Return to Main

Objectives

**Physiology:**
Saggital View
Saggital X-ray
Vocal Cords
Transduction
Spectrogram

**Acoustics:**
Acoustic Theory
Wave Propagation
Helium Speech

**On-Line Resources:**
Spectrograms
Acoustics and Speech
Sound Waves in Tubes
Tube Models
Helium Speech

# LECTURE 02: A BRIEF OVERVIEW OF SPEECH PRODUCTION

- Objectives:
  - Basic speech physiology
  - Speech is a sound pressure wave
  - Transduction to an electrical signal introduces distortion
  - Acoustic analysis follows the same principles used in electromagnetic wave propagation
  - There are many ways to view a speech signal
  - Concatenated tube models (linear acoustics)

This lecture contains material from an excellent textbook on the fundamentals of speech processing:

> J. Deller, et. al., *Discrete-Time Processing of Speech Signals*, MacMillan Publishing Co., ISBN: 0-7803-5386-2, 2000.

as well as information found in the course textbook:

> X. Huang, A. Acero, and H.W. Hon, *Spoken Language Processing - A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Upper Saddle River, New Jersey, USA, ISBN: 0-13-022616-5, 2001.

Return to Main

**Introduction:**

01: Organization
(html, pdf)

**Speech Signals:**

02: Production
(html, pdf)

03: Digital Models
(html, pdf)

04: Perception
(html, pdf)

05: Masking
(html, pdf)

06: Phonetics and Phonology
(html, pdf)

07: Syntax and Semantics
(html, pdf)

**Signal Processing:**

08: Sampling
(html, pdf)

09: Resampling
(html, pdf)

10: Acoustic Transducers
(html, pdf)

11: Temporal Analysis
(html, pdf)

12: Frequency Domain Analysis
(html, pdf)

13: Cepstral Analysis
(html, pdf)

14: **Exam No. 1**
(html, pdf)

15: Linear Prediction
(html, pdf)

16: LP-Based Representations
(html, pdf)

**Parameterization:**

17: Differentiation
(html, pdf)

18: Principal Components
(html, pdf)

# ECE 8463: FUNDAMENTALS OF SPEECH RECOGNITION

Professor Joseph Picone
Department of Electrical and Computer Engineering
Mississippi State University

email: picone@isip.msstate.edu
phone/fax: 601-325-3149; office: 413 Simrall
URL: http://www.isip.msstate.edu/resources/courses/ece_8463

Modern speech understanding systems merge interdisciplinary technologies from Signal Processing, Pattern Recognition, Natural Language, and Linguistics into a unified statistical framework. These systems, which have applications in a wide range of signal processing problems, represent a revolution in Digital Signal Processing (DSP). Once a field dominated by vector-oriented processors and linear algebra-based mathematics, the current generation of DSP-based systems rely on sophisticated statistical models implemented using a complex software paradigm. Such systems are now capable of understanding continuous speech input for vocabularies of hundreds of thousands of words in operational environments.

In this course, we will explore the core components of modern statistically-based speech recognition systems. We will view speech recognition problem in terms of three tasks: signal modeling, network searching, and language understanding. We will conclude our discussion with an overview of state-of-the-art systems, and a review of available resources to support further research and technology development.

Tar files containing a compilation of all the notes are available. However, these files are large and will require a substantial amount of time to download. A tar file of the html version of the notes is available here. These were generated using wget:

```
wget -np -k -m http://www.isip.msstate.edu/publications/courses/ece_8463/lectures/current
```

A pdf file containing the entire set of lecture notes is available here. These were generated using Adobe Acrobat.

Questions or comments about the material presented here can be directed to help@isip.msstate.edu.

19: Linear Discriminant Analysis (html, pdf)

# LECTURE 02: A BRIEF OVERVIEW OF SPEECH PRODUCTION

- Objectives:
  - Basic speech physiology
  - Speech is a sound pressure wave
  - Transduction to an electrical signal introduces distortion
  - Acoustic analysis follows the same principles used in electromagnetic wave propagation
  - There are many ways to view a speech signal
  - Concatenated tube models (linear acoustics)

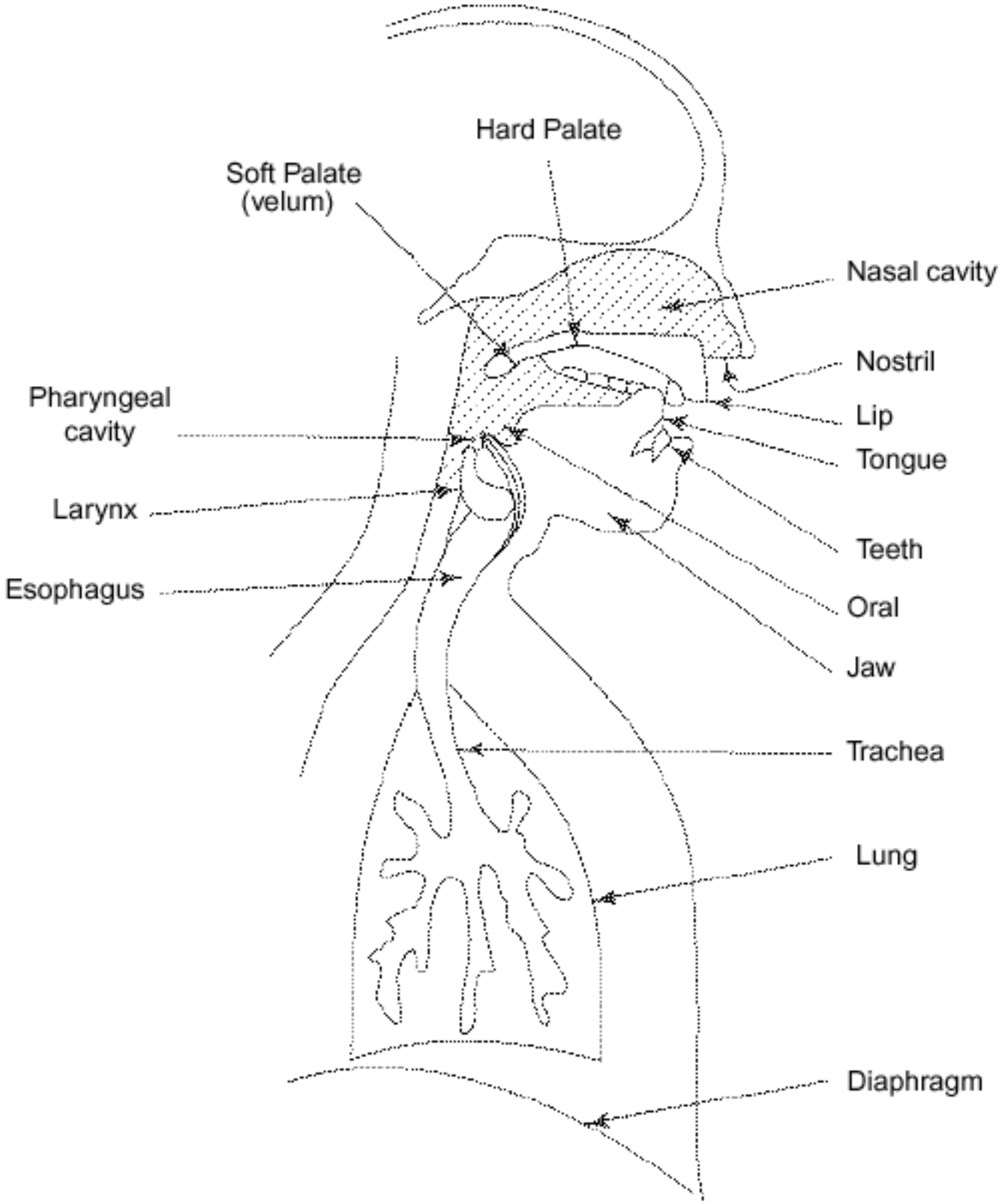
This lecture contains material from an excellent textbook on the fundamentals of speech processing:

> J. Deller, et. al., *Discrete-Time Processing of Speech Signals*, MacMillan Publishing Co., ISBN: 0-7803-5386-2, 2000.

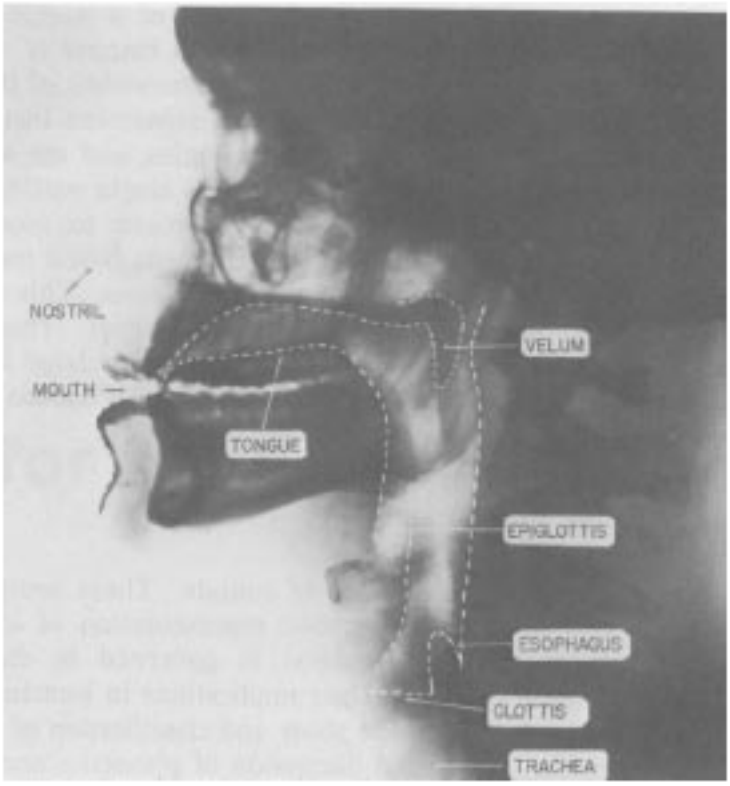
as well as information found in the course textbook:

