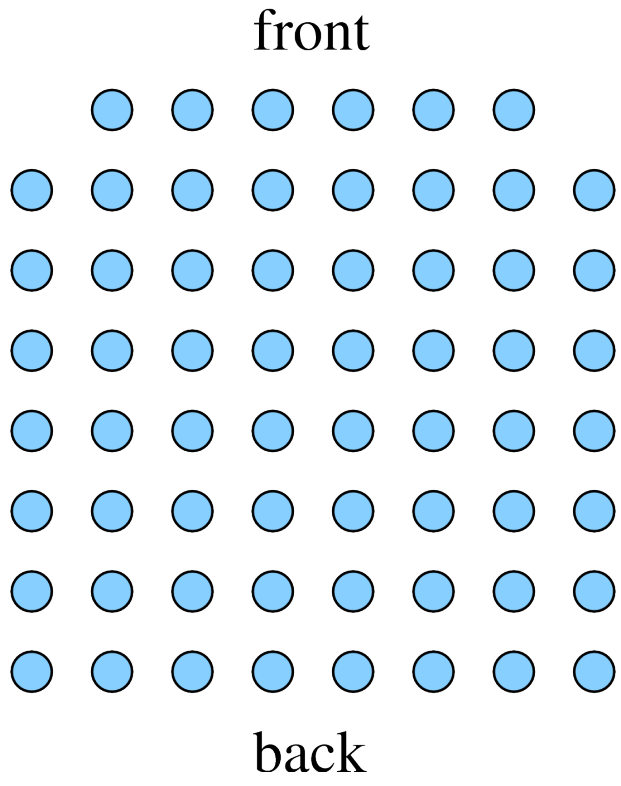
Electropaletography (EPG)

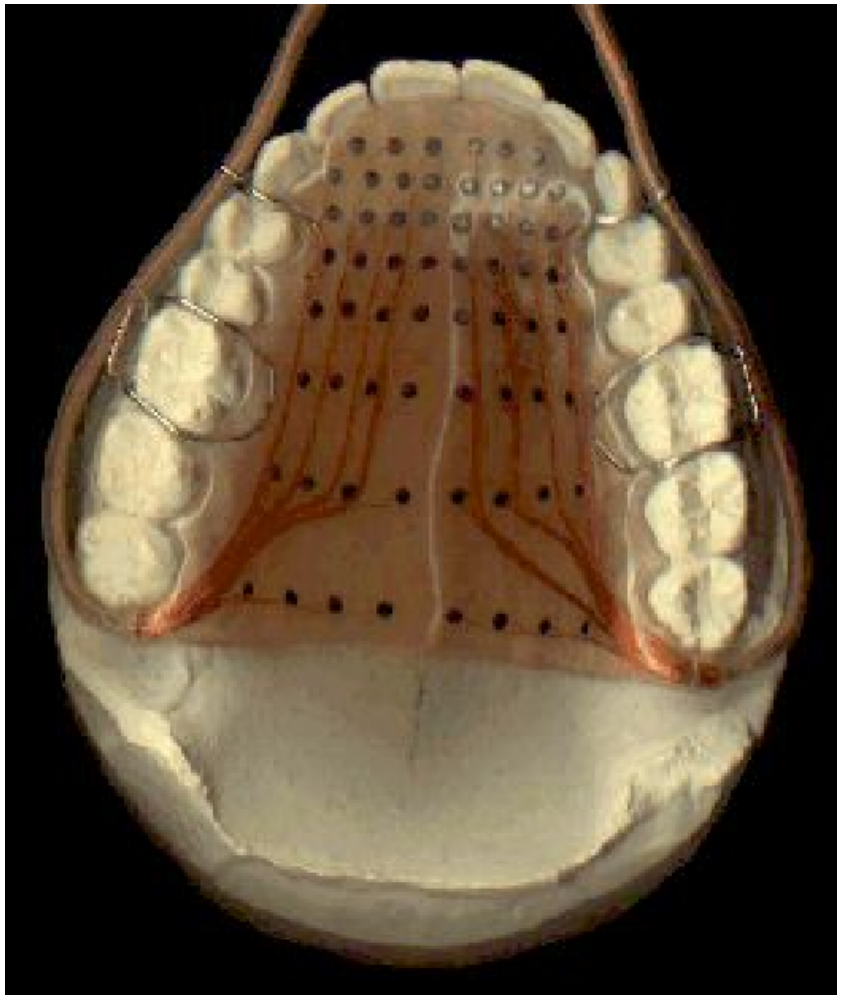
- precisely locates position, size and shape of tongue-palate contact
- sounds such as
  - /t, d, k, g, s, z, S, Z, tS, dZ/
  - the palatal approximant /j/
  - nasals /n, N/
  - lateral /l/
  - relatively close vowels such as /i, I, e/
  - diphthongs with a close vowel component such as /eI, AI, oI/

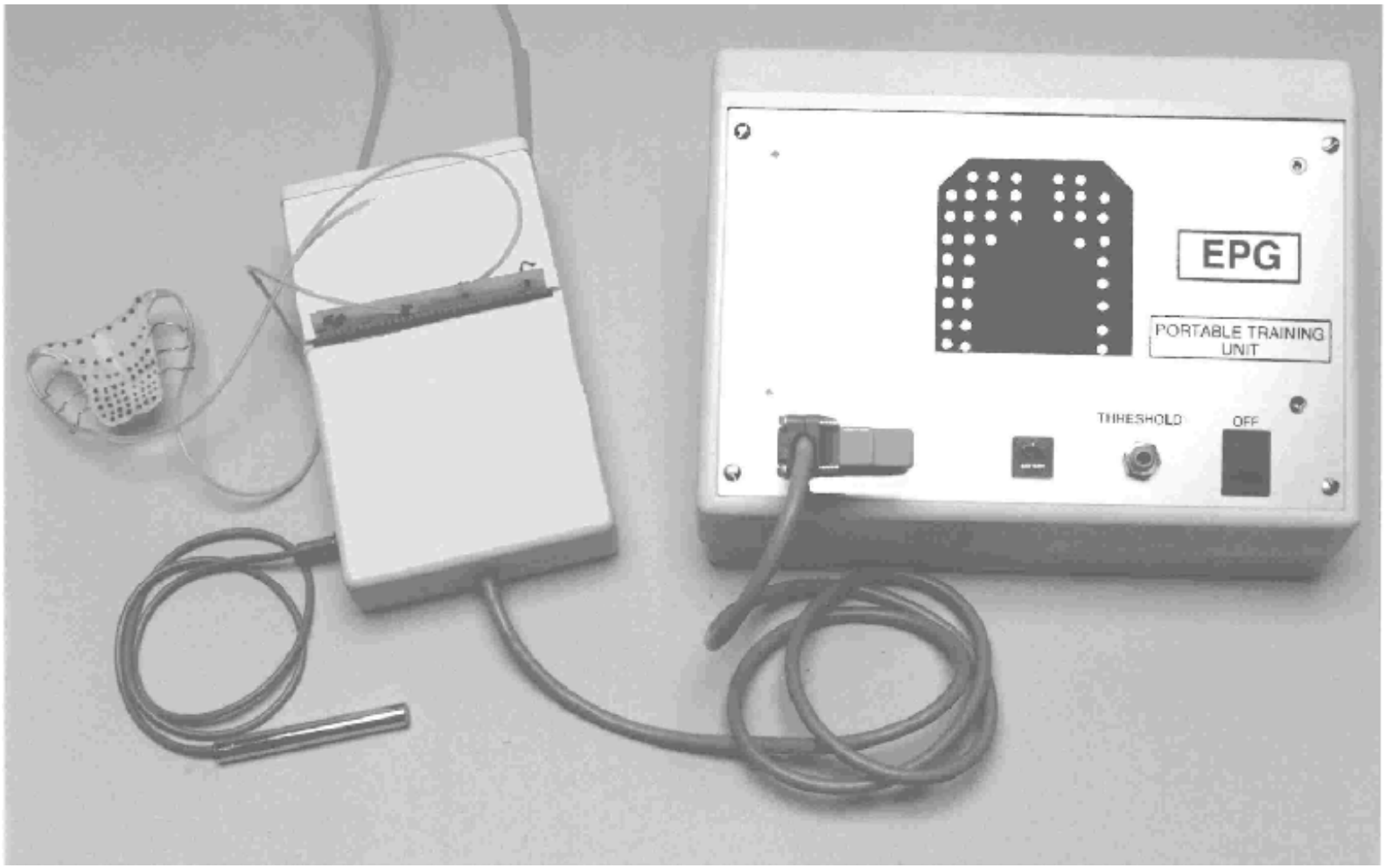
EPG palate

A grid of 62 silver contacts embedded in a hard plastic palate

Must be custom-made for each person







7:6

Laryngograph

- measures vocal fold activity
  - essentially, measures glottal width
- can plot frequency of vocal fold vibration
- i.e. F0

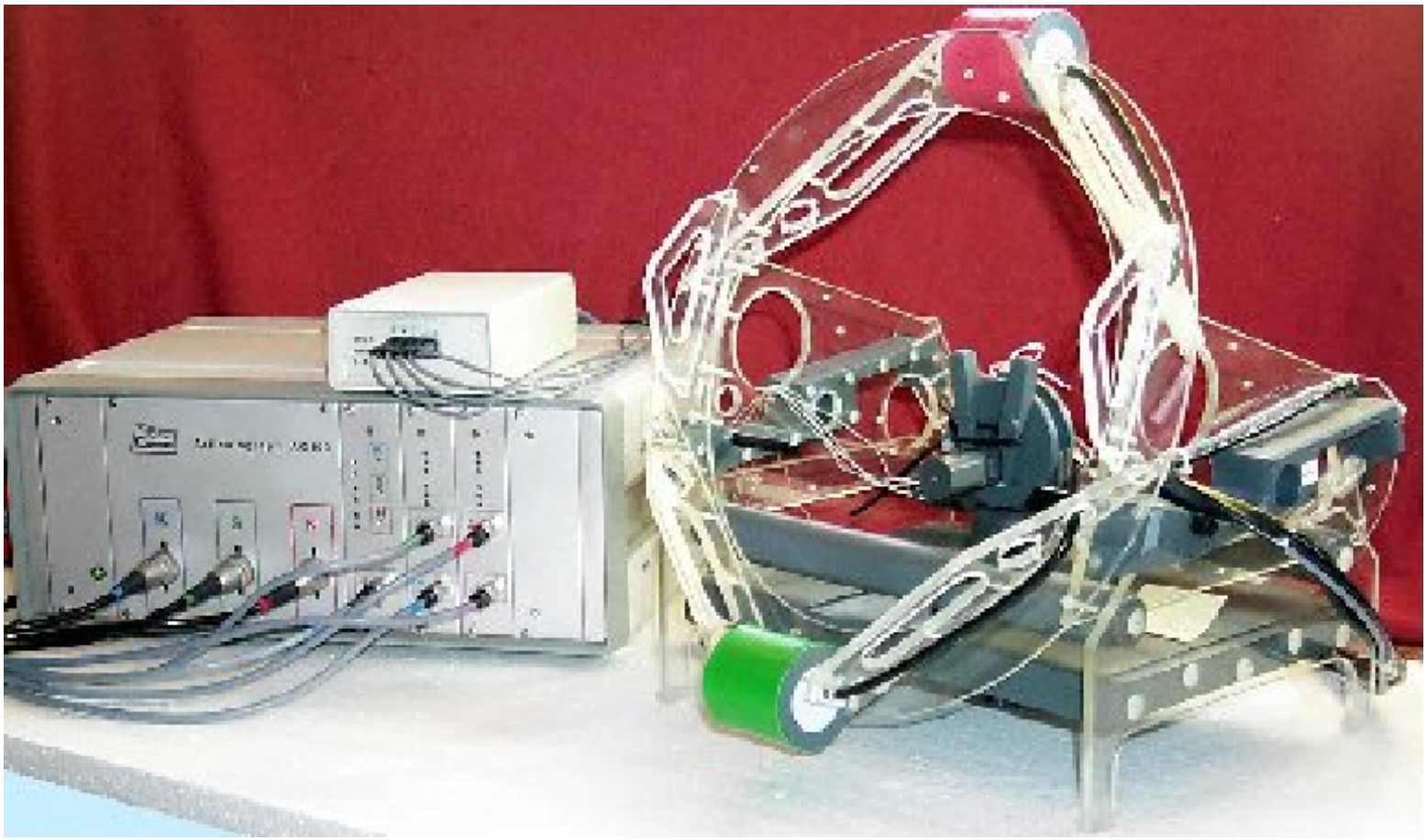
Laryngograph

Also known as an electroglottograph

Measuring articulation: the tongue and velum

- Electromagnetic articulograph (EMA)
- what does it measure?
- how does it work?
- video

Electromagnetic articulograph (EMA)

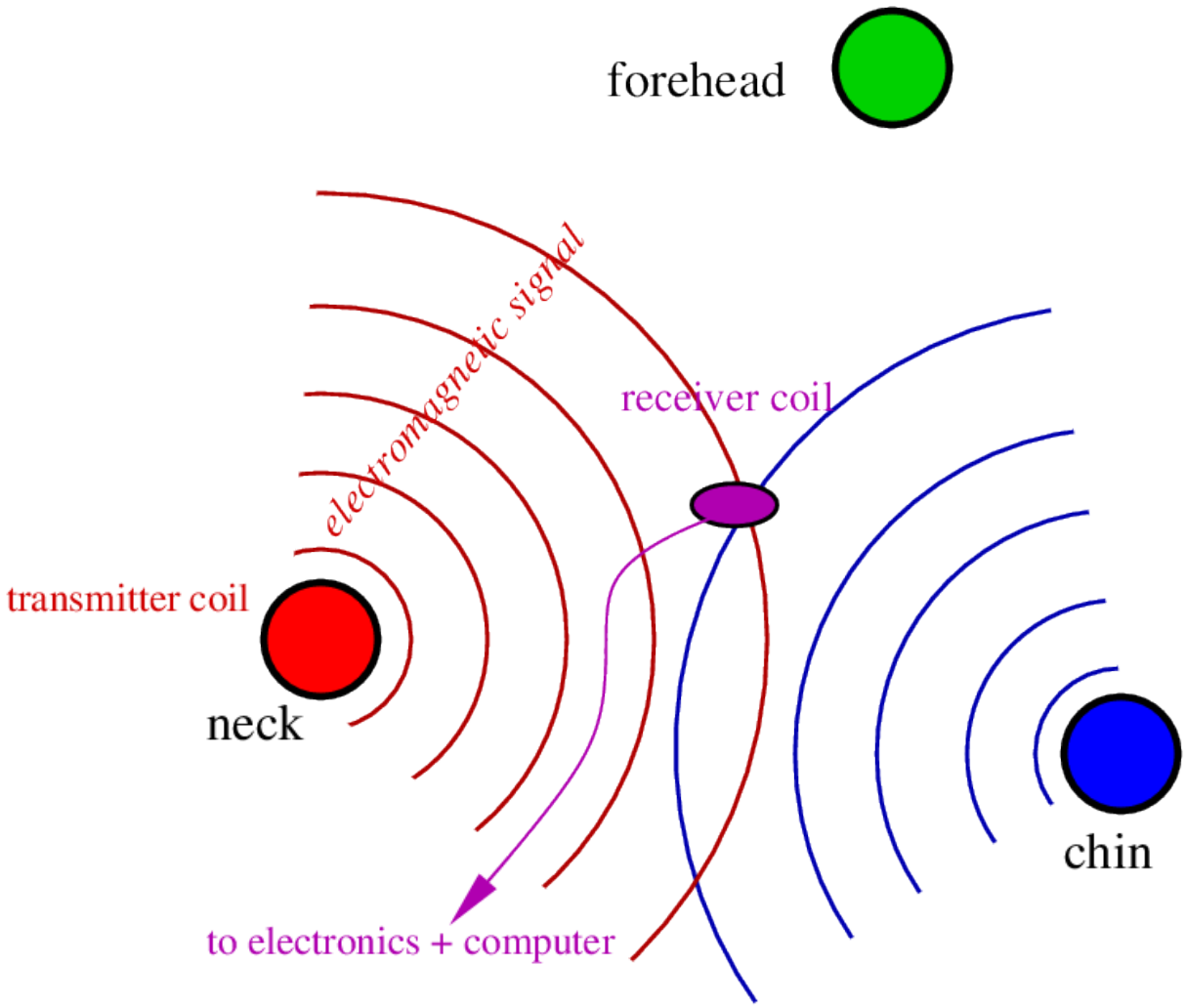


EMA

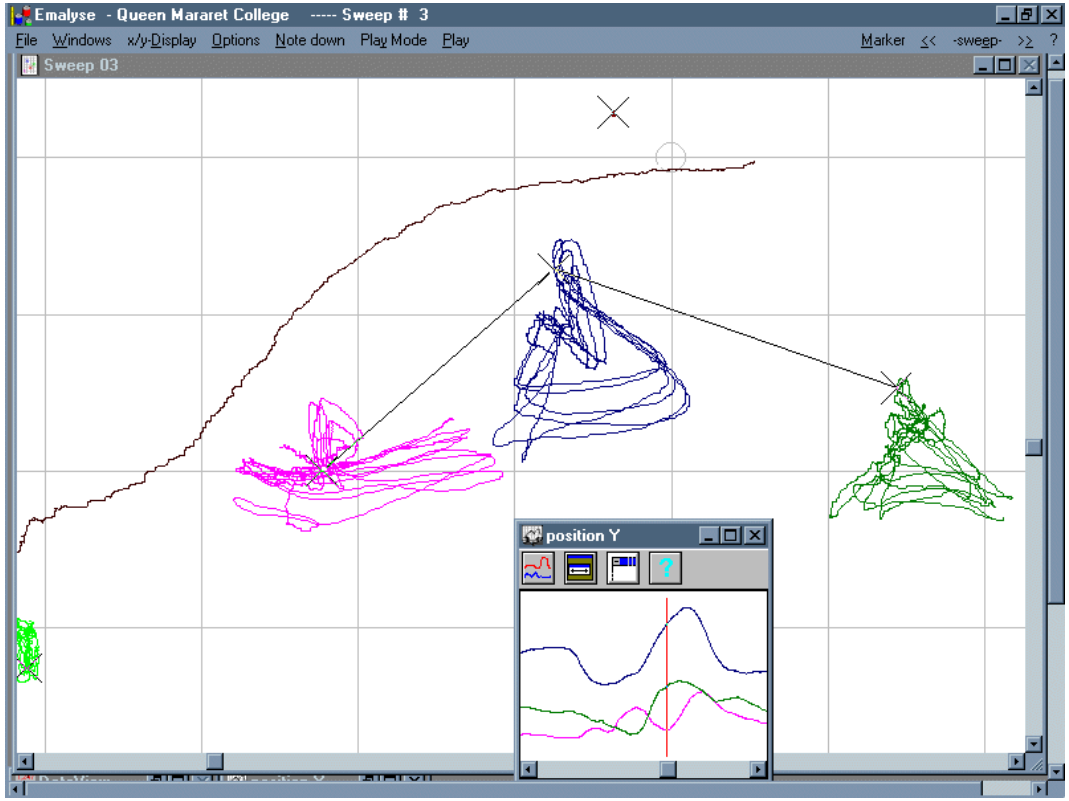
- helmet with three transmitter coils
  - transmit high frequency signals (think of them as radio waves)
- tiny receiver coils attached to points of interest
  - several on tongue
  - lips
  - lower incisors (jaw)
  - lips
  - possibly also velum

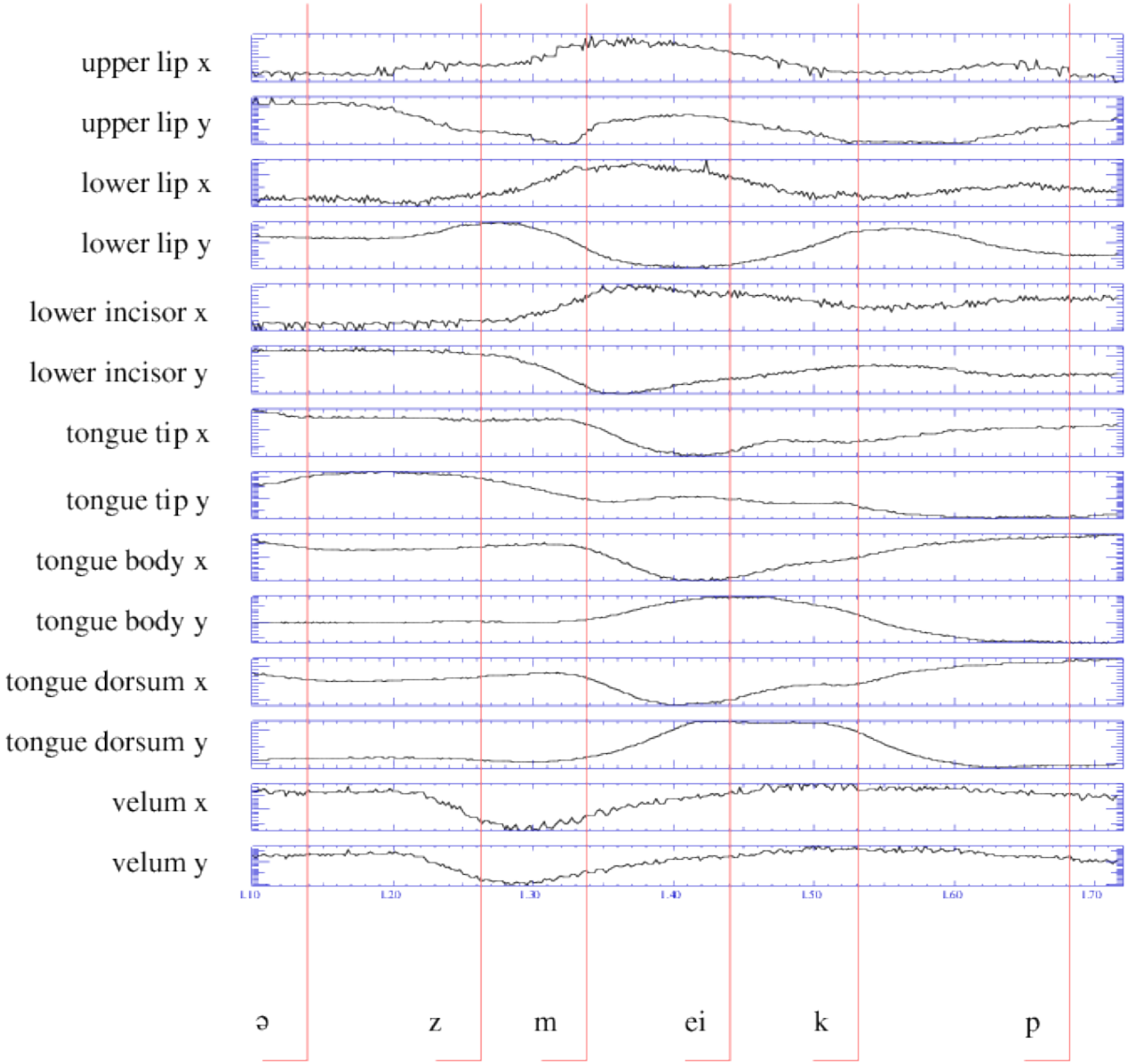
EMA

- distance between receiver and transmitter coils can be calculated automatically (and in real time – typically 500 times per second)
  - so, can plot 2-dimensional position of receiver coils
    - i.e. low–high and front–back
  - accurate to within around 1mm
- receiver coils must be mid-saggital (i.e. in centre)
  - so cannot see tongue grooving, for example
  - but new 3-D version has just been released


forehead
electromagnetic signal
receiver coil
transmitter coil
neck
chin
to electronics + computer

EMA data


Emalyse - Queen Mararet College ----- Sweep # 3
File Windows x/y-Display Options Note down Play Mode Play
Marker <- -sweep- ->
Sweep 03
position Y


upper lip x
upper lip y
lower lip x
lower lip y
lower incisor x
lower incisor y
tongue tip x
tongue tip y
tongue body x
tongue body y
tongue dorsum x
tongue dorsum y
velum x
velum y
1.10
1.20
1.30
1.40
1.50
1.60
1.70
ə
z
m
ei
k
p

X-ray cinematography

- full frame movie
  - therefore relatively large X-ray dosage!
  - now illegal in most countries
- shows all articulators: jaw, velum, tongue, lips
- only shows “cross-section” of head, so impossible to see tongue grooving, for example

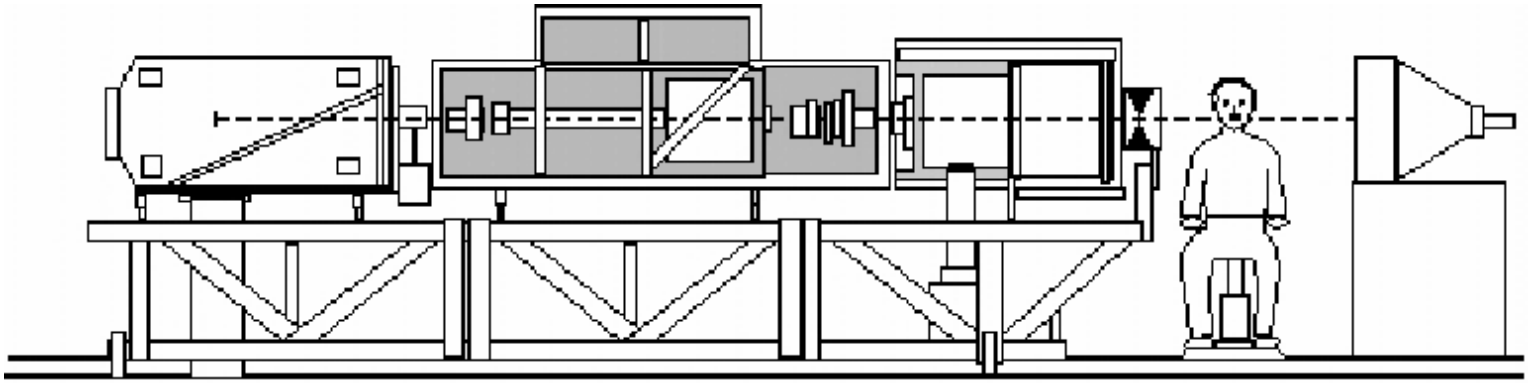
[movies on web]

X-ray microbeam

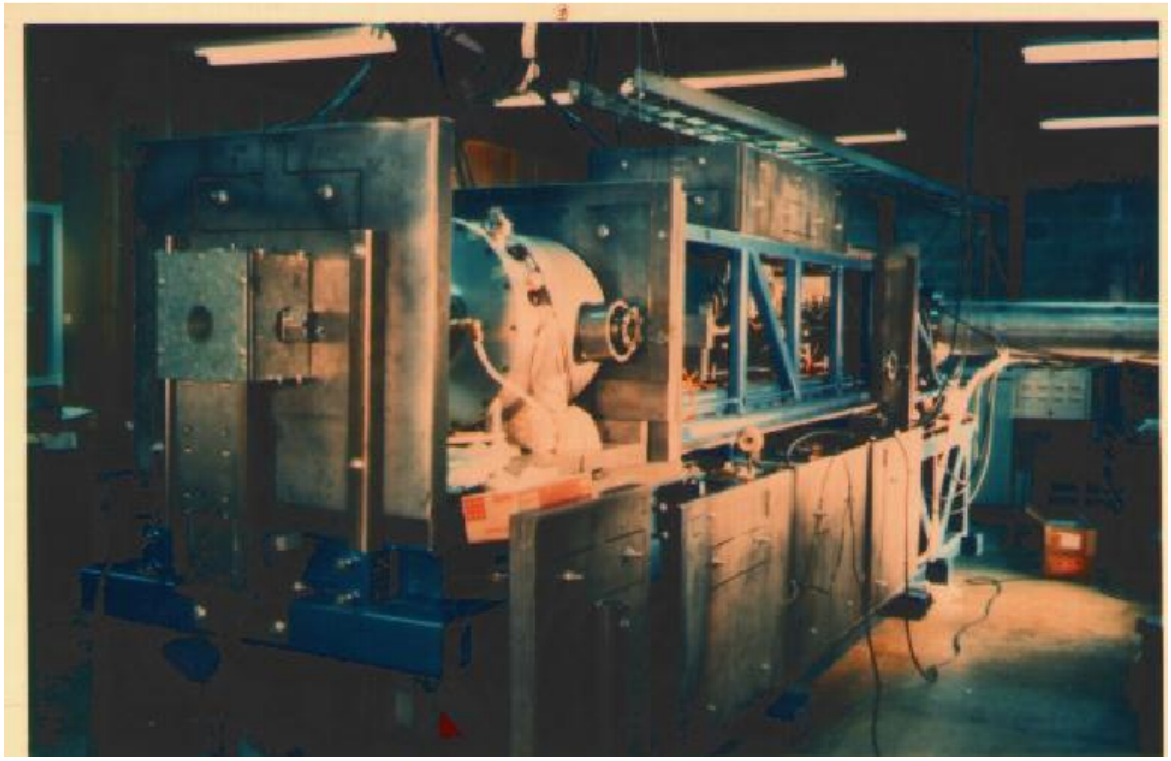
- gold pellets attached to points of interest
- narrow beam of X-rays tracks the pellets
  - relatively low doses of X-ray
  - but subjects can still only record limited amounts of data
- machine quite noisy
  - recorded speech signal not very high quality
  - may even affect way speaker produces speech

