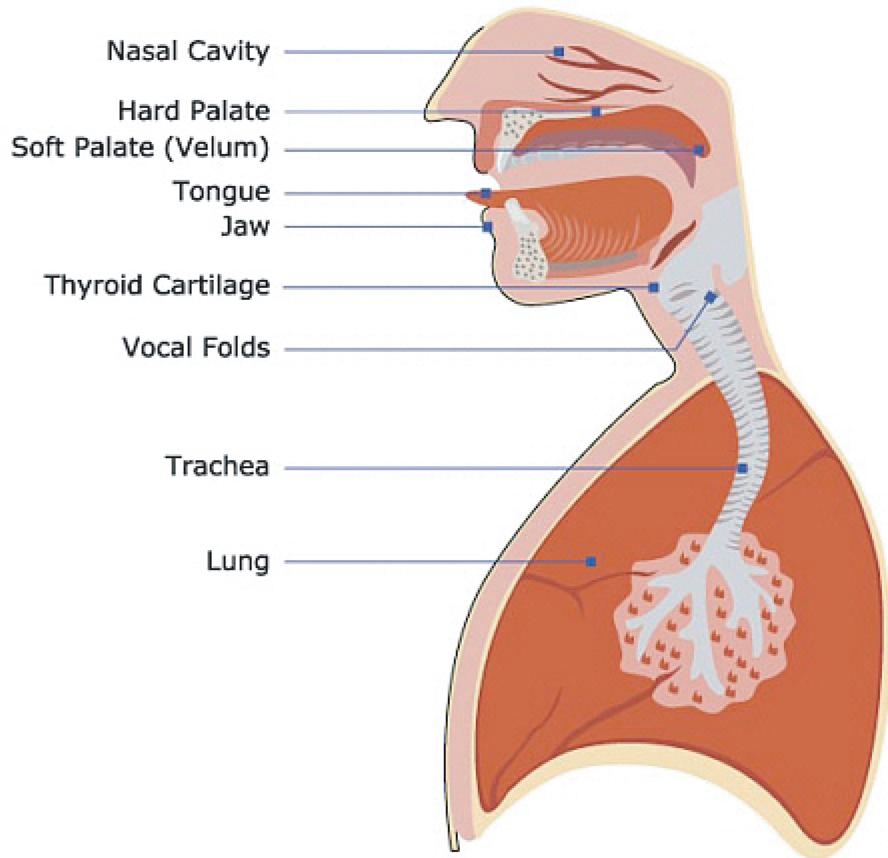
Acoustic Theory of Speech Production

- Overview
- Sound sources
- Vocal tract transfer function
  - Wave equations
  - Sound propagation in a uniform acoustic tube
- Representing the vocal tract with simple acoustic tubes
- Estimating natural frequencies from area functions
- Representing the vocal tract with multiple uniform tubes

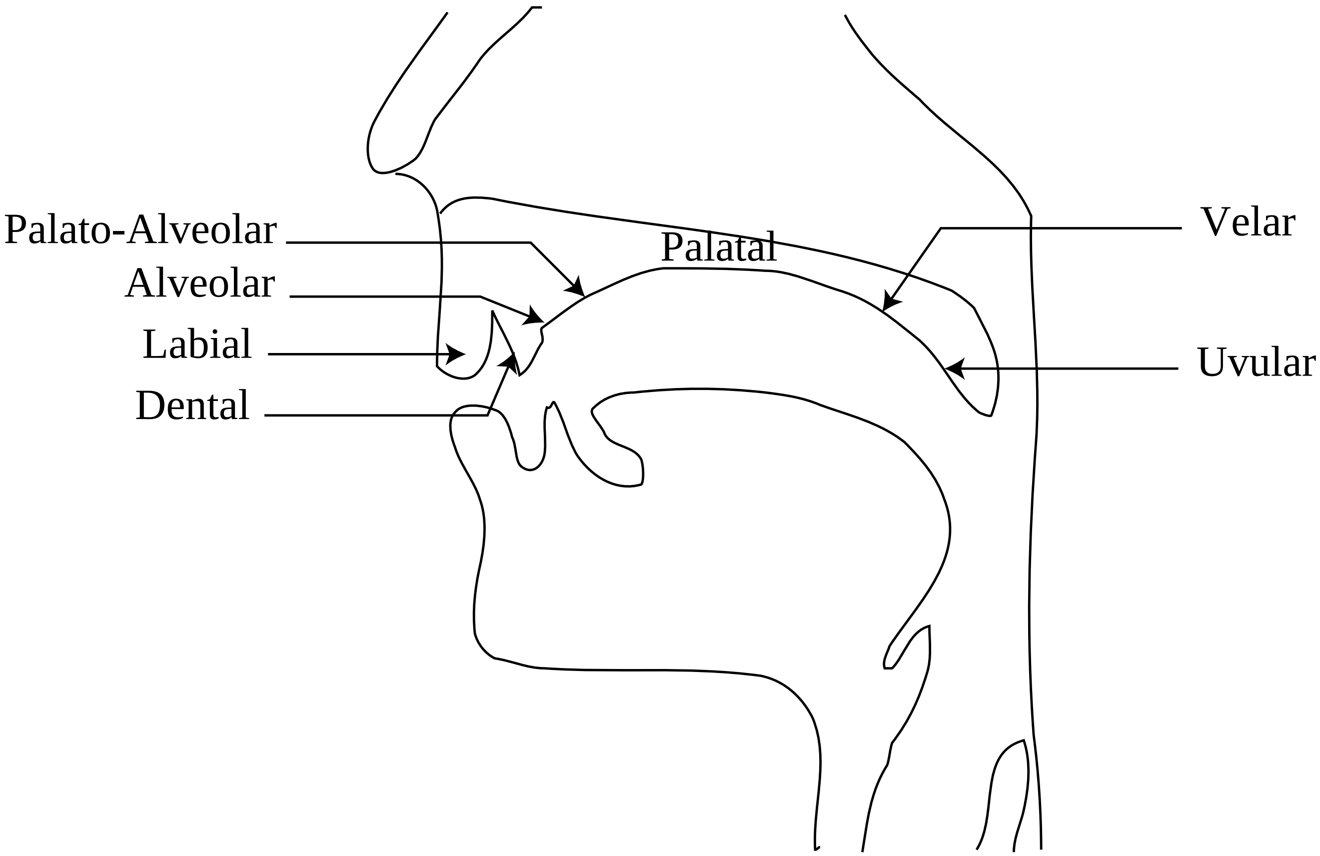
Anatomical Structures for Speech Production



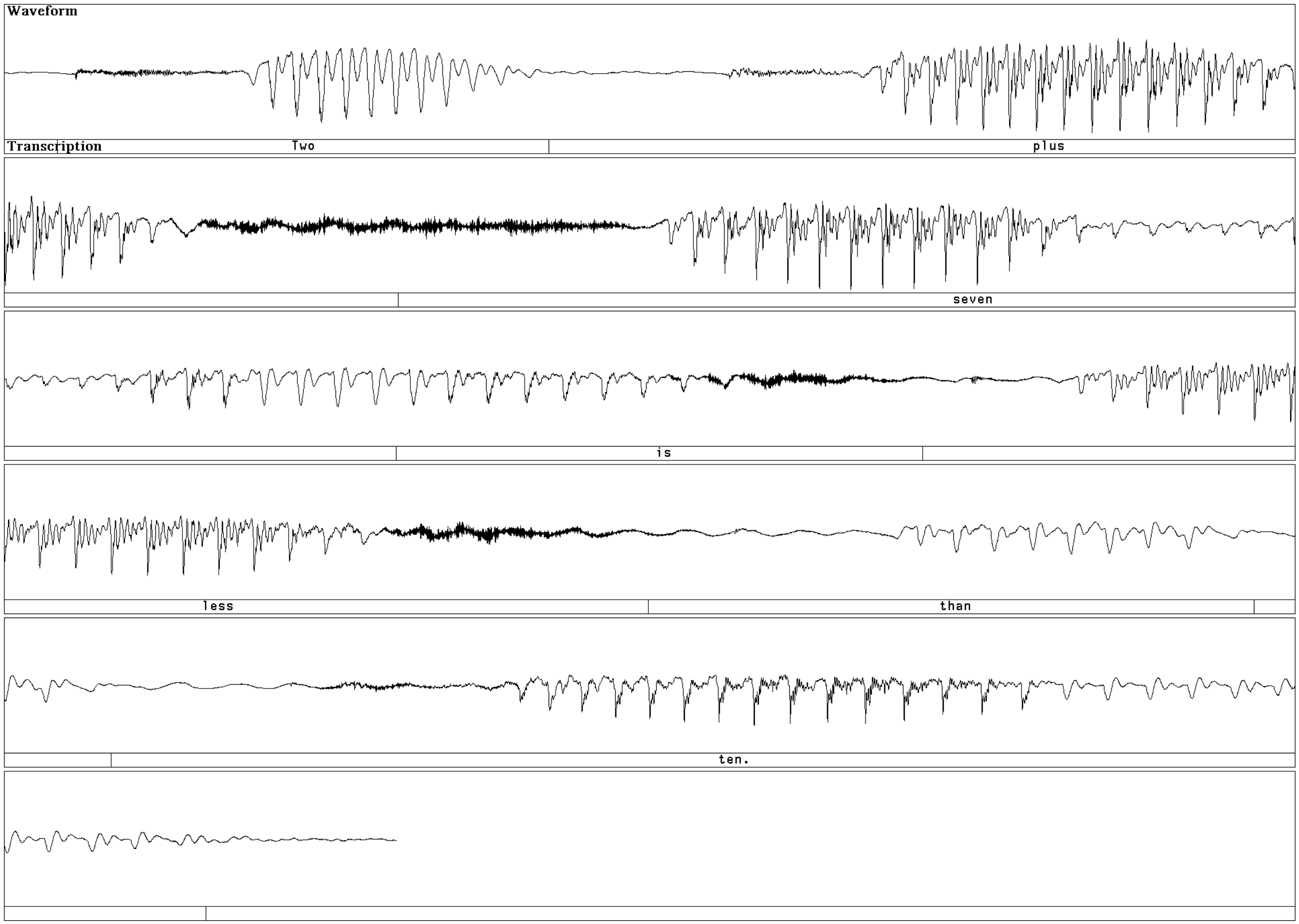
Phonemes in American English

PHONEME	*EXAMPLE*	*PHONEME*	*EXAMPLE*	*PHONEME*	*EXAMPLE*
/i^{y}/	beat	/s/	see	/w/	wet
/ɪ/	bit	/š/	she	/r/	red
/e^{y}/	bait	/f/	fee	/l/	let
/ɛ/	bet	/θ/	thief	/y/	yet
/æ/	bat	/z/	z	/m/	meet
/ɑ/	Bob	/ž/	Gigi	/n/	neat
/ɔ/	bought	/v/	v	/ŋ/	sing
/ʌ/	but	/ð/	thee	/č/	church
/o^{w}/	boat	/p/	pea	/ǰ/	judge
/ʊ/	book	/t/	tea	/h/	heat
/u^{w}/	boot	/k/	key		
/ɝ/	Burt	/b/	bee		
/ɑ^{y}/	bite	/d/	Dee		
/ɔ^{y}/	Boyd	/g/	geese		
/ɑ^{w}/	bout				
/ə/	about				