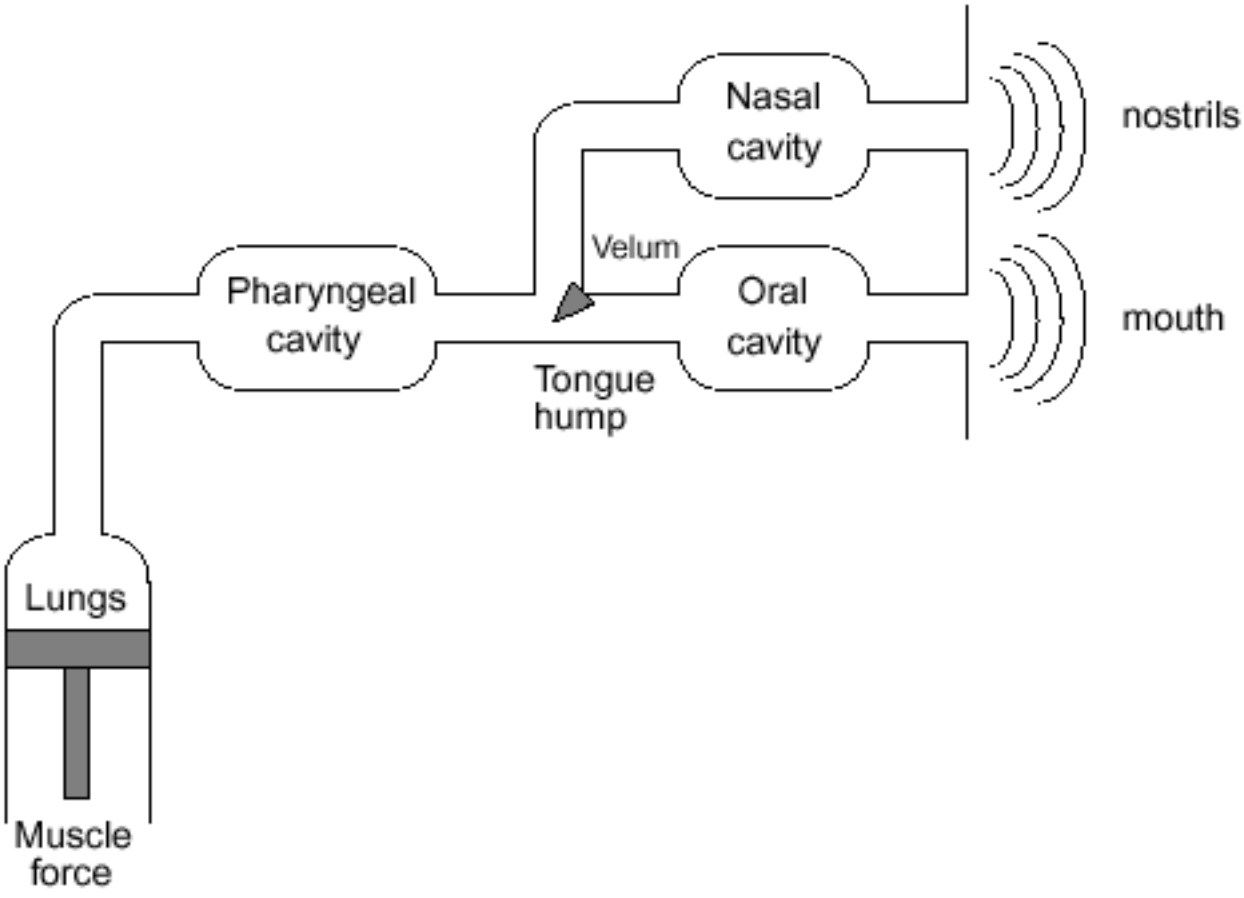
> X. Huang, A. Acero, and H.W. Hon, *Spoken Language Processing - A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Upper Saddle River, New Jersey, USA, ISBN: 0-13-022616-5, 2001.