X-ray microbeam

Large, very expensive equipment

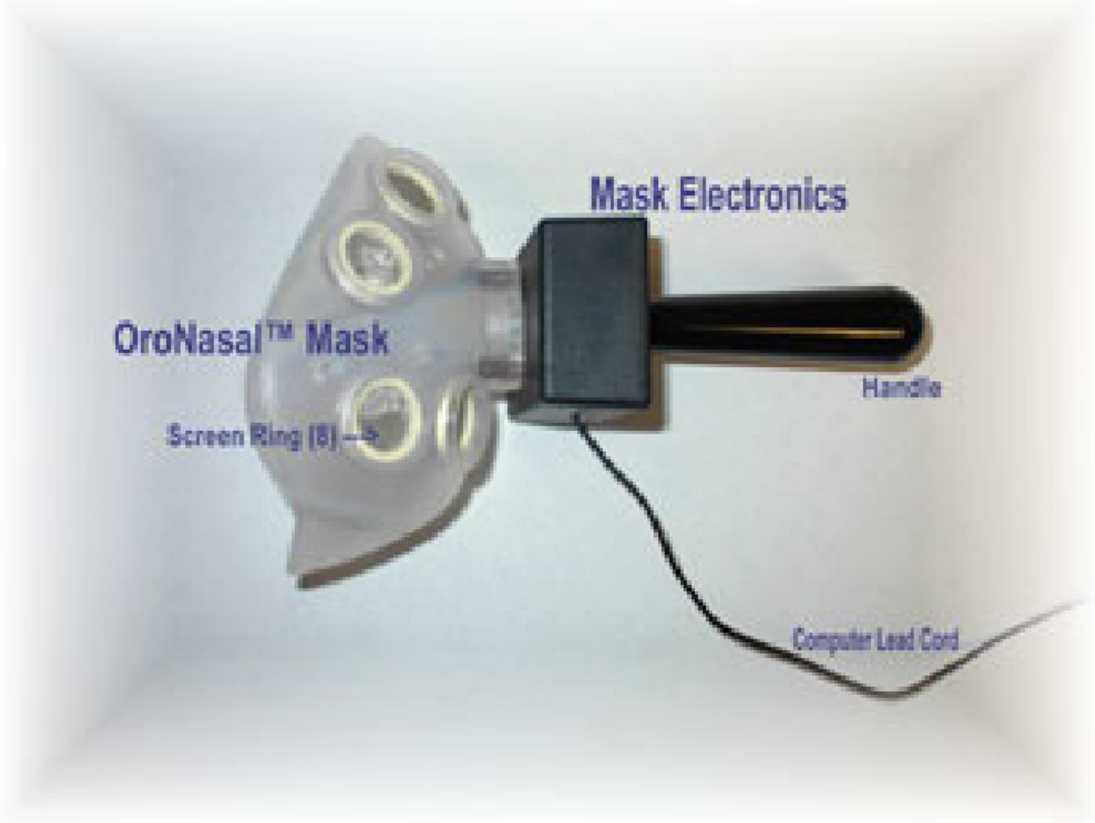


X-ray microbeam



Airflow masks

e.g. Rothenburg mask

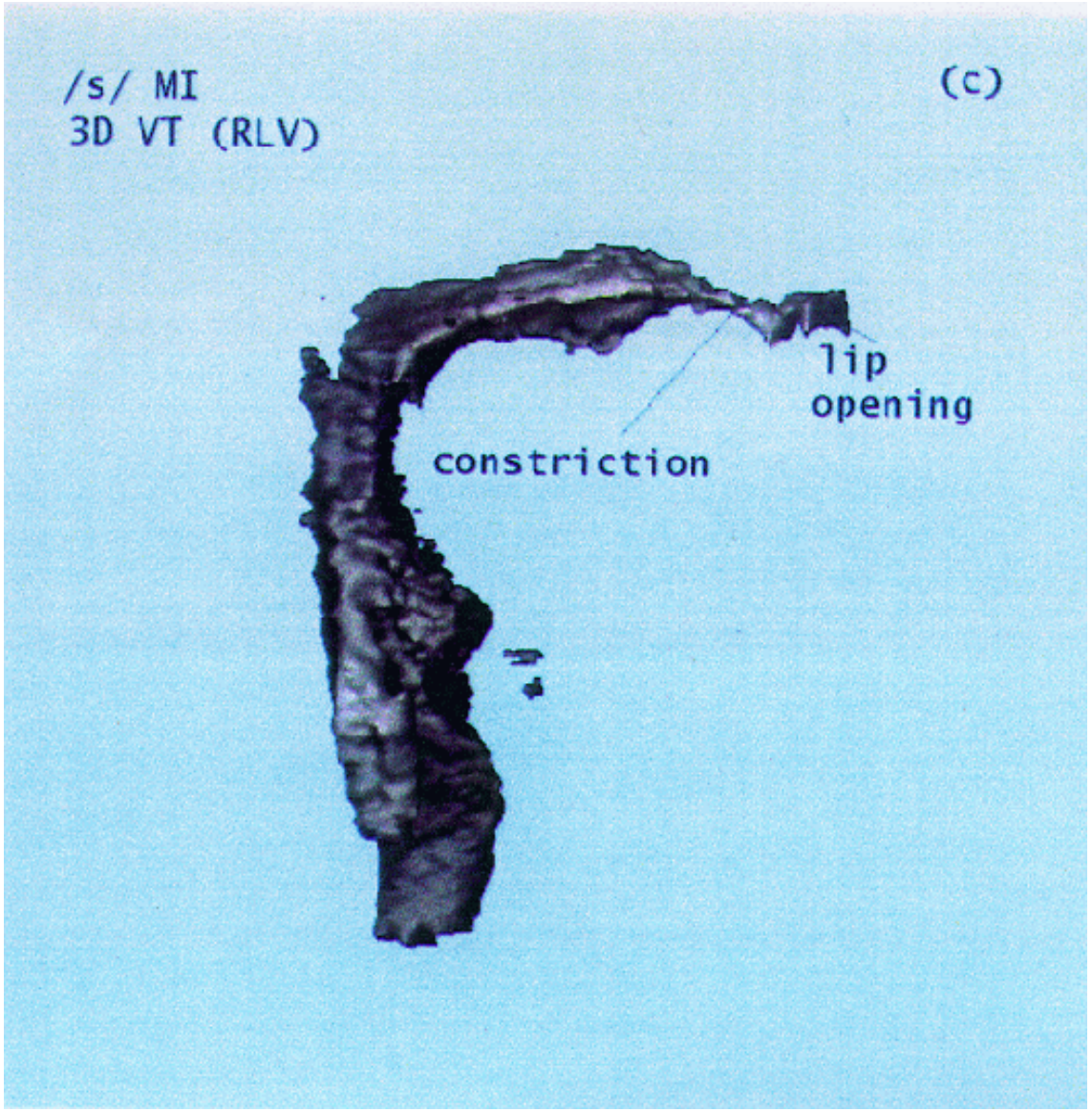


Magnetic-resonance scanning

e.g. fMRI

- full 3-D picture of body organs and cavities (e.g. vocal tract)
- fairly high resolution
- but a complete scan too slow for speech
  - can take several seconds
  - so cannot (with current technology) get moving images
  - user must “freeze” articulator positions during scan
- technology always improving, so one day maybe 3-D full motion images of vocal tract

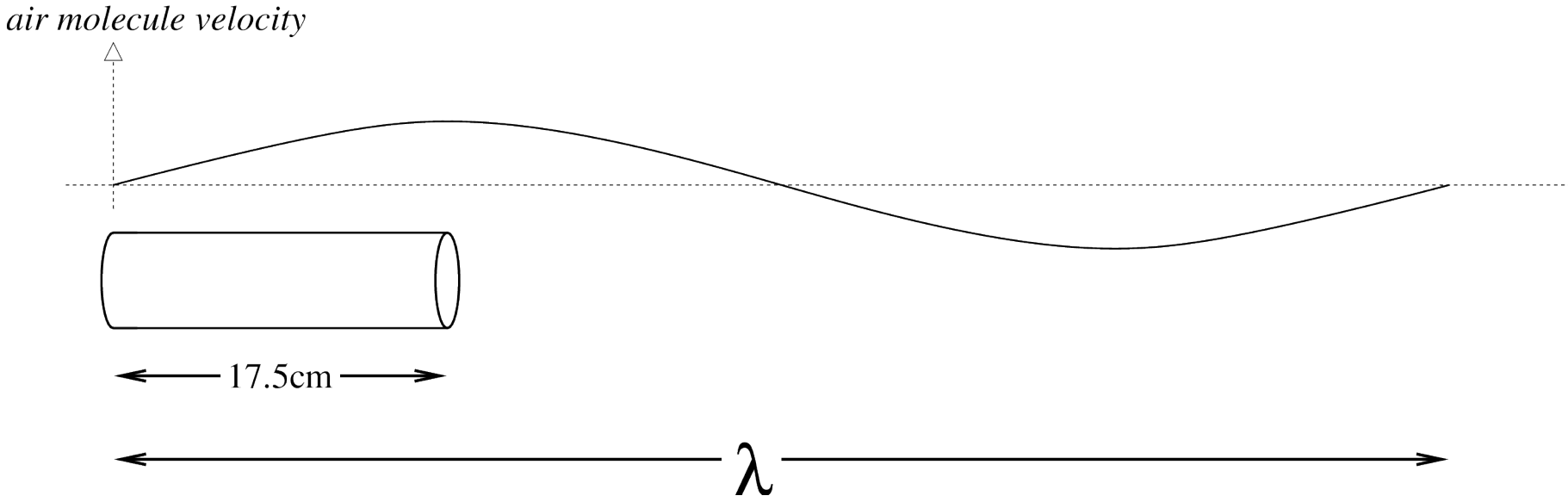
fMRI


/s/ MI
3D VT (RLV)
(c)
lip
opening
constriction

Two-tube models of vowels

- first, recap one-tube model
- then, two-tube models
  - still very simple model
  - each tube has it's own resonance(s)
- predict formant frequencies of some more vowels

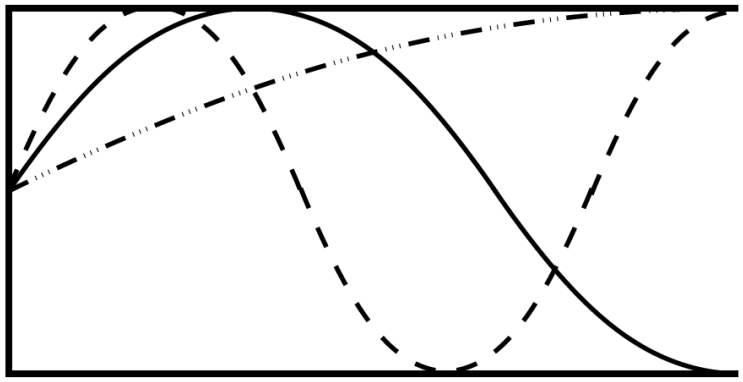
Recap - one-tube model



λ = 4 $\times$ 17.5cm = 70cm = 0.7m

$$f = \frac{c}{\lambda} = \frac{350 ms^{-1}}{0.7m} = 500\,\text{Hz}$$

Recap - multiple resonances in the same tube



This model predicts that the first three formants of [ə] are 500Hz, 1500Hz and 2500Hz.

How accurate is the model?

- if a tube is much longer than it is wide
  - i.e. if length $>$ 4 $\times$ width
- it's resonances depend only on it's length

is this true for the vocal tract?

- length around 17.5cm
- width varies from few mm to 3.5cm

so yes, it's true

Two tubes

- why two tubes?
  - vocal tract shape not uniform for vowels other than [ə]
- still a very simple treatment
  - tubes act *independently*
  - i.e. treat each one like the simple tube for [ə]
- what are the lengths of the tubes?

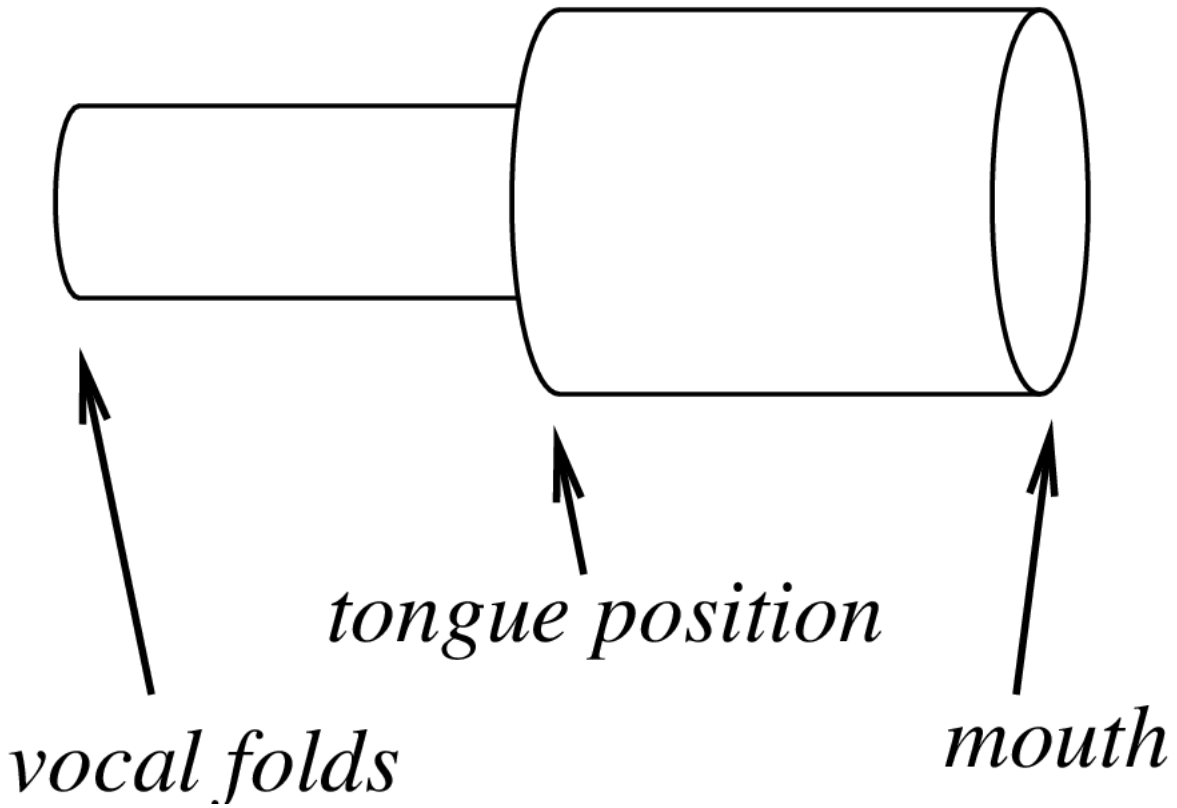
A two-tube model

for [ɑ] as in *father*

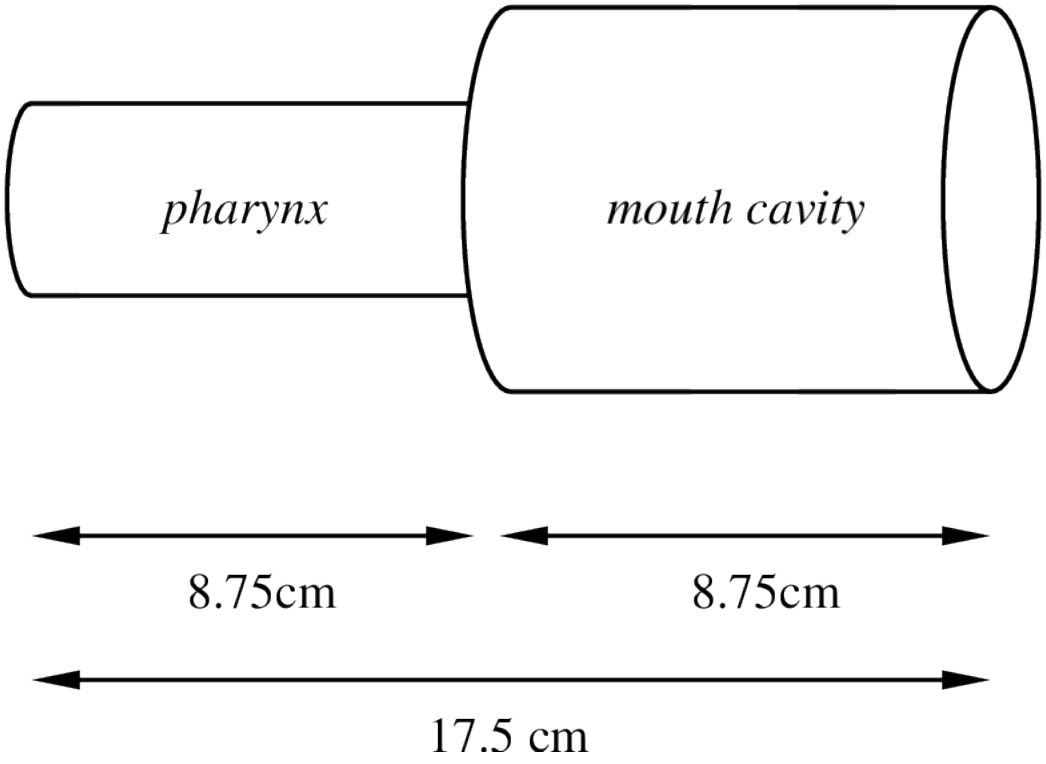
- tongue position
  - low
  - back
- lips
  - unrounded
- velum
  - raised

i.e. it's a **low, back, unrounded vowel**

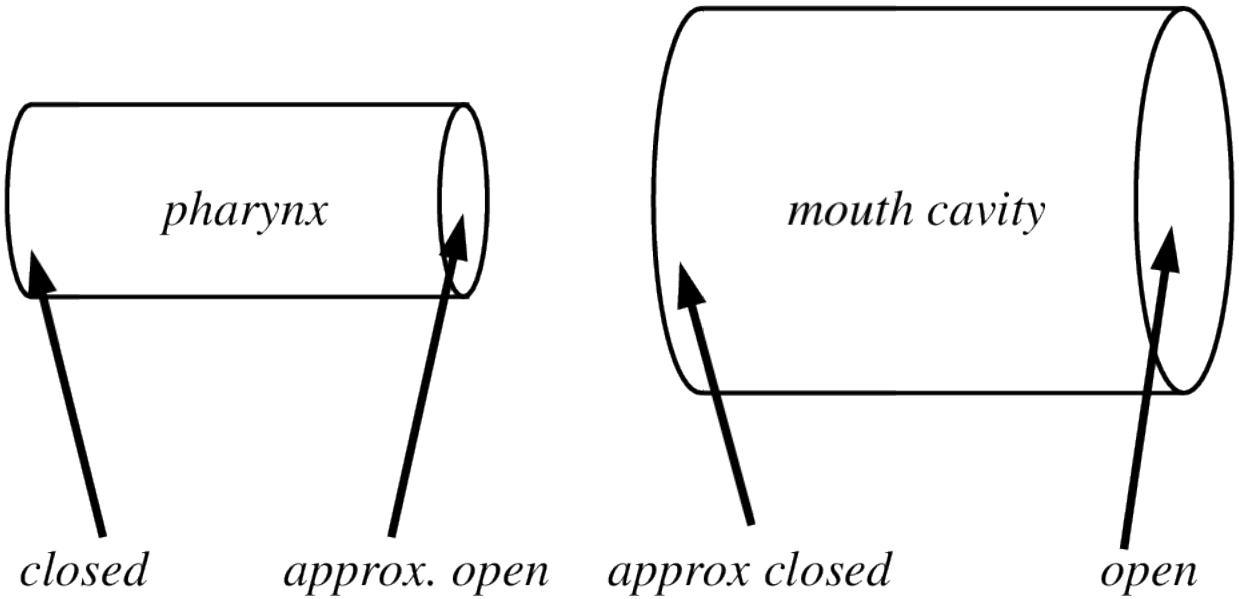
Tube shape



Tube dimensions



Tube types



i.e. both tubes are (approximately) closed at one end and open at the other

Resonances of the two tubes

- both tubes are still much longer than they are wide
  - so the resonance dependes only on the length
- both tubes have same length
- which is half vocal tract length

so resonances of each tube will be twice those for the [ə] model

i.e. 1000Hz, 3000Hz, 5000Hz, ...

which means F_1=1000Hz

A small complication

in reality....

- tubes are not completely independent
  - the pahrynx tube is not *exactly* open at one end - it connects to the mouth tube
  - and the mouth tube is not *exactly* closed at one end - it connects to the pharynx tube
- the effect of this *coupling* is that
  - the pharynx tube resonant frequencies drop slightly
  - the mouth tube resonant frequencies rise slightly

The prediction

So, our prediction becomes:

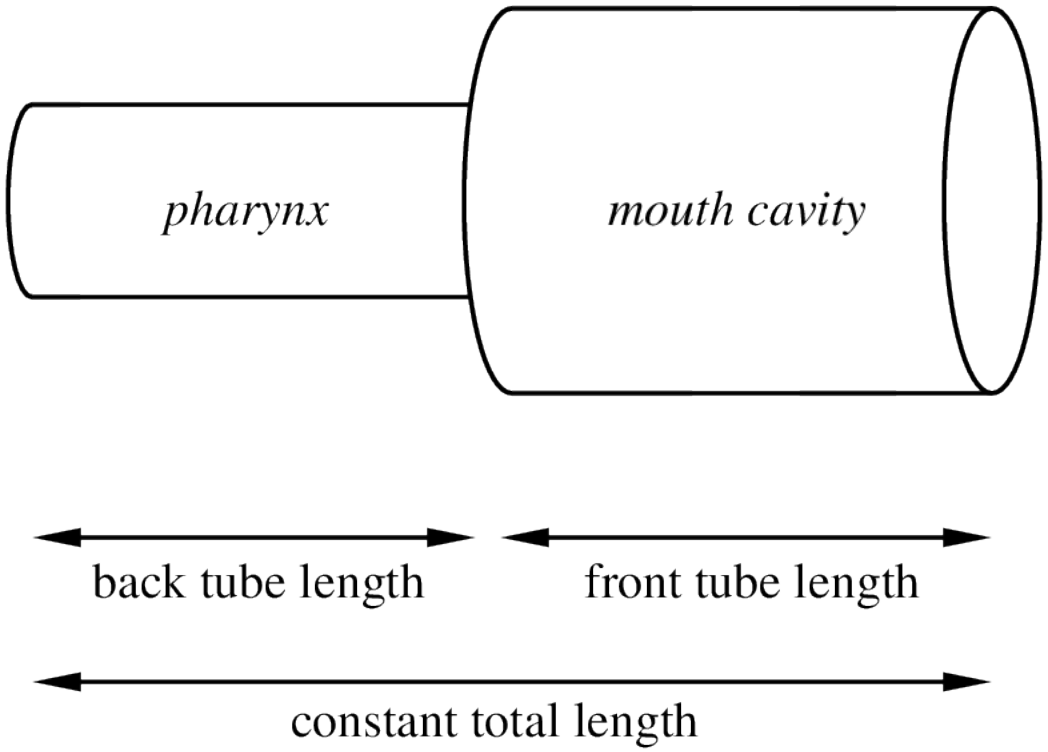
For the low, back, unrounded vowel [ɑ]

- F_1 is produced by the back tube and is about 900Hz
- F_2 is produced by the front tube and is about 1100Hz

How can F_1 and F_2 vary?

- the tube lengths must change
  - because the tongue moves position
- can we predict possible values for F_1 and F_2 ?

Varying length tubes



as back tube gets longer, front tube gets shorter, and vice versa

Back tube resonances

- depends on back tube length
  - for length of 8.75cm we already predicted 1000Hz, 3000Hz, 5000Hz,...

Remember the formula $c = f\lambda$? From this we can write a formula for the resonances of a tube closed at one end, because the wavelengths that fit in such a tube are $\lambda_1 = 4L$, $\lambda_2 = \frac{4}{3}L$, $\lambda_3 = \frac{4}{5}L$,... in other words:

$\lambda_n = \frac{4L}{(2n-1)}$ where $n = 1, 2, 3, ...$

so:

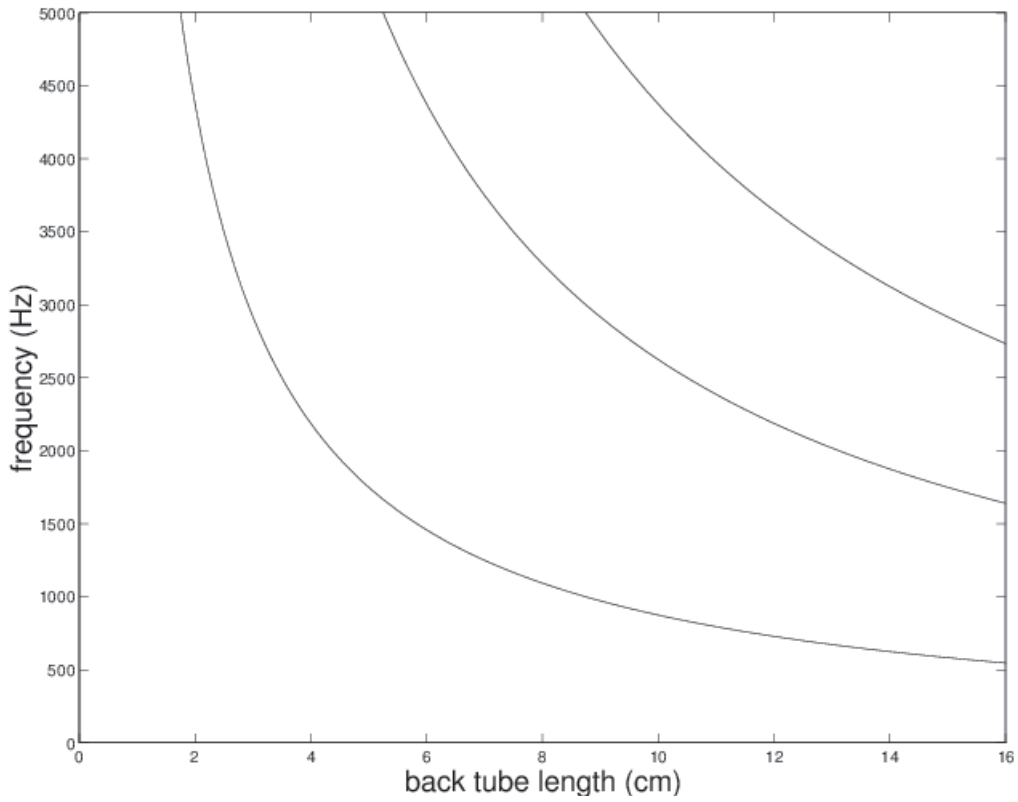
$f_n = \frac{c}{\lambda_n} = \frac{(2n-1)c}{4L}$

Relationship between f and L

$$f_n = \frac{(2n-1)c}{4L}$$

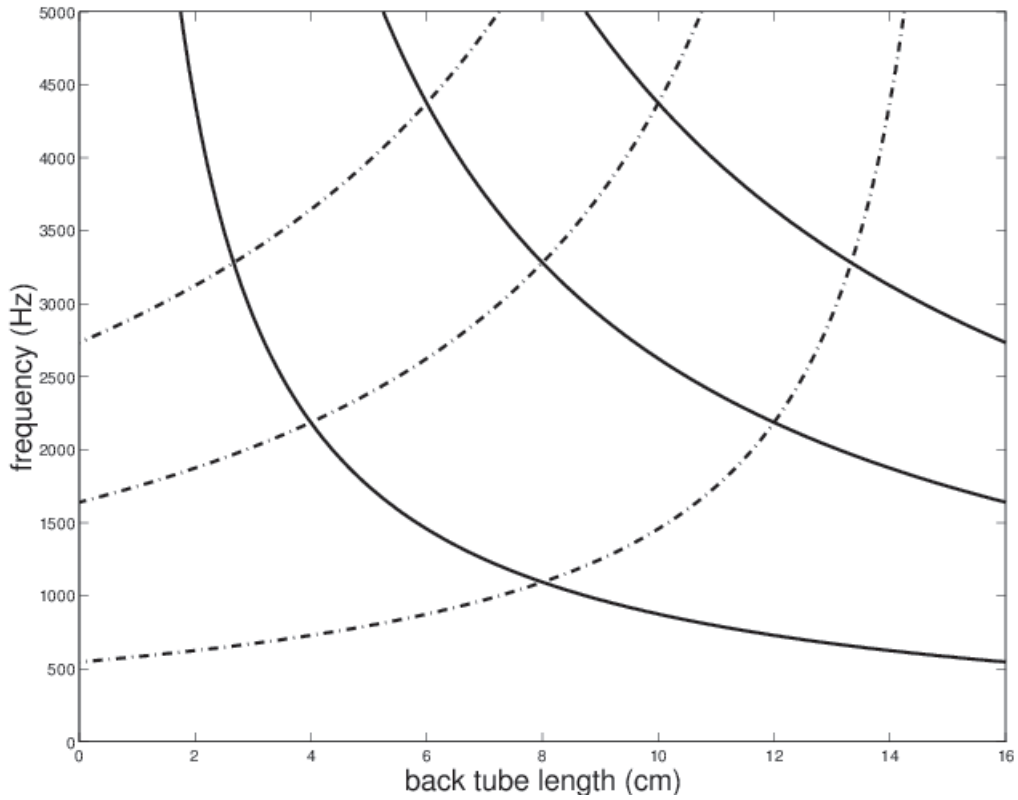
- an inverse relationship
  - as tube length L goes up
  - resonant frequencies f_n all go down
- let's plot f_n against L

Plot of f_n against L



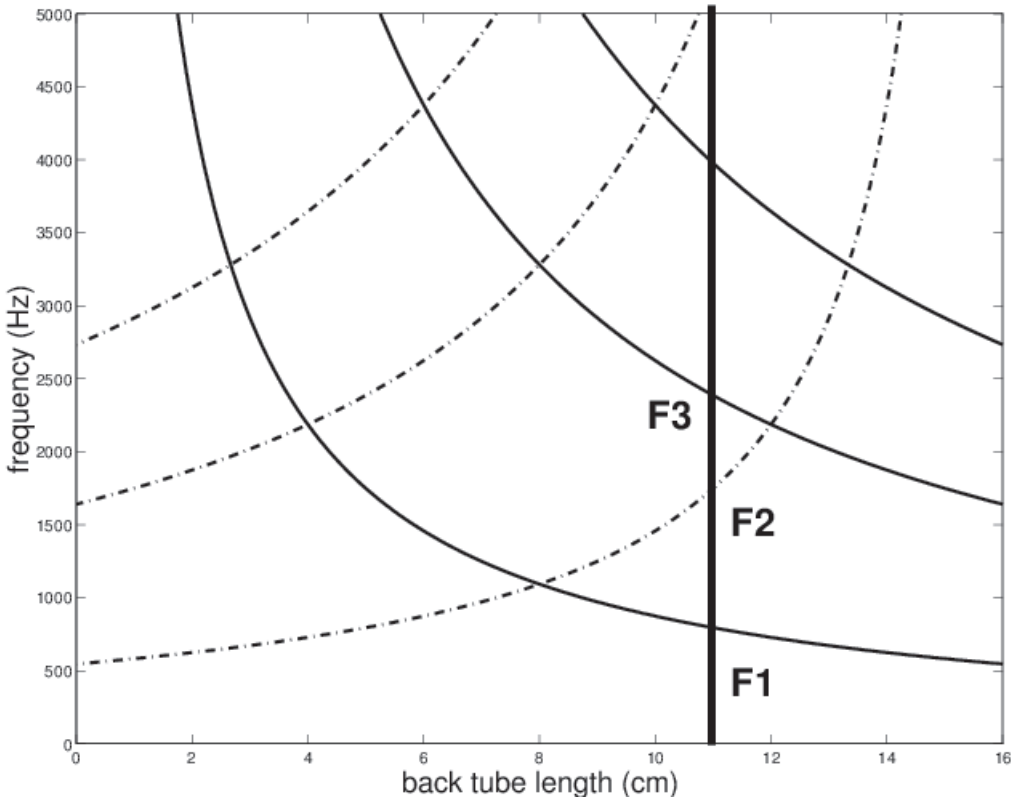
these are *just* the back cavity resonances: counting up from the bottom: f_1, f_2 and f_3. These are NOT the formants ... yet.