Places of Articulation for Speech Sounds

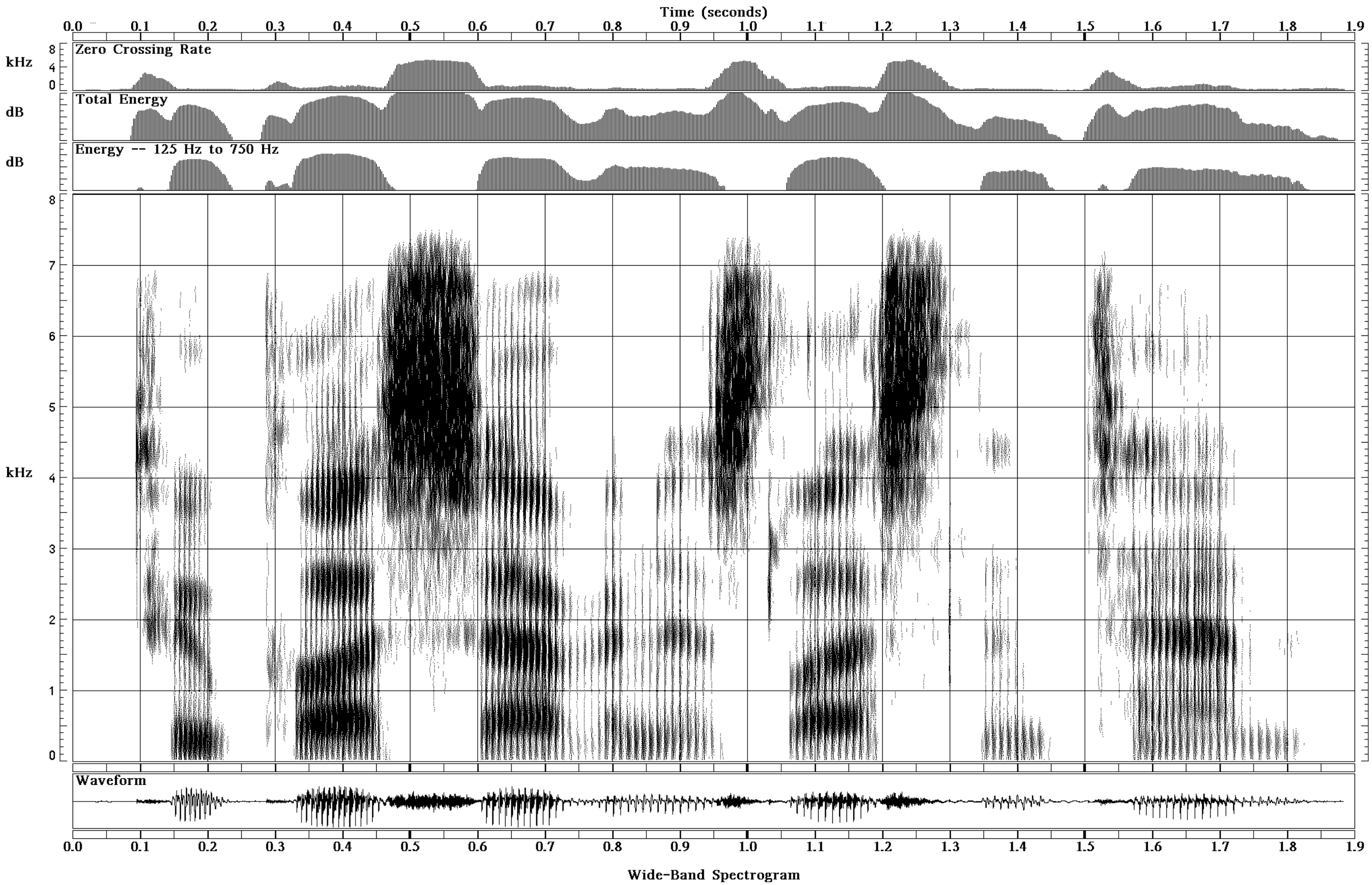


Speech Waveform: An Example



Two plus seven is less than ten

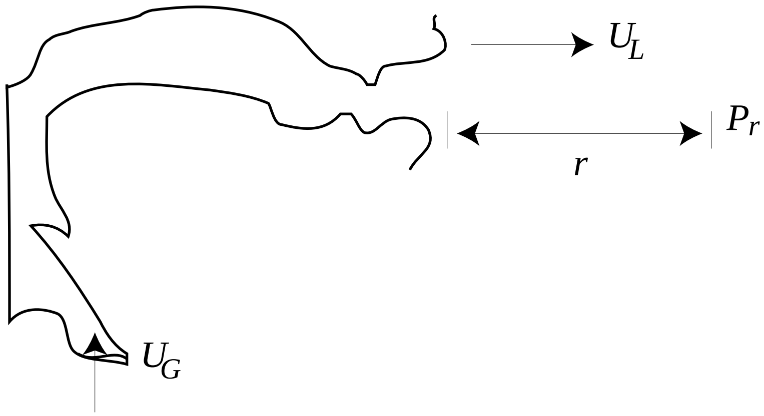
A Wideband Spectrogram



Two plus seven is less than ten

Acoustic Theory of Speech Production

- The acoustic characteristics of speech are usually modelled as a sequence of source, vocal tract filter, and radiation characteristics



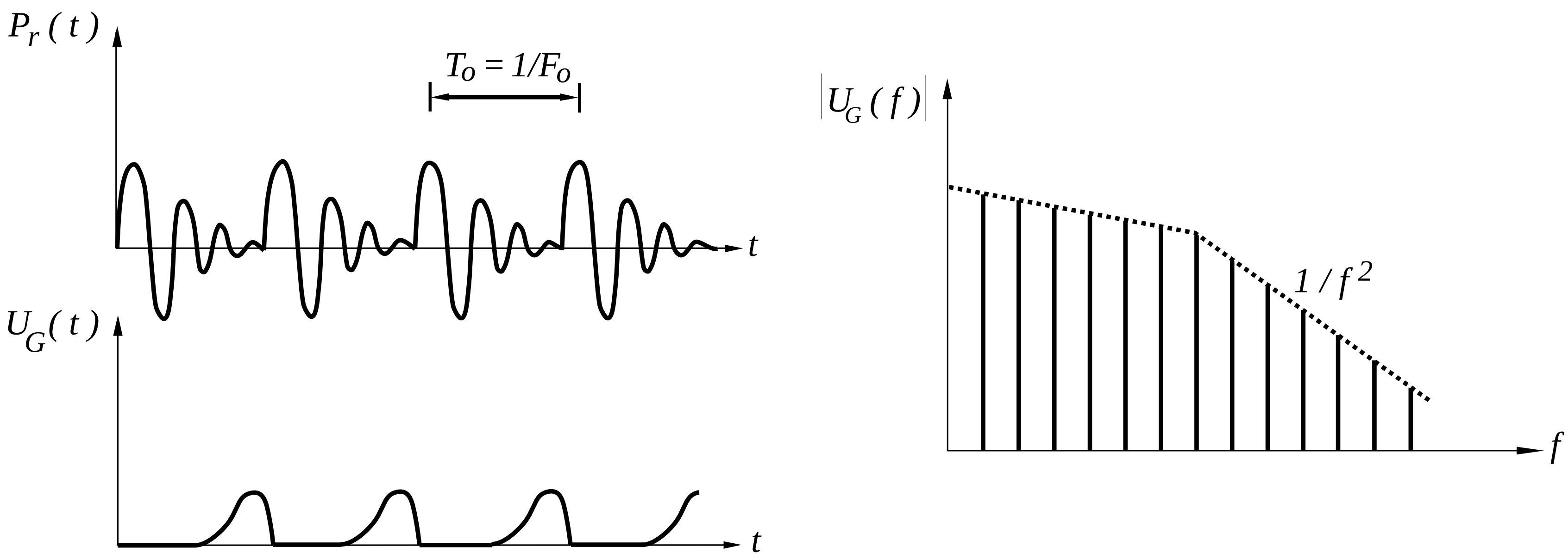
$$P_r(j\Omega) = S(j\Omega)\,T(j\Omega)\,R(j\Omega)$$

- For vowel production:

$$\begin{aligned} S(j\Omega) &= U_G(j\Omega) \\ T(j\Omega) &= U_L(j\Omega)\,/\,U_G(j\Omega) \\ R(j\Omega) &= P_r(j\Omega)\,/\,U_L(j\Omega) \end{aligned}$$

Sound Source: Vocal Fold Vibration

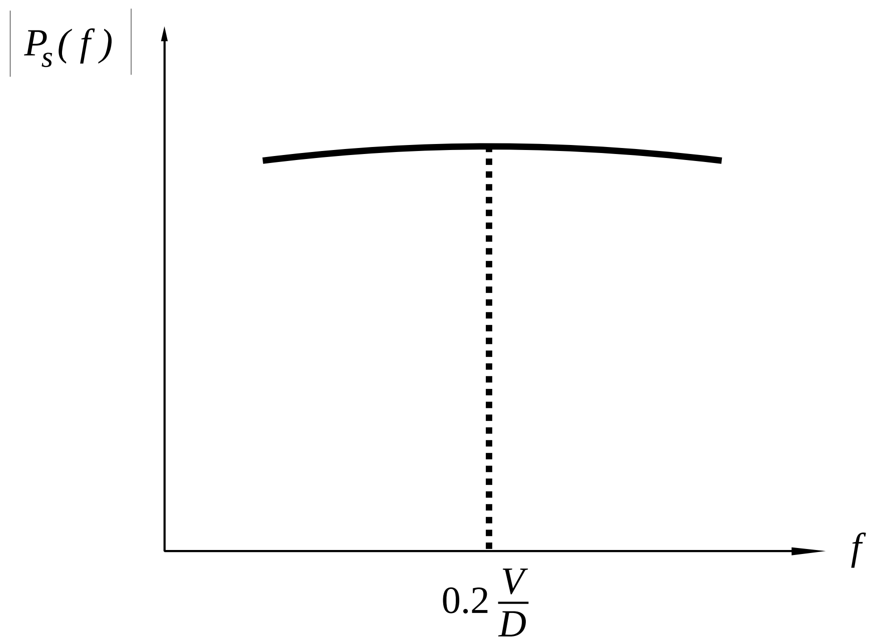
Modelled as a volume velocity source at glottis, $U_G(j\Omega)$



	F_0 ave (Hz)	F_0 min (Hz)	F_0 max (Hz)
Men	125	80	200
Women	225	150	350
Children	300	200	500

Sound Source: Turbulence Noise

- Turbulence noise is produced at a constriction in the vocal tract
  - Aspiration noise is produced at glottis
  - Frication noise is produced above the glottis
- Modelled as series pressure source at constriction, $P_S(j\Omega)$



V: Velocity at constriction $\qquad D$: Critical dimension $= \sqrt{\frac{4A}{\pi}} \approx \sqrt{A}$

Vocal Tract Wave Equations

Define:

$u(x,t)$	$\Longrightarrow$	particle velocity
$U(x,t)$	$\Longrightarrow$	volume velocity ($U = uA$)
$p(x,t)$	$\Longrightarrow$	sound pressure variation ($P = P_O + p$)
ρ	$\Longrightarrow$	density of air
c	$\Longrightarrow$	velocity of sound

- Assuming plane wave propagation (for a cross dimension $\ll \lambda$), and a one-dimensional wave motion, it can be shown that