**SAGGITAL PLANE VIEW**
**OF THE HUMAN VOCAL APPARATUS**

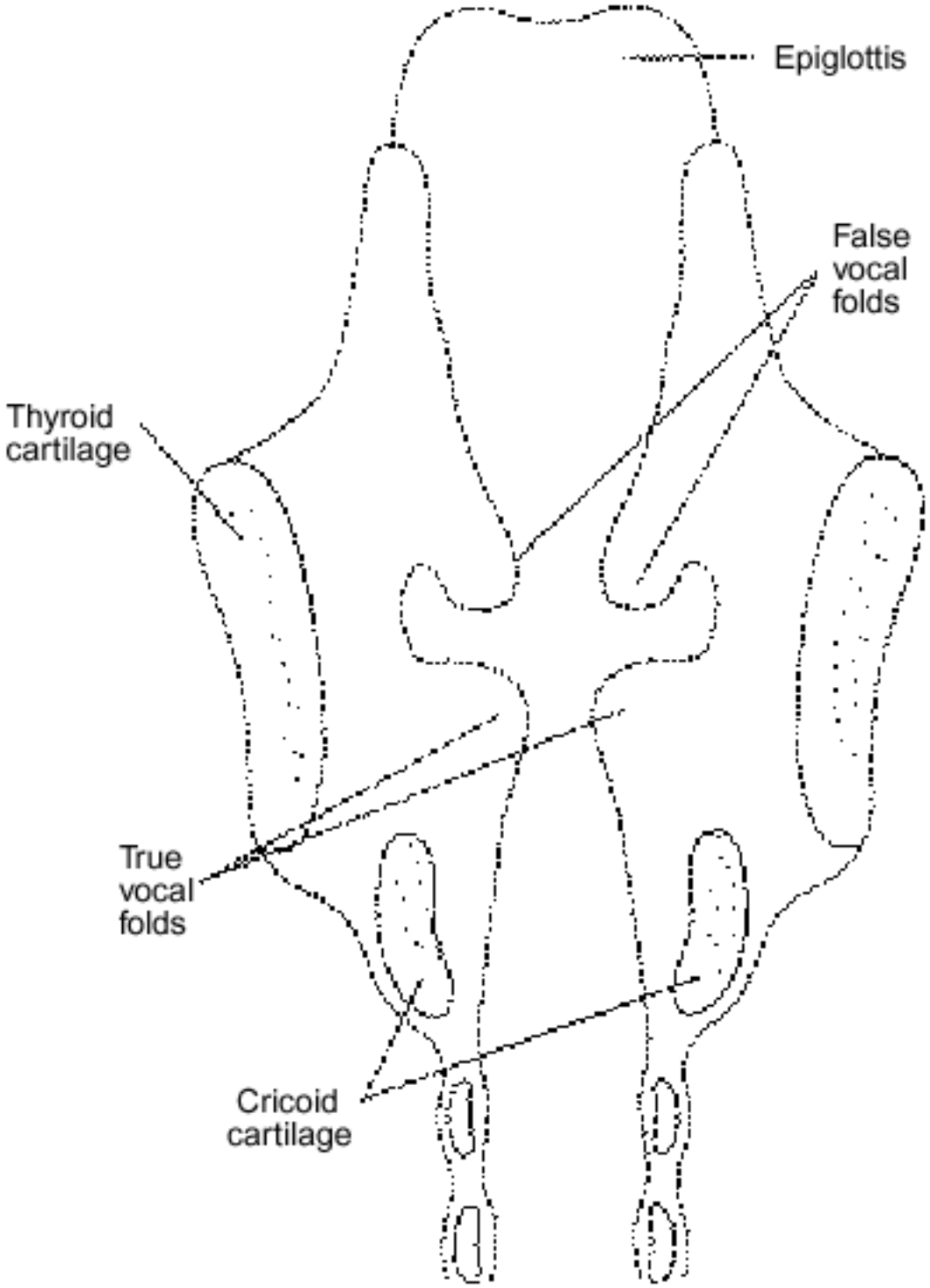


**SAGGITAL X-RAY OF**
**THE HUMAN VOCAL APPARATUS**



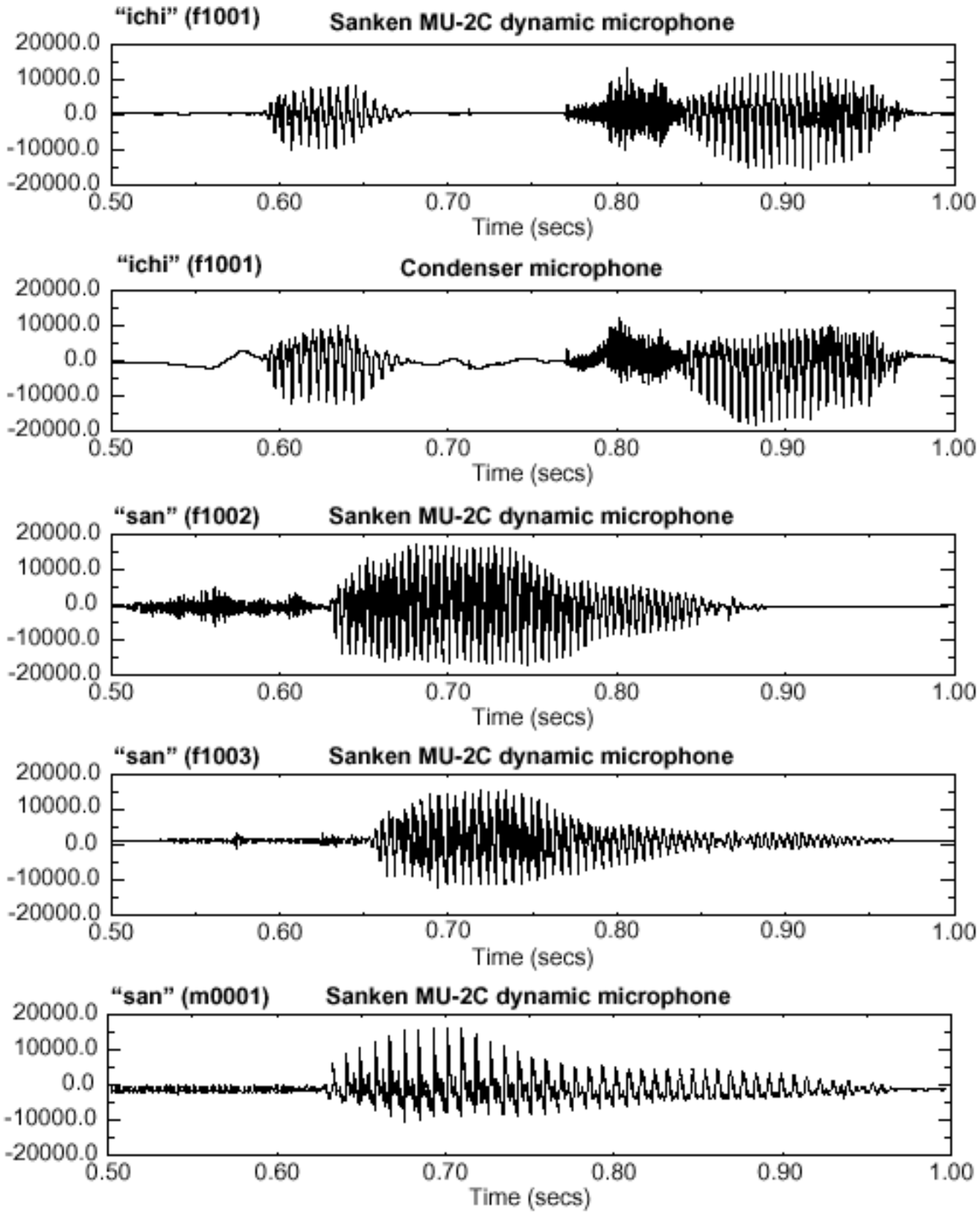



**VOCAL CORDS - SOURCE OF EXCITATION**



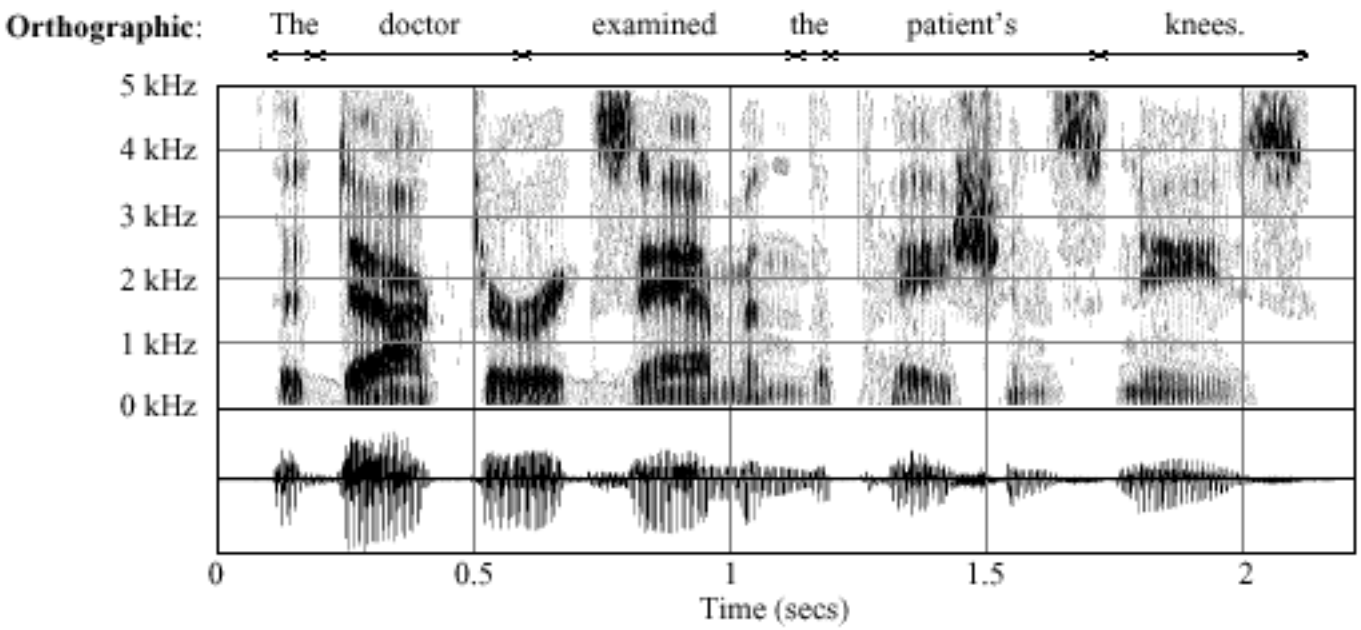
**TRANSDUCTION**

- Speech is a sound pressure wave that must be converted to an electrical signal, and then a digital signal, to be processed. This conversion process introduces distortion (frequency response, nonlinear dynamics, etc.).

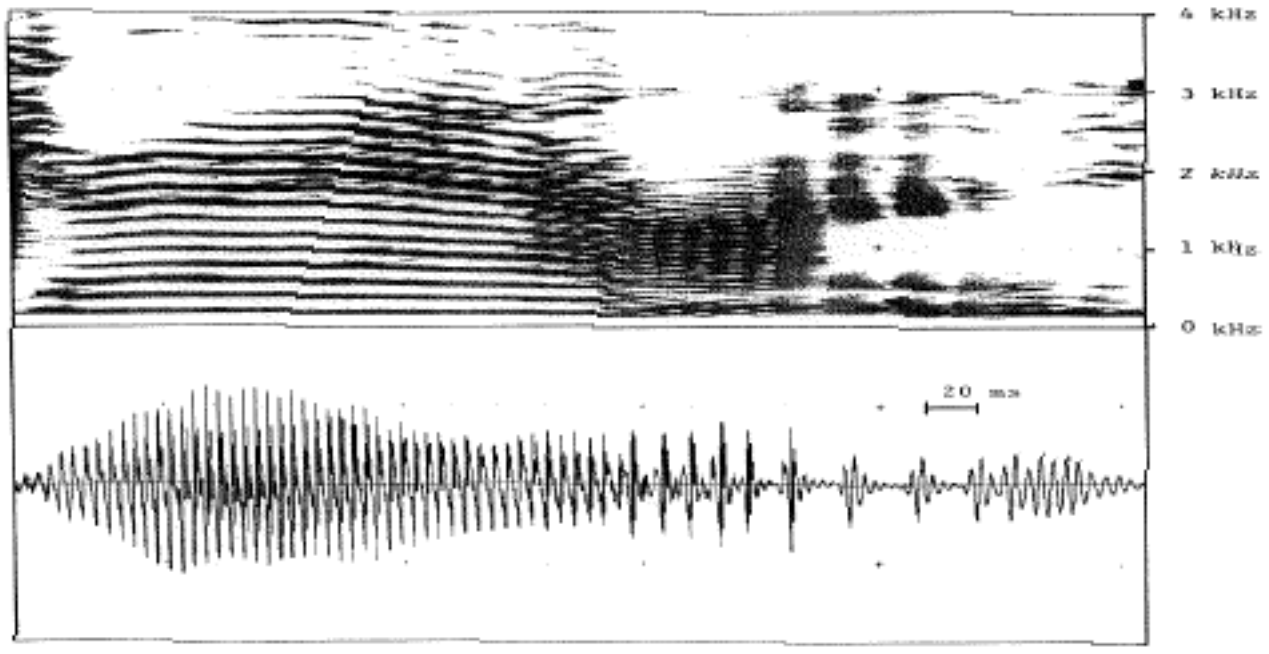


**WHAT DOES A SPEECH SIGNAL LOOK LIKE?**

Standard wideband spectrogram ($f_s = 10\ kHz$, $T_w = 6\ ms$):

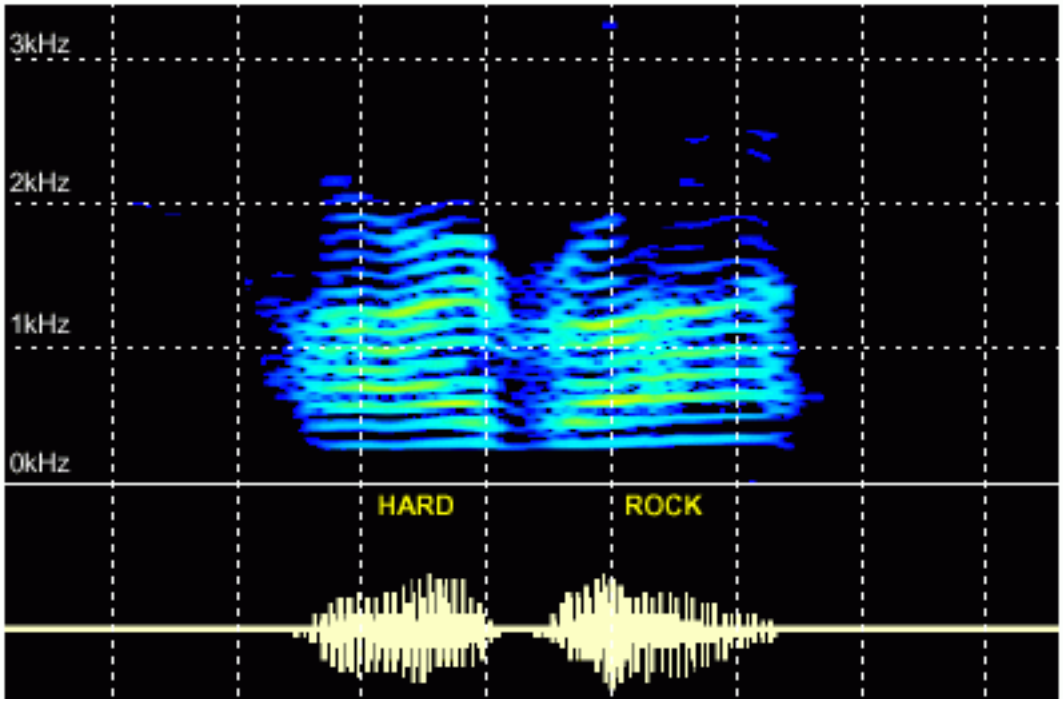



Narrowband Spectrogram ($f_s = 8\ kHz$, $T_w = 30\ ms$):