Plot of back and front tube resonances



dotted lines are front cavity resonances

Working out the formants from this plot



Pick a particular back tube length: formants are counted up in order from lowest frequency to highest

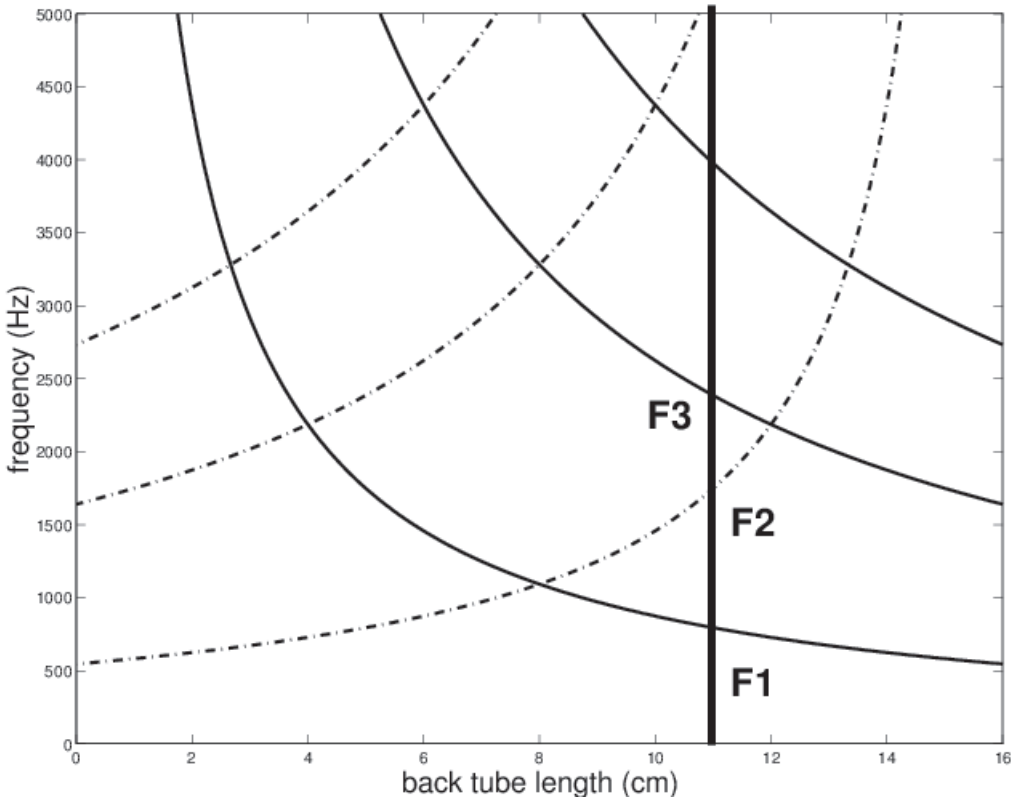
Does this tell us the range of possible formant values?

- it appears on the plot that F_1 cannot go below 500Hz
- but we observe that in real speech, F_1 *can* go lower
  - e.g. [i] as in “see” can have F_1 as low as 250Hz
- how is this possible?
- perhaps this model does not apply to [i] ?

Another type of tube model

- two tube model can predict formant frequencies, e.g. for [ɑ]
  - but appears to predict minimum F_1 of 500Hz
- some vowels such as [i] (as in “see” or “heed”) have lower F_1, maybe as low as 250Hz
- we need a better model (at least, for those vowels)

Recap - two tube model



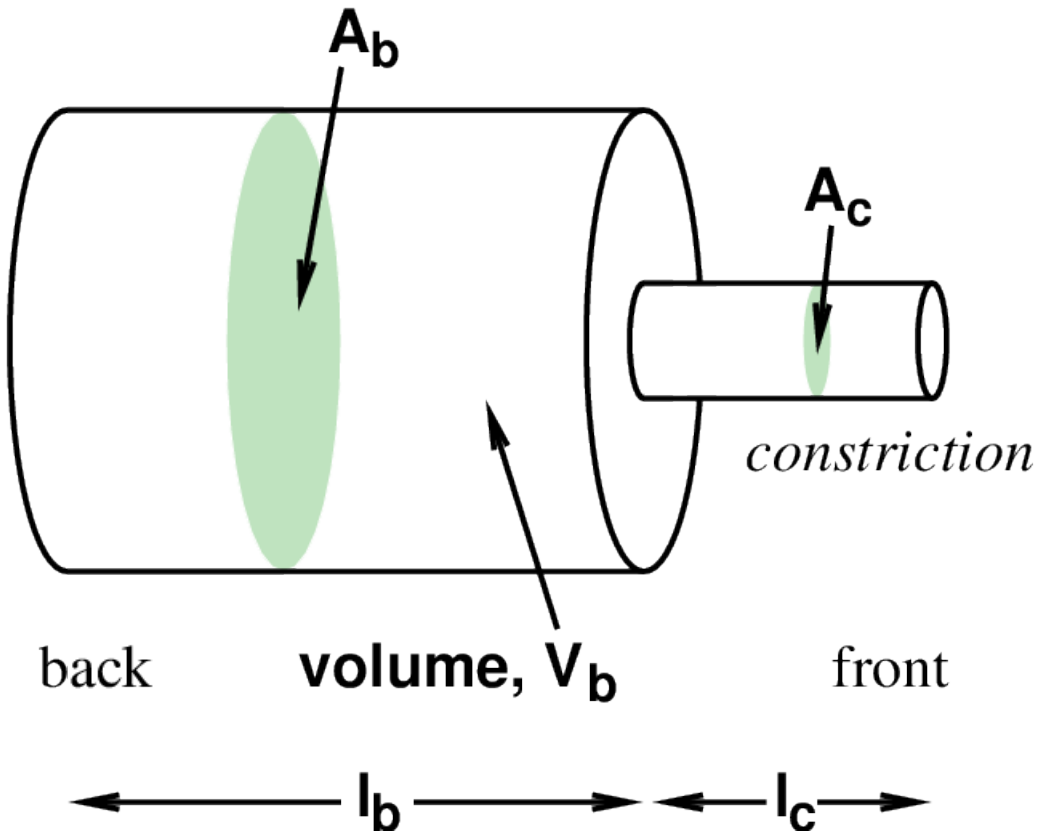
When is this model wrong?

- when the vocal tract is *not* like two tubes, each open at one end
  - but when is that?
- when one of the tubes is better approximated as being closed at both ends
- or one of the tubes is better approximated as being open at both ends
- or the two tubes have another *mode* of resonating

Another tube configuration

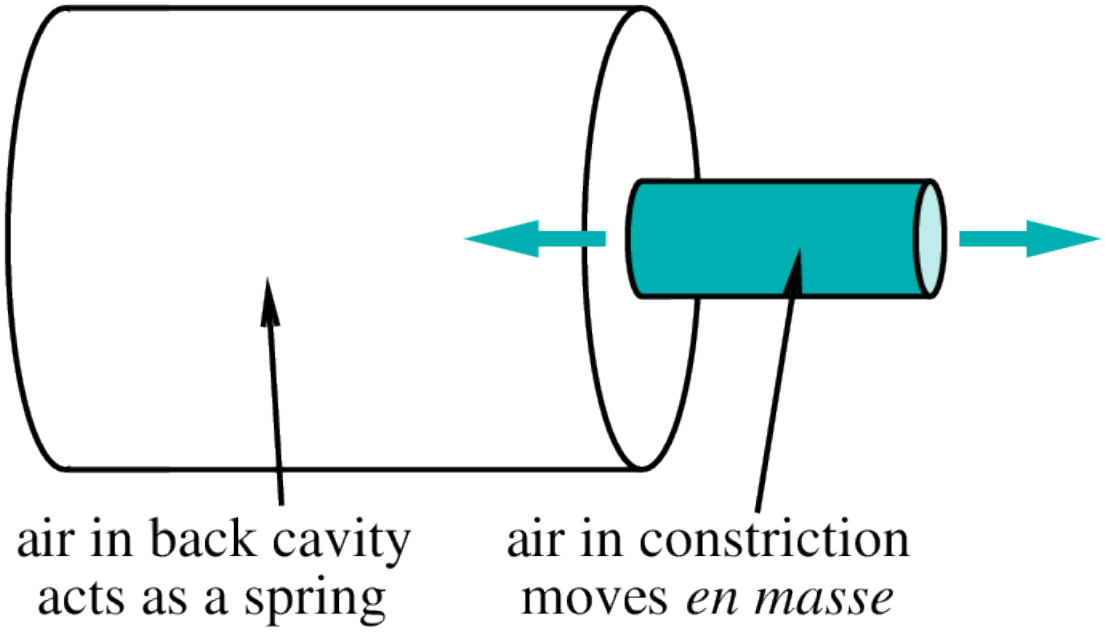
- a narrow constriction near the front of the mouth
  - back tube is approximately closed at *both* ends
  - front tube is open at both ends
  - front tube is short and narrow
- same as a wine bottle
  - bottle body is back cavity
  - bottle neck is front cavity
- this type of tube configuration has a special mode of vibration

Helmholtz resonator



back cavity volume, $V_b = A_b l_b$ constriction volume, $V_c = A_c l_c$

Helmholtz resonator



air in constriction moves back and forth like a piston in a cylinder

The formula

$$f = \frac{c}{2\pi}\sqrt{\frac{A_c}{V_b l_c}}$$

[typo in Johnson eq. 5.2 - missing π]

- where

  f is the resonant frequency of the system

  A_c is the area of the constriction

  l_c is the length of the constriction

  V_b is the volume of the back cavity

  c is the speed of sound in air

The Helmholtz resonance

- example from Ladefoged p.127, worked in proper units

  A_c = 15mm^2 = 15 $\times$ 10^{-6} m^2

  l_c = 1cm = 1 $\times$ 10^{-2} m

  V_b = 60cm^3 = 60 $\times$ 10^{-6} m^3

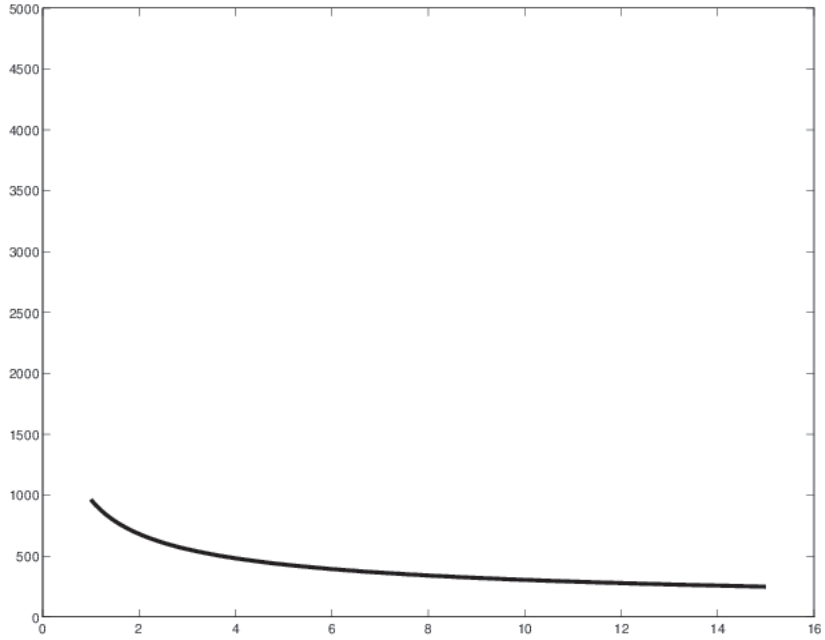
  c = 350 ms^{-1}

$$f = \frac{350}{2\pi}\sqrt{\frac{15 \times 10^{-6}}{60 \times 10^{-6} \times 1 \times 10^{-2}}}$$

$$\approx 280 \text{ Hz}$$

The Helmholtz resonance

- thick line is Helmholtz resonance for parameters as in Ladefoged examples (a constriction length of 1cm)



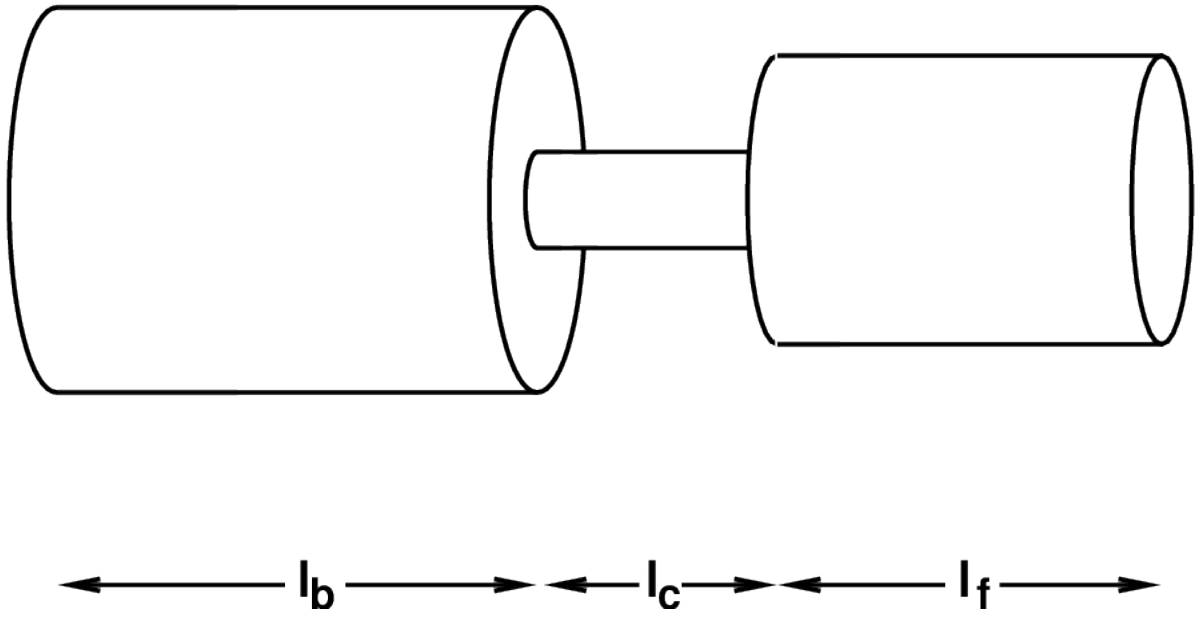
What about other formants?

- Helmholtz resonance is at a low frequency – so it is F_1
  - but how do we explain F_2 and F_3 in this case?
- remember that a system can have *multiple resonances* ?

There must be other resonances of the vocal tract, *in addition to* the Helmholtz resonance

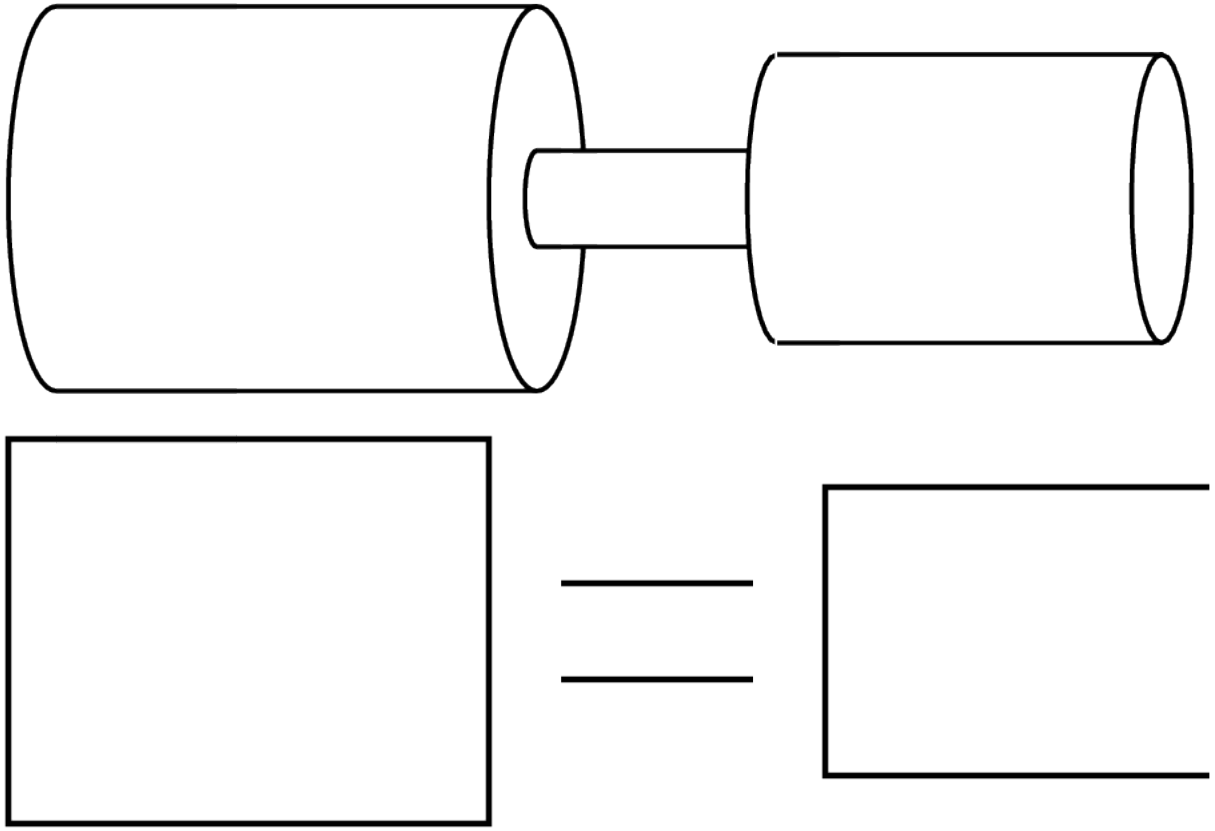
Other resonances

Add a front cavity to complete the vocal tract model



Other resonances

Approximate each tube as either completely closed or open at each end



Just one more simple formula

- we already know about a tube that is open at one end, closed at the other. It's resonances are at frequencies f_1, f_2,.... where

$$f_n = \frac{(2n-1)c}{4L}$$

- now for a tube open at both ends

$$f_n = \frac{nc}{2L}$$

- and for a tube closed at both ends

$$f_n = \frac{nc}{2L}$$

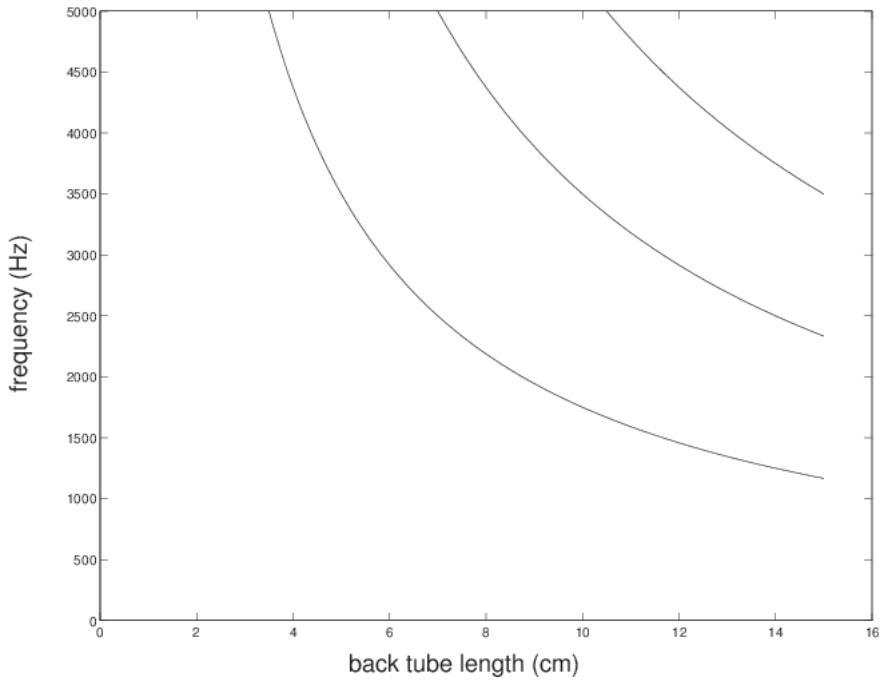
homework: revisit the animation on the web and convince yourself the above formulae are true

The non-Helmholtz resonances

- three tubes: back and front cavities, and constriction
  - constriction is very short $\longrightarrow$ very high resonant frequencies
  - so we'll neglect it
- back and front cavity length varies as position of constriction moves
  - depends on tongue position

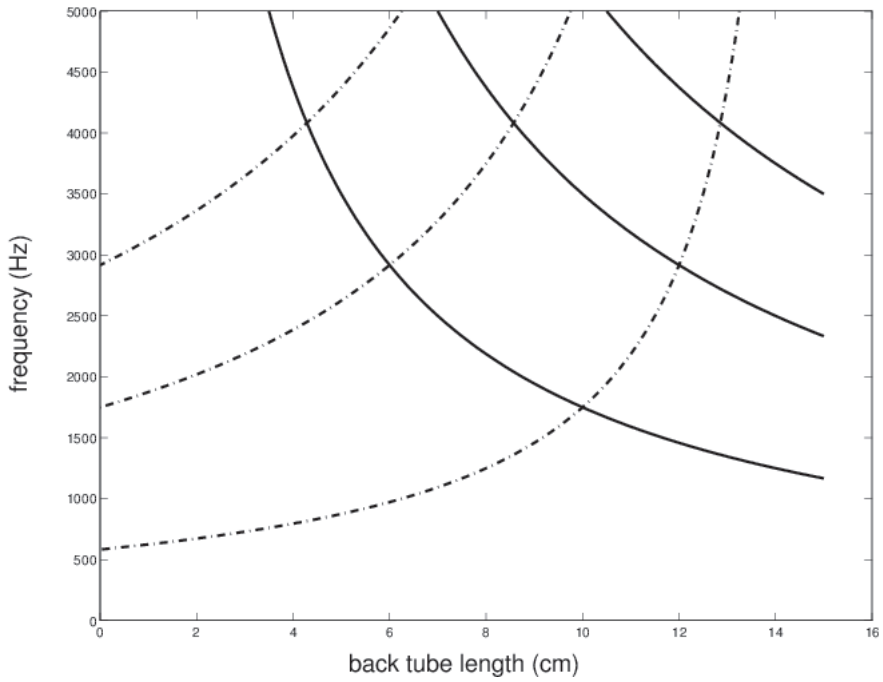
Back cavity resonances

- back cavity length depends on position of constriction



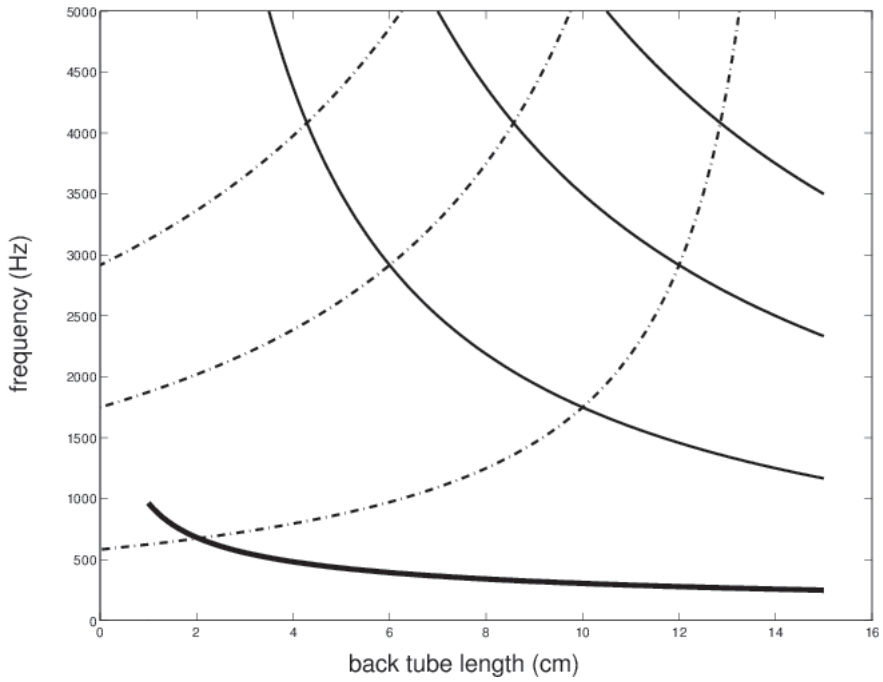
Add in the front cavity resonances

- dotted lines are front cavity resonances



All three resonances together

- thick line is Helmholtz resonance



A prediction

- for the high-front vowel [i]
  - back tube length of around 11-12cm

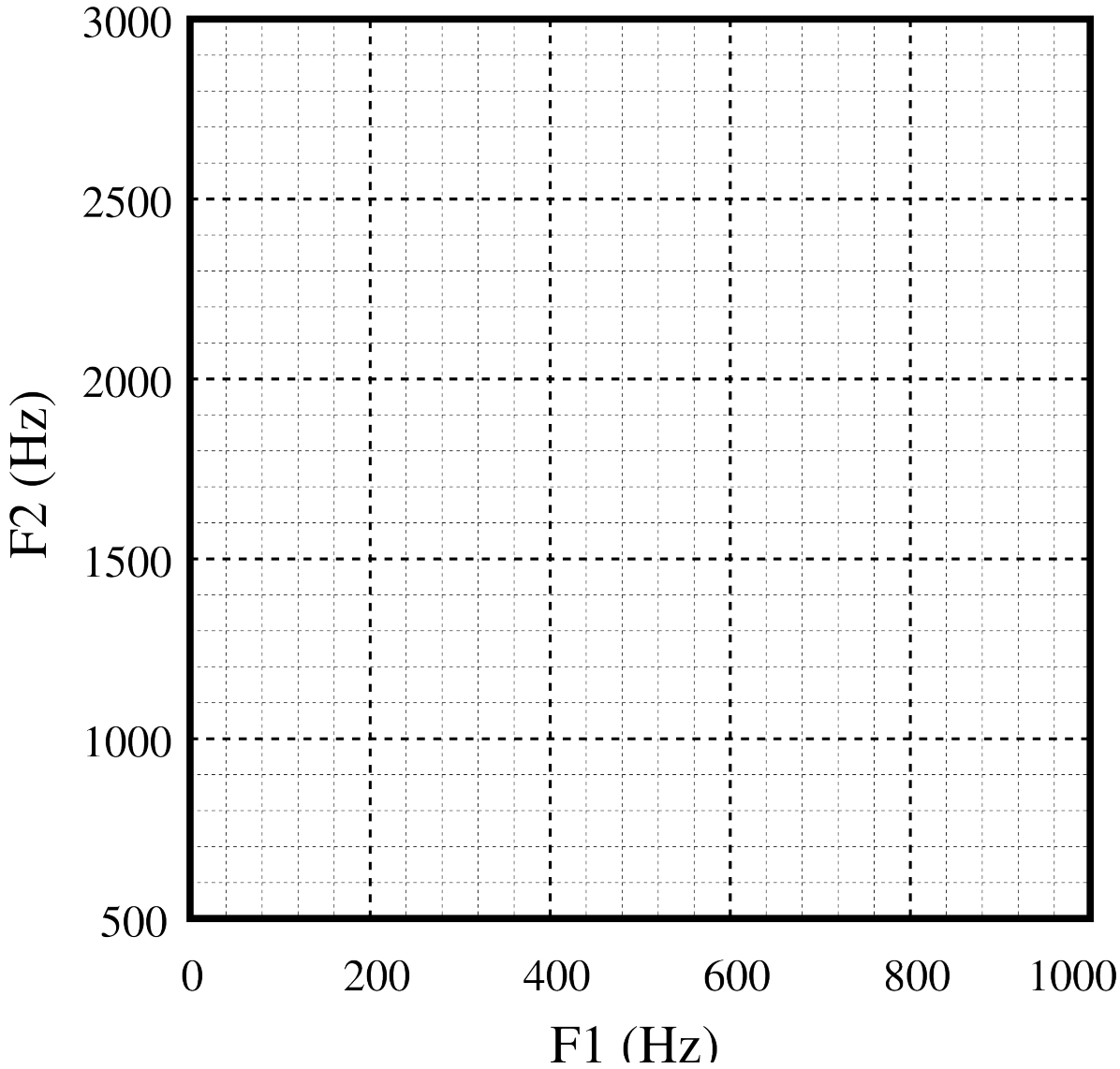
Plot shows

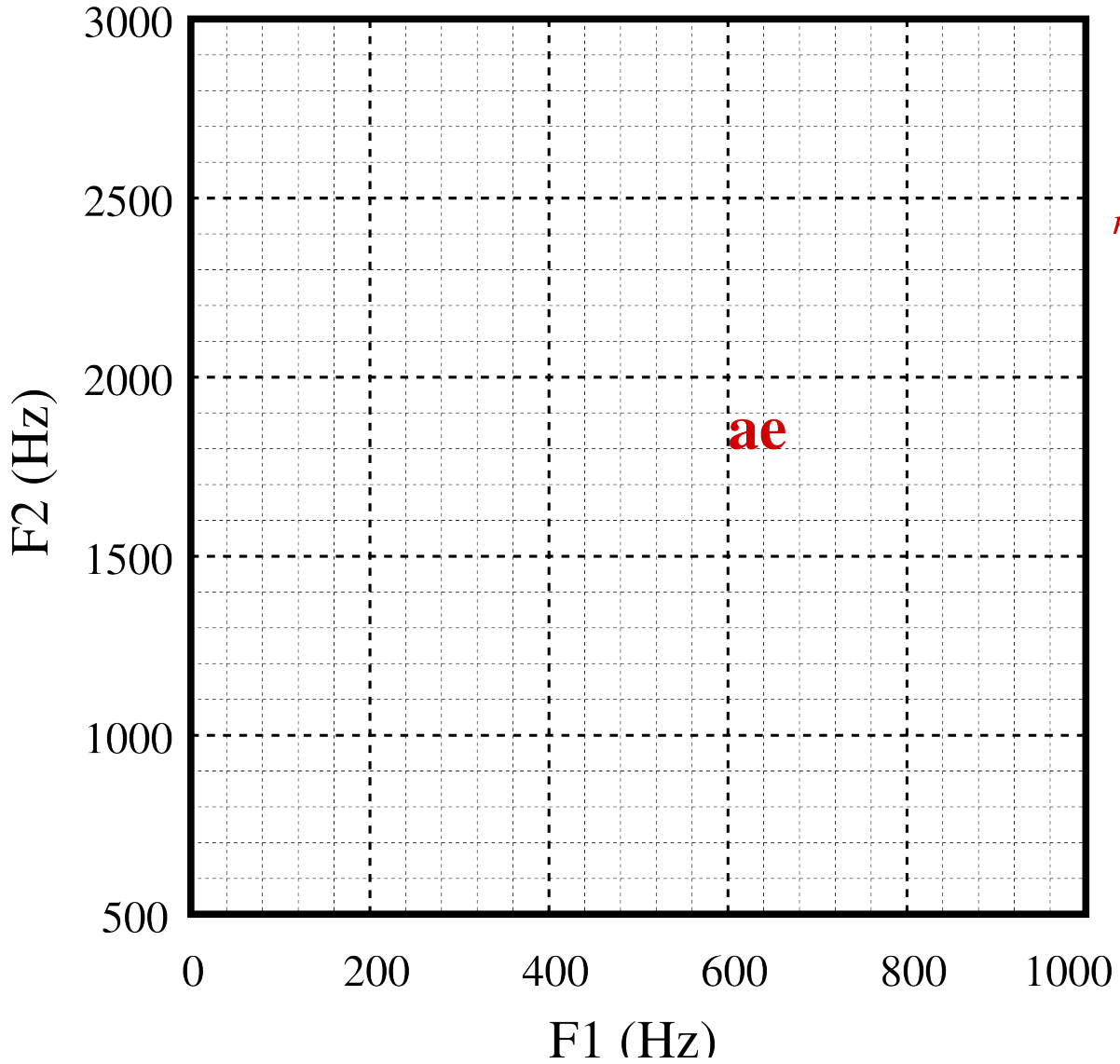
- $F_1 \approx$ 280Hz
- $F_2 \approx$ 1700Hz
- $F_3 \approx$ 2200Hz

which is in the right ball-park for high vowels, which have low F_1 and front vowels which have high F_2

What is this vowel?