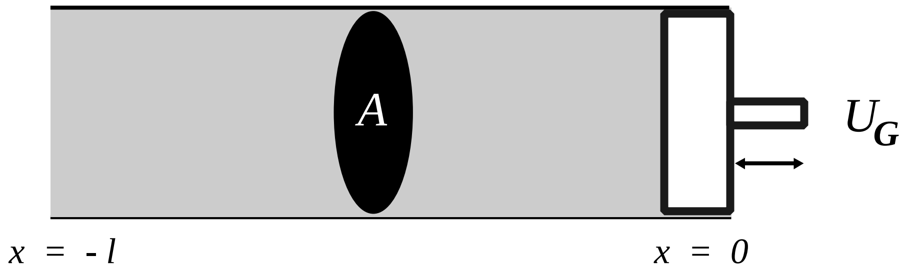
$$-\frac{\partial p}{\partial x} = \rho \frac{\partial u}{\partial t} \qquad -\frac{\partial u}{\partial x} = \frac{1}{\rho c^2}\frac{\partial p}{\partial t} \qquad \frac{\partial^2 u}{\partial x^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2}$$

- Time and frequency domain solutions are of the form

$$u(x,t) = u^+(t - \frac{x}{c}) - u^-(t + \frac{x}{c}) \qquad u(x,s) = \frac{1}{\rho c}\left[P_+ e^{-sx/c} - P_- e^{sx/c}\right]$$

$$p(x,t) = \rho c\left[u^+(t - \frac{x}{c}) + u^-(t + \frac{x}{c})\right] \qquad p(x,s) = P_+ e^{-sx/c} + P_- e^{sx/c}$$

Propagation of Sound in a Uniform Tube



- The vocal tract transfer function of volume velocities is

$$T(j\Omega) = \frac{U_L(j\Omega)}{U_G(j\Omega)} = \frac{U(-\ell, j\Omega)}{U(0, j\Omega)}$$

- Using the boundary conditions $U(0,s) = U_G(s)$ and $P(-\ell, s) = 0$

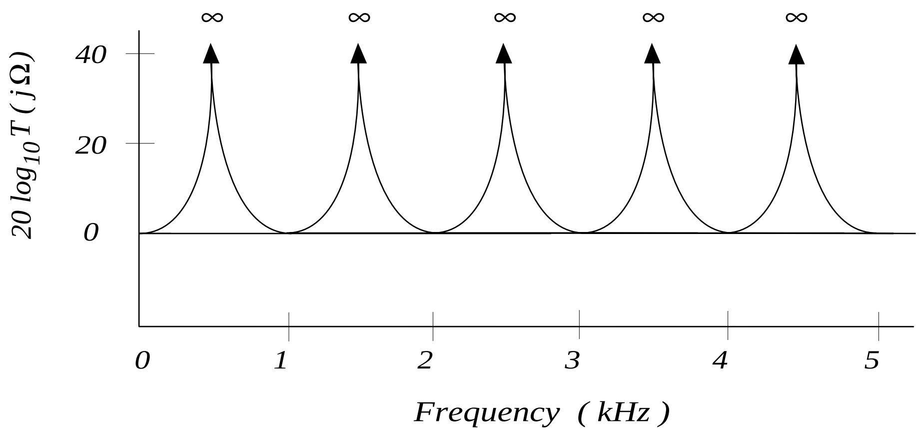
$$T(s) = \frac{2}{e^{s\ell/c} + e^{-s\ell/c}} \qquad T(j\Omega) = \frac{1}{\cos(\Omega\ell/c)}$$

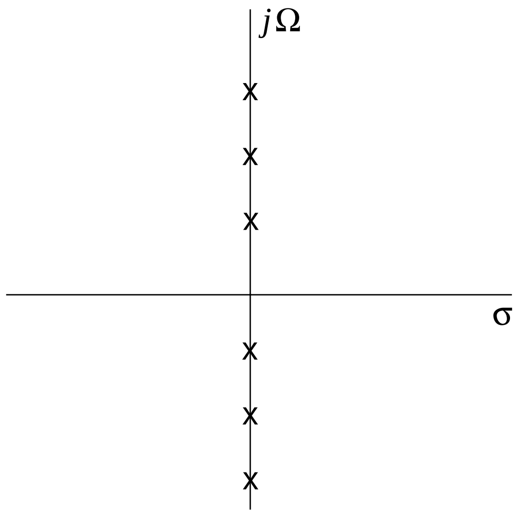
- The poles of the transfer function $T(j\Omega)$ are where $\cos(\Omega\ell/c) = 0$

$$\frac{(2\pi f_n)\ell}{c} = \frac{(2n-1)}{2}\pi \qquad f_n = \frac{c}{4\ell}(2n-1) \qquad \lambda_n = \frac{4\ell}{(2n-1)} \qquad n = 1, 2, \ldots$$

Propagation of Sound in a Uniform Tube (con't)

- For $c = 34,000\,\text{cm/sec}$, $\ell = 17\,\text{cm}$, the natural frequencies (also called the *formants*) are at 500 Hz, 1500 Hz, 2500 Hz, ...

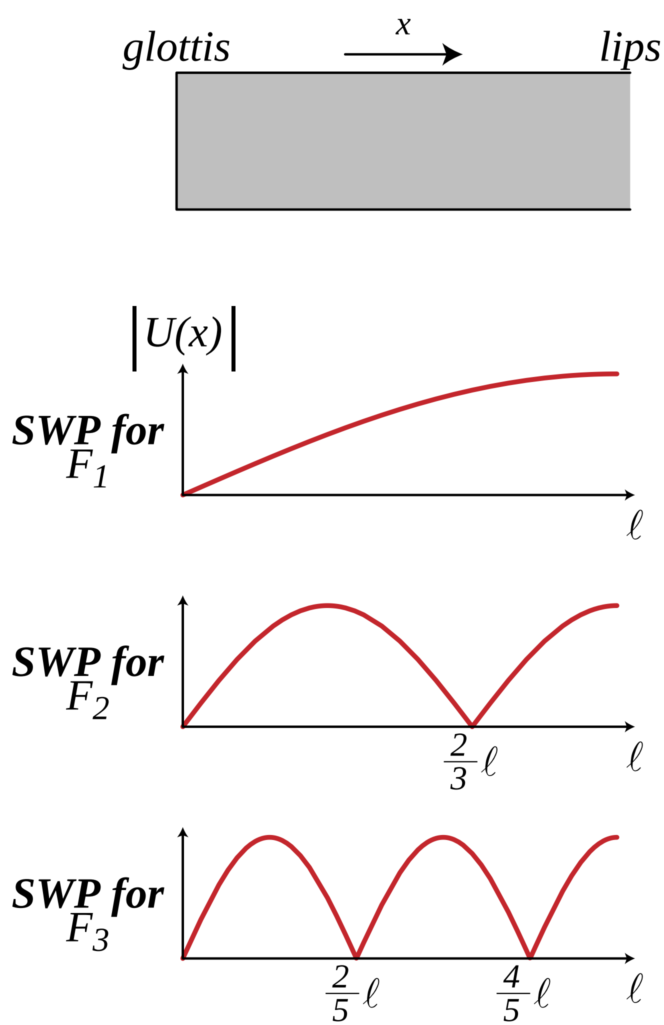




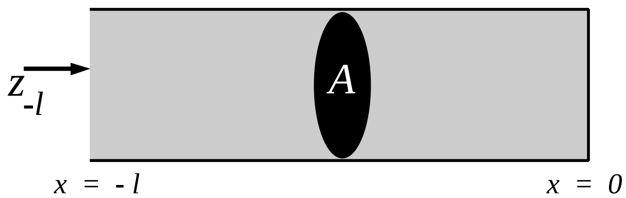
- The transfer function of a tube with no side branches, excited at one end and response measured at another, only has poles
- The formant frequencies will have finite bandwidth when vocal tract losses are considered (e.g., radiation, walls, viscosity, heat)
- The length of the vocal tract, ℓ, corresponds to $\frac{1}{4}\lambda_1, \frac{3}{4}\lambda_2, \frac{5}{4}\lambda_3$, ..., where λ_i is the wavelength of the i^{th} natural frequency

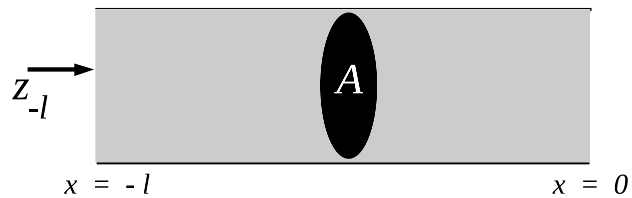
Standing Wave Patterns in a Uniform Tube

A uniform tube closed at one end and open at the other is often referred to as a quarter wavelength resonator



Natural Frequencies of Simple Acoustic Tubes





Quarter wavelength resonator

$$P(x, j\Omega) = 2P_+ \cos \frac{\Omega x}{c}$$

$$U(x, j\Omega) = -j \frac{A}{\rho c} 2P_+ \sin \frac{\Omega x}{c}$$

$$Y_{-\ell} = j \frac{A}{\rho c} \tan \frac{\Omega \ell}{c}$$

$$\approx j\Omega \frac{A\ell}{\rho c^2} = j\Omega C_A \quad \Omega\ell/c \ll 1$$