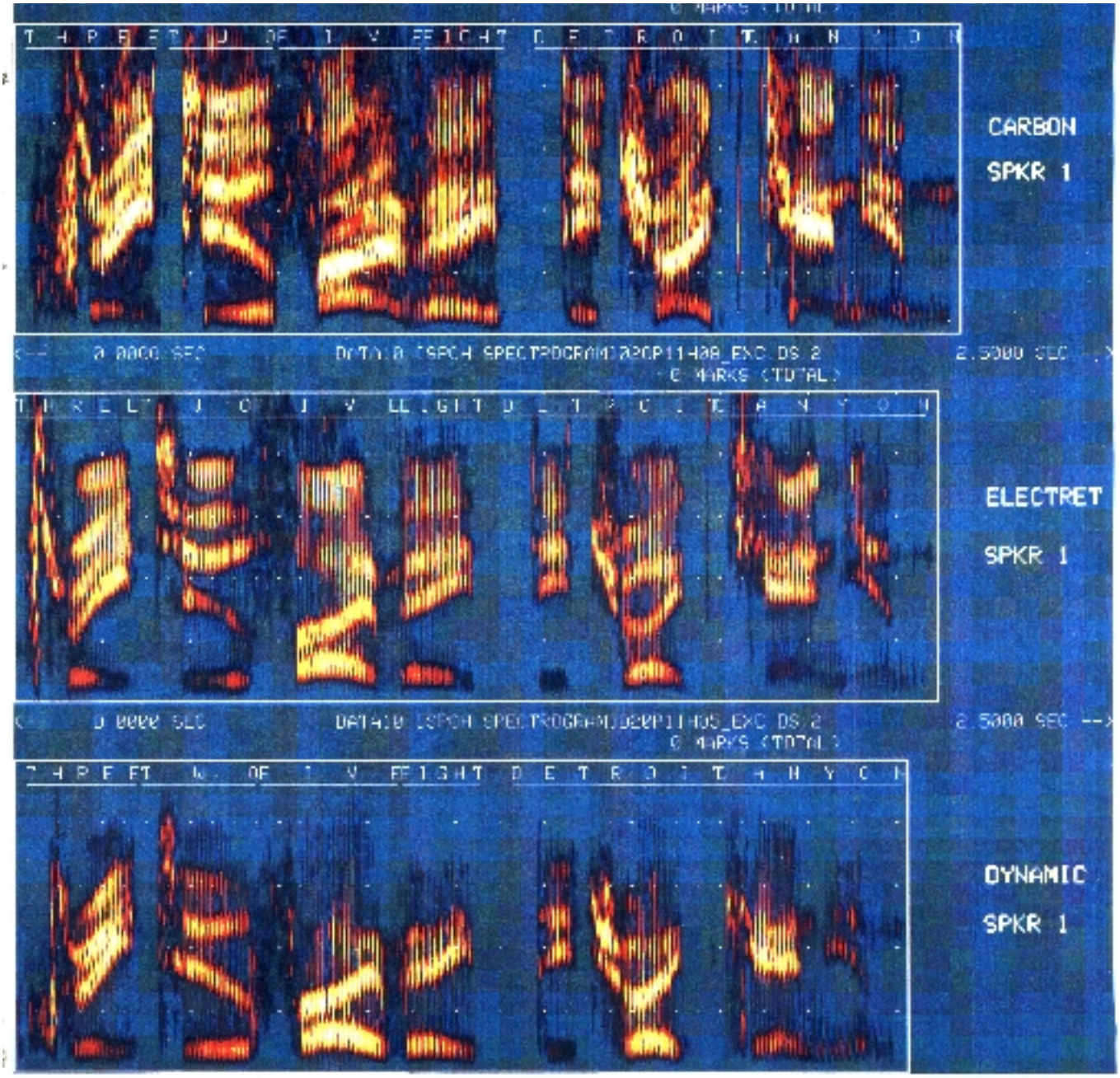


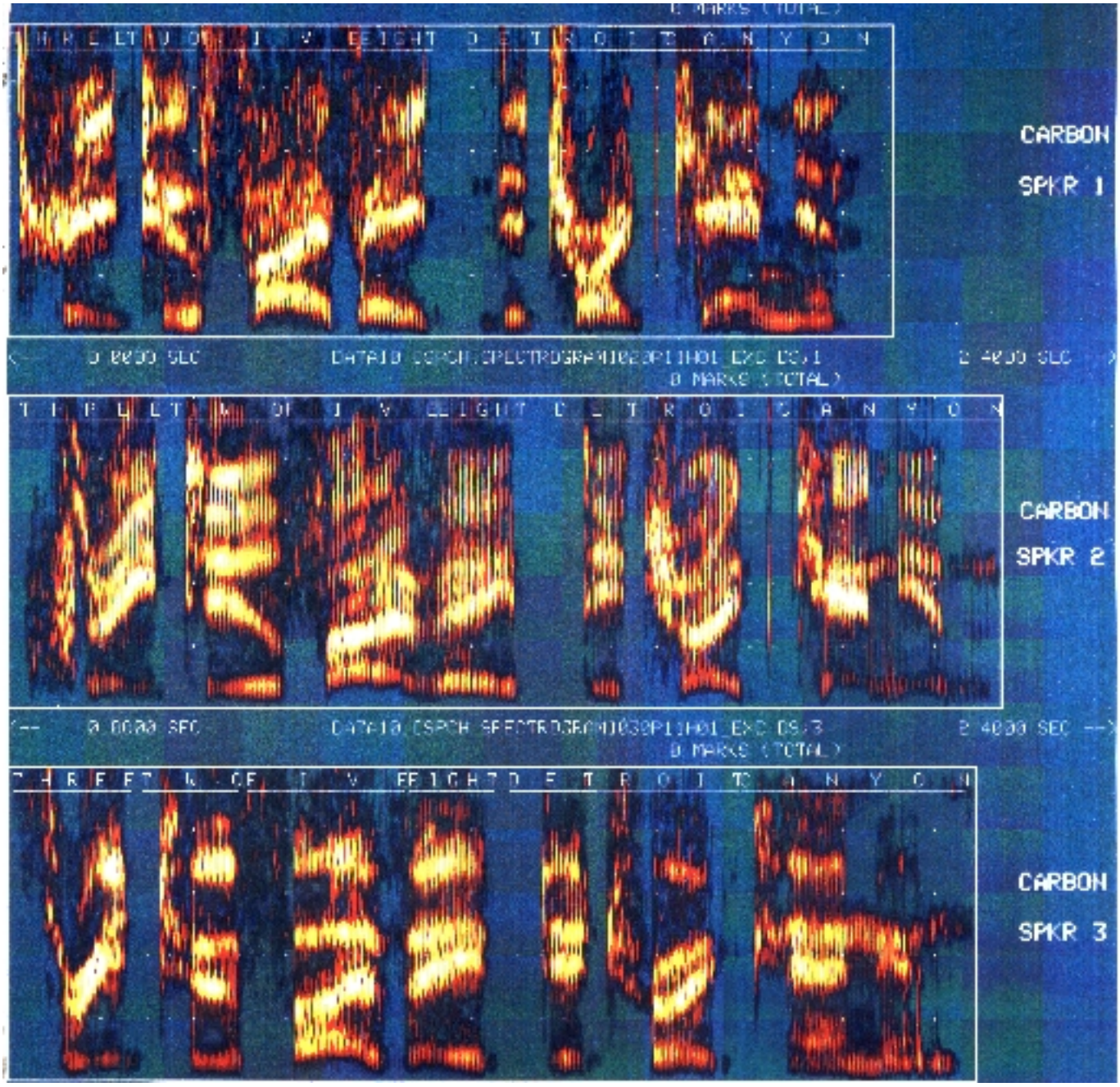

"Drown" (female)

- We often prefer to view a spectrogram using a color visualization in which spectral log magnitude is mapped to "temperature" (the color that emanates from a steel bar when it is heated):



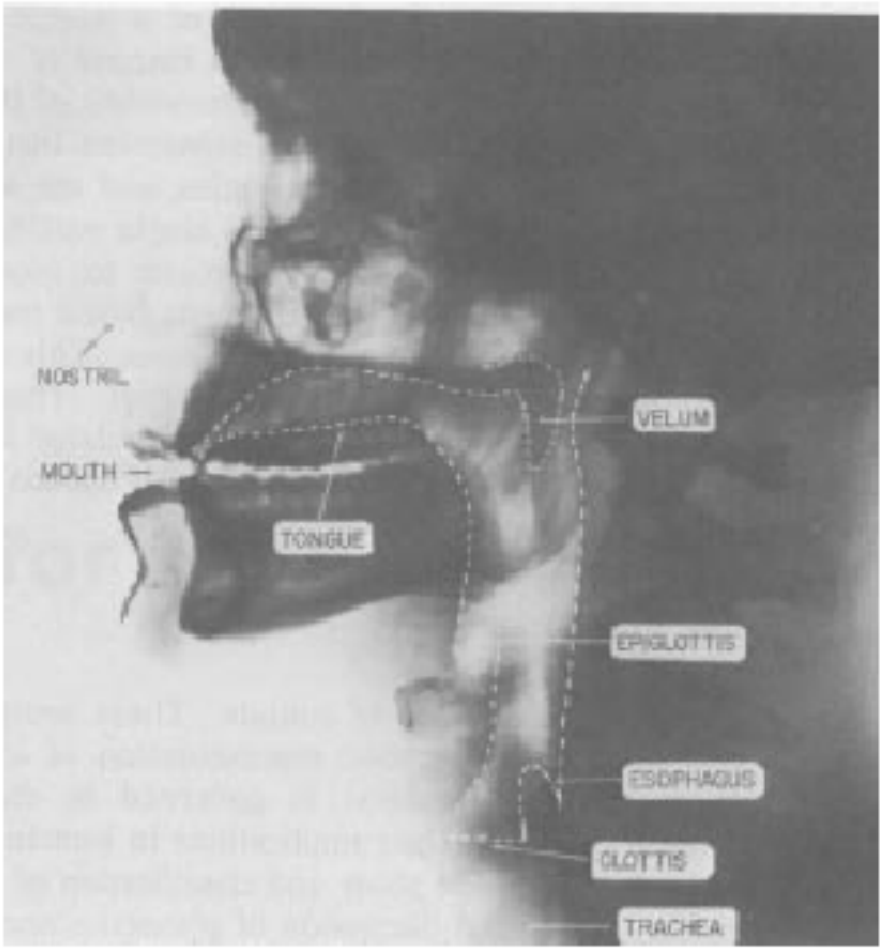

- Here are more examples of color spectrograms using the ever-popular Texas Instruments color map:


CARBON
SPKR 1
ELECTRET
SPKR 1
DYNAMIC
SPKR 1

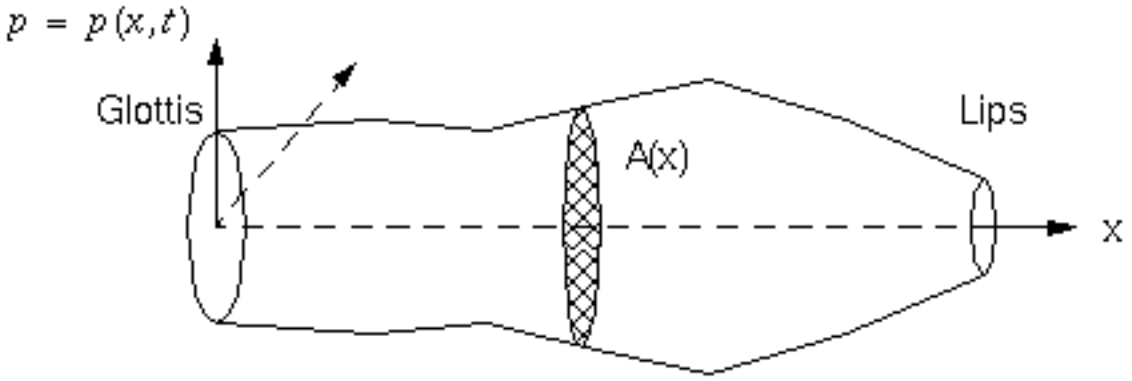

CARBON
SPKR 1
CARBON
SPKR 2
CARBON
SPKR 3

A detailed acoustic theory must consider the effects of the following:

- Time variation of the vocal tract shape
- Losses due to heat conduction and viscous friction at the vocal tract walls
- Softness of the vocal tract walls
- Radiation of sound at the lips
- Nasal coupling
- Excitation of sound in the vocal tract




- Let us consider a simple case of a lossless tube:


$p = p(x,t)$
Glottis
A(x)
Lips
x

**WAVE PROPAGATION**

For frequencies that are long compared to the dimensions of the vocal tract (less than about 4000 Hz, which implies a wavelength of 8.5 cm), sound waves satisfy the following pair of equations:

$$\rho \frac{\partial (u/A)}{\partial t} + grad\ p = 0 \qquad \text{or} \qquad -\frac{\partial p}{\partial x} = \rho \frac{\partial (u/A)}{\partial t}$$

$$\frac{1}{\rho c^2}\frac{\partial p}{\partial t} + \frac{\partial A}{\partial t} + div\ u = 0 \qquad -\frac{\partial u}{\partial x} = \frac{1}{\rho c^2}\frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t}$$

where

| | |
|---|---|
| $p = p(x, t)$ | is the variation of the sound pressure in the tube |
| $u = u(x, t)$ | is the variation in the volume velocity |
| $\rho$ | is the density of air in the tube (1.2 mg/cc) |
| $c$ | is the velocity of sound (35000 cm/s) |
| $A = A(x, t)$ | is the area function (about 17.5 cm long) |

**HELIUM SPEECH:**
**RELATIONSHIP BETWEEN FREQUENCY AND DENSITY**

- Deep-sea diving to depths exceeding about 140 feet of sea water requires the use of heliox (a mixture of helium and oxygen) as a breathing gas, rather than compressed air.
- Heliox eliminates the danger of nitrogen narcosis and reduces the risk of decompression sickness which would otherwise be present.
- Heliox presents another risk. The diver's speech is rendered unintelligible because the higher velocity of sound in the diver's vocal tract shifts the frequency components of the diver's speech to much higher frequencies - an effect that has been likened to the "Donald Duck" voice.
- Heliox is less dense than air or pure oxygen. Hence, the speed of sound is greater, so the resonances occur at higher frequencies.
- The excitation remains largely unchanged since flesh in your vocal folds still vibrates at the same frequency, so the harmonics occur at the same frequency. (There could be a small change because the less dense Helium loads the vocal folds a bit less than the air, but this effect is slight.)
- Examples of helium speech are always fun to listen to.
- Descramblers are available that will perform real-time spectral shifting.
- Such systems use real-time spectral shifting.

The information on this page comes from two sources:

K. Bryden and J. Hothi
Communications Research Centre
3701 Carling Avenue
P.O. Box 11490, Stn. H
Ottawa, ON K2H 8S2
Tel: (613) 998-2515
Fax: (613) 990-7987
Email: karen.bryden@crc.ca
URL: http://www.crc.ca/en/html/crc/tech_transfer/10085

and,

J. Wolfe
School of Physics
The University of New South Wales
SYDNEY 2052
Australia
Tel: 61 2 9385 4954
Fax: 61 2 9385 6060
Email: J.Wolfe@unsw.edu.au

URL: http://www.phys.unsw.edu.au/STAFF/ACADEMIC/wolfe.html

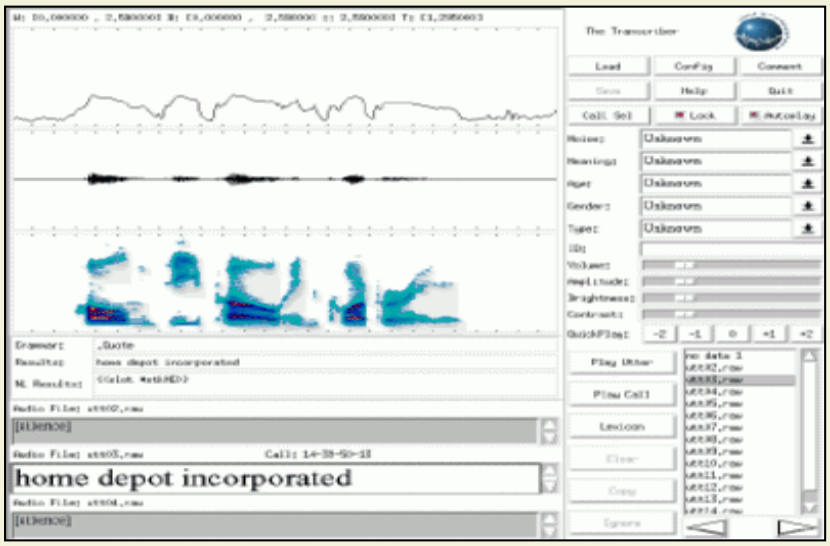
Work on real-time frequency scaling can be found in several journals including the *IEEE Transactions on Speech and Audio Processing* (formerly *Acoustics, Speech, and Signal Processsing*), and the *Journal of the Acoustical Society of America*.

# Transcriber

This is a graphical user interface tool for speech segmentation and speech transcription. The tool provides spectrograms and energy plots, speech selection, and audio playback capabilities. The tool is a single channel version which is specifically designed for quick access to multiple files from a single speaker (mono). It is written partly in object-oriented C++ (using GNU's gcc compiler) which is interfaced to Tcl-Tk (v8.0) utilities.

To download the current version of the Transcriber tool click here. Also please feel free to send us comments or suggestions regarding the tool.

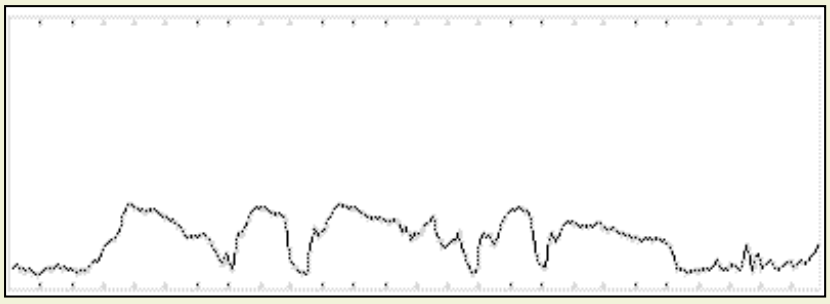
**The interface:**




The main features include

- displays signal waveforms
- displays spectrograms
- displays energy plots
- zoom in/out on the displays
- ability to set time marks
- automatic completion of words listed in a lexicon file
- ability to set attributes of each individual transcription
- ability to quickly switch between audio files from a list
- plays audio on mono channels
- modify/enter transcriptions
- set user-defined configurations

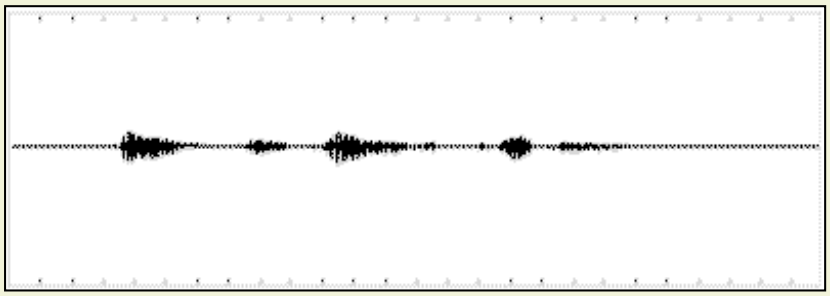
**The energy plot display:**



The energy plot features include

- zoom in/out on the energy plot
- ability to set the canvas size
- ability to change the amplitude
- ability to set the frame length
- ability to set the window length
- ability to set the RMS scale factor
- ability to set the preemphasis coefficient
- ability to set the window function