- $F_1 \approx$ 620Hz
- $F_2 \approx$ 1850Hz
- plot on chart and find out


3000
2500
2000
1500
1000
500
F2 (Hz)
0
200
400
600
800
1000
F1 (Hz)

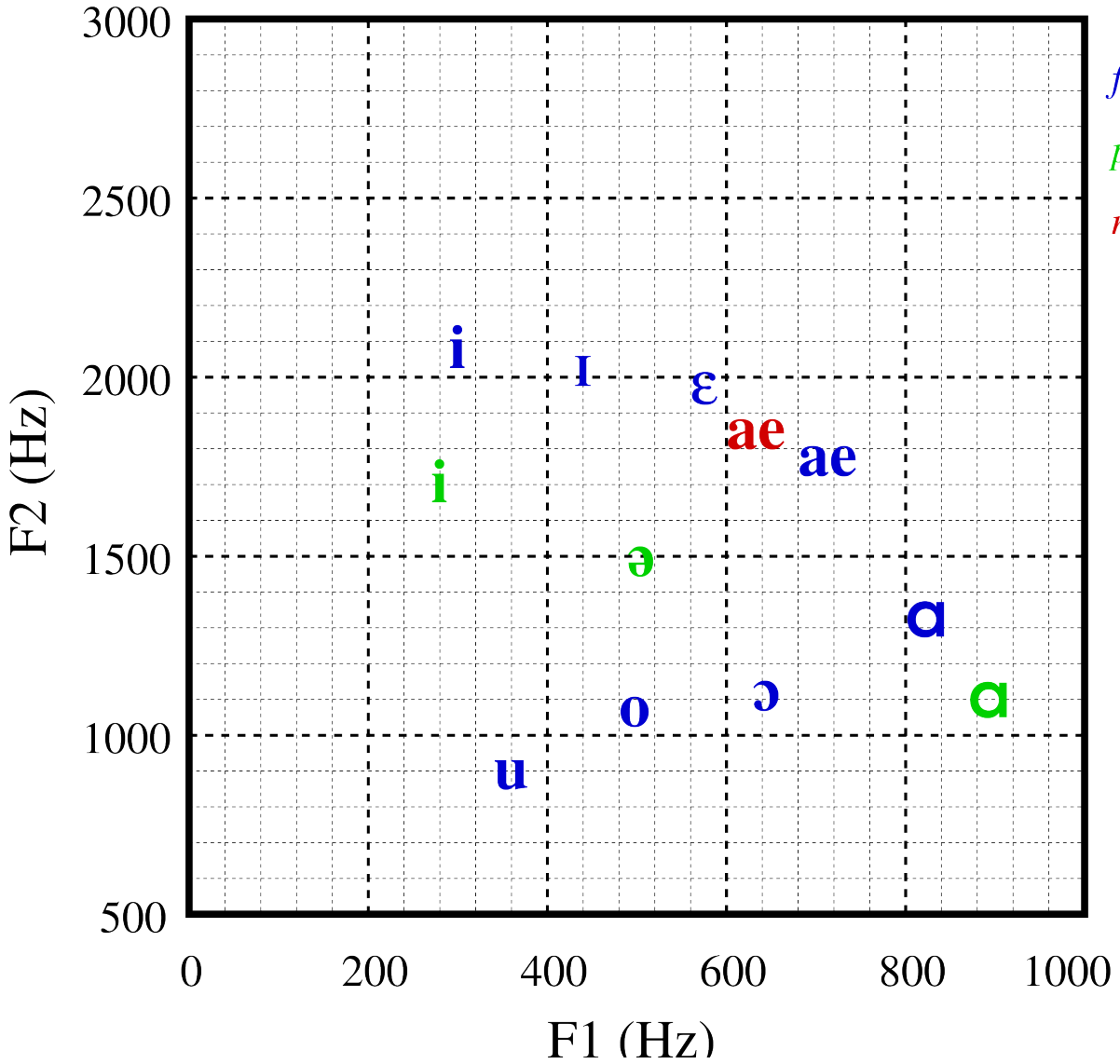


measured by Simon

10:12

The predictions

- neutral vowel [ə] F_1 = 500Hz, F_2 = 1500Hz
- low, back, unrounded vowel [ɑ] F_1 = 900Hz, F_2 = 1100Hz
- high-front vowel [i] F_1 = 280Hz, F_2 = 1700Hz



from textbooks

predicted in lectures

measured by Simon

10:14

Vowels in stop contexts

- so far, only considered isolated vowels
  - formant values constant throughout the vowel
- now we look at vowels before and after stops
  - stop identity will change formant values
  - because articulators will move position

Formants depend on articulator positions

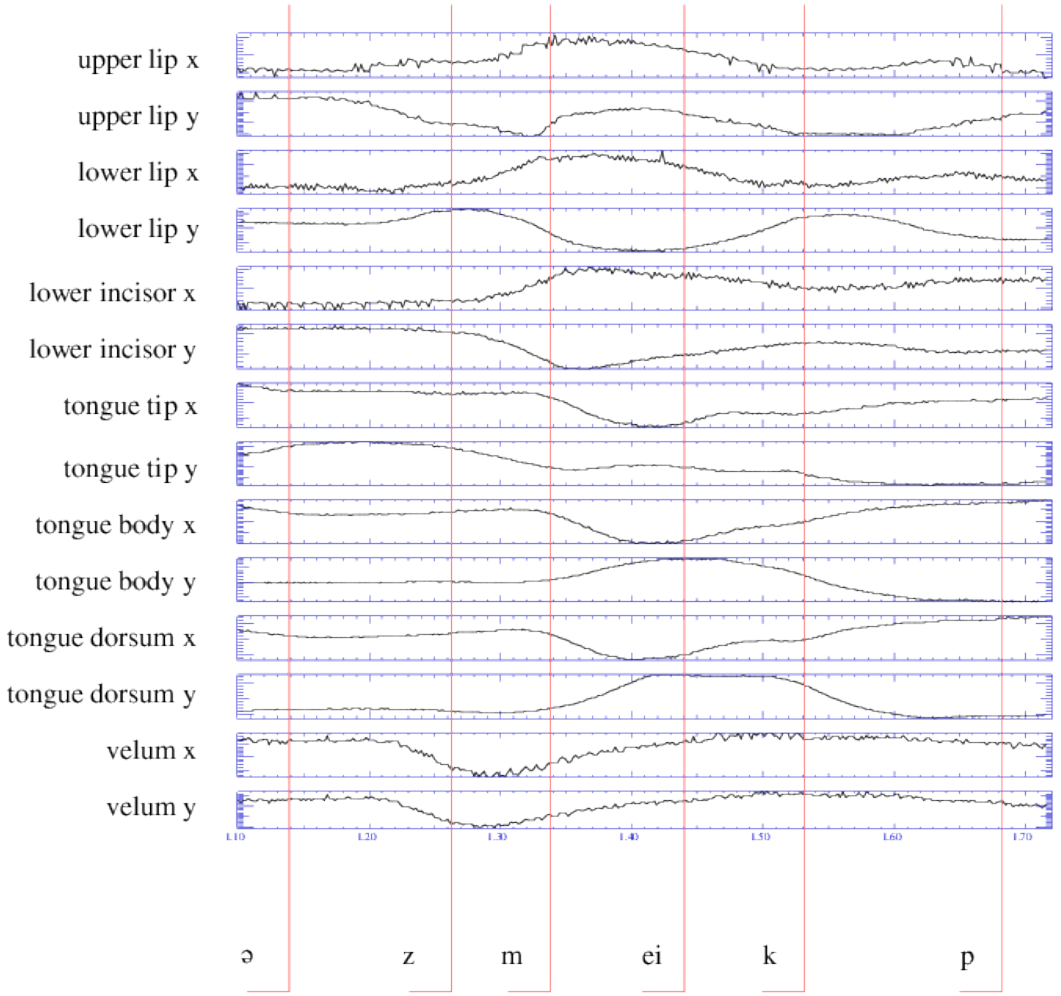
- in particular, on the tongue position
  - since it controls the vocal tract shape/size
- but also the lips (rounding)
  - since rounded lips extend the overall vocal tract length

Vowels after stops

- articulators must first be positioned to make the stop
  - different stops have different places of articulation
  - in other words, articulators are in different positions (they have different *settings*)
- then articulators must move to settings required to make the vowel
  - this movement *cannot be instantaneous*
  - articulators will be moving *whilst speech is being produced*

moving articulators $\rightarrow$ changing vocal tract shape $\rightarrow$ changing formants

Articulators move with limited velocity



Formant trajectories

- we talk about formant *trajectories*
  - formants are *moving* from one value to another
  - following some path, or *trajectory*
- for vowels preceded by stops
  - trajectories will be heading towards some *target* frequency
  - and will reach (or get close to) the target at some point within the vowel
- but at what frequency do the trajectories start?
  - depends on the stop identity

[do...]

or [dɔ...]

- alveolar voiced stop [d]
- followed by back mid-high [o] or back mid-low [Ɔ]
  - somewhere in the middle of the vowel, formants will approach

    $F_1 \approx$ 500-600Hz

    $F_2 \approx$ 1000-1100Hz
- but what shape will the formant trajectories be?
  - that is, where will they be *coming from*

Formants of [d]

- not a vowel
- does it even *have* formants?
  - formants simply arise from resonances
  - if there are resonances, then there must be formants
- but during the stop there is (almost) no sound
- are there still formants?
  - yes, but they may not be audible (or visible on spectrogram)

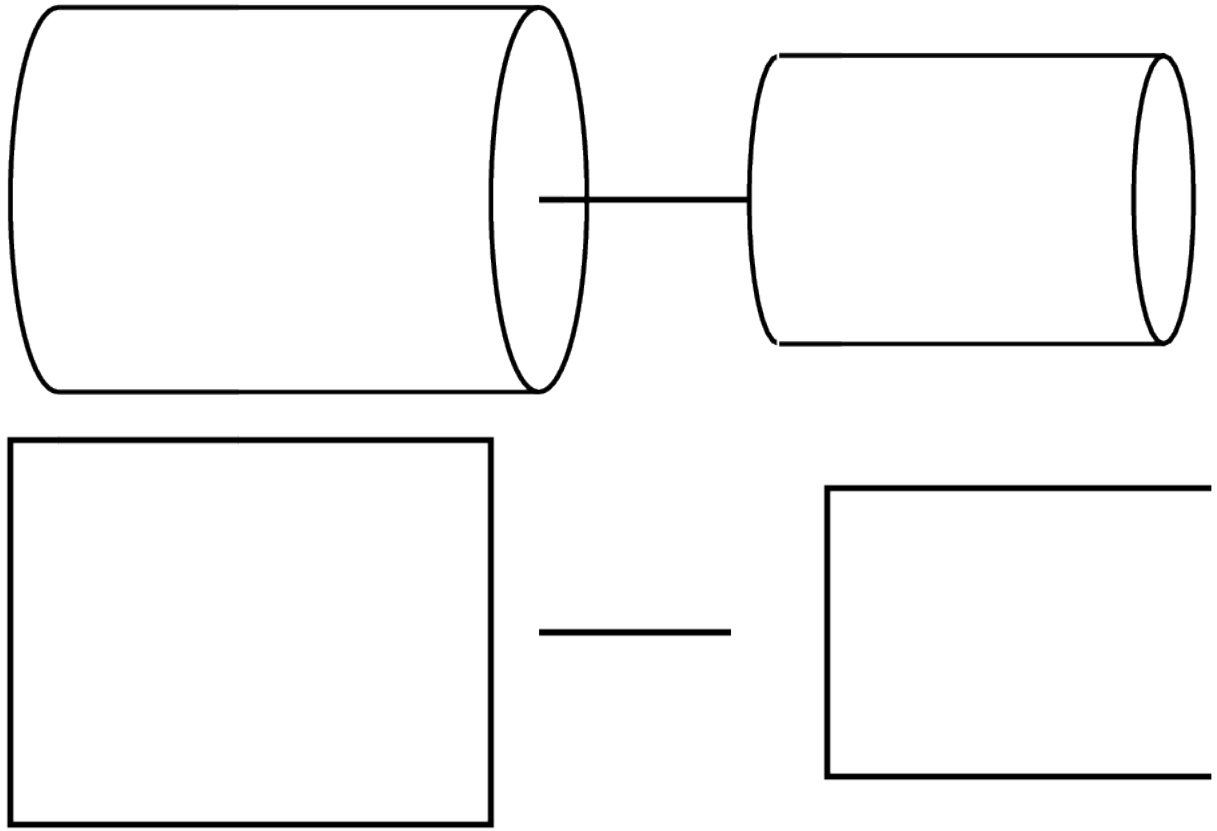
Formant locus

- formants have values during the stop
  - may not be able to directly measure this though because the resonator (cavity) is not being excited
- let's think about where the formant values would be, if we could hear/see them
- need to think about tube models

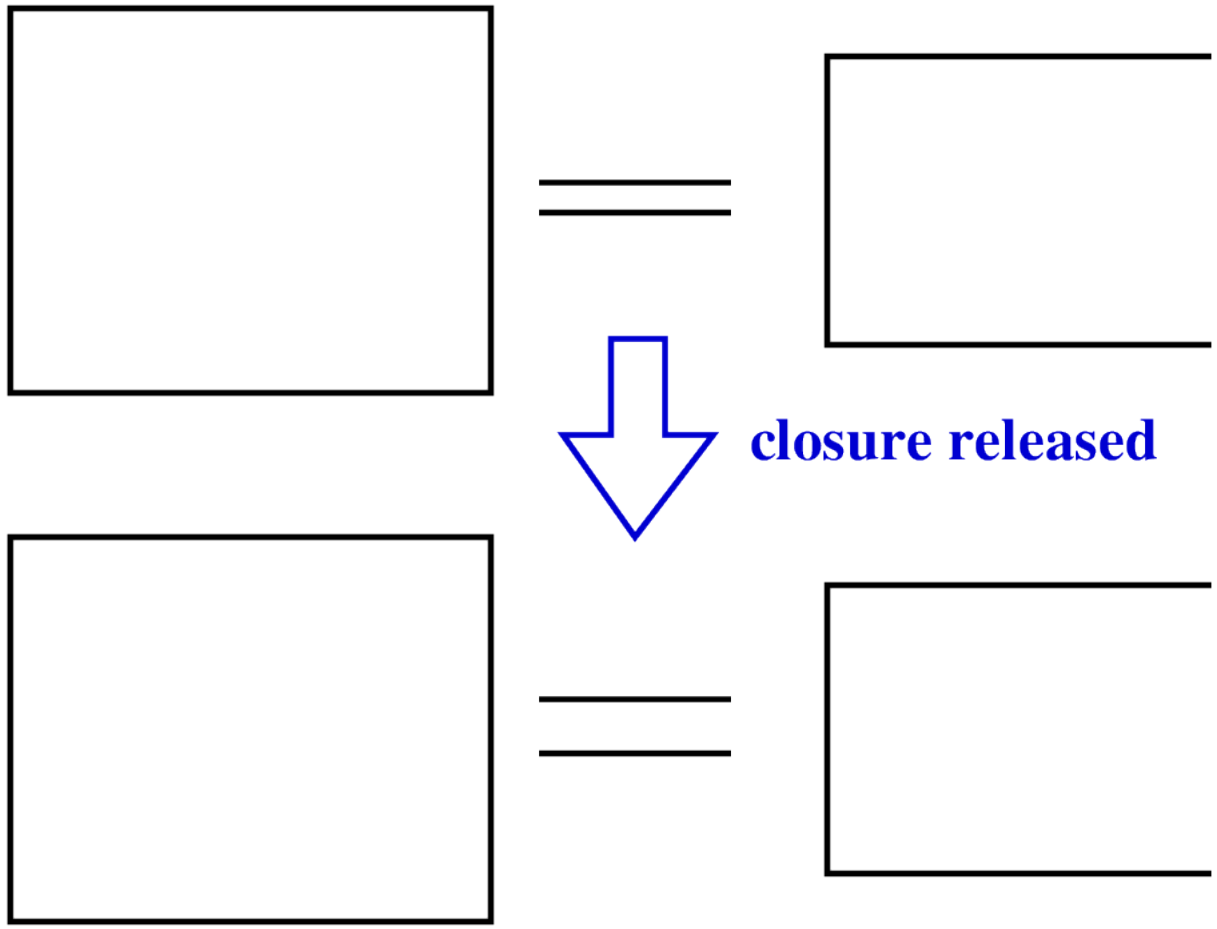
Tube model of [dV] where V is a vowel

- what is the articulator setting for [d]?
  - i.e. what shape is the tube?
  - what are the resonances going to be?
- tongue tip touches alveolar ridge
  - making complete closure
- as we release the closure and move into the vowel
  - short, narrow opening between tongue tip and alveolar ridge
  - sound familiar?

Tube configuration during stop closure



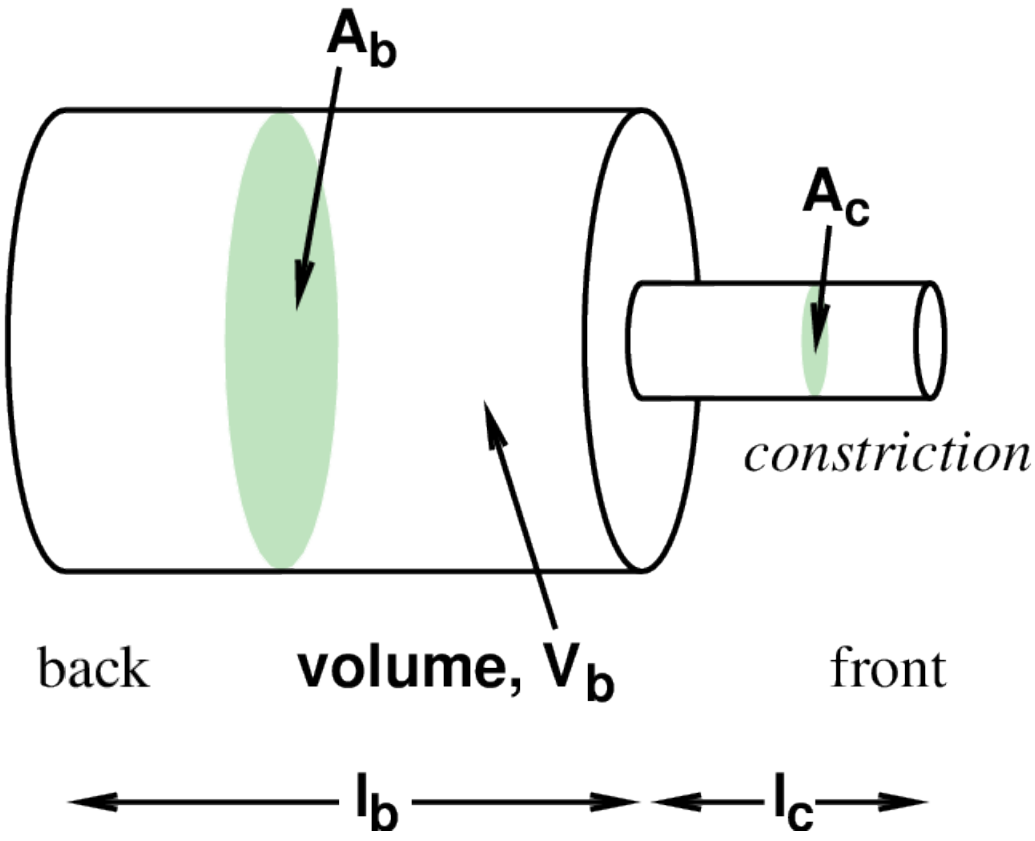
Tube configuration during stop release



During release of closure

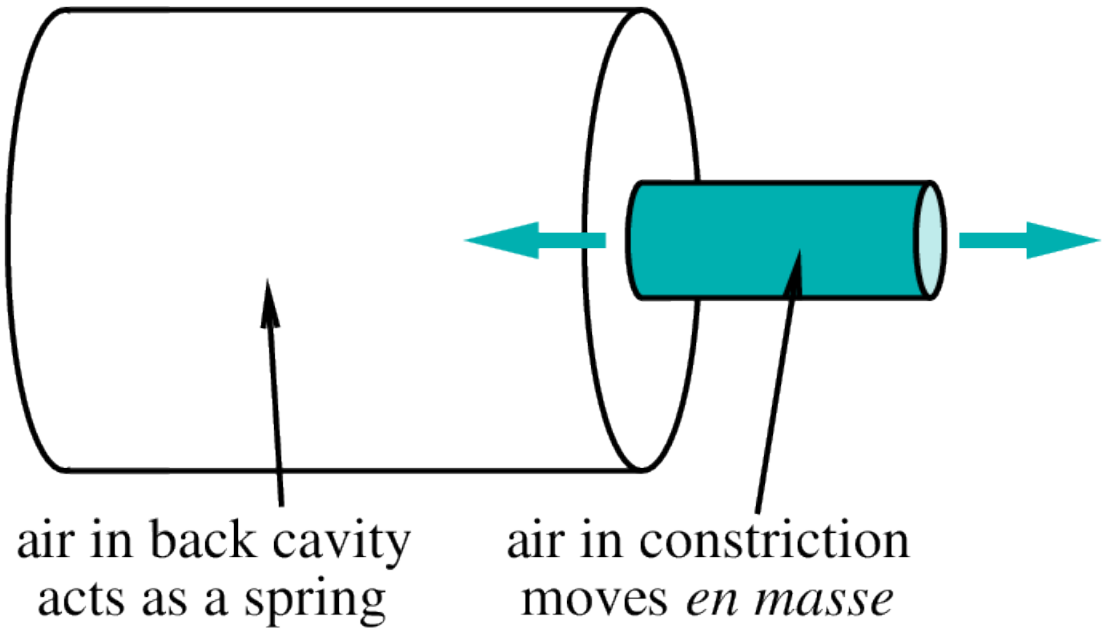
- narrow constriction opens up
  - cross-sectional area grows
- what does this mean in terms of resonance?
- what mode of resonance will produce F_1 ?
  - F_1 is whatever the lowest resonance in the system is
  - in this configuration, that will be a Helmholtz resonance

Recap - Helmholtz resonance



back cavity volume, $V_b = A_b l_b$ constriction volume, $V_c = A_c l_c$

Recap - Helmholtz resonance



air in constriction moves back and forth like a piston in a cylinder

Recap - Helmholtz resonance

$$f = \frac{c}{2\pi}\sqrt{\frac{A_c}{V_b l_c}}$$

- where

  f is the resonant frequency of the system

  A_c is the area of the constriction

  l_c is the length of the constriction

  V_b is the volume of the back cavity

  c is the speed of sound in air

What happens as A_c increases?

- for simplicity, assume all other dimensions remain constant
  - only cross-sectional area of constriction, A_c, varies

$$f = \frac{c}{2\pi}\sqrt{\frac{A_c}{V_b l_c}}$$

- A_c is initially very small

  f is very low

- as A_c increases

  f increases

Locus of F_1

- *locus* of a formant is where it *originates* (i.e. where it starts)
  - this will be at the point of closure in the stop
- cannot actually observe this *locus*
  - since closure means (almost) no excitation of the resonances
- but will see the formant transition into the following vowel
- what is the locus of F_1 in [d] ?

Locus of F_1

- produced by the Helmholtz resonance
  - since during release, a very narrow constriction is present
- will therefore be **very low**
- furthermore, F_1 will increase during transition, since

  area of constriction increases, and

  target value in following vowel is higher than locus

What about other stops?

- pattern will be the same:
  - stop closure formed at some point in vocal tract

    [p] and [b] – lips meet (bilabial)

    [t] and [d] – tongue tip touches alveolar ridge

    [k] and [g] – closure is made by tongue against velar region

    etc.
  - closure released
    - short narrow tube, with wider cavity behind (“wine bottle”)
    - Helmholtz resonance $\longrightarrow$ low F_1 locus

What about other vowels?

- pattern will be the same
  - Helmholtz resonance is very low
    - so F_1 will always rise going into the vowel
  - the target value that F_1 is heading for will depend on vowel identity

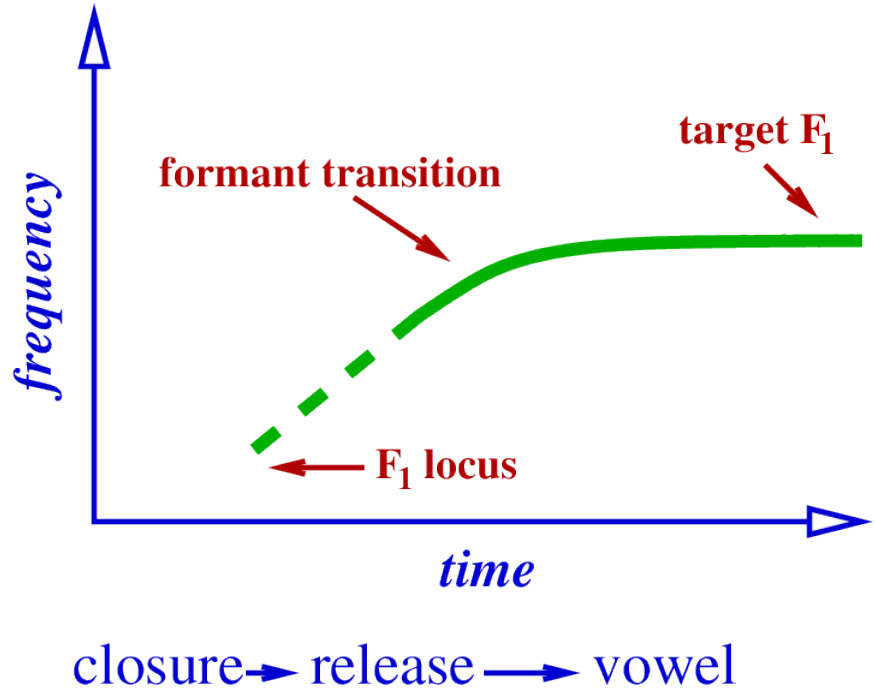
    [ɑ] F_1 high

    [ɔ] and [ɛ] F_1 mid

    [u] and [i] F_1 low

    etc.

Trajectory of F_1 during stop-vowel transitions



- locus will always be low frequency
- actual value of F_1 target will depend on vowel identity

What about F_2 ?