$C_A = A\ell/\rho c^2 =$ acoustic compliance

$$f_n = \frac{c}{4\ell}(2n-1) \quad n = 1, 2, \ldots$$

Half-wavelength resonator

$$P(x, j\Omega) = -j2P_+ \sin \frac{\Omega x}{c}$$

$$U(x, j\Omega) = \frac{A}{\rho c} 2P_+ \cos \frac{\Omega x}{c}$$

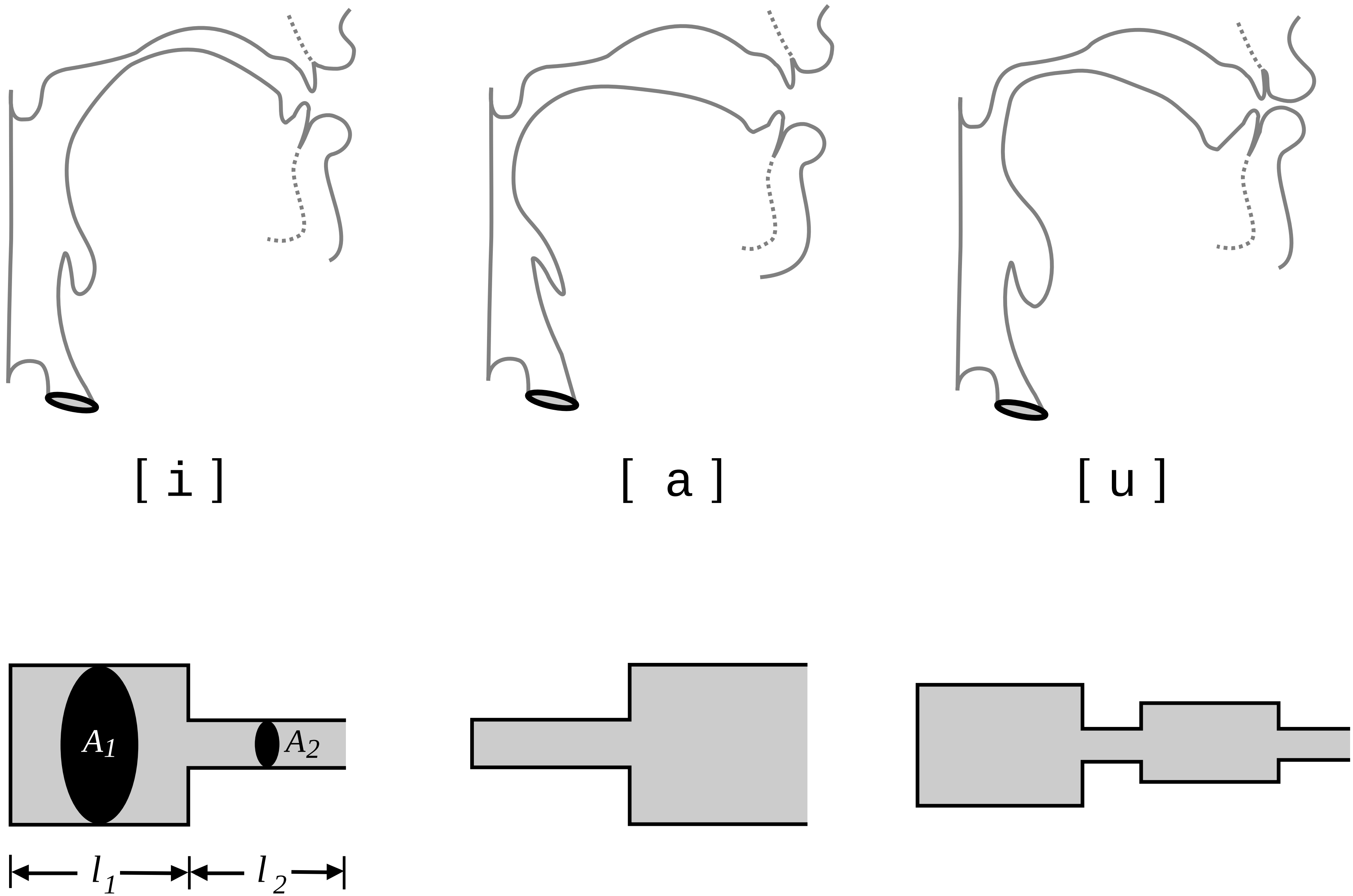
$$Y_{-\ell} = -j \frac{A}{\rho c} \cot \frac{\Omega \ell}{c}$$

$$\approx -j \frac{A}{\Omega \rho \ell} = -j \frac{1}{\Omega M_A} \quad \Omega\ell/c \ll 1$$

$M_A = \rho\ell/A =$ acoustic mass

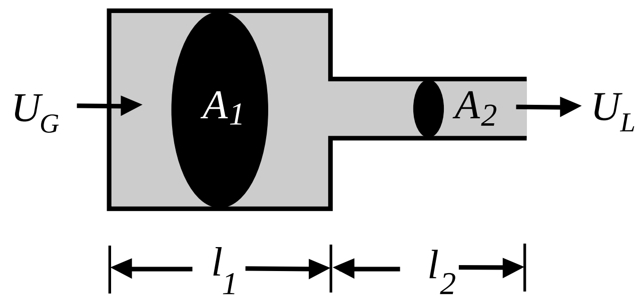
$$f_n = \frac{c}{2\ell} n \quad n = 0, 1, 2, \ldots$$

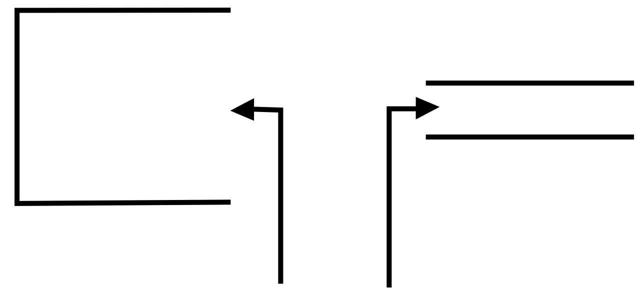
Approximating Vocal Tract Shapes



Estimating Natural Resonance Frequencies

- Resonance frequencies occur where impedance (or admittance) function equals natural (e.g., open circuit) boundary conditions





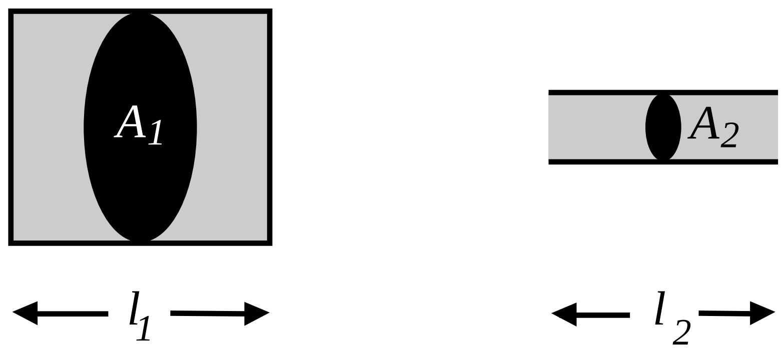
- For a two tube approximation it is easiest to solve for $Y_1 + Y_2 = 0$

$$j\frac{A_1}{\rho c}\tan\frac{\Omega\ell_1}{c} - j\frac{A_2}{\rho c}\cot\frac{\Omega\ell_2}{c} = 0$$

$$\sin\frac{\Omega\ell_1}{c}\sin\frac{\Omega\ell_2}{c} - \frac{A_2}{A_1}\cos\frac{\Omega\ell_1}{c}\cos\frac{\Omega\ell_2}{c} = 0$$

Decoupling Simple Tube Approximations

- If $A_1 \gg A_2$, or $A_1 \ll A_2$, the tubes can be decoupled and natural frequencies of each tube can be computed independently
- For the vowel $/\mathrm{i}^{y}/$, the formant frequencies are obtained from:



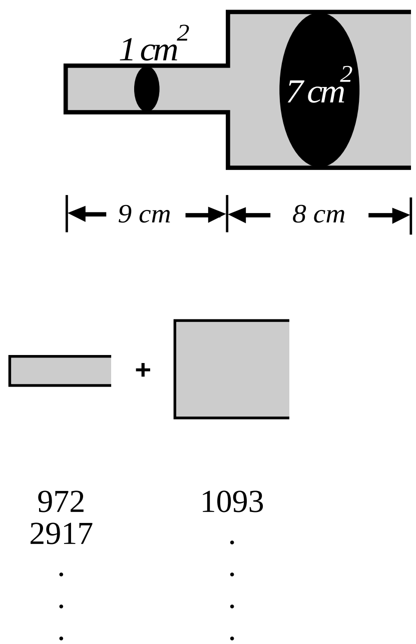
$$f_n = \frac{c}{2\ell_1} n \qquad \text{plus} \qquad f_n = \frac{c}{2\ell_2} n$$

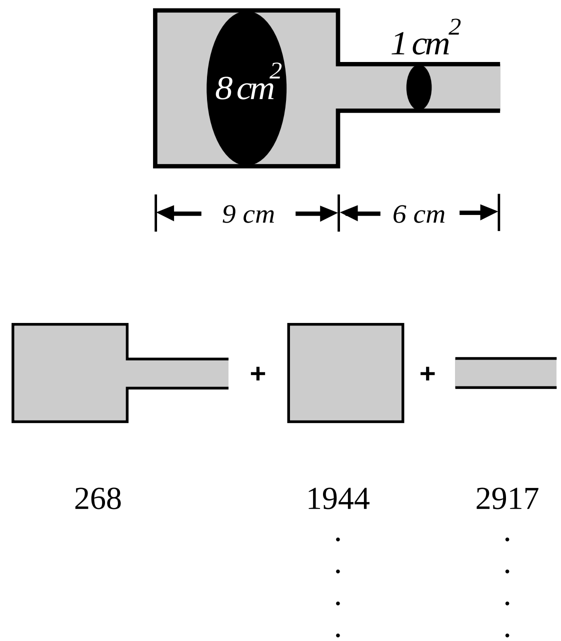
- At low frequencies:

$$f = \frac{c}{2\pi} \left[\frac{A_2}{A_1 \ell_1 \ell_2} \right]^{1/2} = \frac{1}{2\pi} \left[\frac{1}{C_{A_1} M_{A_2}} \right]^{1/2}$$