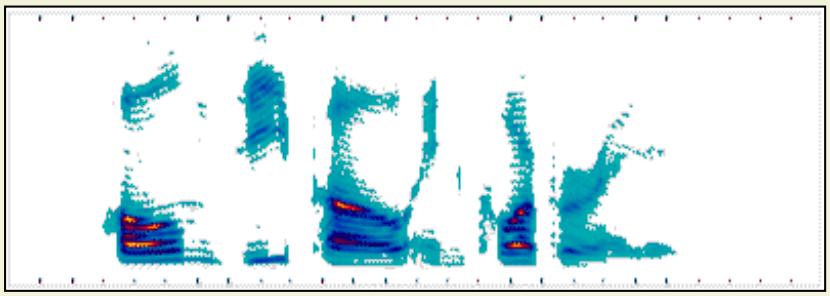
**The signal plot display:**



The signal plot features include

- zoom in/out on the signal plot
- ability to change the amplitude
- ability to change the volume
- ability to set the audio device and server

**The spectrogram display:**



The spectrogram features include

- zoom in/out on the spectrogram
- ability to change brightness
- ability to change the contrast
- ability to set the preemphasis coefficient
- ability to set the window function

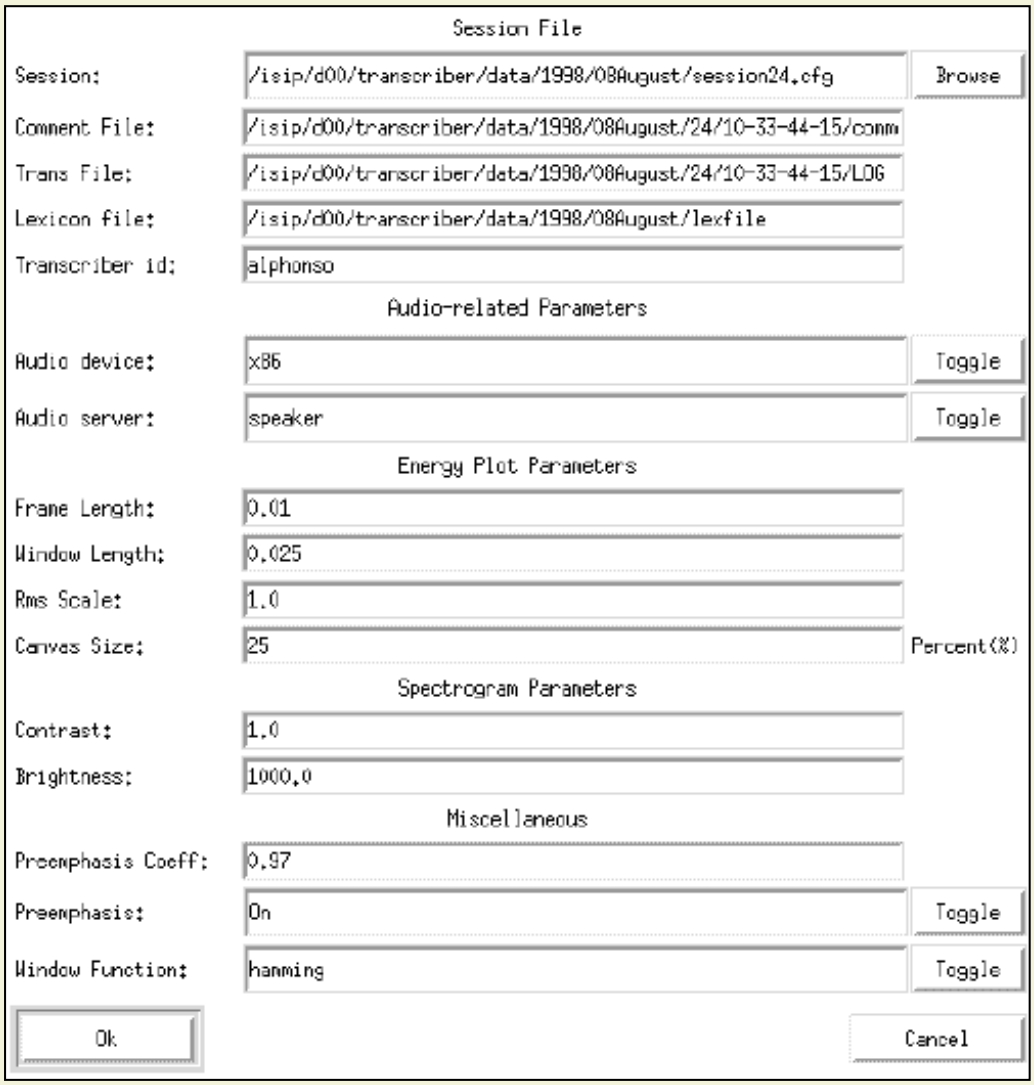
The Transcriber tool currently supports only 16 bit single channel linear data (RAW). In order to use the Transcriber with other types of data you will need to use the NIST SPHERE tools to convert your data to RAW format.

In order to use your own data with the Transcriber you will need to set up a configuration file with parameters like the audio device, audio server, sample frequency, sample number of bytes etc. You will also need to specify the lexicon file path (lexfile) and the call file path (callfile). The lexicon file for all purposes is a user defied reference dictionary that can be viewed, searched, and modified according to one's preference. The call file contains the location of the transcription file, audio list and comment file. Each of the three previous parameters are significant in which the transcription file contains a set key value pairs that describe each entry in the file. The comment file on the other hand contains a set of bookmarks that tells you the start and stop time along with the duration of the transcription process. Finally, the audio list contains the location of all the audio data that is associated with the given transcription file.

An example directory structure of the Transcriber follows:

```
data

  1998

    08August
      lexfile              (lexicon file)
      session24.cfg        (configuration file)

      24
        callfile           (location of the transcription file, comment file and audio list)

        10-33-44-15
          10-33-44-15
          LOG              (transcription file)
          audiolist        (audio list)

          utt01.raw        (audio files)
          utt02.raw
          utt03.raw
```

There are several options that are available for using the display and audio facilities. These options are accessible by clicking on the Config button on the main screen.




In the first section under Session File the current configuration file, comment file, transcription file and lexicon file are listed. You can even browse through and select another configuration file via the Browse button.

In the second section under Audio-Related Parameters you have the option of setting the audio device (sparc, dat, ncd, x86) to you system. You can also select the audio server (speaker, headphone, line) from the options offered.

In the third section under Energy Plot Parameters you have the option of changing the energy plot parameters. The options include changing the frame length, window length and the RMS scale factor of the energy plot. You can even enlarge or diminish the size of the energy plot canvas by setting the Canvas Size option to your preference.

In the fourth section under Spectrogram Parameters you have the option of setting the brightness and contrast of the spectrogram to any specific value instead of using the slider bars on the main screen.

Finally in the last section under Miscellaneous you have the option of preemphasizing the data for the energy plot and spectrogram. You can either set the preemphasis on or off depending on your preference. You can also set the preemphasis coefficient to any desired value.

The user also has the option of windowing the data using the standard window functions like Hamming, Hanning, rectangular, Bartlett and Blackman.
The Transcriber also has a very nifty auto fill facility which automatically completes the word by hitting the Tab key. However, the auto fill facility will not work if the word to be completed is not in the Lexicon file. Also if a word completion has several possible outcomes, a pop up box will be generated which will list all possible completion for that word. You can then select the desired word from the generated list.

The Transcriber can be used not only for speech segmentation and speech transcription but also for viewing signals. In order to view the signal without having to deal with the transcription protocols just click on the the Lock button on the screen. Apart form viewing the signal there are several display options that are available. Some of these displays options include having the ability to zoom into a section of the signal, change the amplitude of the signal and set the brightness and contrast of the spectrogram.

# Index of /publications/courses/ase_6713

| Name | Last modified | Size | Description |
|---|---|---|---|
| Parent Directory | 27-Dec-2001 14:47 | - | |
| lecture_02.fm | 15-May-1998 12:16 | 2.7M | |
| lecture_02.pdf | 15-May-1998 17:34 | 224k | |

## Lecture 2

## Sound Waves in a Tube

- Derive a theoretical model of how sound waves are affected by the vocal tract
- Describe a model for lip radiation
- Describe a model for the pulsating glottal waveform during voiced speech
- Assemble the components of a simple speech synthesiser

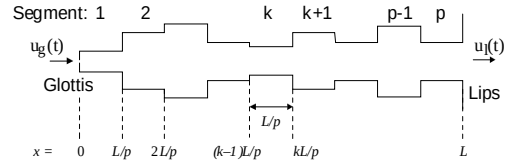
**Appendix (not examinable)**

- The physics of 1-dimensional sound waves



## Multi-Tube Model of Vocal Tract

We model the vocal tract as a tube that has $p$ segments:




$u_g$ and $u_l$ are the volume flows of air at the glottis and lips respectively (measured in litres per second).

Vocal tract is of length $L$ (typically 15-17 cm in adults)

Length of each segment is the distance sound travels in half a sample period = $0.5cT$ : 1.5 cm @ 11 kHz

- $c$ = speed of sound in air

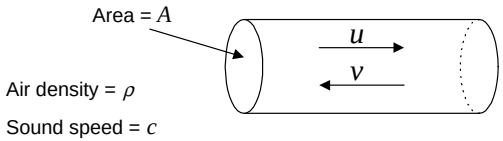
  $\approx 20\sqrt{\text{Absolute Temperature}} \approx 340$ m/s

- $T$ = sample period = $1/f_{samp}$

Number of tube segments needed = $2L/cT \approx 0.001\, f_{samp}$

## Sound Waves in a Tube

Acoustic signal is the superposition of two waves: $u$ in the forward direction and $v$ in the reverse direction:




Total volume flow = $u–v$

Total acoustic pressure = $(u+v) \times \rho c/A$

Exactly analogous to transmission lines:

- Volume flow ≈ Current, Pressure ≈ Voltage
- Acoustic Impedance of tube = $\rho c/A$

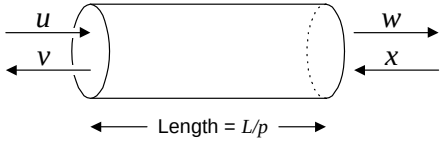
**Assumptions:**

- Sound waves are 1-dimensional: true for frequencies < 3 kHz whose wavelengths are long compared to the tube width
- No frictional or wall-vibration energy losses

See appendix for a non-examinable derivation.