- need to find the locus of F_2 during stop closure
  - this time it **won't** be a Helmholtz resonance since
    - there is **one** Helmholtz resonance in the system
    - it is always (very) low frequency
    - so it produces F_1
  - locus of F_2 must be one of the other resonances
    - front cavity
    - back cavity

Locus of F_2

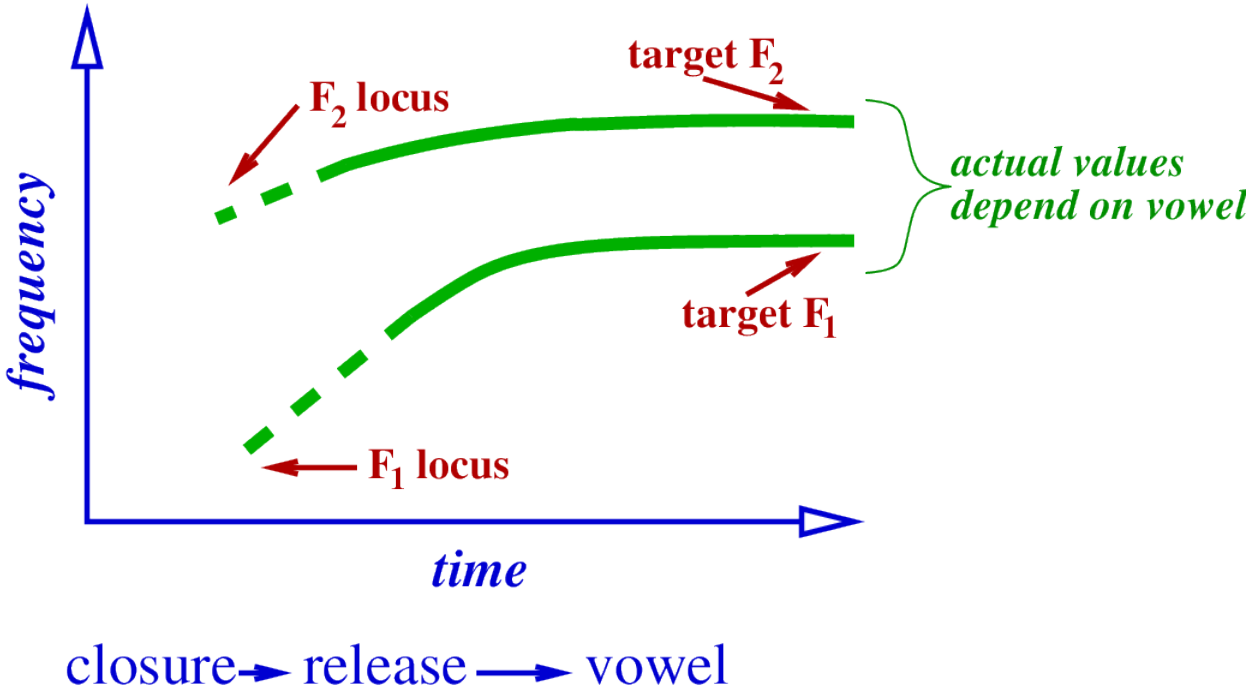
- depends on stop identity
  - stops have differing locations of the closure
  - and therefore differing front/back cavity sizes
- let's just consider [b], [d] and [g]

  [p], [t] and [k] have same places (respectively)
  they only differing in voicing

Locus of F_2 for [b]

- bilabial – lips form closure
  - no front cavity and a large (i.e. long) back cavity
- back cavity resonance will be F_2
- cavity is as long as it can be (entire vocal tract)
- so locus of F_2 in [b] will be **very low**
  - **lower** than the F_2 of **any** vowel
  - because back cavity is longer than for any vowel
- therefore trajectory of F_2 will always be rising coming out of a [b] into a vowel

Trajectory of F_2 during [bV]

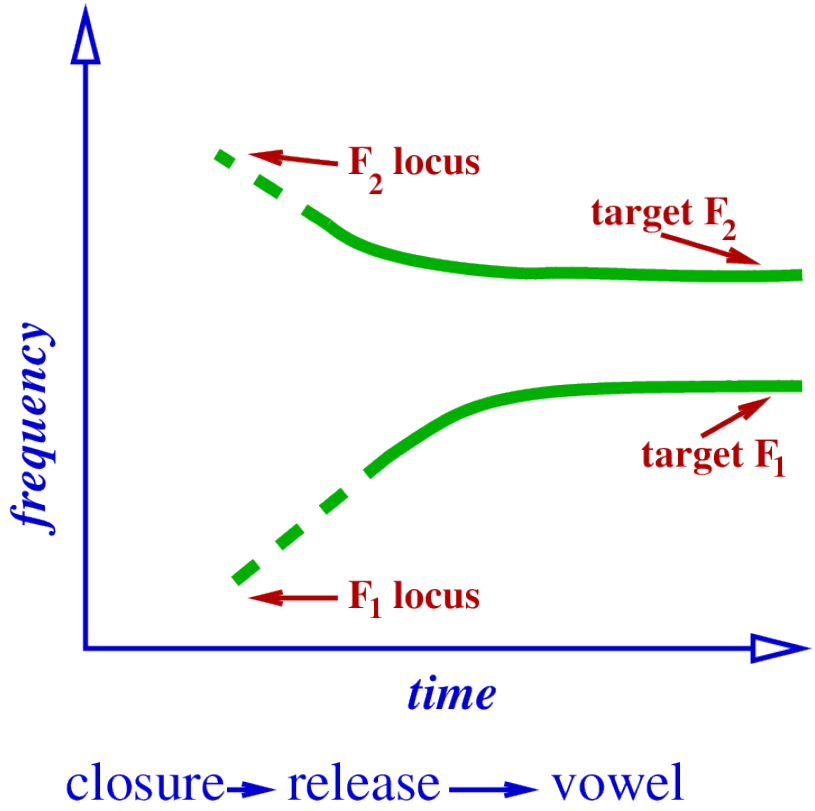


- F_1 locus as before – very low
- F_2 locus also very low

Locus of F_2 for [d]

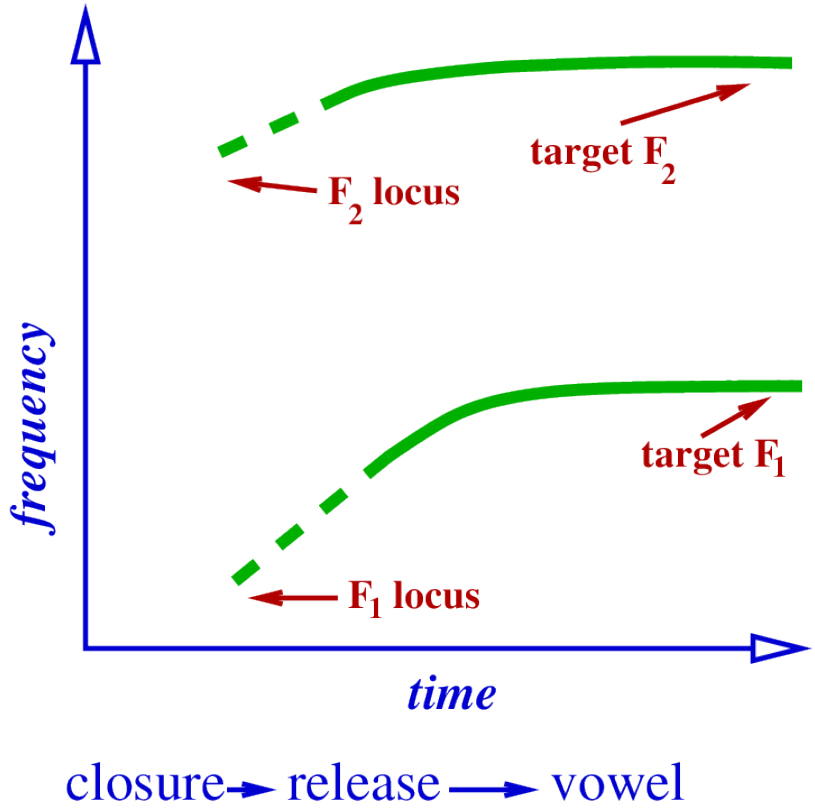
- alveolar (tongue touches alveolar ridge)
  - very short front cavity and a long back cavity
- so F_2 is produced by back cavity during stop and transition
- cavity is not as long as in [b]
- so locus of F_2 in [d] will be **low**, but not as low as for [b]
  - **lower** than the F_2 of some vowels and higher than the F_2 of **other** vowels
- therefore trajectory of F_2 will sometimes be rising coming out of a [d] into a vowel, and sometimes falling, depending on the vowel F_2

Trajectory of F_2 during [dV] for V with low F_2



- locus of F_2 now around 1800Hz, target is lower than 1800Hz

Trajectory of F_2 during [dV] for V with high F_2

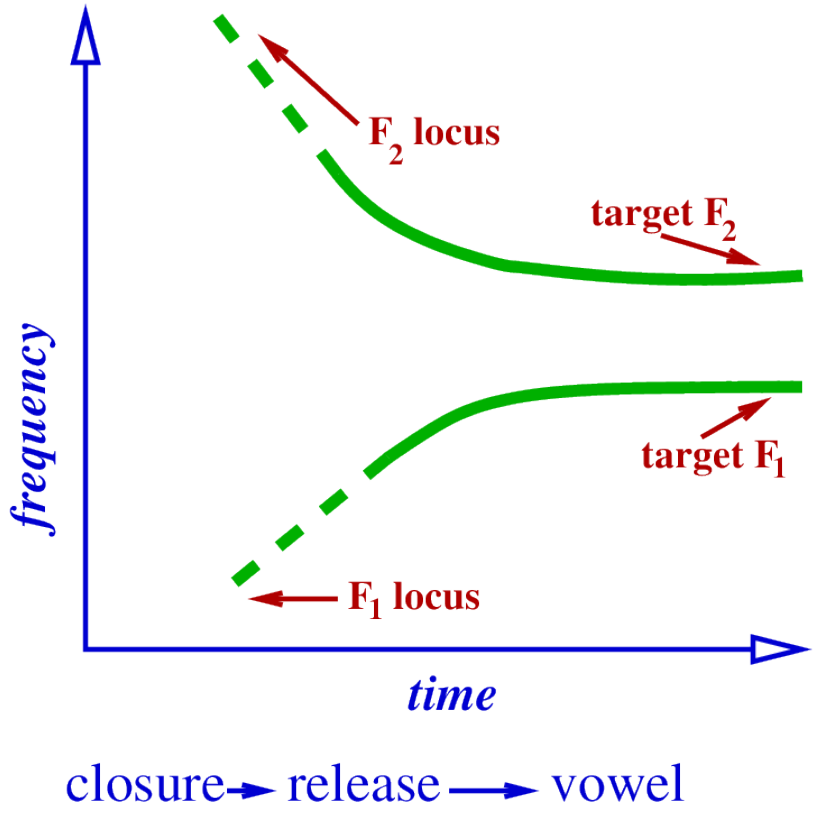


- locus of F_2 still around 1800Hz, target is higher than 1800Hz

Locus of F_2 for [g]

- velar
  - medium length front cavity and a shorter back cavity
- so F_2 is now produced by front cavity during stop and transition
- locus of F_2 in [g] will be **high**
  - **higher** than the F_2 of any vowel
- therefore trajectory of F_2 will always be falling coming out of a [g] into a vowel

Trajectory of F_2 during [gV]



- locus of F_2 now always higher than target

Summary

both F_1 and F_2 have a locus and a target for stop-vowel transitions

- locus
  - is frequency of formant in the stop closure
  - may not be actually present in the speech signal
  - frequency of locus depends on stop identity
- target
  - is the frequency that the formant approaches in the vowel
  - may not be reached, only approximated
  - frequency of target depends on vowel identity

Looking ahead

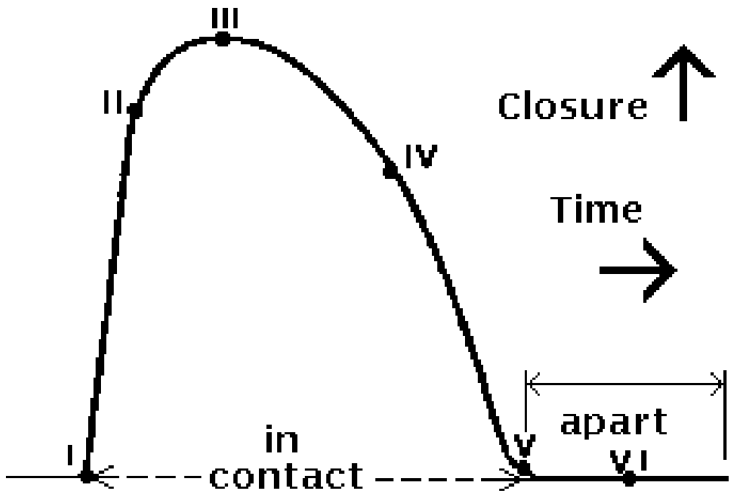
- lecture 13 (today)
  - a proper analysis of sound sources
- lecture 14
  - presentation of formant trajectories of vowels in contexts: bring along your measurements, plus sketches of formant trajectories, just for the vowels in stop-vowel-stop contexts
- lecture 15
  - the source-filter theory of speech production, and some things we can use it for (including computer speech synthesis)

Sources of sound revisited

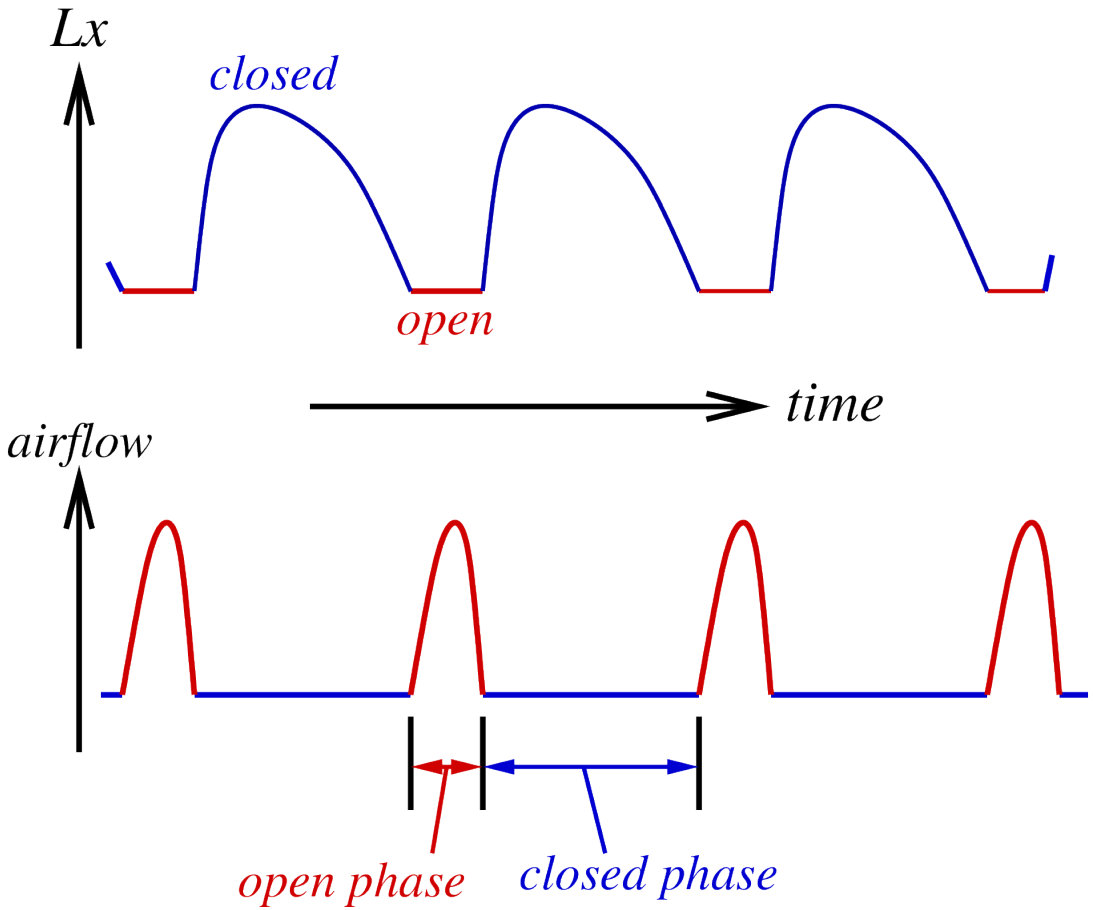
- a proper analysis of the sound produced by
  - the vocal folds
  - including non-periodic sounds
- which will lead us on to understanding
  - the harmonics in the spectrum of a voiced sound
  - the reasons behind different voice qualities
- and ultimately to
  - a source-filter model of speech production

Recap – vocal fold action

Laryngograph waveform (Lx):



Lx vs. glottal airflow



The open phase

- sub-glottal pressure builds up during closed phase
  - vocal folds remain closed
- sub-glottal pressure eventually forces folds apart
  - airflow starts, and then rises as folds move apart
  - sub-glottal pressure eventually drops
  - then vocal folds start to close due to tension (caused by muscles)

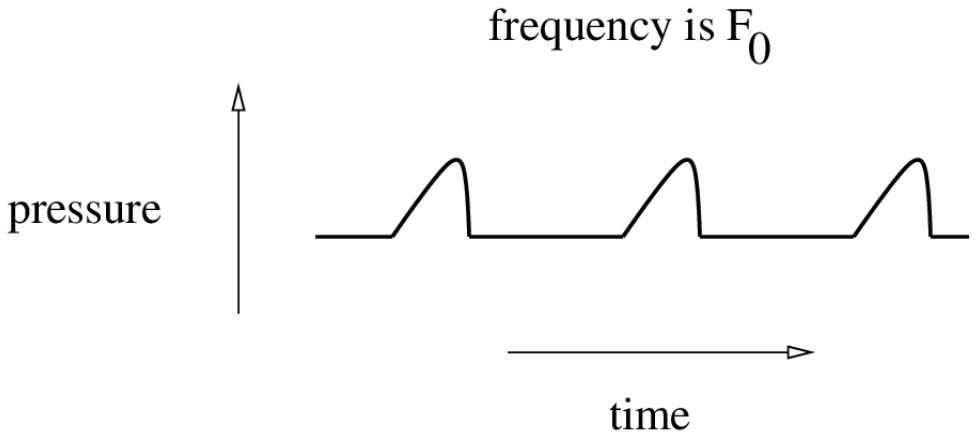
- sub-glottal pressure has dropped
  - airflow reduces
  - glottis (gap between folds) eventually becomes small
  - Bernoulli effect causes pressure *in* glottis to drop
  - vocal folds drawn into closed position very rapidly (“sucked together”)

The Bernoulli effect means that as flow velocity increases,
pressure decreases

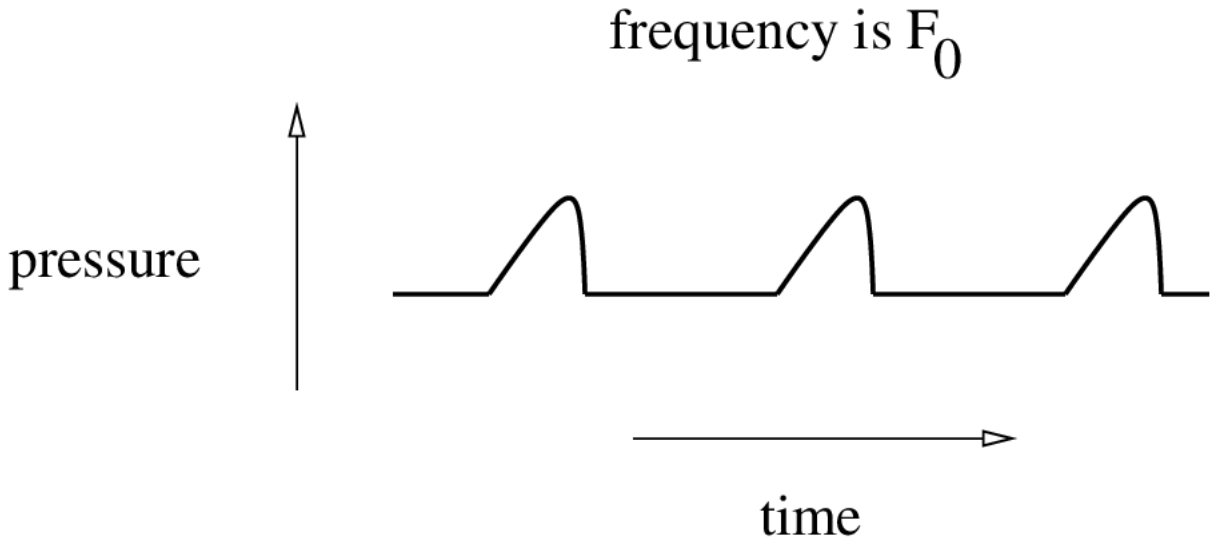
When the glottis becomes narrow, flow velocity through it must increase (air is flowing through a smaller area)

The sound pressure wave just above the folds

- open phase releases a pulse of higher pressure air into the vocal tract
  - pressure pulse travels upwards through air in vocal tract
  - a moving pressure wave = a sound wave
- so sound pressure waveform just above vocal folds is



Fourier analysis of glottal pressure wave



- contains energy at frequency F0
- and at every multiple of F0: 2×F0, 3×F0, 4×F0, ...
- these are the *harmonics* of F0

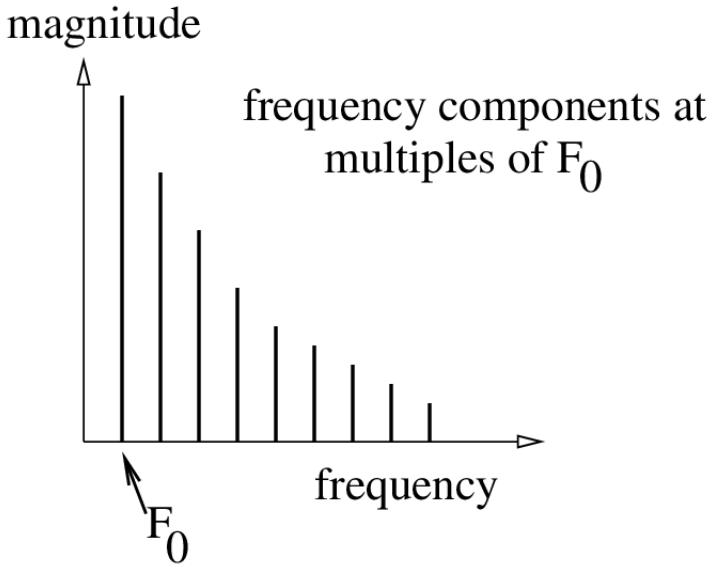
Harmonics of F0

- fundamental frequency is F0
  - also known as the *first harmonic*
- the component at 2×F0 is the *second harmonic*
- the energy of the harmonics decreases as frequency increases

This is the modern terminology, and is compatible with Ladefoged.

[Older books may call 2×F0 the first harmonic, or first overtone.]

The spectrum of the glottal pressure wave



- magnitude drops off as frequency increases

[homework: prove to yourself that the spectrum of a voiced sound *tilts* – that is, slopes downward with increasing frequency]

Voice quality

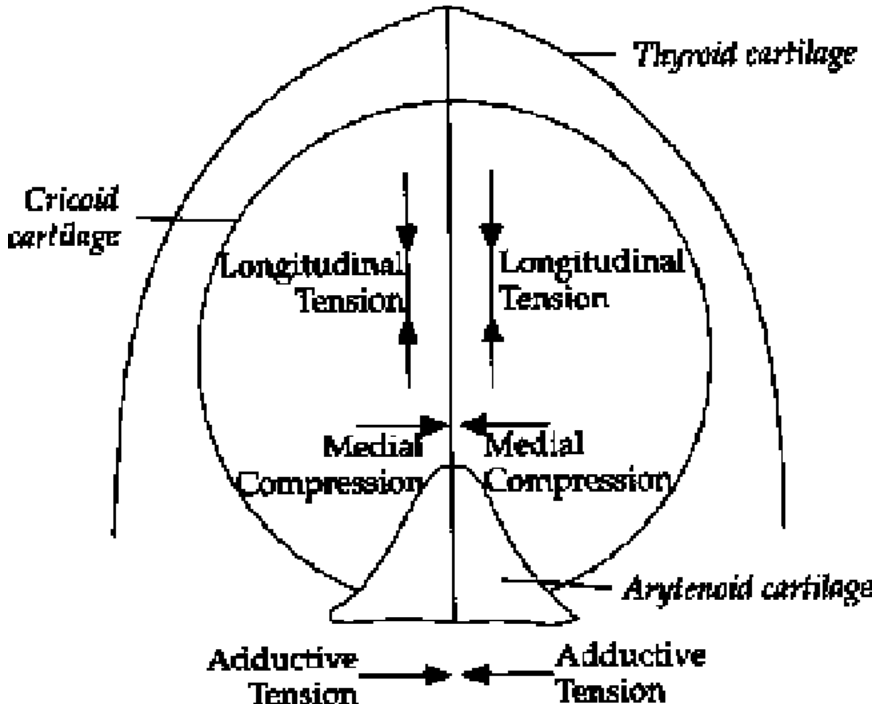
- vocal tract can be modelled as a tube
  - can only change shape/area
  - which just controls formant frequencies
- what about voice *quality* differences?
  - what exactly do we mean by voice quality?
- and what determines voice quality?

Glottal waveform can vary

- ratio of
  - open phase to closed phase
- degree of closure during phonation
  - complete or partial (“leaky”)
- lack of phonation
  - vocal folds do not vibrate at all
  - unvoiced sounds
  - whispering

The vocal folds

[Posterior at bottom]



Modal voice

- normal speaking voice
- enough tension in vocal folds to allow vibration
  - vocal folds vibrate during voiced sounds
  - complete closure is achieved during closed phase
- F0 in central region of speaker's pitch range

Breathy voice

- even during closed phase
  - complete closure is not achieved
- a *chink* remains open, allowing air to escape throughout the glottal cycle
  - escaping air through narrow opening
  - means turbulent air flow
  - which produces a breathy quality to the voice

Smoker's voice

or....how to give up smoking

- smoking causes vocal folds and surrounding tissue
  - to become irritated, swollen and dry
  - healthy folds are normally wet (lubricated by mucus)
- which has these effects
  - folds do not close fully during closed phase (breathy quality)
  - vibration efficiency reduced (F0 lower)
- result: low pitch and breathy voice quality

Creak / creaky voice

- long closed phase
- relatively short open phase
  - vocal folds tightly closed during closed phase
- tends to occur at bottom of speaker's pitch range
- period of vocal fold vibration tends to become irregular

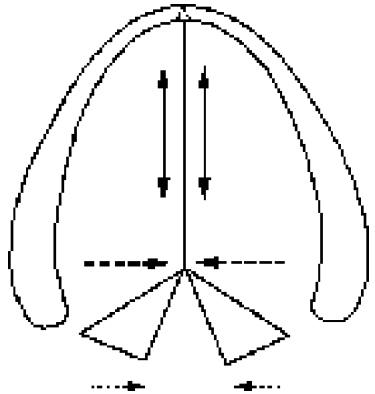
Intuitively this is less like a sine wave than modal voice, so we expect it to sound further away from a sine wave sound

Sine wave = pure tone = "smooth" sound

Creaky voice = impure tone = harsher or sharper sound

Whispering

- no vocal fold vibration
  - folds are closed, apart from a small chink
  - like in breathy voice



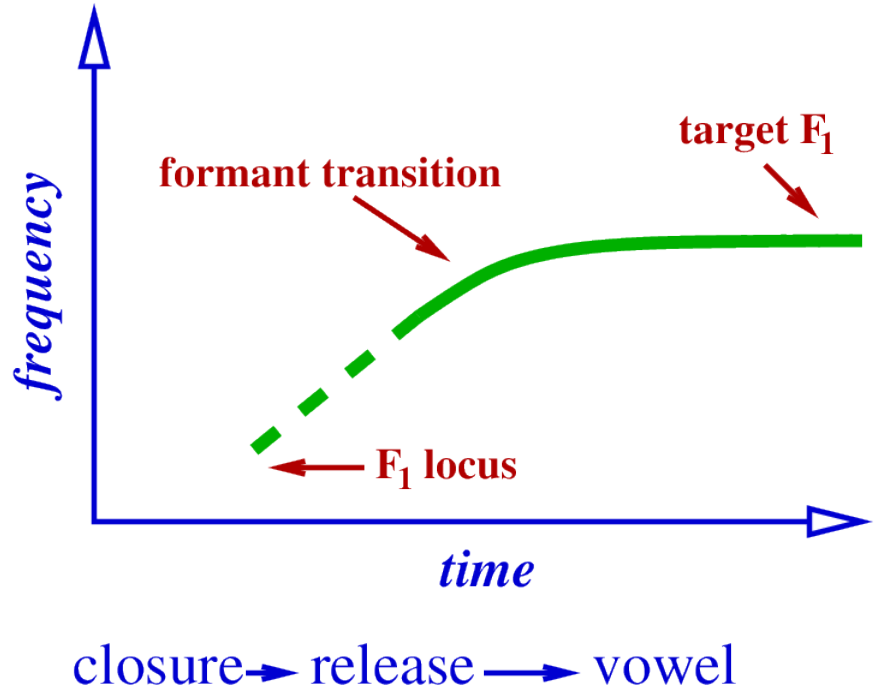
- noise-like sounds wave produced, which is then filtered by the vocal tract in the usual way

- despite lack of phonation, whispering is intelligible

Falsetto

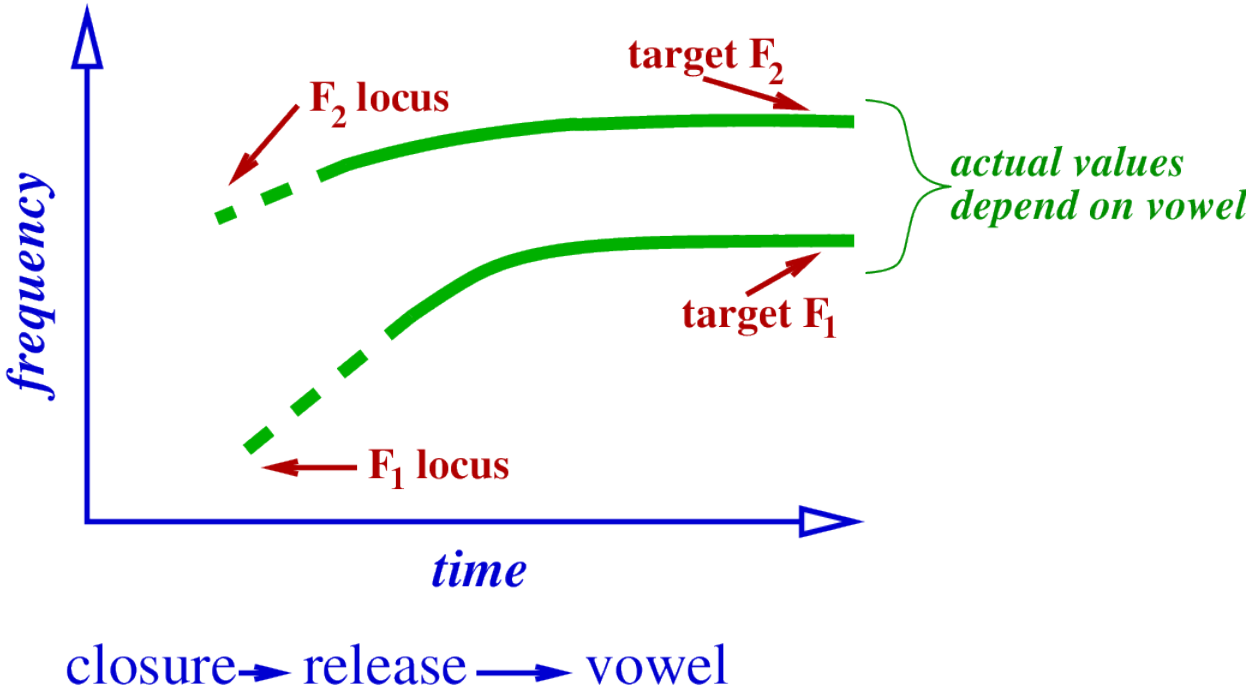
- very high vocal fold tension
- only internal edges of folds are involved in vibration
  - frequency of vibration is high
  - amplitude of vibration is small
  - speech is relatively quiet

Trajectory of F_1 during stop-vowel transitions



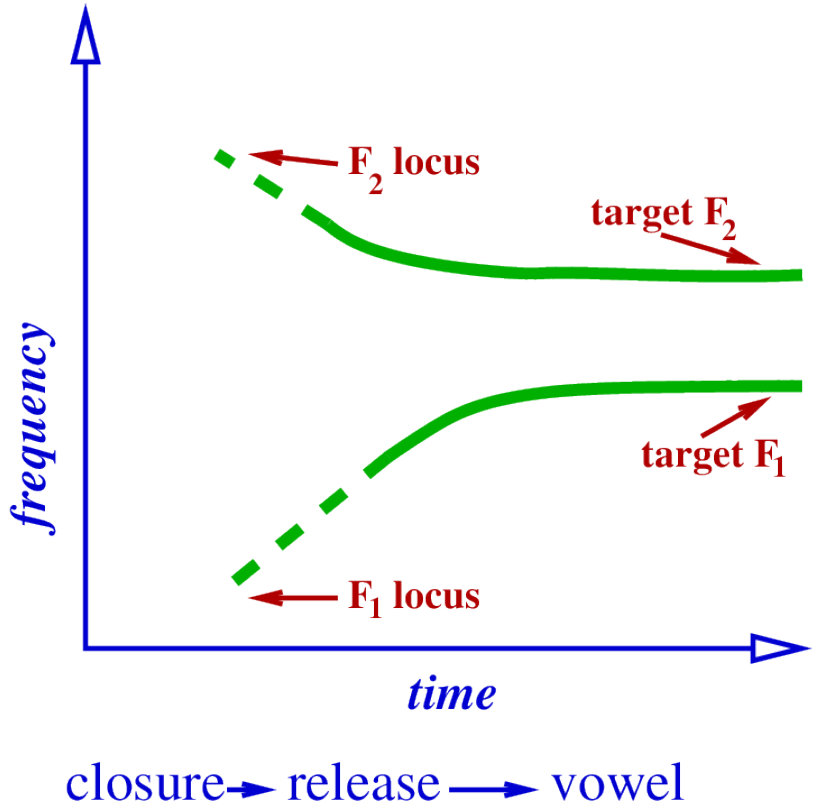
- locus will always be low frequency
- actual value of F_1 target will depend on vowel identity