- This low resonance frequency is called the Helmholtz resonance

Vowel Production Example

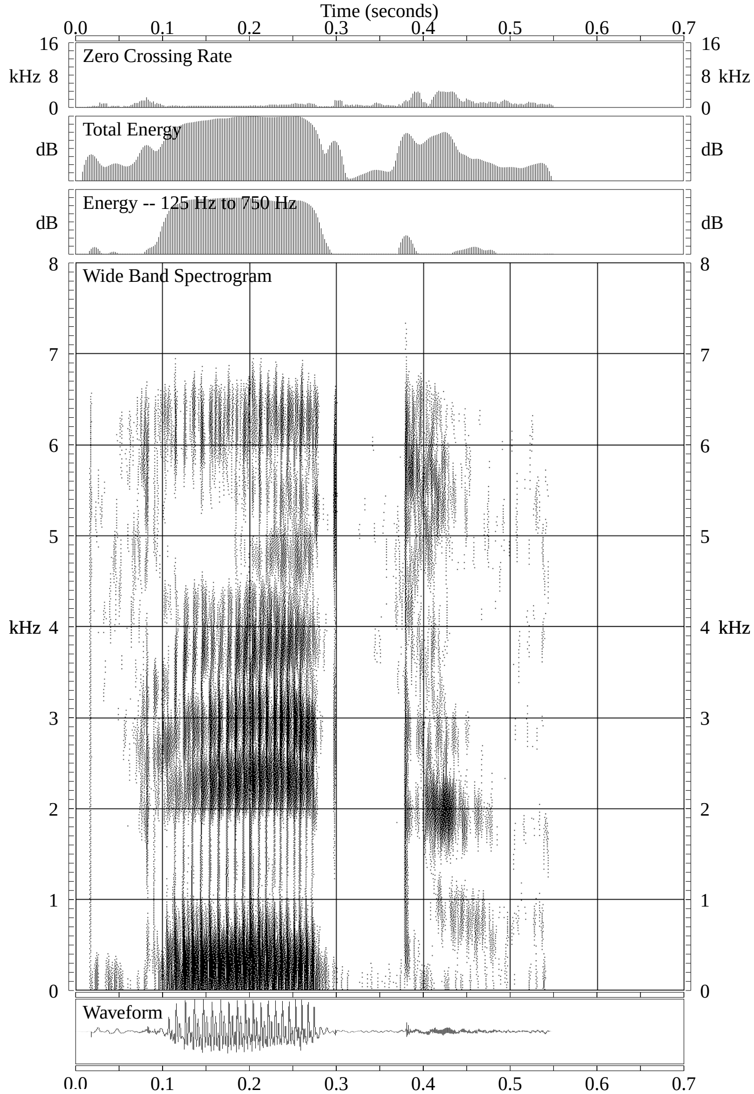




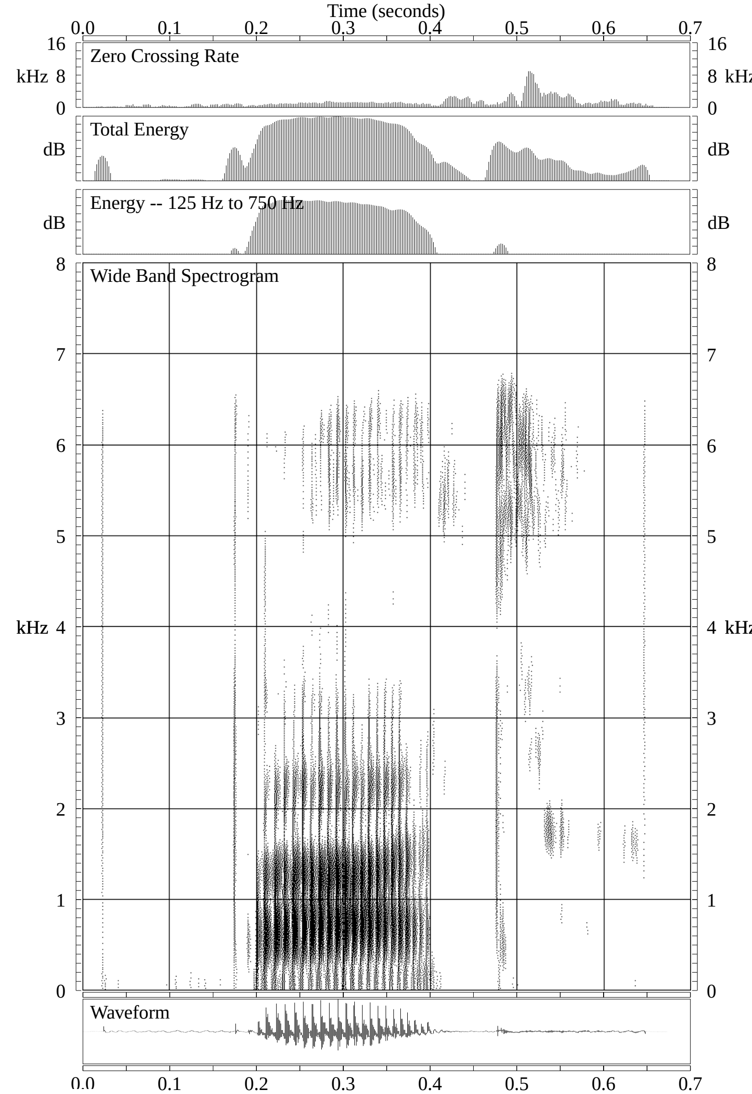
Formant	*Actual*	*Estimated*
F1	789	972
F2	1276	1093
F3	2808	2917
.	.	.
.	.	.

Formant	*Actual*	*Estimated*
F1	256	268
F2	1905	1944
F3	2917	2917
.	.	.
.	.	.

Example of Vowel Spectrograms



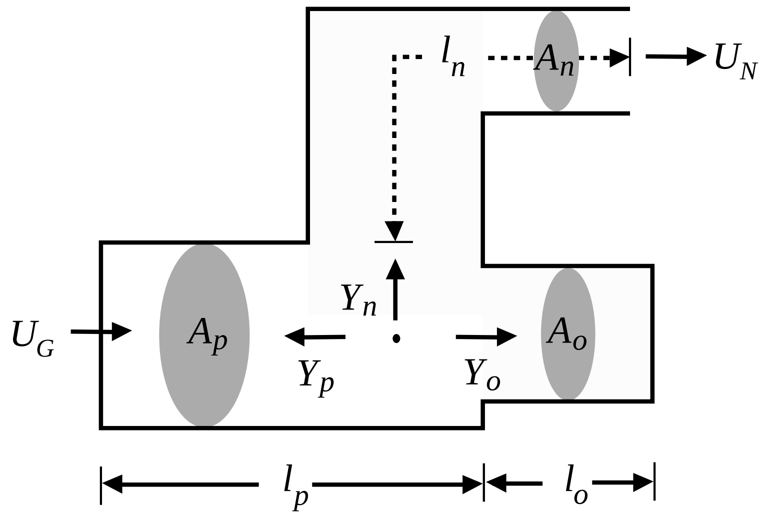
/bit/

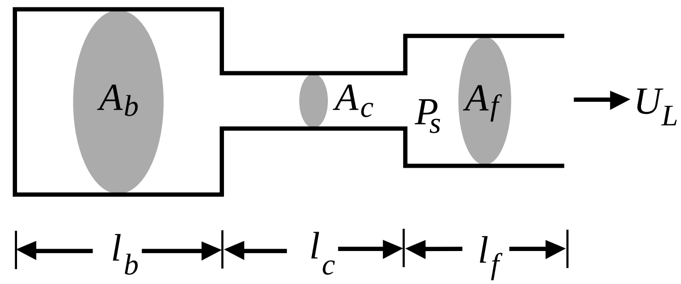


/bat/

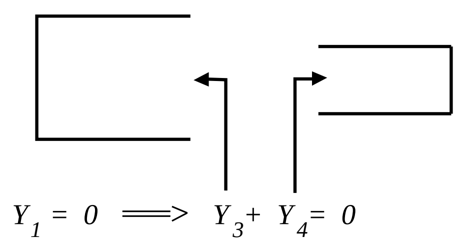
Estimating Anti-Resonance Frequencies (Zeros)

Zeros occur at frequencies where there is no measurable output



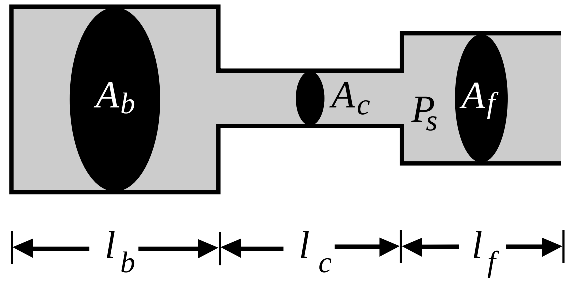


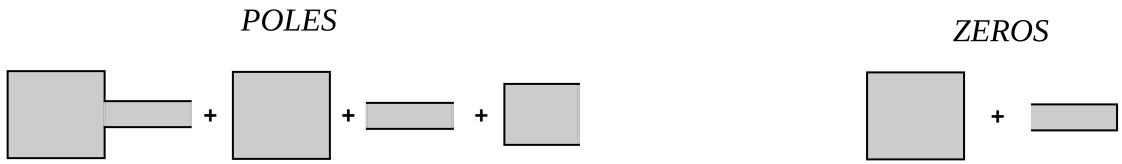
- For nasal consonants, zeros in U_N occur where $Y_O = \infty$
- For fricatives or stop consonants, zeros in U_L occur where the impedance behind source is infinite (i.e., a hard wall at source)



- Zeros occur when measurements are made in vocal tract interior

Consonant Production

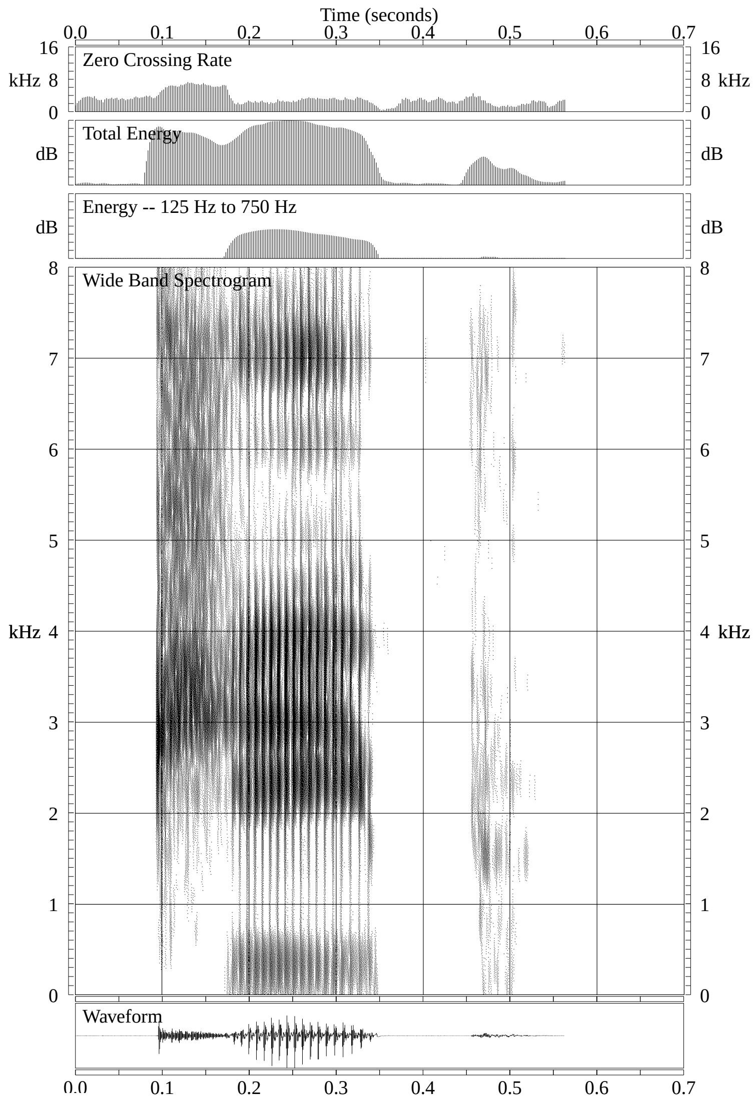




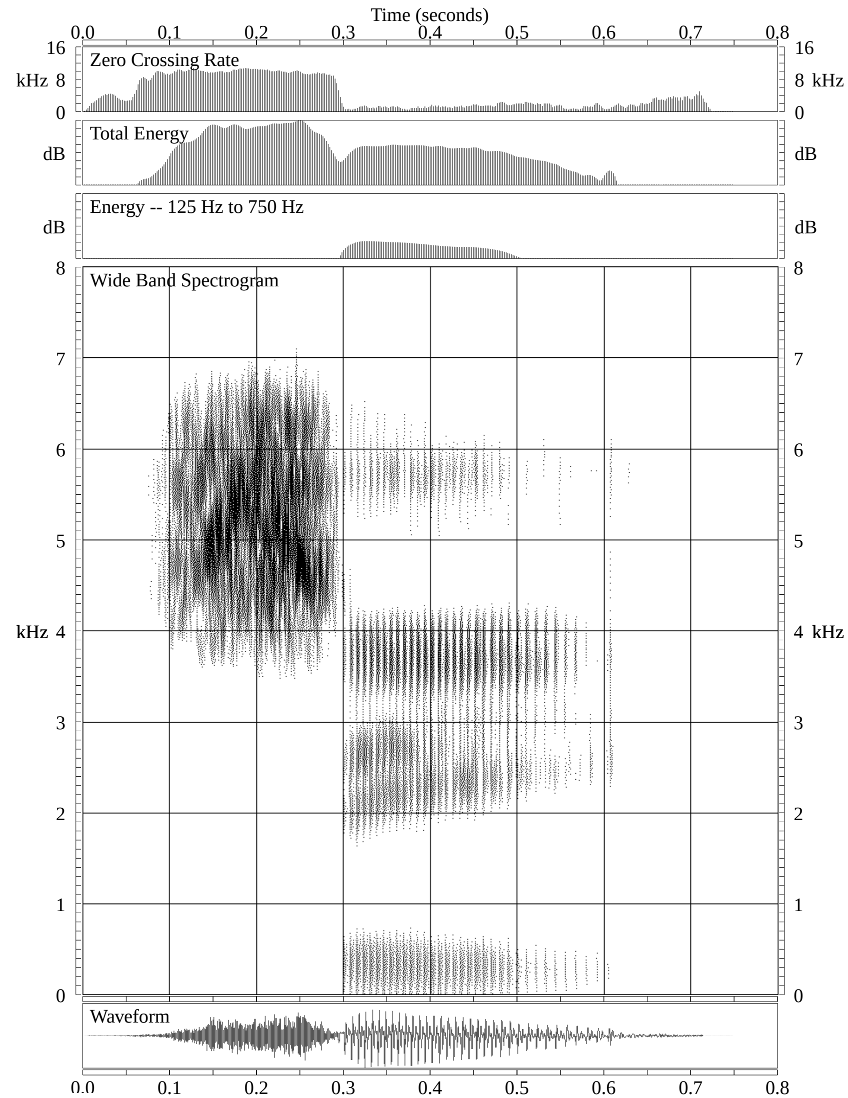
	A_b	A_c	A_f	ℓ_b	ℓ_c	ℓ_f
[g]	5	0.2	4	9	3	5
[s]	5	0.5	4	11	3	2.5

[g] *poles*	[g] *zeros*	[s] *poles*	[s] *zeros*
215	0	306	0
1750	1944	1590	1590
1944	2916	3180	2916
3888	3888	3500	3180
.	.	.	.
.	.	.	.