## Segment Delays




Time for sound to travel along segment = $L/cp$

Hence: $v(t) = x\left(t - \frac{L}{cp}\right)$ and $u(t) = w\left(t + \frac{L}{cp}\right)$

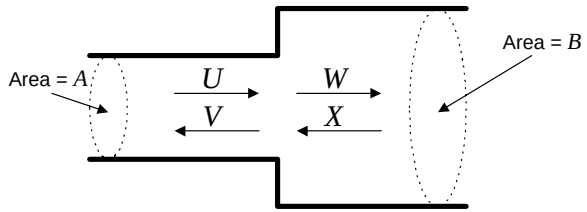
Segment length chosen to correspond to half a sample period. If we take $z$-transforms, this time delay corresponds to multiplying by $z^{-½}$ :

$$V(z) = z^{-½}X(z) \quad \text{and} \quad U(z) = z^{+½}W(z)$$

In matrix form:

$$\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} z^{+½} & 0 \\ 0 & z^{-½} \end{pmatrix} \begin{pmatrix} W \\ X \end{pmatrix} = z^{+½} \begin{pmatrix} 1 & 0 \\ 0 & z^{-1} \end{pmatrix} \begin{pmatrix} W \\ X \end{pmatrix}$$

## Segment Junction




Flow Continuity: $(U - V) = (W - X)$

Pressure Continuity: $\frac{\rho c}{A}(U + V) = \frac{\rho c}{B}(W + X)$

In matrix form: $\begin{pmatrix} 1 & -1 \\ B & B \end{pmatrix}\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ A & A \end{pmatrix}\begin{pmatrix} W \\ X \end{pmatrix}$

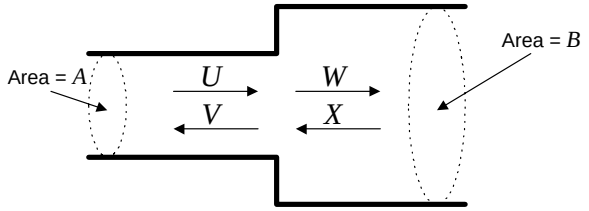
Hence:

$$\begin{pmatrix} U \\ V \end{pmatrix} = \frac{1}{2B}\begin{pmatrix} B & 1 \\ -B & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 \\ A & A \end{pmatrix}\begin{pmatrix} W \\ X \end{pmatrix}$$

$$= \frac{1}{2B}\begin{pmatrix} A+B & A-B \\ A-B & A+B \end{pmatrix}\begin{pmatrix} W \\ X \end{pmatrix}$$



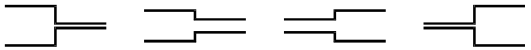
## Reflection Coefficients



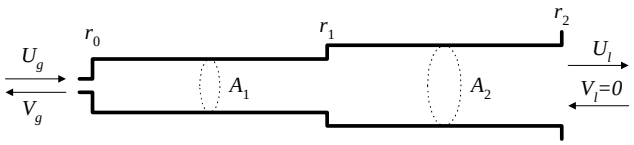

Define the reflection coefficient to be $r = \frac{B - A}{B + A}$

$$\begin{pmatrix} U \\ V \end{pmatrix} = \frac{1}{2B}\begin{pmatrix} A+B & A-B \\ A-B & A+B \end{pmatrix}\begin{pmatrix} W \\ X \end{pmatrix} = \frac{1}{1+r}\begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}\begin{pmatrix} W \\ X \end{pmatrix}$$

Reflection coefficients always lie in the range ±1:



## 2-Segment Vocal Tract




$$r_k = \frac{A_{k+1} - A_k}{A_{k+1} + A_k}$$

$$\frac{1}{1+r_2}\begin{pmatrix} 1 & -r_2 \\ -r_2 & 1 \end{pmatrix}\begin{pmatrix} U_l \\ 0 \end{pmatrix}$$

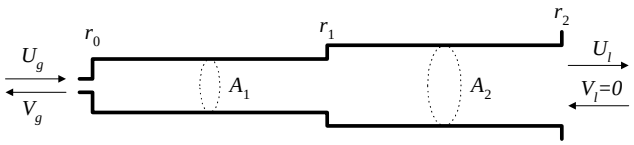
$$\frac{1}{1+r_1}\begin{pmatrix} 1 & -r_1 \\ -r_1 & 1 \end{pmatrix} \times z^{-\frac{1}{2}}\begin{pmatrix} 1 & 0 \\ 0 & z^{-1} \end{pmatrix} \times$$

$$\begin{pmatrix} U_g \\ V_g \end{pmatrix} = \frac{1}{1+r_0}\begin{pmatrix} 1 & -r_0 \\ -r_0 & 1 \end{pmatrix} \times z^{-\frac{1}{2}}\begin{pmatrix} 1 & 0 \\ 0 & z^{-1} \end{pmatrix} \times$$

- Assume $V_l = 0$: no sound reflected back into mouth
- Work backwards from lips towards glottis:
  - Junction: use the reflection matrix
  - Tube segment: use the delay matrix
- $A_3$ is large but not infinite: assumption of narrow tube breaks down at this point
- $A_0$ is approximately zero: area of glottis opening



## Vocal Tract Transfer Function




Multiplying out the matrices gives:

$$\begin{pmatrix} U_g \\ V_g \end{pmatrix} = \frac{z^{+1}}{\prod_{k=0}^{2}(1+r_k)}\begin{pmatrix} 1+(r_0r_1+r_1r_2)z^{-1}+r_0r_2z^{-2} \\ -r_0-(r_1+r_0r_1r_2)z^{-1}-r_2z^{-2} \end{pmatrix} U_l$$

We can ignore $V_g$: it gets absorbed in the lungs.

The vocal tract transfer function is given by the ratio of $U_l$ to $U_g$:

$$\frac{U_l}{U_g} = \frac{\prod_{k=0}^{2}(1+r_k)\times z^{-1}}{1+(r_0r_1+r_1r_2)z^{-1}+r_0r_2z^{-2}}$$

$$= \frac{Gz^{-1}}{1+(r_0r_1+r_1r_2)z^{-1}+r_0r_2z^{-2}}$$

$$= \frac{Gz^{-1}}{1-a_1z^{-1}-a_2z^{-2}}$$

## p-segment Vocal Tract

Note that: $\frac{1}{1+r}\begin{pmatrix}1 & -r\\ -r & 1\end{pmatrix}\times z^{-½}\begin{pmatrix}1 & 0\\ 0 & z^{-1}\end{pmatrix}=\frac{z^{-½}}{1+r}\begin{pmatrix}1 & -rz^{-1}\\ -r & z^{-1}\end{pmatrix}$

Multiplying together all the matrices for a *p*-segment vocal tract gives:

$$\begin{pmatrix}U_g\\ V_g\end{pmatrix}=\frac{z^{-½p}}{\prod_{k=0}^{p}(1+r_k)}\prod_{k=0}^{p-1}\begin{pmatrix}1 & -r_k z^{-1}\\ -r_k & z^{-1}\end{pmatrix}\times\begin{pmatrix}1\\ -r_p\end{pmatrix}U_l$$

This results in a transfer function of the form:

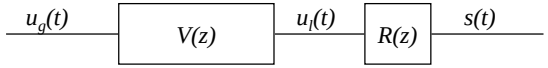
$$V(z)=\frac{U_l}{U_g}=\frac{Gz^{-½p}}{1-a_1z^{-1}-a_2z^{-2}-\ldots-a_pz^{-p}}$$

Where:

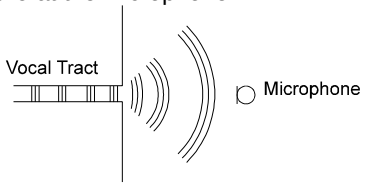
- $G$ is a gain term
- $z^{-½p}$ is the acoustic time delay along the vocal tract
- The denominator represents a $p^{th}$ order all-pole filter



## Lip Radiation




R(z) is the transfer function between *airflow* at the lips and *pressure* at the microphone.



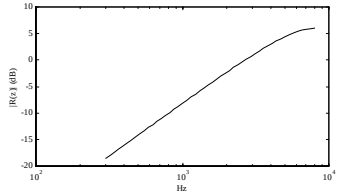

For a lip-opening area of A, acoustic theory predicts a 1st-order high-pass response with a corner frequency of:

$$\frac{c}{\sqrt{4A}}\text{ Hz}\approx 5\text{ kHz}$$

For $f_{samp}$ < 20 kHz, a good approximation is:

$$R(z)=\frac{S(z)}{U_l(z)}=1-z^{-1}$$

$$\Rightarrow\left|R(z)\right|=2\sin\left(\frac{\omega T}{2}\right)$$

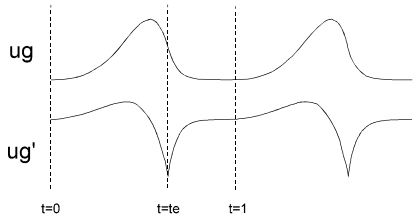


## Spectrum of Glottal Waveform

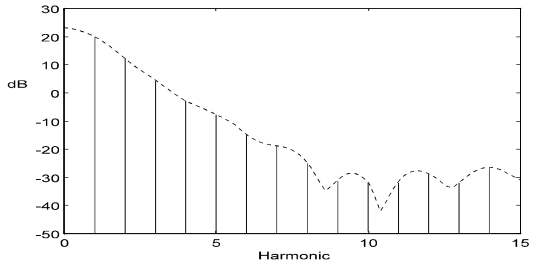
"LF Model" (Liljencrants & Fant)

$$u_g'(t) = \begin{cases} e^{at}\sin(bt) & 0 \le t < t_e \\ c + de^{-ft} & t_e \le t < 1 \end{cases}$$

with $u_g(0) = u_g(1) = 0$; $u_g(t)$ and $u_g'(t)$ continuous at $t_e$



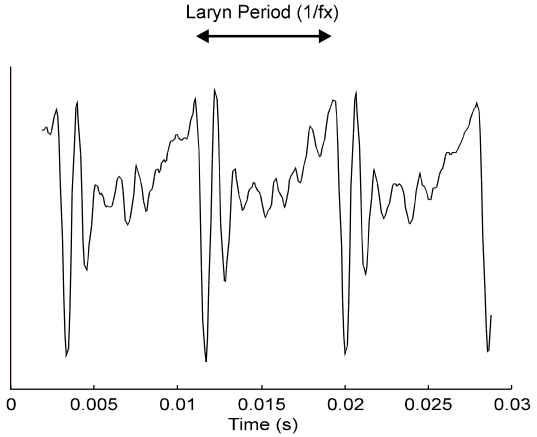

Line Spectrum of ug (approx –12 dB/octave):