Trajectory of F_2 during [bV]



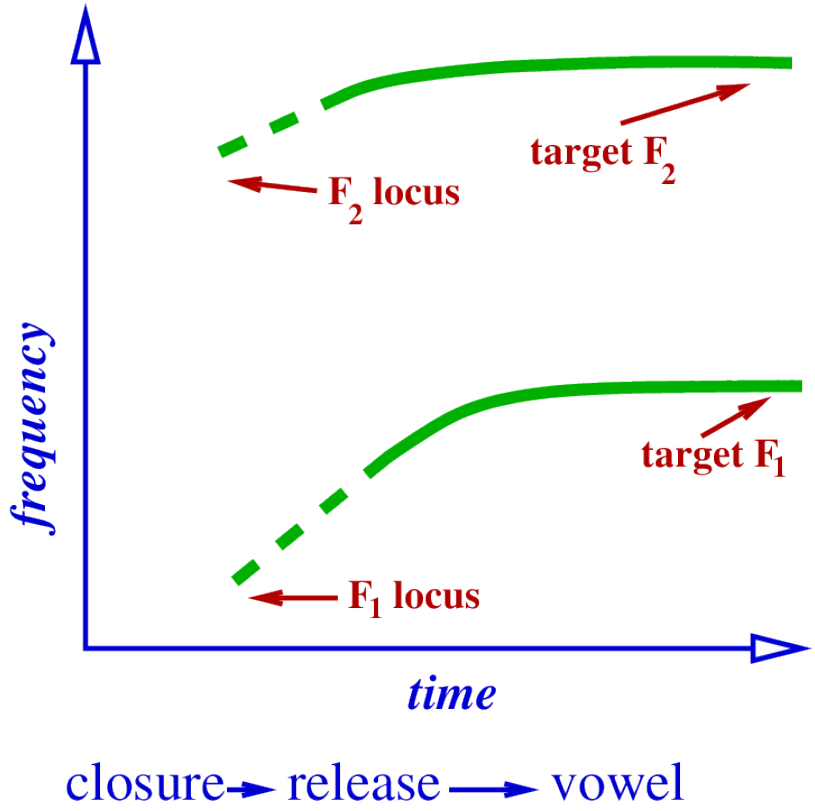
- F_2 locus very low

Trajectory of F_2 during [dV] for V with low F_2



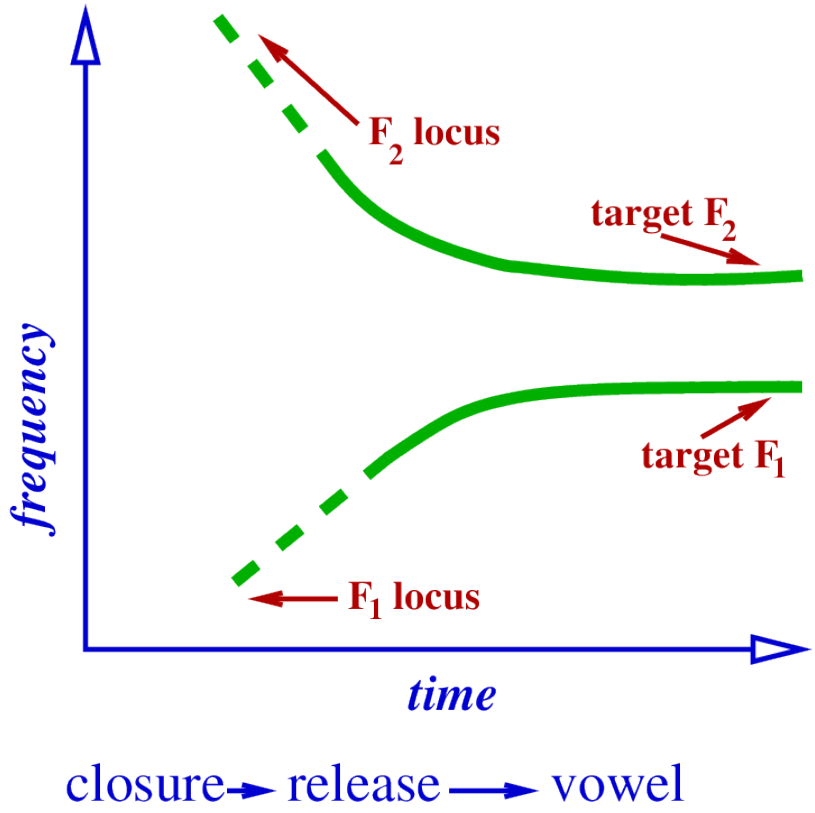
- locus of F_2 now around 1800Hz, target is lower than 1800Hz

Trajectory of F_2 during [dV] for V with high F_2



- locus of F_2 still around 1800Hz, target is higher than 1800Hz

Trajectory of F_2 during [gV]



- locus of F_2 now always higher than target

An aside: formant bandwidths

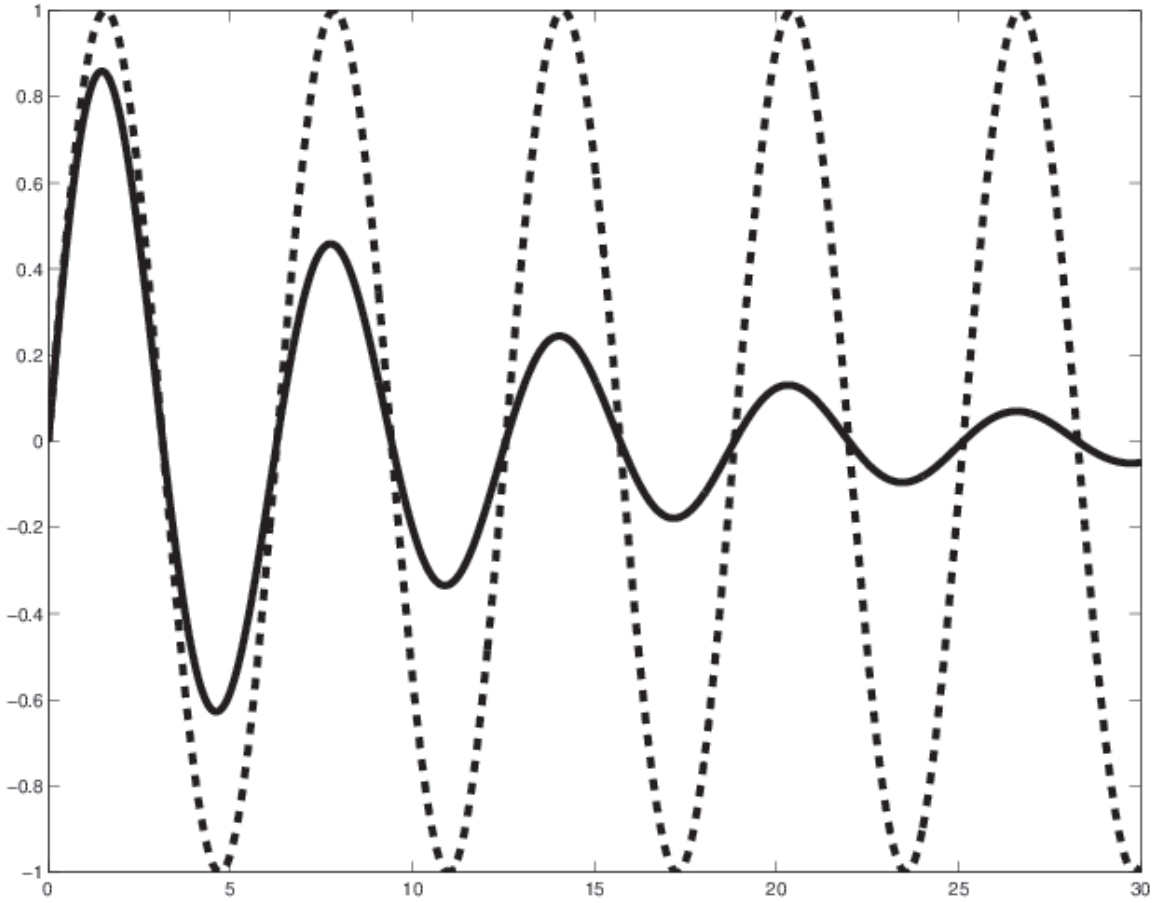
- formant peaks might be
  - narrow, tall and sharply peaked
  - broad, not so tall and flatter on top
- each formant therefore has
  - a frequency, measured in Hertz
  - a bandwidth, also measured in Hertz
- what controls bandwidth?

Bandwidth and damping

- a bit harder to get intuitions about than simple resonance
  - when you give one push to a child's swing
  - it starts to swing (resonate)
  - amplitude gradually decays (decreases)
- why does it decay?

the answer is *damping* – energy is being dissipated (lost) from the system (due to air resistance, for example)

Undamped vs. damped oscillations

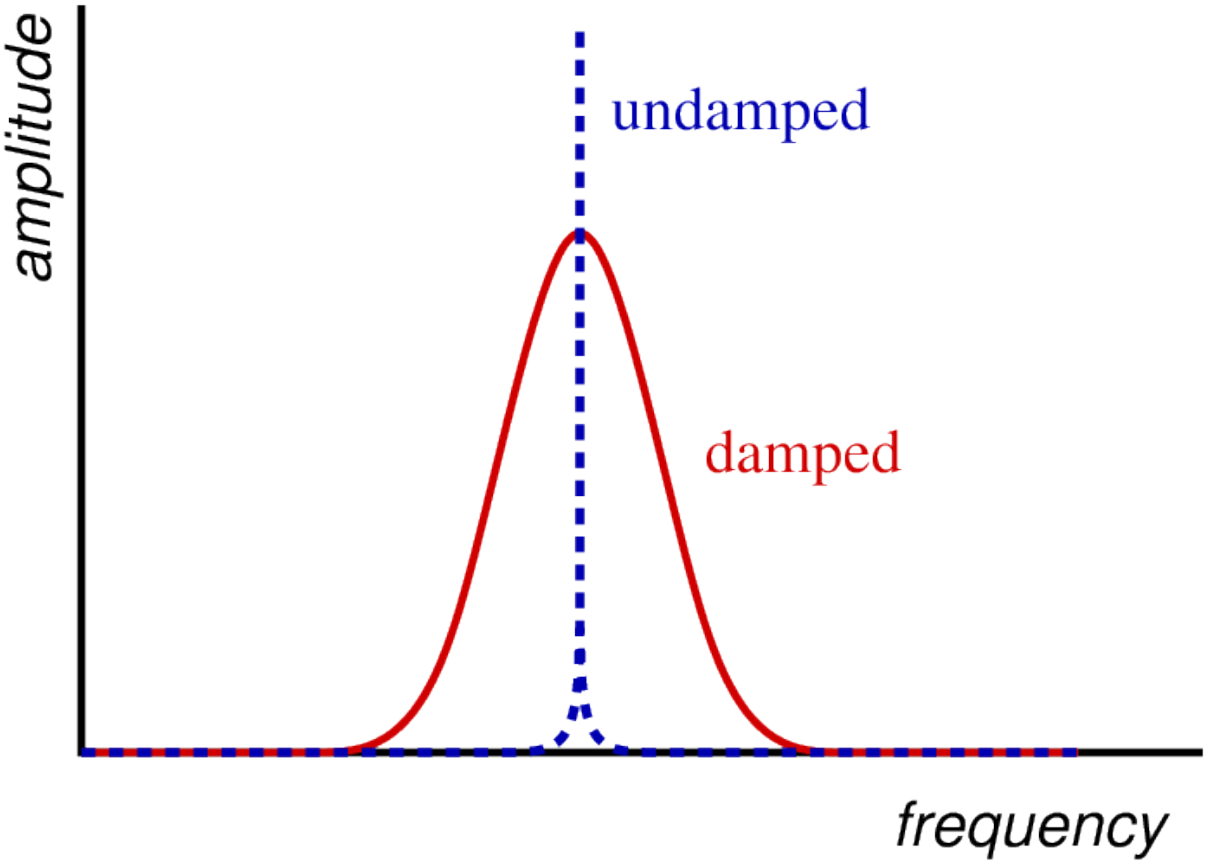


15:14

Undamped vs. damped oscillations

- undamped oscillations
  - pure sine wave
  - spectrum contains a single line: energy at exactly one frequency
- damped oscillations
  - decaying sine wave – not pure
  - spectrum contains energy at more than one frequency

Undamped vs. damped oscillations



Damping in the vocal tract

- walls of vocal tract are soft
  - absorb sound energy
  - resonating sound waves are slowly dissipated
- which is why formant peaks are not tall and narrow

Fricative sounds

- a very brief look at what makes fricatives
  - sound the way they do
  - sound different from one another (e.g. [s] vs. [ʃ])

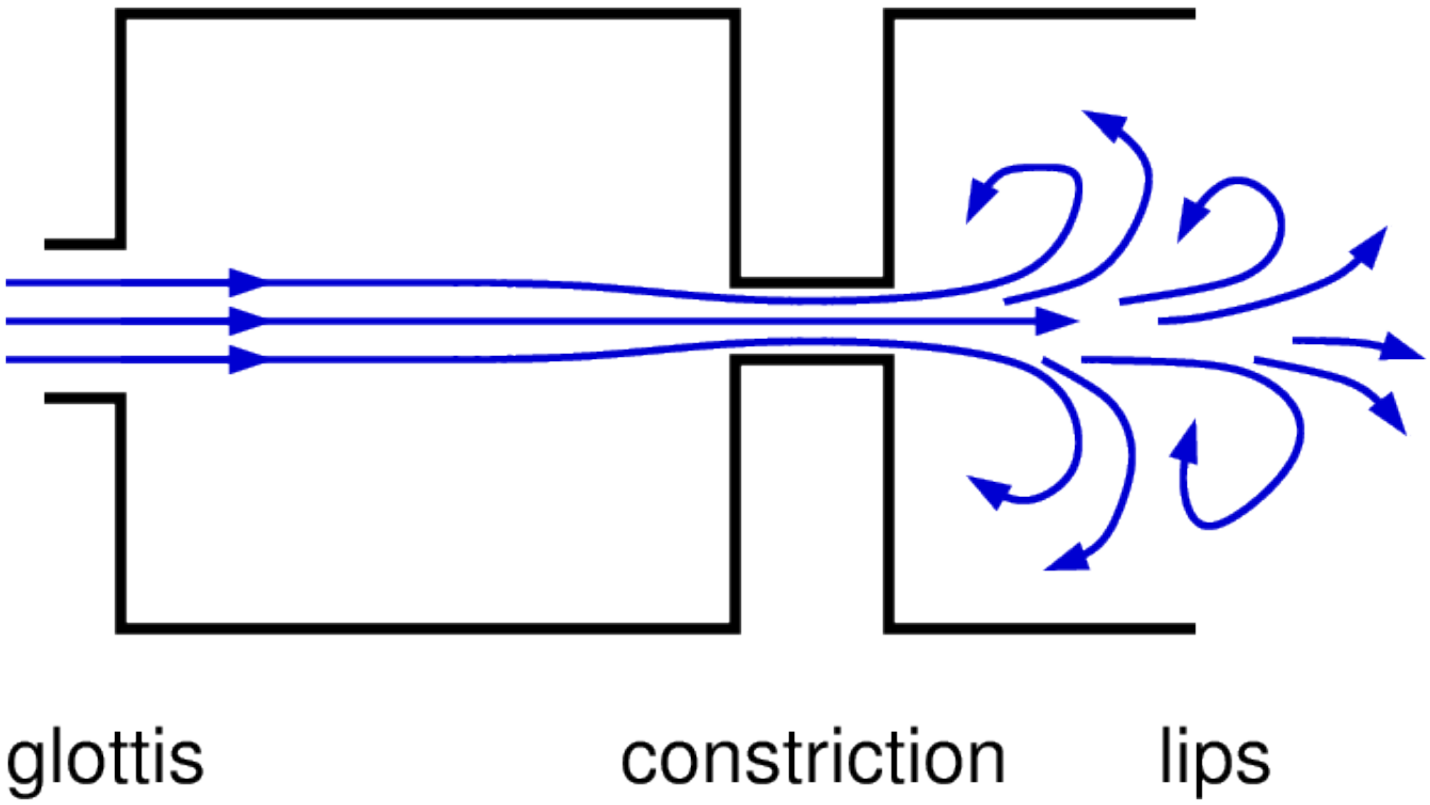
Fricatives: a source-filter approach

- we have already mentioned the source of sound
  - turbulent airflow
  - caused by
    - narrow constriction and/or obstacle in airflow
  - turbulent airflow is essentially random
    - sound wave produced is therefore random
    - sounds like hissing or white noise

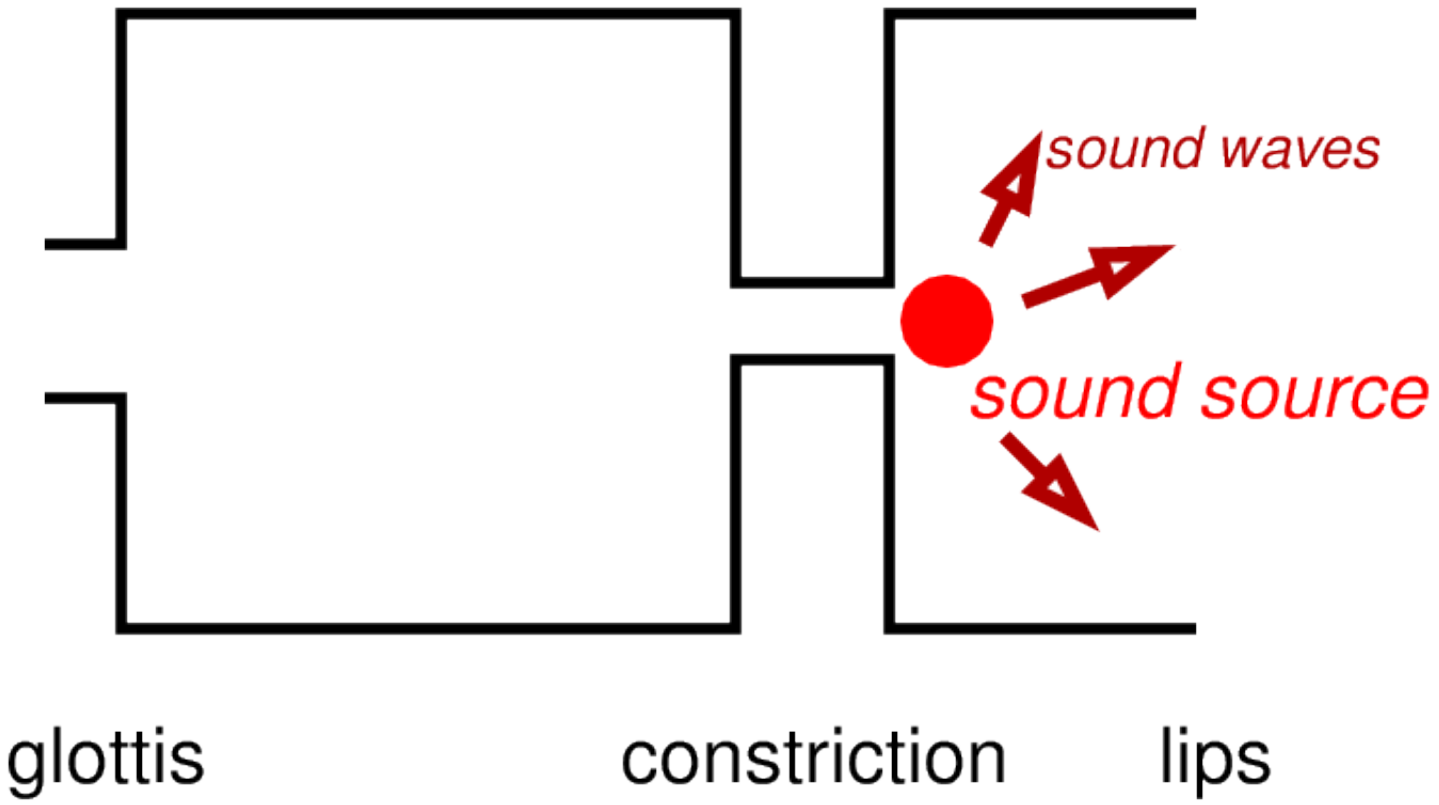
What shapes the sound?

- the vocal tract
  - source of frication located somewhere in vocal tract
  - sound wave filtered by front cavity
- so, can we explain the spectra of fricatives?
  - in terms of
    - location of frication and therefore
    - size of front cavity

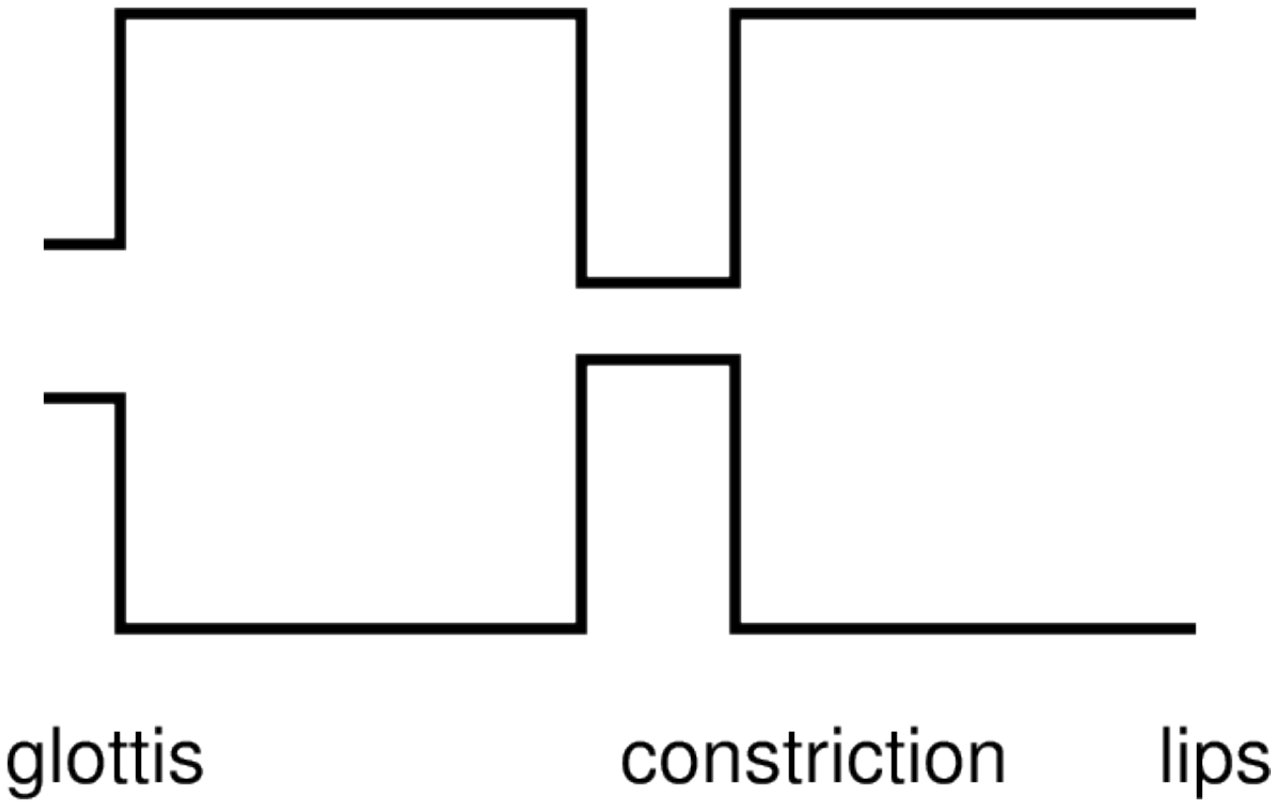
Turbulence



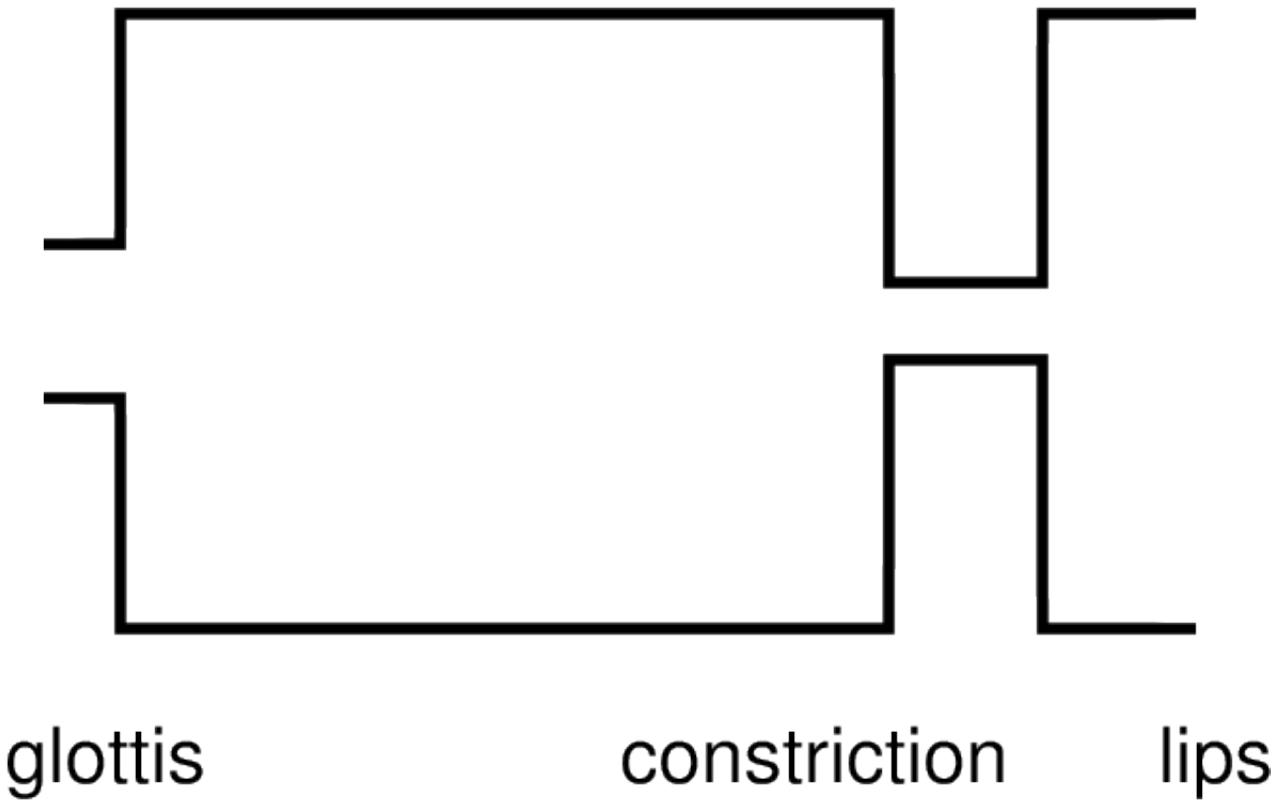
Tube model for fricatives



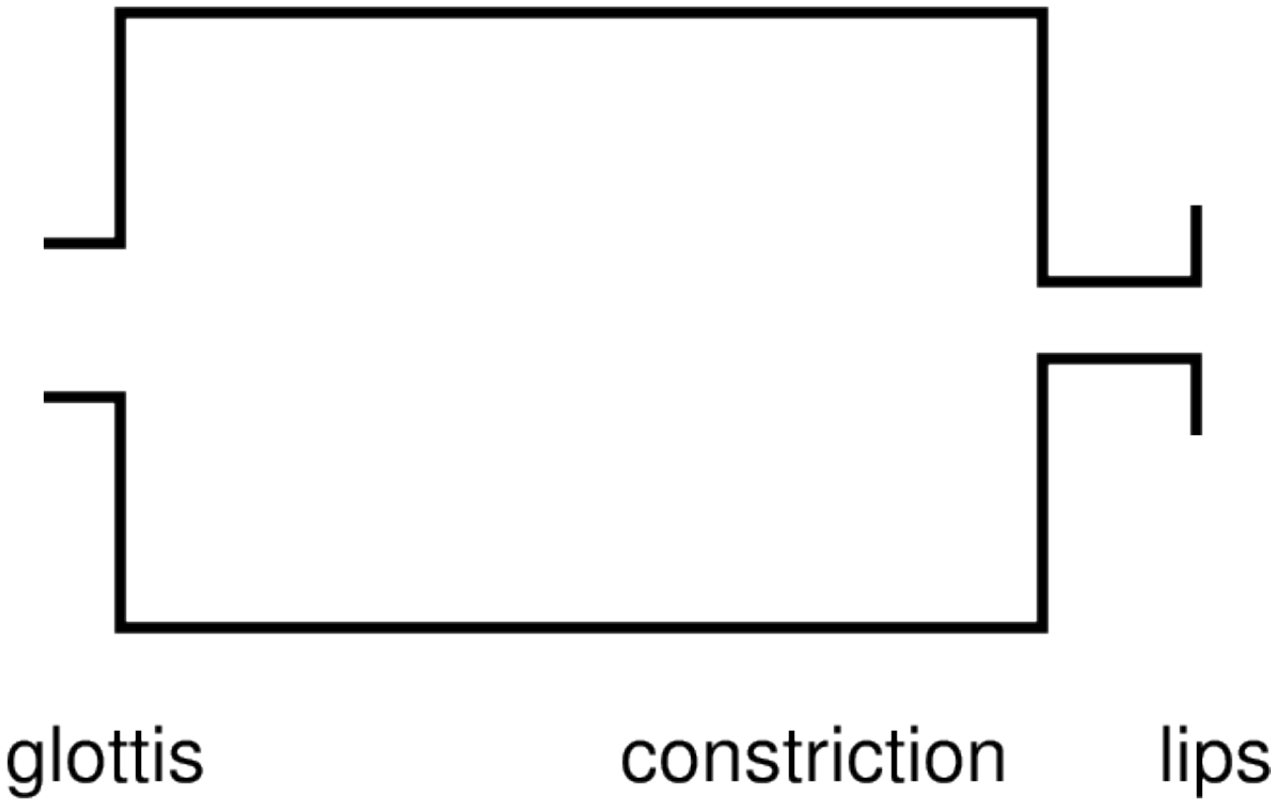
Tube model for fricatives



Tube model for fricatives



Tube model for fricatives



Front cavity resonance

- size of front cavity
  - determines resonance(s)
- large front cavity means
  - low resonant frequency
  - fricative spectrum has energy peak at lower frequency
- small front cavity means
  - high resonant frequency
  - fricative spectrum has energy peak at higher frequency

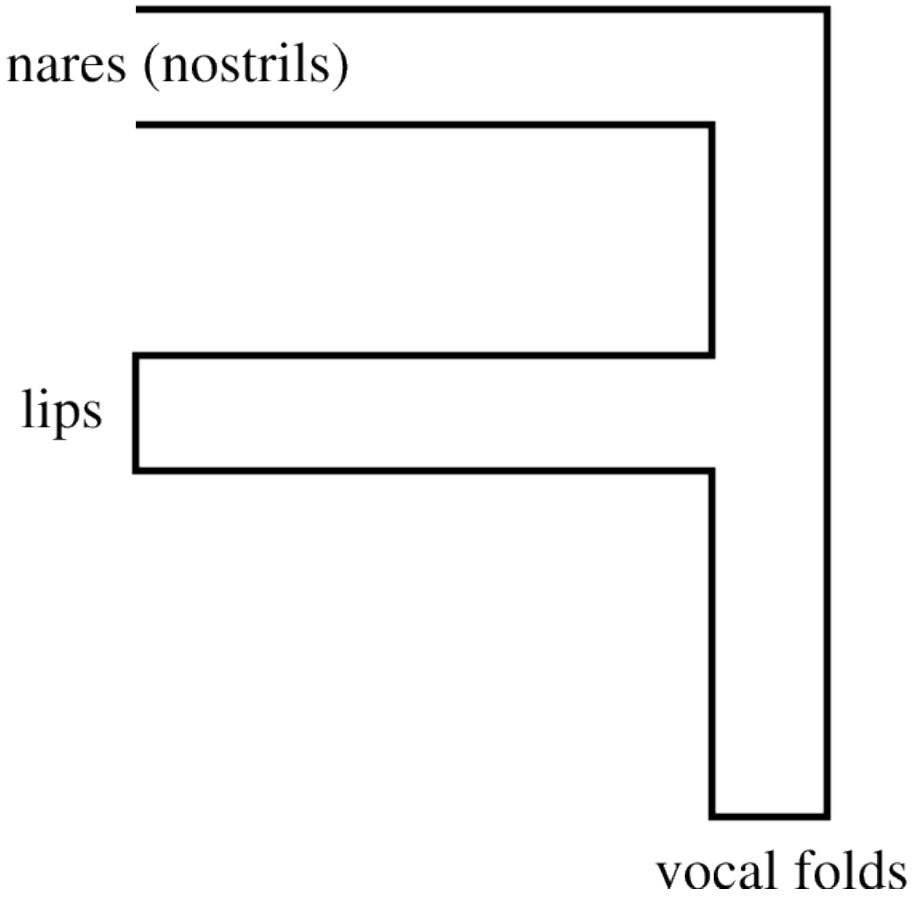
Tube models for nasals

- nasal passage becomes connected to vocal tract
  - tube has side branch
- what effect does this have on the spectrum?