Example of Consonant Spectrograms

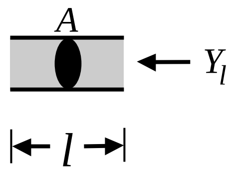


/kiʸp/



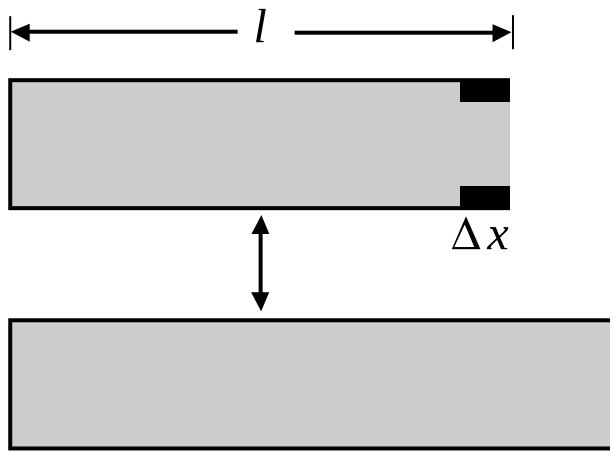
/siʸ/

Perturbation Theory



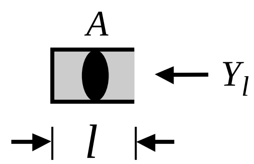
$$Y_\ell \simeq -j\frac{A}{\Omega\rho\ell} \text{ for small } \ell$$

- Consider a uniform tube, closed at one end and open at the other

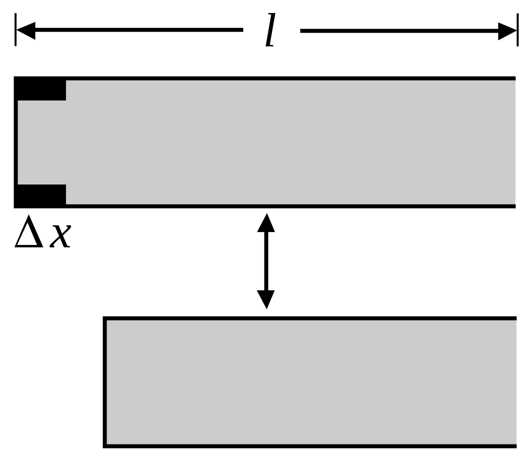


- Reducing the area of a small piece of the tube near the opening (where U is max) has the same effect as keeping the area fixed and lengthening the tube
- Since lengthening the tube lowers the resonant frequencies, narrowing the tube near points where $U(x)$ is maximum in the standing wave pattern for a given formant decreases the value of that formant

Perturbation Theory (cont'd)



$$Y_\ell \simeq j\Omega \frac{A\ell}{\rho c^2} \quad \text{for small } \ell$$



- Reducing the area of a small piece of the tube near the closure (where p is max) has the same effect as keeping the area fixed and shortening the tube
- Since shortening the tube will increase the values of the formants, narrowing the tube near points where $p(x)$ is maximum in the standing wave pattern for a given formant will increase the value of that formant