Tube model of [m]

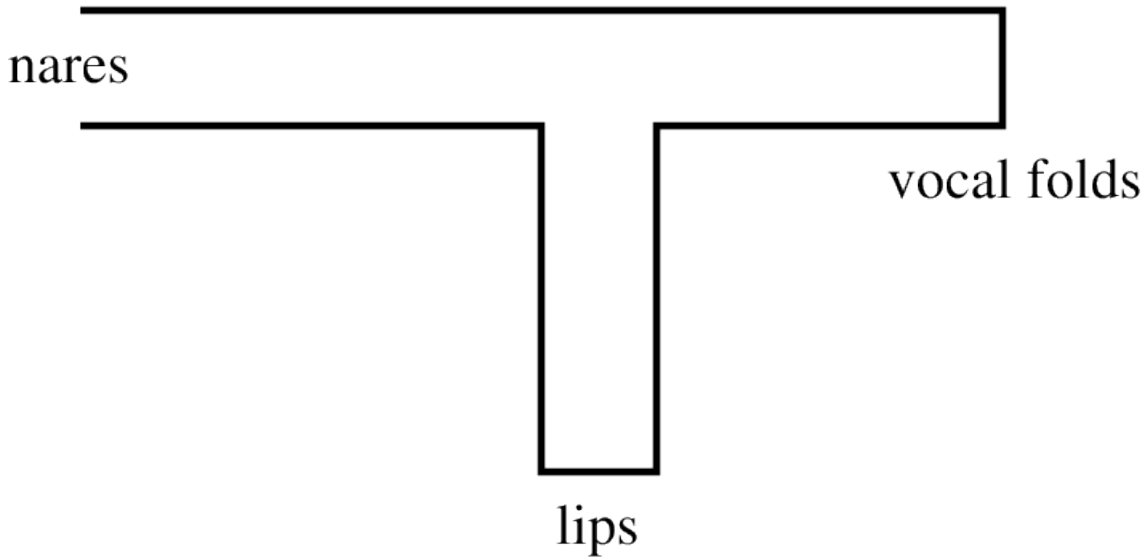
- velum lowered
  - nasal passage connected
- lips closed
  - oral cavity closed at one end
  - connected to pharynx and nasal cavity at other end
- voiced sound: vocal folds are vibrating

Tube model of [m]

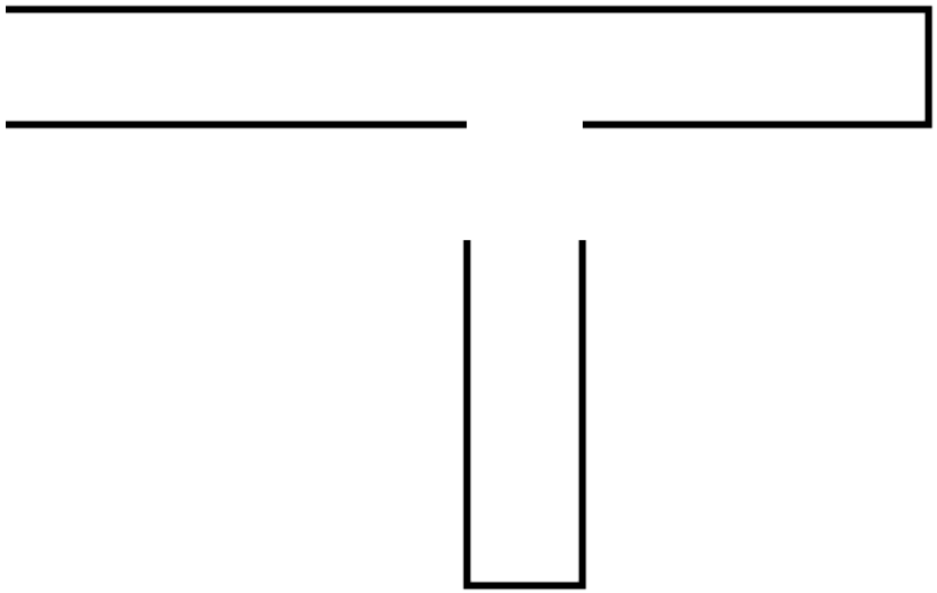


Equivalent tube model of [m]

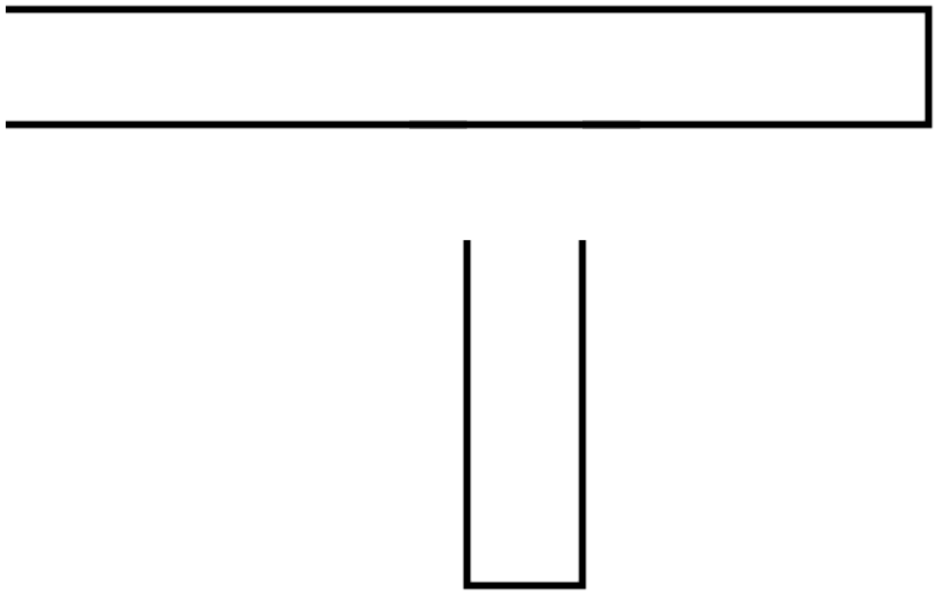
Remember - we can treat curved tubes the same as straight ones



Two parts to the tube



Approximation



Resonances

- main tube
  - resonates in the usual way
  - resonances are formants of nasal
- side tube
  - resonates too
  - but is a “dead end” – not connected directly to outside world
  - therefore **absorbs** energy at the resonant frequency/frequencies
  - produces **antiformants** (known as A_1, A_2, ...)

Frequency of antiformant(s)

- can easily work out from length of side branch
  - it is the oral cavity
  - lips close one end off
  - so use formula for tube open at one end, closed at other
- for [m], assume length is 8cm (0.08m) then antiformants are at around A_1=1100Hz and A_2=3300Hz

[Johnson chapter 6]

Anti-resonance

- this is **not** simply a lack of resonance
  - it is an actual *removal* (suppression) of frequencies from the spectrum
- technical terms:
  - *pole*: another way of describing a resonance
  - *zero*: another way of describing an anti-resonance

Vowels only have resonances – i.e. just poles.
Nasals have resonances *and* antiresonances – i.e. poles *and* zeros.

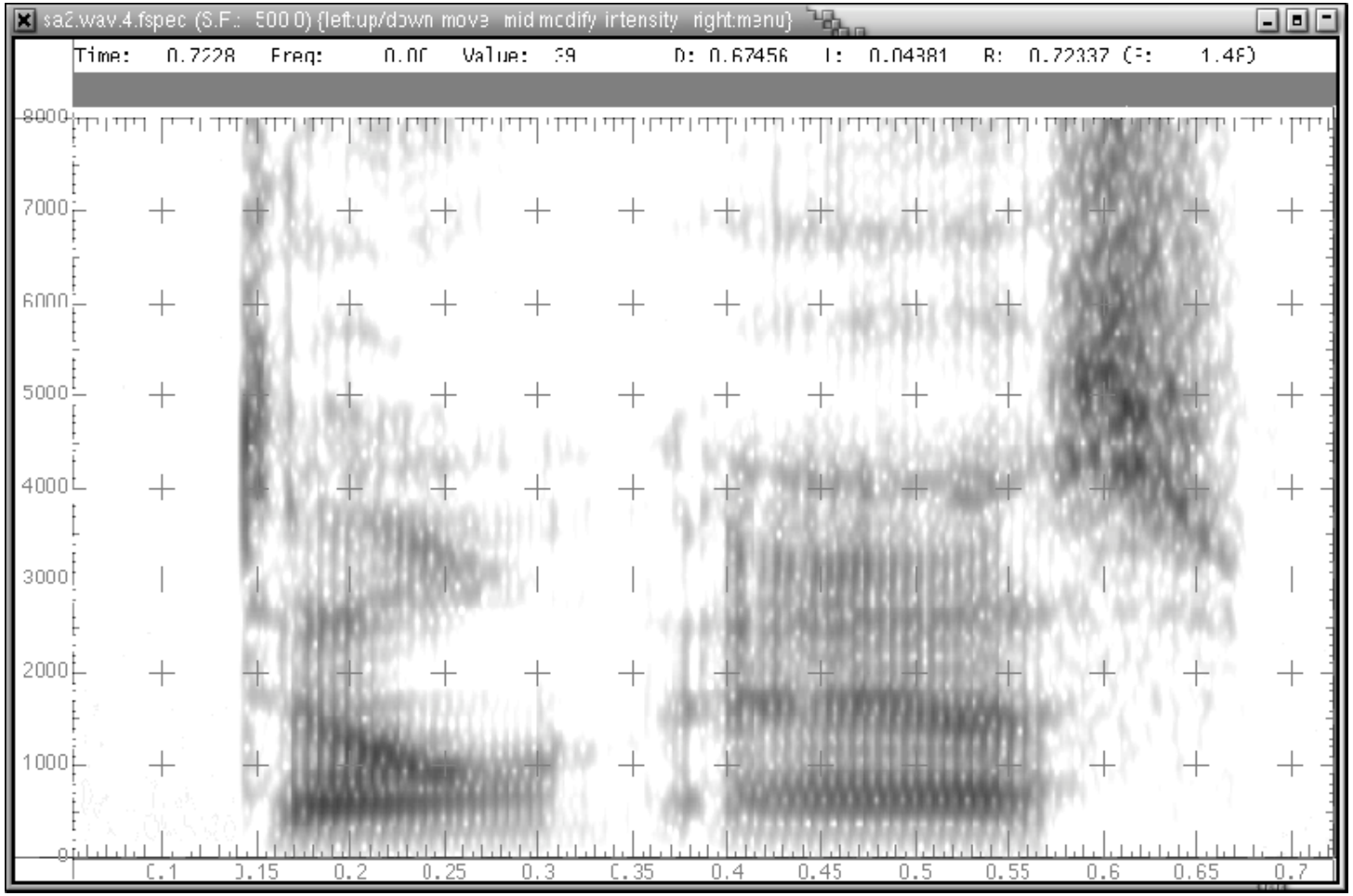
Other nasals - [n]

- oral branch still closed
  - but not by lips, as in [m]
  - but further back
- oral branch no shorter than in [m]
- assume length is 5.5cm (0.055m) then antiformants are at around A_1=1600Hz and A_2=4800Hz

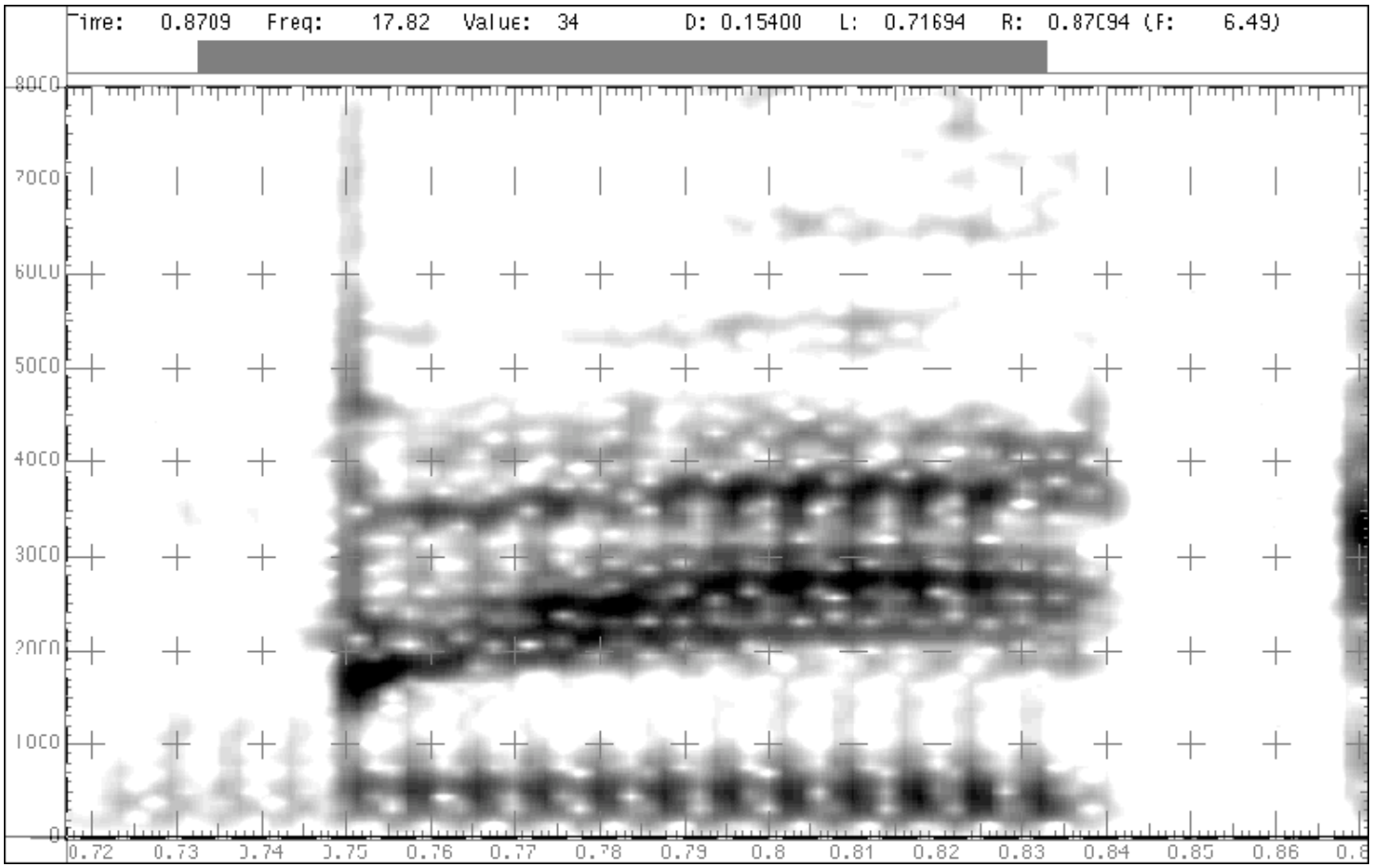
What do zeros look like in the spectrum?

- poles (resonances) appear as peaks in the spectrum
- zeros (anti-resonances) appear as dips
  - on a spectrogram: an area with low energy (a “gap”)
- the lowest antiformant (A_1) is easiest to see
  - for [m] – we predict area of low energy around 1100Hz
  - for [n] – we predict area of low energy around 1600Hz

Spectrogram of “...don’t ask...”



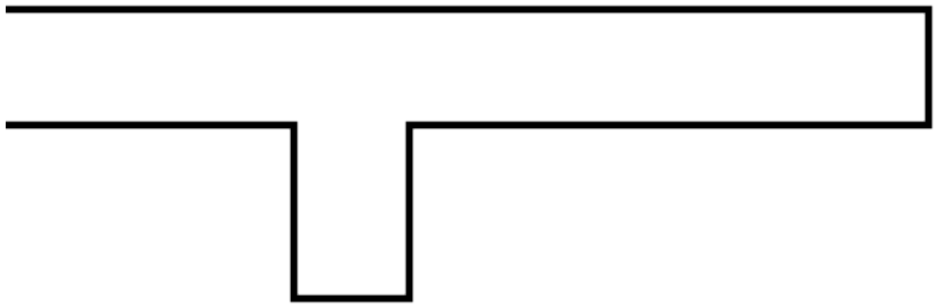
Spectrogram of “[...as]k me t[o...]”



Laterals

- can use a similar model to nasals [n] and [m]
- for [l]
  - tongue tip makes contact with palate
  - but vocal tract remains open around edges (that's why it's called a lateral)
  - a pocket of air remains above the tongue
  - a simplified tube model for this will have
    - tube for vocal tract
    - side branch for air pocket

Tube model for [l]



- very simplified model
  - but main prediction is correct: an antiformant due to the side-branch

Formant bandwidths of nasals

- bandwidth is due to damping
  - more damping means larger bandwidth
- in vocal tract
  - soft tube walls provide damping
- so, a greater area of tube wall means more damping
  - during nasals, nasal cavity adds extra tube wall area
  - therefore, nasals have greater formant bandwidths

Lip rounding

- main effect of rounding is that
  - lips protrude more
- thus making the vocal tract longer
  - and affecting the formant values: they are lowered
  - particularly those formants associated with the front cavity
- precisely which formant is most affected depends on the vocal tract configuration
- often F_3 is most strongly lowered

[Ladefoged chapter 8, last 5 pages]

What we are not covering

- *nasalised vowels*
  - *both nasal and oral cavities open*
  - *more complex system of formants*
    - *and anti-formants*
  - *formants also affected by acoustic coupling, so simple tube model predictions are less accurate than before*

Quantal theory

- before we start, a caveat:
  - this theory is not entirely uncontroversial
  - you may choose to believe only parts of it
- what is quantal theory?

  there are certain regions of “articulatory space” where making small movements has little or no effect on the acoustic output

  $\longrightarrow$ there are certain articulator settings which produce stable sounds, even when the articulators are *not precisely positioned*

First aspect of quantal theory: articulatory “slop” is allowable

- for a range of articulatory configurations
  - acoustic output varies only a very small amount

This is good news, since

- speakers do not have infinitely precise control over their articulators
- despite this, listeners will still receive relatively stable speech signals

Quantal theory: vocal folds

- there are three main ways the folds can behave
  1. apart, laminar airflow, no sound (e.g. breathing, unvoiced sounds except [h])
  2. close enough to phonate (all voiced sounds)
  3. completely closed (glottal stops)
- and, **most importantly**, these three behaviours can easily be acheived without precise muscle control
- that is, we can get away with a little articulatory “slop”

Quantal theory: fricatives

- require a constriction to create turbulent airflow
  - articulators too far apart: laminar flow
  - too close: a stop (closure)
- but for a (small) range of constriction sizes
  - turbulence is caused
  - acoustic signal essentially independent of constriction area
- so, although constriction area must be fairly accurately acheived, some articulatory “slop” is allowable

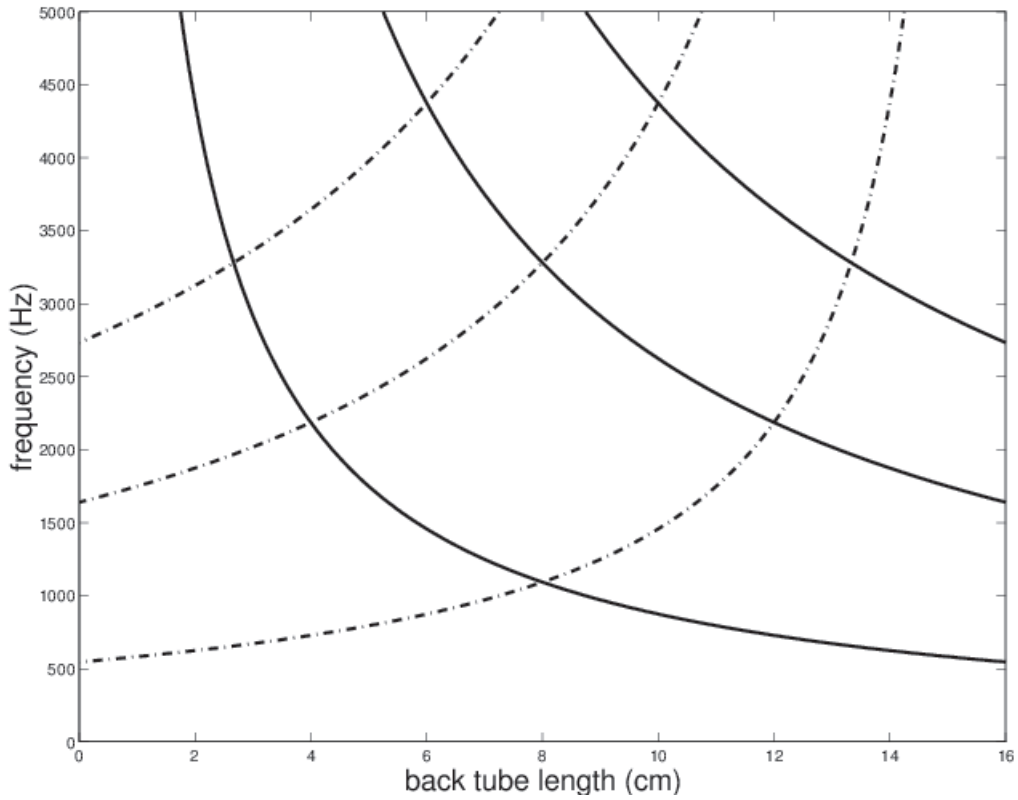
Second aspect of quantal theory: abrubt changes

- regions of stability
  - e.g. vocal folds have three distinct modes of behaviour
- abrubt transitions between these regions
  - onset of phonation as folds come together is relatively abrubt, not gradual
  - turbulence during fricatives created suddenly as constriction area becomes small enough

Quantal theory: vowels

- across the languages of the world
  - some vowels are more common than others
  - why?
- quantal theory provides one possible explanation
- we'll see an alternative explanation also

Resonances: two tube model



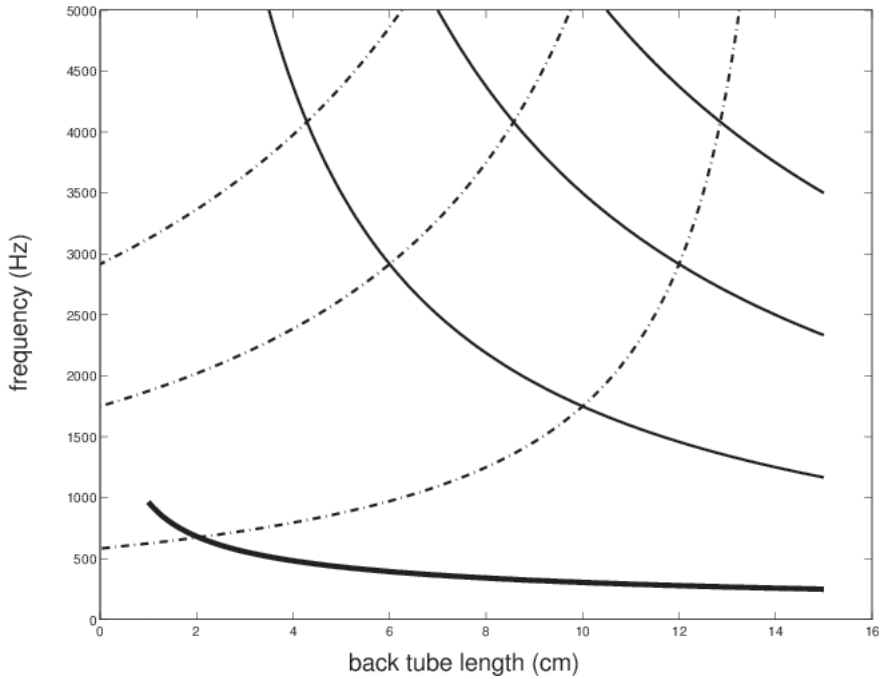
dotted lines are front cavity resonances

16:26

Quantal theory: preferred vowel [ɑ]

- certain regions of the diagrams (i.e. certain vocal tract shapes)
  - produce essentially constant acoustic output (i.e. filter has same frequency response)
  - even when small variations are made about that vocal tract shape
- e.g. for back cavity length of around 7-9cm, F_1 and F_2 do not vary very much
- which would predict that [ɑ] is preferred vowel

Resonances: tubes including a Helmholtz resonance



thick line is Helmholtz resonance

16:28

Quantal theory: preferred vowel [i]

- for back cavity length of 9-11cm, there is a stable F_2 region
  - which predicts that [i] will also be a preferred vowel
- similar reasoning [see Johnson 5.3] can be used to predict another preferred vowel: [u]

Does anybody speak a language without these vowels: [ɑ, i, u] ?

Alternative explanation for preferred vowels: adaptive dispersion

- a fancy name for an intuitively very obvious thing
  - preferred vowels are widely spaced in formant space
  - therefore easy to produce so they sound different from one another
  - and also easy to perceive
- so vowels with extreme F_1 and F_2 are preferred: [ɑ, i, u]
- *noting that we need to look at F_1 and F_2 on a perceptual frequency scale – next lecture!*

This theory only talks about vowels, whereas quantal theory attempts to make predictions about other (all?) sound classes

Psychoacoustics

- two lectures on properties of human auditory system

  1: biological basis

  2: consequences for speech perception

- human auditory system *non-linear* in various ways
  - e.g. frequency scale, amplitude sensitivity

- we will look at
  - why this is
  - what this means for speech

Psychoacoustics, lecture 1: biological basis

- human auditory system
  - anatomy
  - physiology
  - a functional model
  - non-linear properties

Anatomy of the auditory system: overall

Peripheral auditory system (i.e. the part not in the brain)

- outer ear
  - collects sound
- middle ear
  - system of bones transmits sound to inner ear
- inner ear
  - converts sound waves into nerve impulses

Anatomy of the auditory system: outer ear

- directional sound gathering device
  - having two ears allows us to localise sounds
  - can face sound source to “focus”
  - can distinguish sounds in front / behind us

Anatomy of the auditory system: middle ear

- function: to transmit sound from outer ear (air filled) to inner ear (fluid filled)
  - sound waves hit tympanic membrane (eardrum), causing it to vibrate
  - system of three bones conducts vibrations from eardrum to inner ear

[Perkins & Kent figures 9-1, 9-2, 9-4]

Anatomy of the auditory system: cochlea

- cochlea: a fluid filled tube
  - converts sound waves to nerve impulses
  - decomposes signal into frequency components, a bit like a Fourier transform (more detail in a moment)

[Perkins & Kent figures 9-8, 9-10]

Anatomy of the auditory system: hair cells in the cochlea

- convert movement (due to sound waves travelling in the cochlea) to nerve impulses
  - tiny hairs caused to move by sound wave
  - hair movement triggers nerve firing
  - nerve signal sent along auditory nerve to brain

[Perkins & Kent figures 9-11]

Physiology of the auditory system: cochlea

i.e. what function does it perform, and how?

- takes complex sound wave
  - and decomposes it into frequency components
- inside cochela is the *basilar membrane*
- thin and narrow at one end, thick and wide at the other

[Perkins & Kent figures 10-1]

Physiology of the auditory system: basilar membrane

- remember: small things resonate at high frequencies
  - thin and narrow end of membrane responds to higher frequencies
  - causing membrane to displace (move) – actual amount of movement is very small
  - movement is sensed by hair cells

[Perkins & Kent figures 10-4]

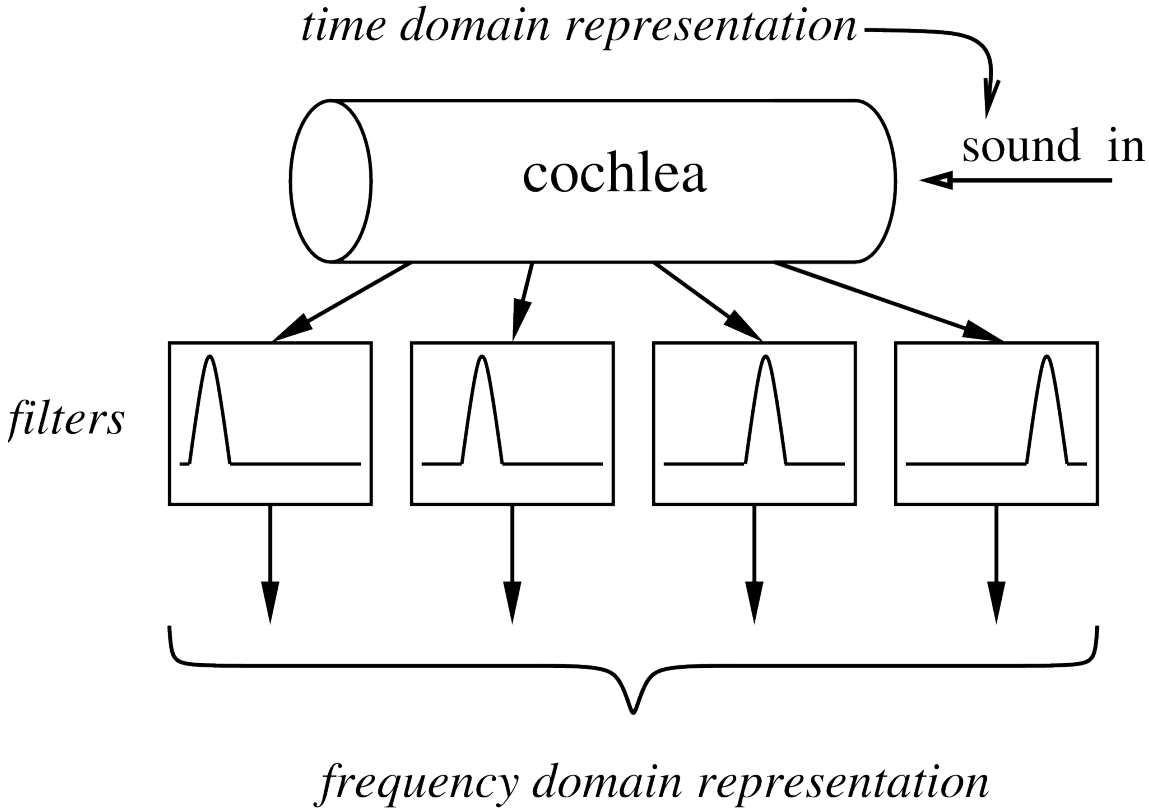
Physiology of the auditory system: hair cells

- movement of basilar membrane causes
  - shearing (sideways) force on hair cells

[Perkins & Kent figures 10-5]

Cochlea functional model

Acts much like a bank of filters (i.e. resonators)



Frequency sensitivity/discrimination

- resonances are spaced out along the basilar membrane which runs the length of the cochlea
  - but not linearly (evenly) spaced on a Hertz scale
- what scale are they spaced on? (this lecture)
- what are the consequences for speech perception? (next lecture)

Non-linear frequency scales

- two aspects to the frequency scale that go hand-in-hand
  - ability to distinguish two similar tones varies with absolute frequency
  - perceived difference of two tones depends on absolute frequency as well as relative difference in frequencies
- both due to the same underlying physiological reason
  - the positions along the basilar membrane that respond to different frequencies are not spaced linearly

Just-noticable-difference (JND)

- minimum frequency difference required for perceptual difference
  - varies with frequency
  - exact figures depend on experimental method used

For example, using sequential presentation of pure tones:

- up to around 1000Hz (1kHz): JND is about 2Hz
- above 1000Hz: JND is around 0.2% of the frequency
  - 6Hz at 3kHz, 14Hz at 7kHz, ...

Non-linear frequency scale

- based on perceptual experiment
- roughly speaking:
  - frequency scale linear up to 1kHz
  - then logarithmic

What does a logarithmic scale actually mean?

- perceived frequency difference depends on *ratio* of frequencies in Hz
- e.g. perceptual distance between 2kHz $\leftrightarrow$ 4kHz is same as between 4kHz $\leftrightarrow$ 8kHz and between 8kHz $\leftrightarrow$ 16kHz

Perceptual frequency scales

- often, more useful to use perceptual scale than a linear Hertz scale
  - various scales have been proposed, all very similar
  - three most popular: Mel, Bark and ERB (equivalent rectangular bandwidth)
- all have roughly same properties:
  - frequency scale linear up to 1kHz
  - then approx. logarithmic

Non-linear amplitude sensitivity

- we can perceive very quiet sounds
  - but also stand very loud sounds (at least briefly) without pain/damage
- can conduct perceptual experiments
  - e.g. ask subjects to adjust one sound to be half/same/twice loudness as a reference sound
- for a given frequency, perceptual loudness scale is essentially logarithmic
- across different frequencies, it is very non-linear

Bels and decibels (dB)

- define a logarithmic scale for sound pressure level (SLP)
  - such that a factor of 10 change in **intensity** is 1 Bel or 10dB
- **intensity** is proportional to the *square* of sound pressure level (**SPL**)
  - a factor of 10 change in **SPL** is 20dB
- the dB SLP scale measures physical sound pressure
- human perception caries with frequency, so there is a perceptual version of the scale: dB SL (sensation level)

Non-linear amplitude sensitivity

- across different frequencies
  - signal amplitude must vary to keep constant perceptual loudness

[3 figures showing equal loudness curves; first one like Ladefoged 6.7]

Most sensitive across frequencies found in speech

- auditory system is particularly sensitive across the frequency range
  - 500 – 5000Hz
- which co-incides with the frequencies containing the information in speech signals

Next lecture

- using non-linear frequency scales to analyse speech
  - reveals F_1 and F_2 more easily than linear scale
- other aspects of human audition affecting speech perception
  - limited frequency resolution
  - masking in time and frequency

Psychoacoustics 2 : implications for the perception of speech (and other sounds)

we can only scratch the surface – this is a big subject area

- non-linear frequency scales
  - and critical bands
- higher level (more central) processing
  - masking
    - some sounds can obscure others
  - binaural hearing
    - cocktail party effect

Perceptual correlates of acoustic properties

acoustic property	perceptual phenomenon
intensity	loudness
frequency	pitch
duration	perceived duration
spectral properties	timbre

These can interact

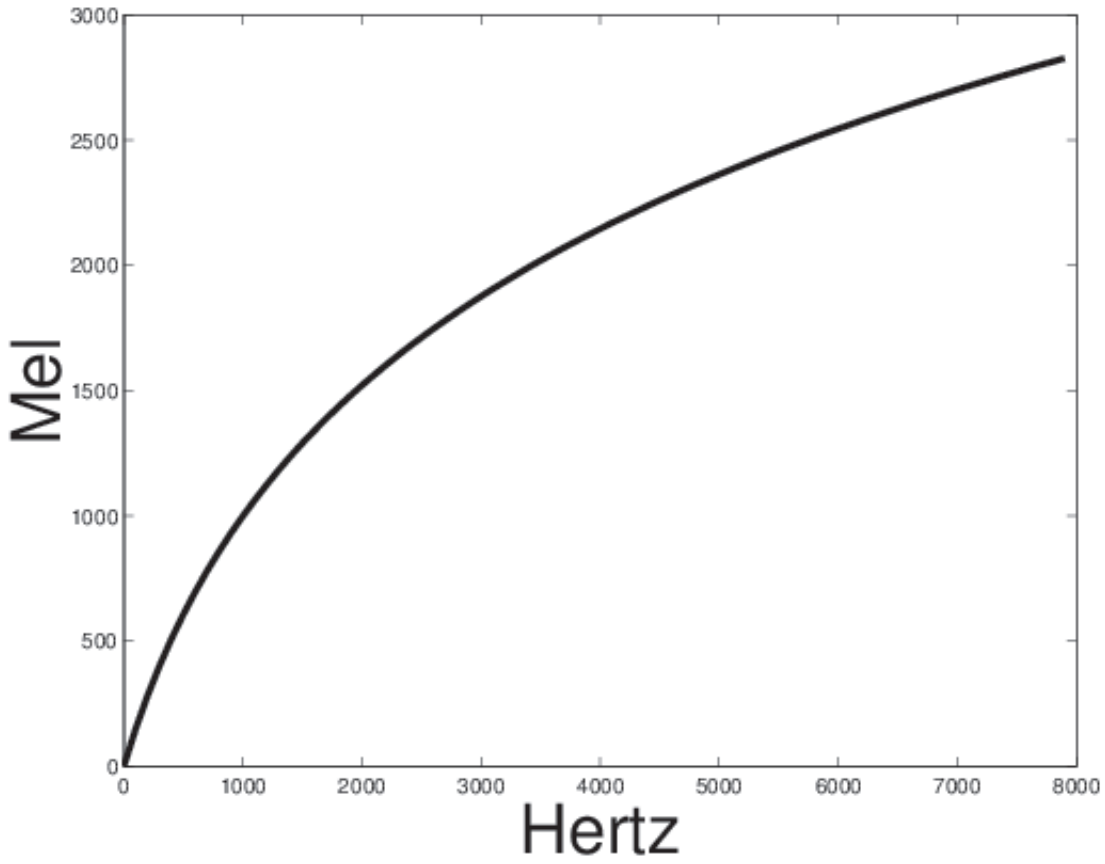
e.g. for short sounds, longer duration gives perceptually greater loudness

Recap: loudness

- loudness is the perceptual correlate of intensity
- loudness of sounds
  - roughly corresponds to logarithm (log) of *intensity*
- intensity is a physical property of the sound wave
  - intensity is the mean *squared* amplitude of the sound pressure

Recap: non-linear frequency scales

e.g. the Mel scale



Critical bands

- frequency resolution of cochlea is limited
  - two sounds of slightly different frequency may have same perceived pitch
- width (in Hz) of critical bands increases with frequency (in Hz)

Auditory spectrograms

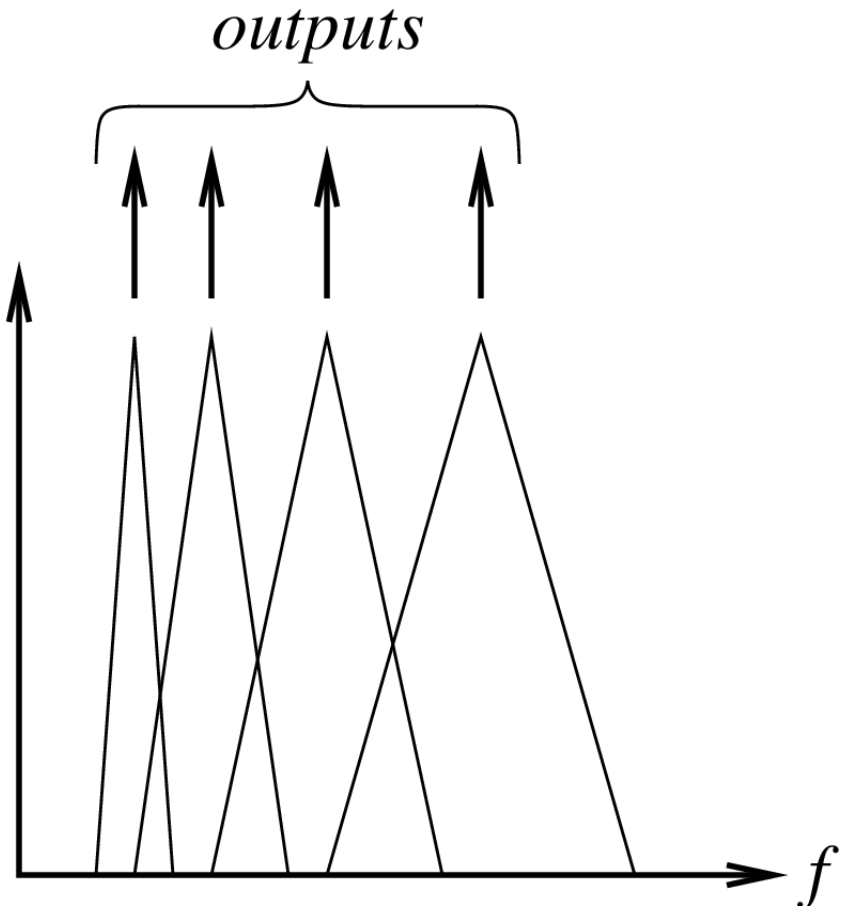
- vertical axis uses a perceptually-motivated frequency scale
  - lower frequencies more spread out – see more detail
  - for speech: formants more separated and easy to distinguish
- sometimes also model other aspects of the cochlea, e.g. critical bands
  - then called **cochleagrams**

[Johnson figures 3.8 and 7.9]

Cochlea: functions as set of overlapping filters

- filters spaced non-linearly along the frequency scale
- each filter corresponds to one critical band
- accounts for:
  - limited frequency resolution
  - non-linear relationship between frequency and perceived pitch
  - some aspects of masking

Cochlea as a filterbank



Loudness of complex tones

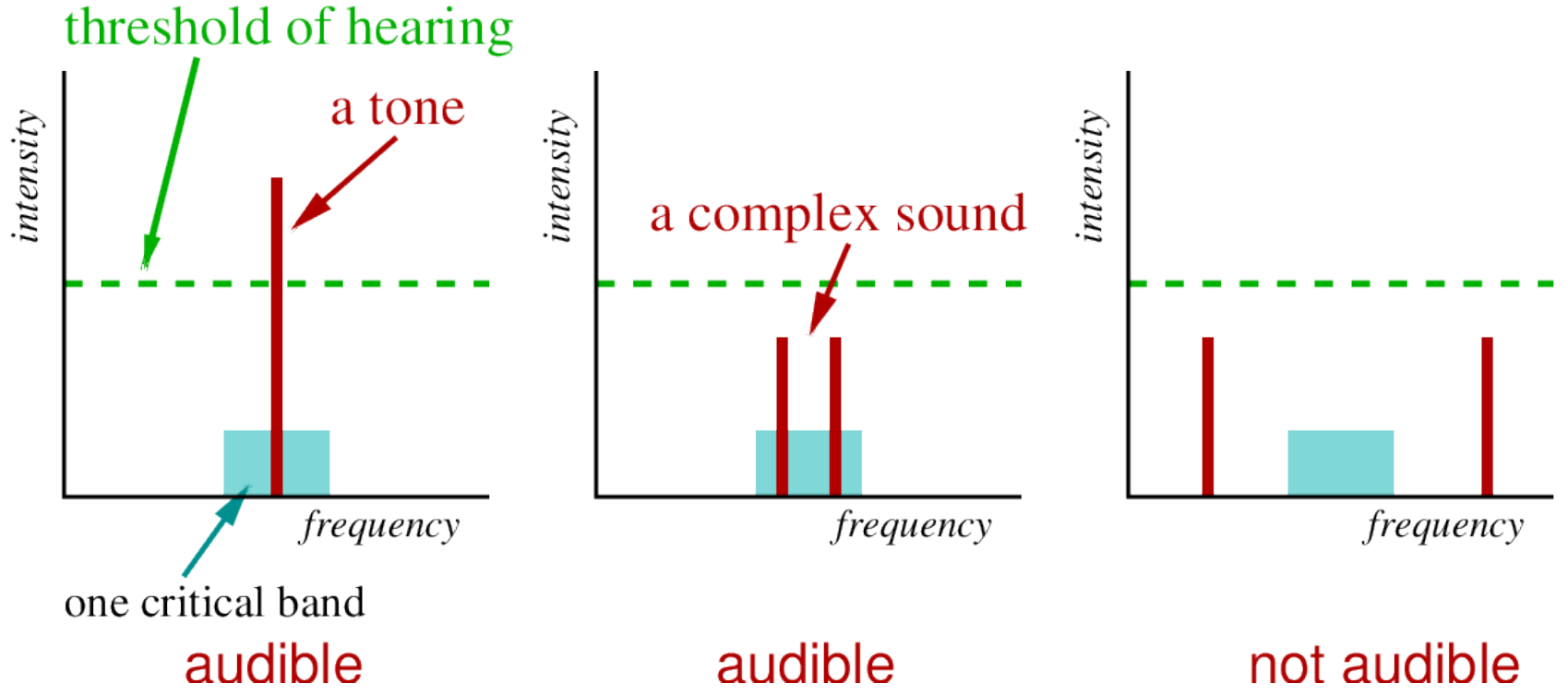
remember, loudness is approx. **log (intensity)**

how loud is a sound consisting of two tones?

- if within a single critical band
  - intensity is summed
  - then “converted” to loudness
- otherwise
  - loudnesses are summed

Implications for perception of complex sounds

If two tones fall within a single critical band, intensity of each tone required for detection of this complex tone is **half** that for a single tone



Masking

Masking means reducing our ability to detect some acoustic event

- one tone can inhibit perception of another tone
  - when tones are presented together
  - or even when presented separately
- can measure this using perceptual experiments

Masking in the basilar membrane

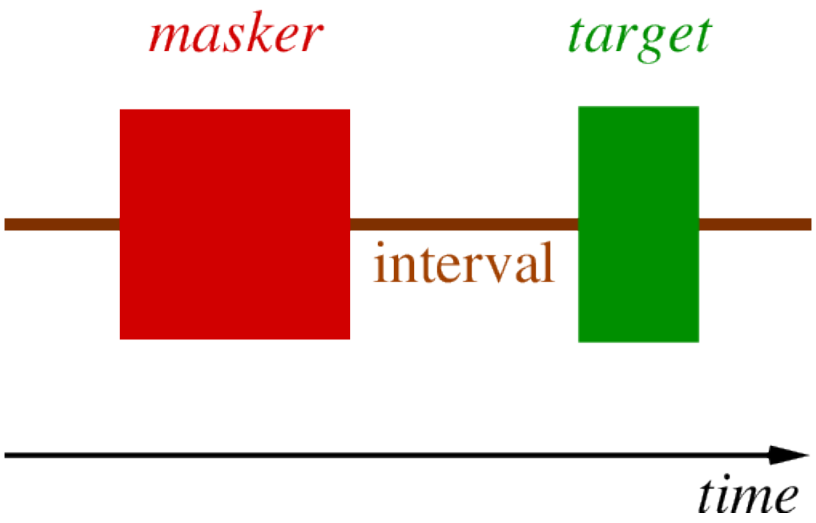
- presenting two tones of different frequency simultaneously
  - louder tone can **mask** quieter one
- can explain in terms of critical bands
- this process may take place in the basilar membrane

However, masking can occur in ways that cannot be explained in terms of the peripheral auditory system

Masking in more central processes

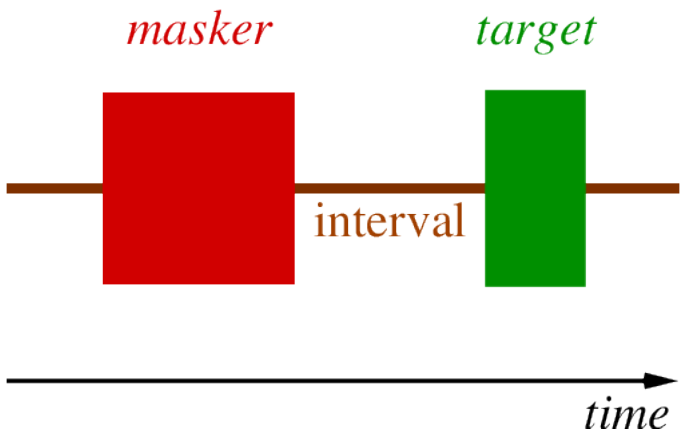
- masking effect can happen even if
  - two tones not presented simultaneously
  - e.g. in sequence or binaural presentation

Masking



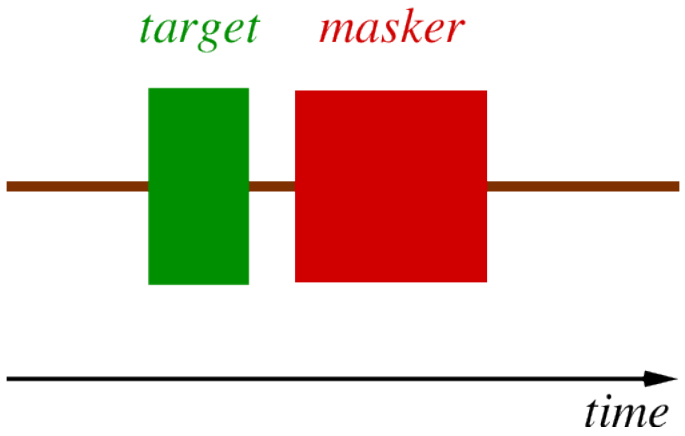
- masking tone or narrow-band noise
- target tone

Forward masking



- first present masking tone
- then target tone
  - if interval is less than around 300ms, masking occurs

Backward masking



- first present target tone
- then masking tone
  - can still get masking, if masking tone is presented less than 40ms after target tone

Central masking

- present target tone to one ear
- and masking tone to other ear
  - masking effect observed
- therefore masking doesn't (only) happen in the peripheral auditory system, there must be some central interaction

Masking and complex sounds

- complex sounds (e.g. speech)
  - essentially composed of a number of pure tones
  - so masking can occur between these tones
- implies that
  - we cannot perceive all the component sounds of speech
  - or indeed of other sounds

  This knowledge is very useful for audio coding/compression, e.g. mp3

Speech (and other audio) coding

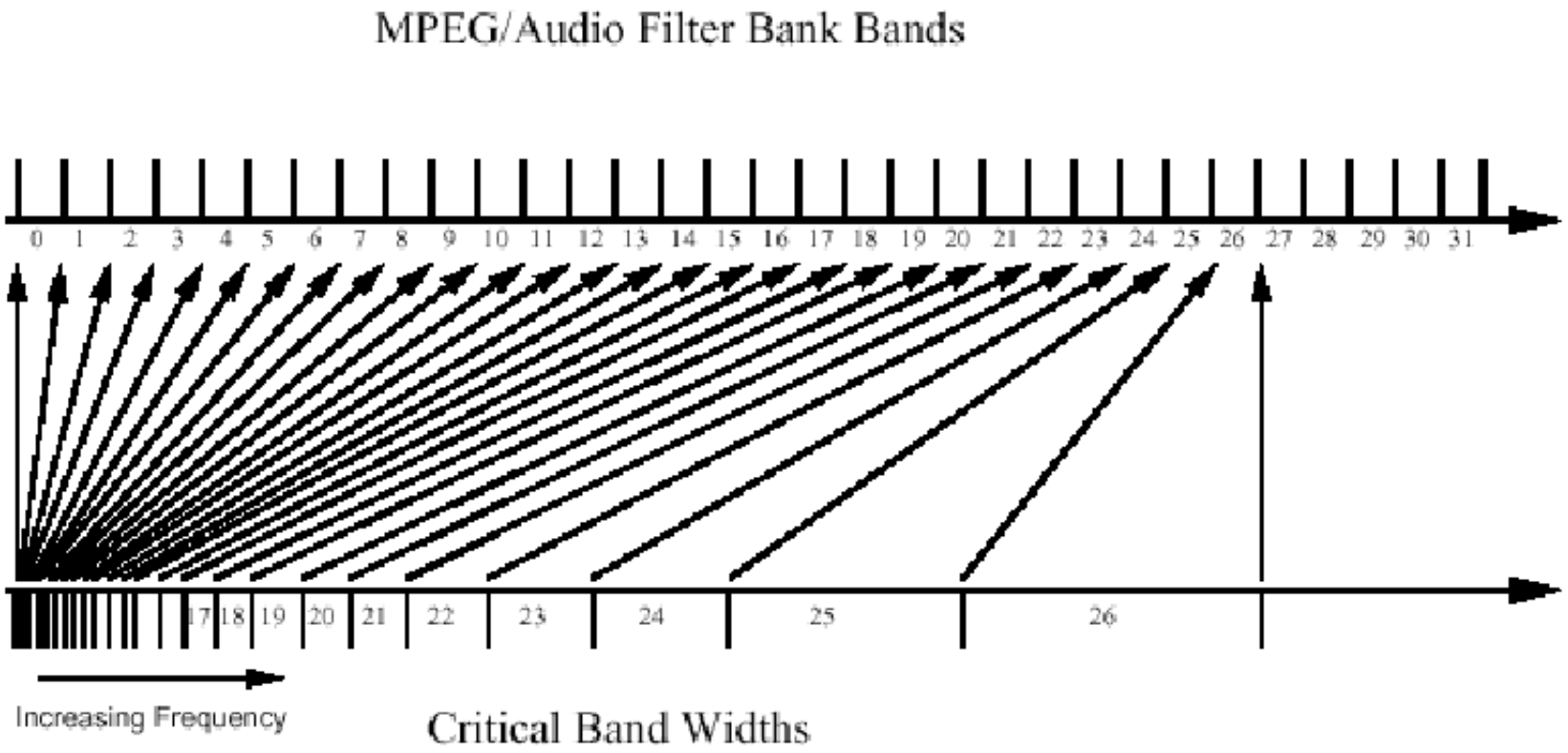
- often, want to compress audio signals
  - for storage (e.g. minidisc, mp3 coding)
  - or transmission (e.g. mobile phones, digital radio, satellite links, transatlantic phone lines, internet, etc.)
- for maximum compression
  - need to discard some of the signal
  - i.e. *lossy* compression
- knowledge of properties of human auditory perception
  - helps us choose what to discard

mp3 audio compression

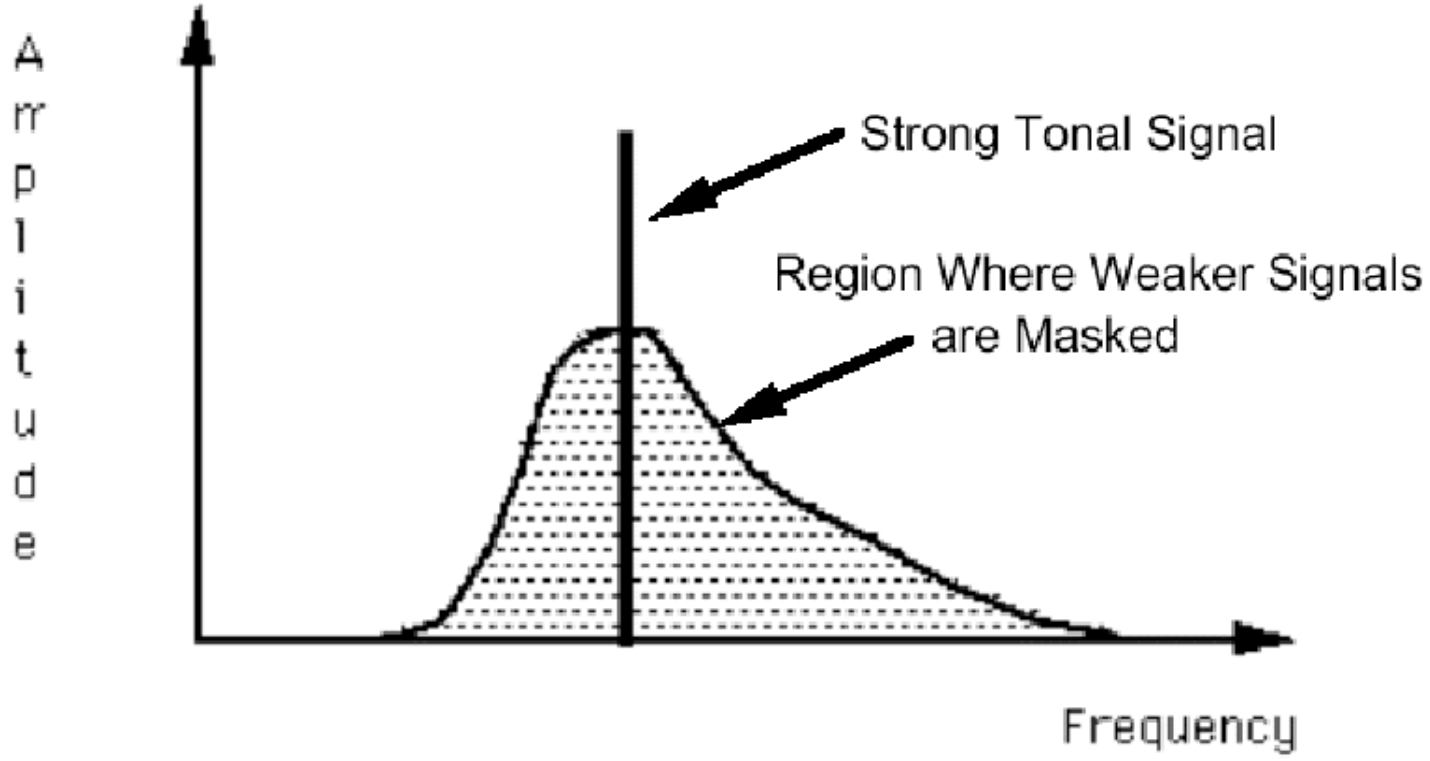
- normally, 1 minute of CD-quality sound takes about 10Mb (70 minutes of music on one CD = around 700Mb)
  - with mp3 compression, we can reduce the space required by a factor of over 10
  - so one CD can hold over 10 hours of near-CD-quality music
- to acheive this compression, some information must be discarded
- psychoacoustic models are used to determine what parts of the signal can be removed with little or no perceptual difference

mp3 auditory filterbank

Transforms signal to frequency-domain representation (like a Fourier transform, except on a non-linear frequency scale)



mp3 masking model



Cocktail party effect

- binaural hearing
  - allows us to localise sound (determine direction)
  - makes it easier to pick out one particular sound in a noisy environment
- we already saw how masking can inhibit detection of a sound
  - binaural hearing helps separate masker and sound of interest

Binaural unmasking

Present a masker and a target tone

- if both presented to just one or to both ears
  - the target tone is not detected

Now present

- masker + target to left ear
- just the masker to the right ear
  - this makes it possible to detect the tone

Binaural unmasking

- the target is perceived as being close to the left ear
- the masker is perceived as being directly in front
- the separation of mask and tone in space results in unmasking
  - the target tone is now detected

Only some central processing can be doing this – the signals from both ears are required.

Summary

- some peripheral processing in the middle ear and cochlea
  - gain control and amplitude compression (logarithmic)
  - non-linear frequency scale (compresses higher frequencies)
  - critical bands
- and other, more central processes
  - masking and un-masking / cocktail party effect
- one application of this knowledge
  - audio compression