Summary of Perturbation Theory Results

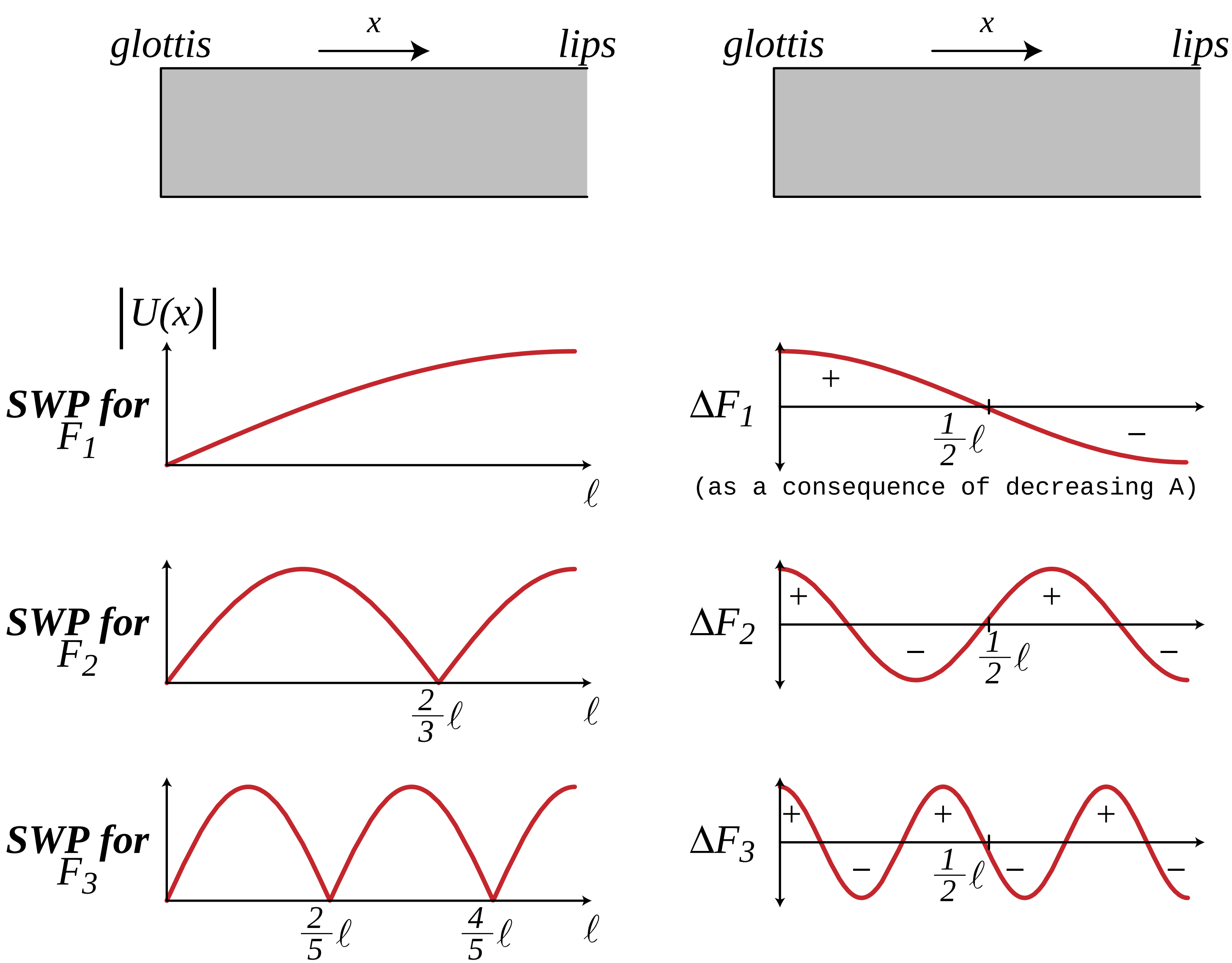


Illustration of Perturbation Theory

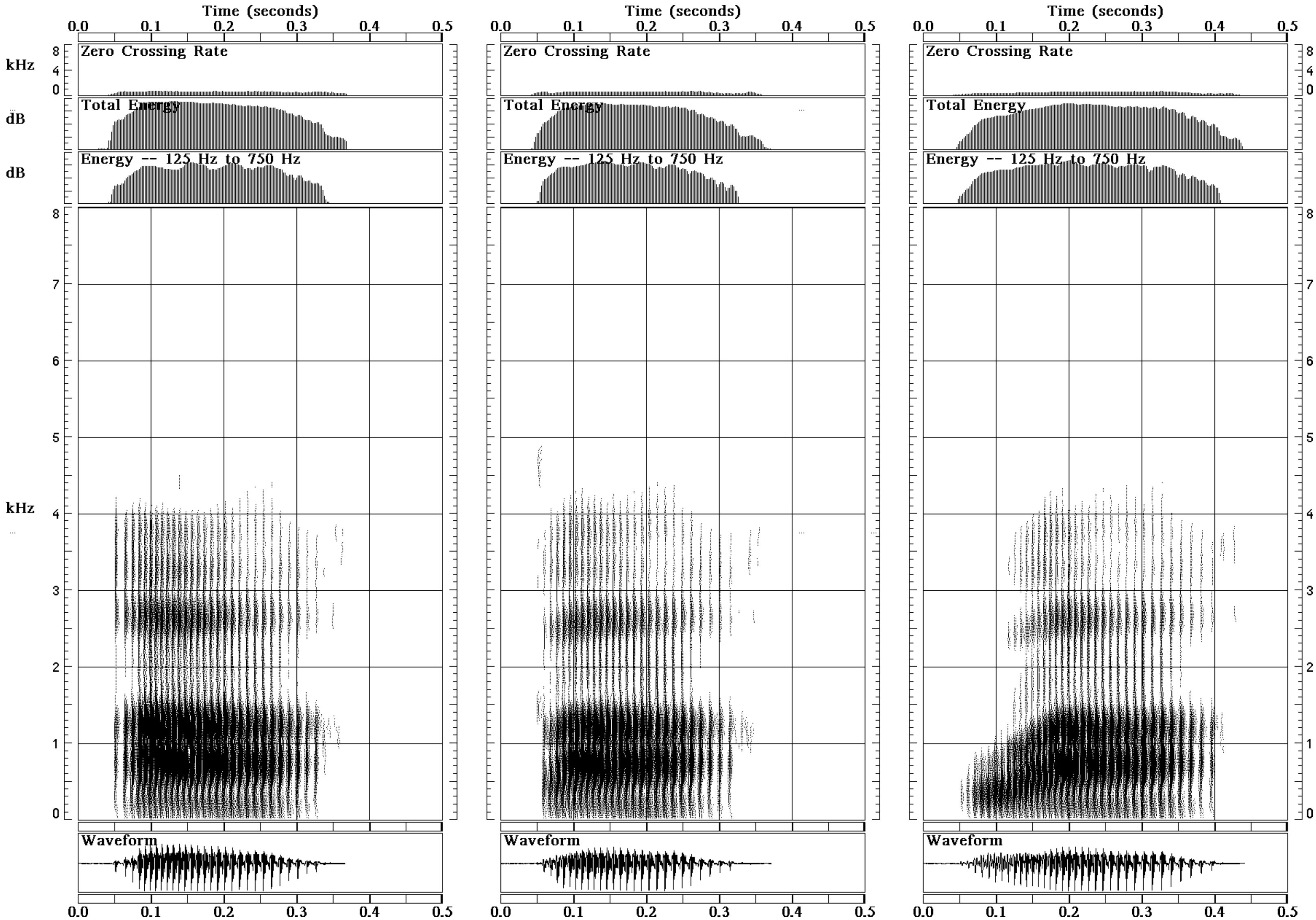
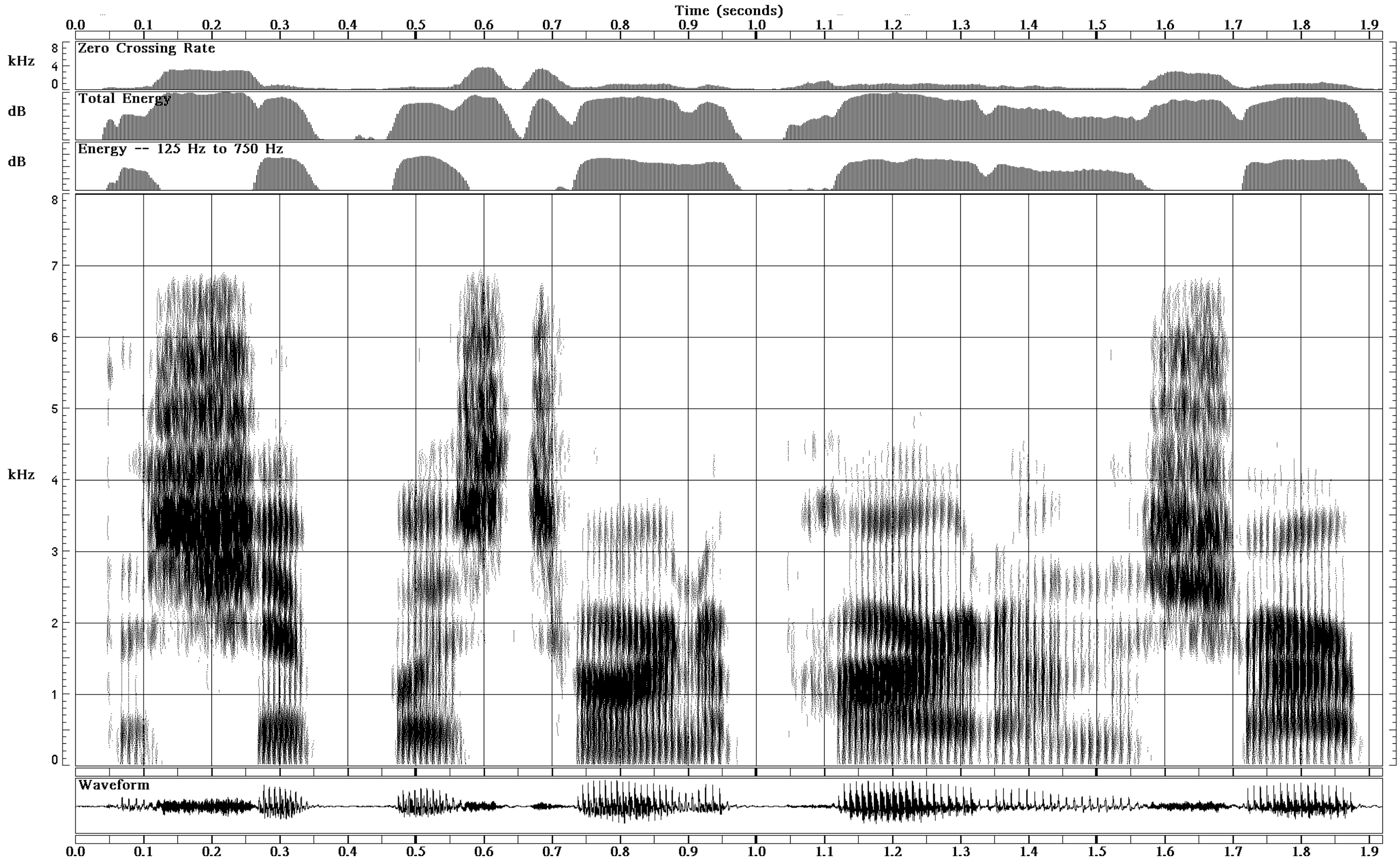
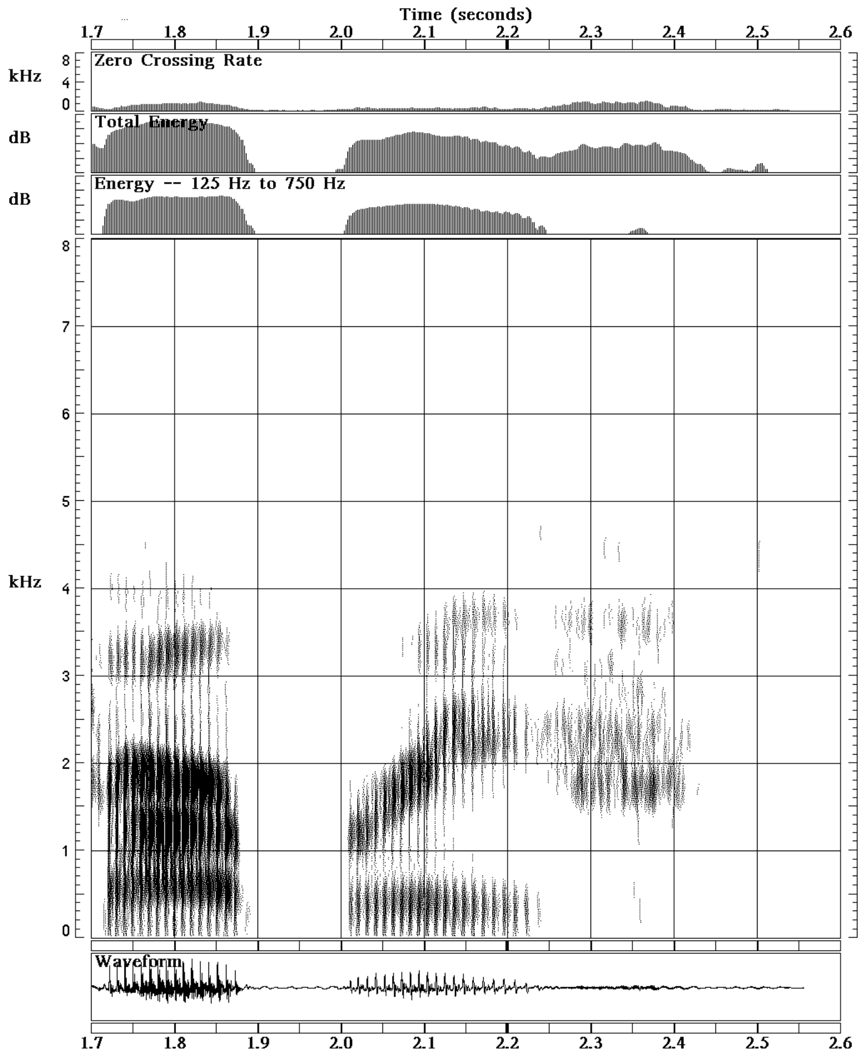


Illustration of Perturbation Theory



The ship was torn apart on the sharp (reef)

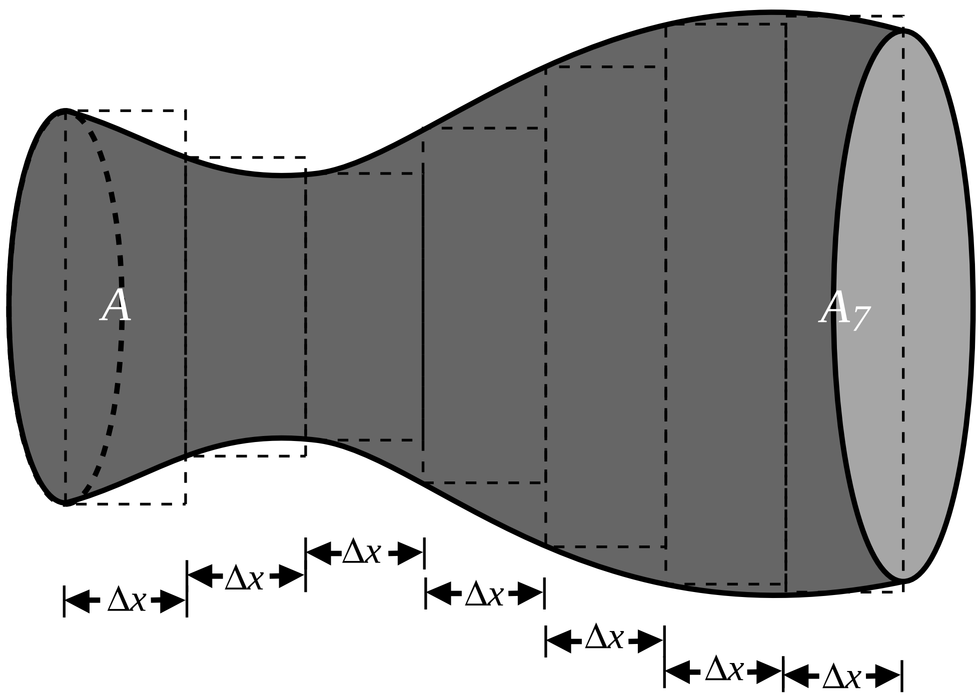
Illustration of Perturbation Theory



(The ship was torn apart on the sh)arp reef

Multi-Tube Approximation of the Vocal Tract

- We can represent the vocal tract as a concatenation of N lossless tubes with constant area $\{A_k\}$ and equal length $\Delta x = \ell / N$
- The wave propagation time through each tube is $\tau = \frac{\Delta x}{c} = \frac{\ell}{Nc}$



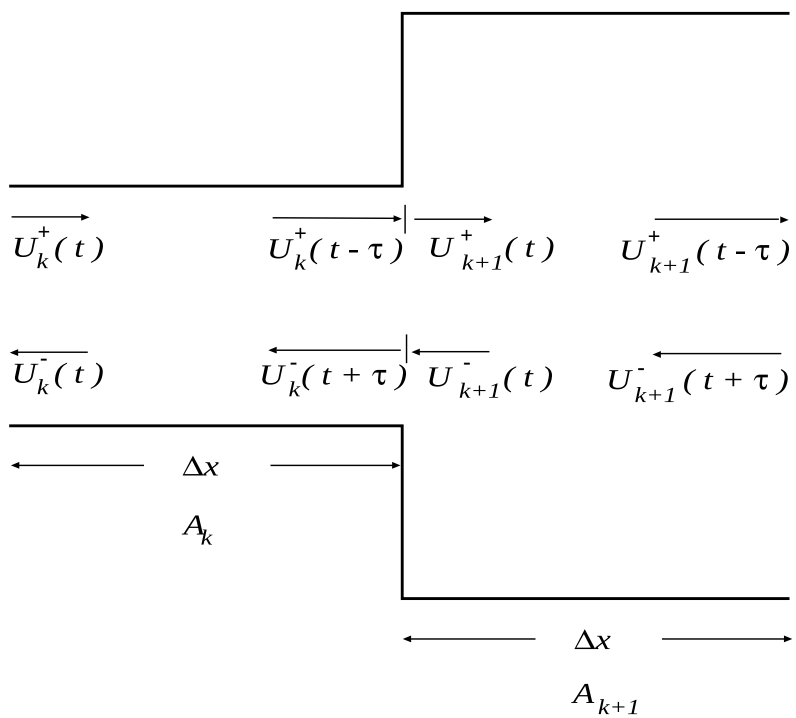
Wave Equations for Individual Tube

The wave equations for the k^{th} tube have the form

$$p_k(x,t) = \frac{\rho c}{A_k}[U_k^+(t - \frac{x}{c}) + U_k^-(t + \frac{x}{c})]$$

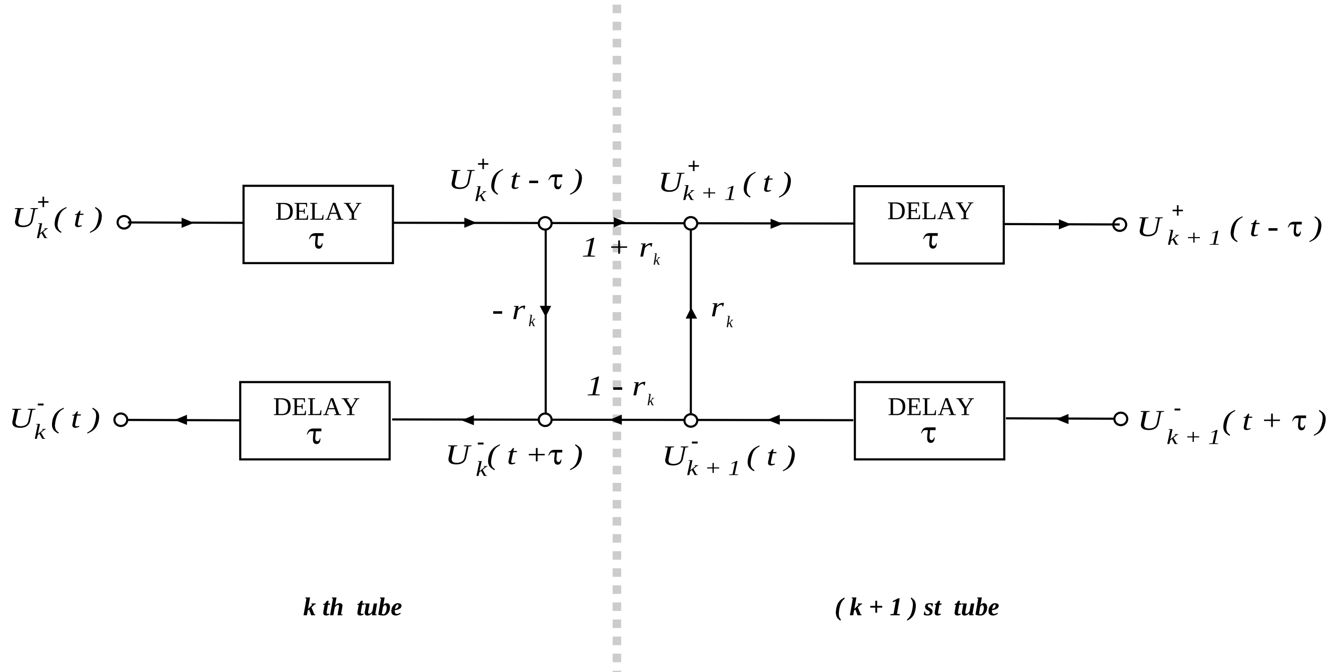
$$U_k(x,t) = U_k^+(t - \tfrac{x}{c}) - U_k^-(t + \tfrac{x}{c})$$

where x is measured from the left-hand side ($0 \le x \le \Delta x$)



Update Expression at Tube Boundaries

We can solve update expressions using continuity constraints at tube boundaries e.g., $p_k(\Delta x, t) = p_{k+1}(0, t)$, and $U_k(\Delta x, t) = U_{k+1}(0, t)$



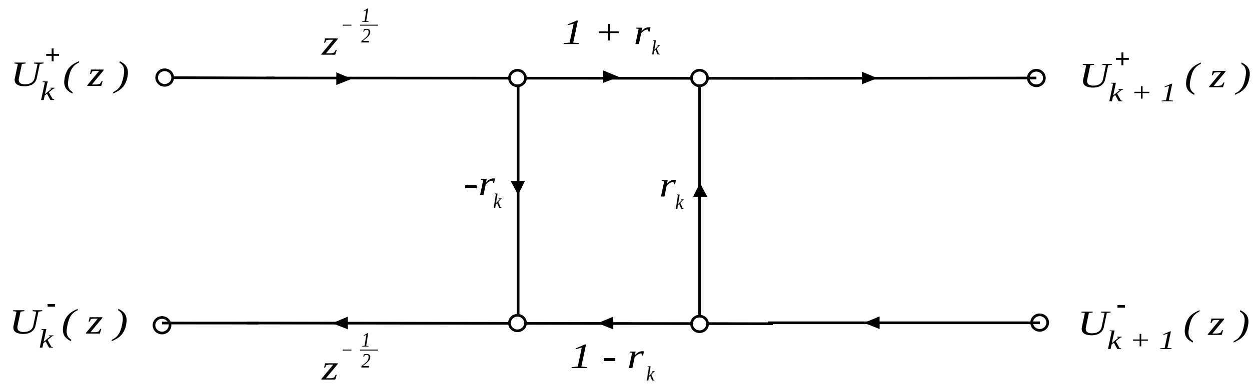
$$U_{k+1}^+(t) = (1 + r_k)U_k^+(t - \tau) + r_k U_{k+1}^-(t)$$

$$U_k^-(t + \tau) = -r_k U_k^+(t - \tau) + (1 - r_k)U_{k+1}^-(t)$$

$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k} \qquad \text{note } |r_k| \le 1$$

Digital Model of Multi-Tube Vocal Tract

- Updates at tube boundaries occur synchronously every 2τ
- If excitation is band-limited, inputs can be sampled every $T = 2\tau$
- Each tube section has a delay of $z^{-1/2}$

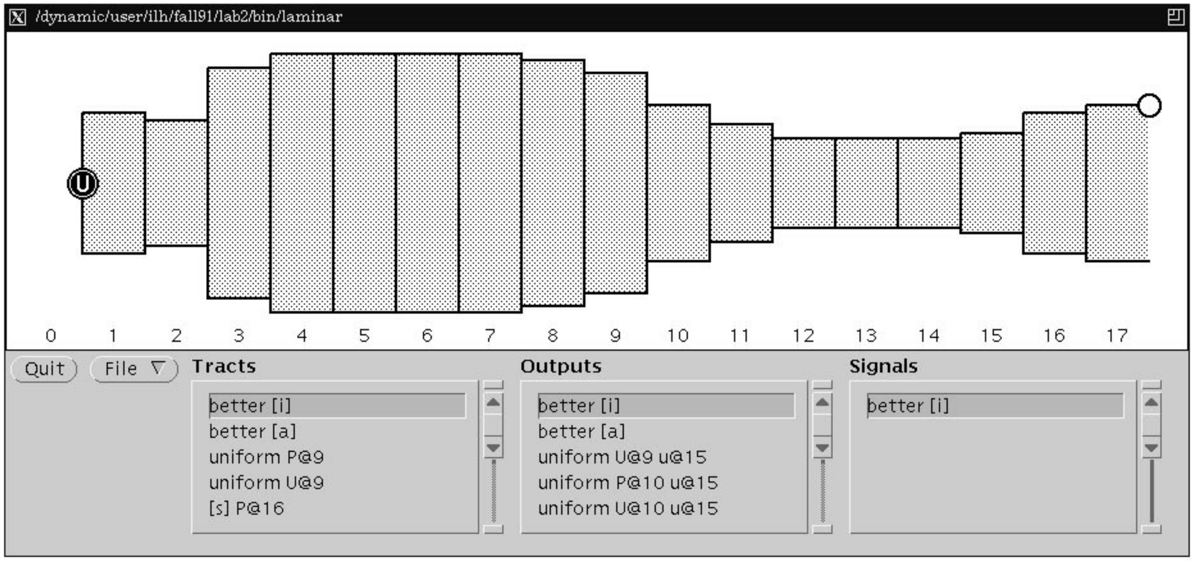


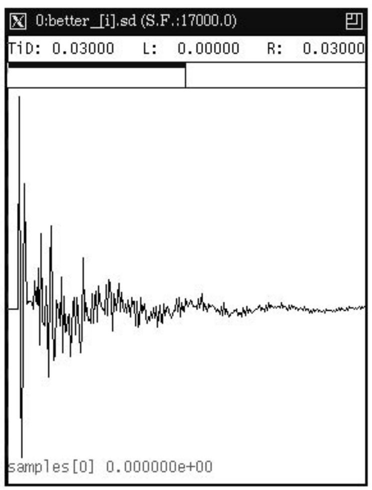
- The choice of N depends on the sampling rate T

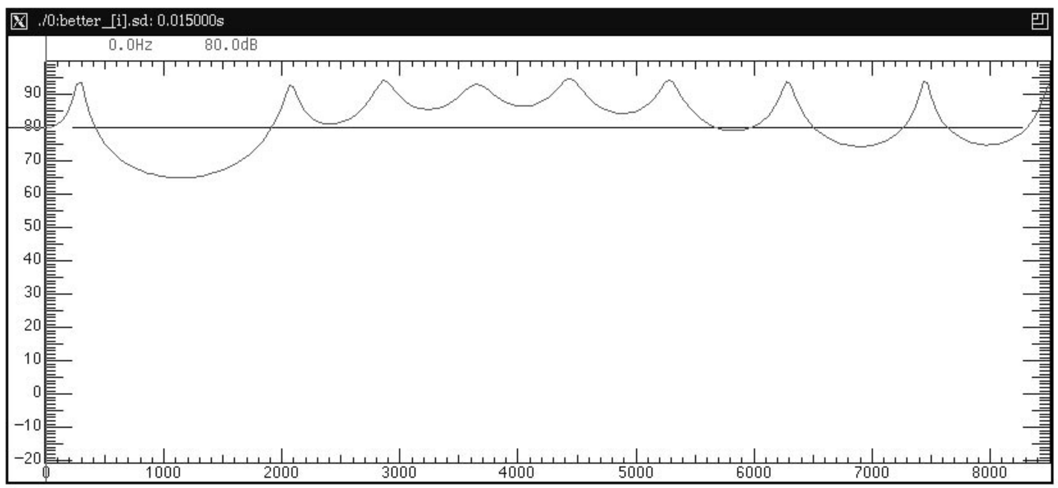
$$T = 2\tau = 2\frac{\ell}{Nc} \quad \Longrightarrow \quad N = \frac{2\ell}{cT}$$

- Series and shunt losses can also be introduced at tube junctions
  - Bandwidths are proportional to energy loss to storage ratio
  - Stored energy is proportional to tube length

Assignment 1







References

- Zue, *6.345 Course Notes*
- Stevens, *Acoustic Phonetics*, MIT Press, 1998.
- Rabiner & Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.