## Vowel Waveform

Vowel /ɑ/ from "part"



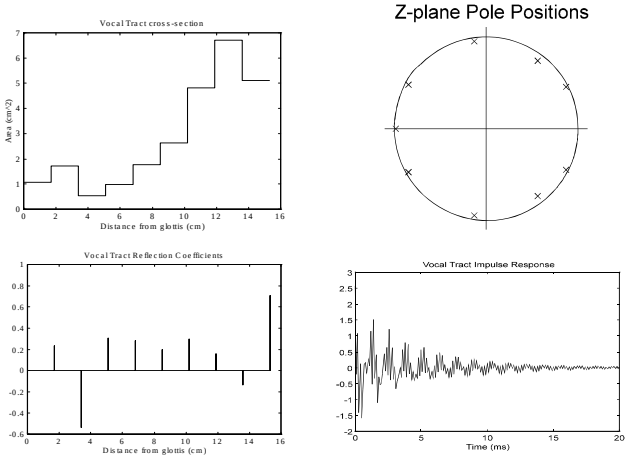

- Larynx Frequency ≈ 130 Hz
- First Vocal tract resonance (formant) ≈ 1 kHz

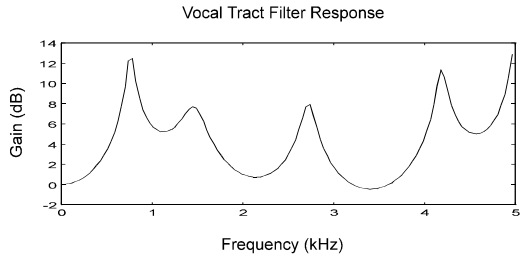
There is not necessarily any relation between the larynx frequency and the vocal tract resonances.

Resonances at a multiple of the larynx frequency will be louder (good for singers)

## Vocal Tract Shape and Response

Example: /ɑ/ vowel (“part”)





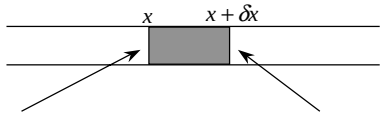

## Appendix

### Theoretical Derivation of Sound Waves

This section is non-examinable

## 1-Dimensional Sound Waves

Consider a small chunk of air in a tube with a uniform cross-sectional area A:




Pressure $= p$

Velocity $= v = \dfrac{u}{A}$

Pressure $= p + \delta x \dfrac{\partial p}{\partial x}$

Velocity $= v + \delta v = \dfrac{1}{A}\left(u + \delta x \dfrac{\partial u}{\partial x}\right)$

$$\Rightarrow \quad \delta v = \frac{\delta x}{A}\frac{\partial u}{\partial x}$$

Volume of air chunk: $V = A \times \delta x$

Hence: $$\frac{\partial V}{\partial t} = A \times \delta v = A \times \frac{\delta x}{A}\frac{\partial u}{\partial x} = \frac{V}{A} \times \frac{\partial u}{\partial x} \qquad ①$$

Net force on air chunk:

$$F = Ap - A\left(p + \delta x \frac{\partial p}{\partial x}\right) = -A\,\delta x \frac{\partial p}{\partial x}$$



## Gas Laws

**Ideal Gas Law :**

We can express the pressure in terms of the density:

$$pV = nRT$$
$$= \frac{\rho V}{M} RT$$
$$\Rightarrow \quad p = \rho \times \frac{RT}{M}$$

$n$ = moles of air = molecules ÷ $(6 \times 10^{23})$
$R$ = gas constant = 8.314 J / (K · mol)
$T$ = Temperature (° K)
$\rho$ = density ($\approx 1.225$ kg / m$^3$)
$M$ = molecular weight of air = 0.029 kg / mol
$\gamma$ = specific heat ratio of air = 1.4

We define $$c^2 = \frac{\gamma RT}{M} \approx (340 \text{ m/s})^2 \quad \Rightarrow p\gamma = \rho c^2 \qquad ②$$

$c$ will turn out to be the speed of sound and depends only on T.

**Adiabatic Gas Law:** For pressure changes too rapid for heat conduction to occur (e.g. sound vibrations):

$$\frac{d}{dt}\left(pV^\gamma\right) = 0 \quad \Rightarrow \quad V^\gamma \frac{\partial p}{\partial t} + p\gamma V^{\gamma-1}\frac{\partial V}{\partial t} = 0$$

using ① and ② $$\Rightarrow \quad V^\gamma \frac{\partial p}{\partial t} = -\rho c^2 \times \frac{V^\gamma}{A} \times \frac{\partial u}{\partial x}$$

$$\Rightarrow \quad A\frac{\partial p}{\partial t} = -\rho c^2 \frac{\partial u}{\partial x} \qquad ③$$

## Wave Equations

**Mass × Acceleration = Force:**

$$\rho V \times \frac{1}{A}\frac{\partial u}{\partial t} = -A\delta x \frac{\partial p}{\partial x} = -V\frac{\partial p}{\partial x} \quad \Rightarrow \quad \rho\frac{\partial u}{\partial t} = -A\frac{\partial p}{\partial x} \qquad ④$$

**Wave Equations:**

Equations ③ and ④ are known as the *wave equations*:

$$\rho\frac{\partial u}{\partial t} = -A\frac{\partial p}{\partial x} \quad \text{and} \quad A\frac{\partial p}{\partial t} = -\rho c^2\frac{\partial u}{\partial x}$$

Solution:

$$u(x,t) = u^+(t - x/c) - u^-(t + x/c)$$

$$p(x,t) = \frac{\rho c}{A} \times \left\{u^+(t - x/c) + u^-(t + x/c)\right\}$$

It is easily verified that this solution satisfies the wave equations for any differentiable functions $u^+$ and $u^-$.

The two functions $u^+$ and $u^-$ represent waves travelling in +ve and –ve $x$ directions at velocity $c$. The actual values of the waves are determined by the boundary conditions at the end of the tube section.

The equations are the same as for a transmission line with $u \approx$ current, $p \approx$ voltage and $\rho c/A \approx$ impedance.

University of California
Berkeley

College of Engineering
Department of Electrical Engineering
and Computer Sciences

Professors : N.Morgan / B.Gold

EE225D

Spring,1999

# Acoustic Tube Models

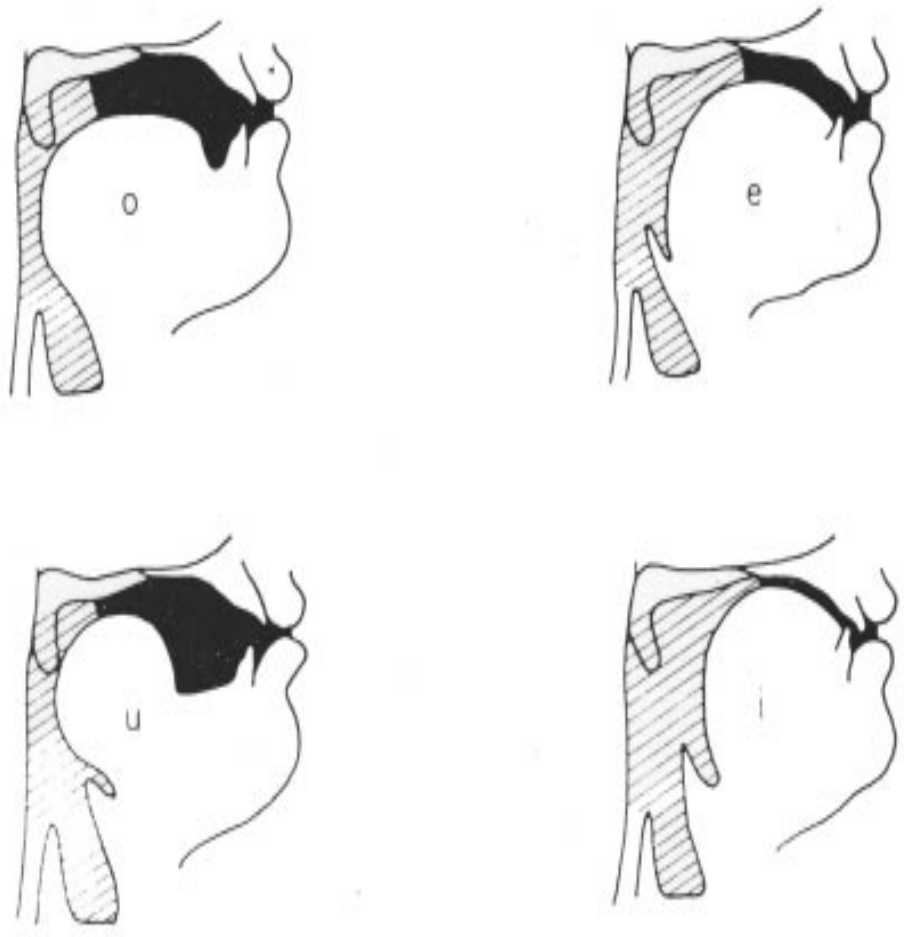
# Lecture 13

## Introduction :

Acoustic Tube Models of English Phonemes → 2 tube model.

## Assumptions :

- Lossless tubes
- Plane waves
- Rigid walls
- Friction
- Thermal effect




Vocal tract area for four vowel sounds

Vocal tract areas for four vowel sounds.

i - Tongue is High.

e - Tongue is a little Lower.

u - Tongue is very Low.

o - Togue is somewhat low.

1. Tube response vs. area function.
2. Discrete-time-space version.
3. Example - 2 tube representation of vowels.

**Problem for Today :**

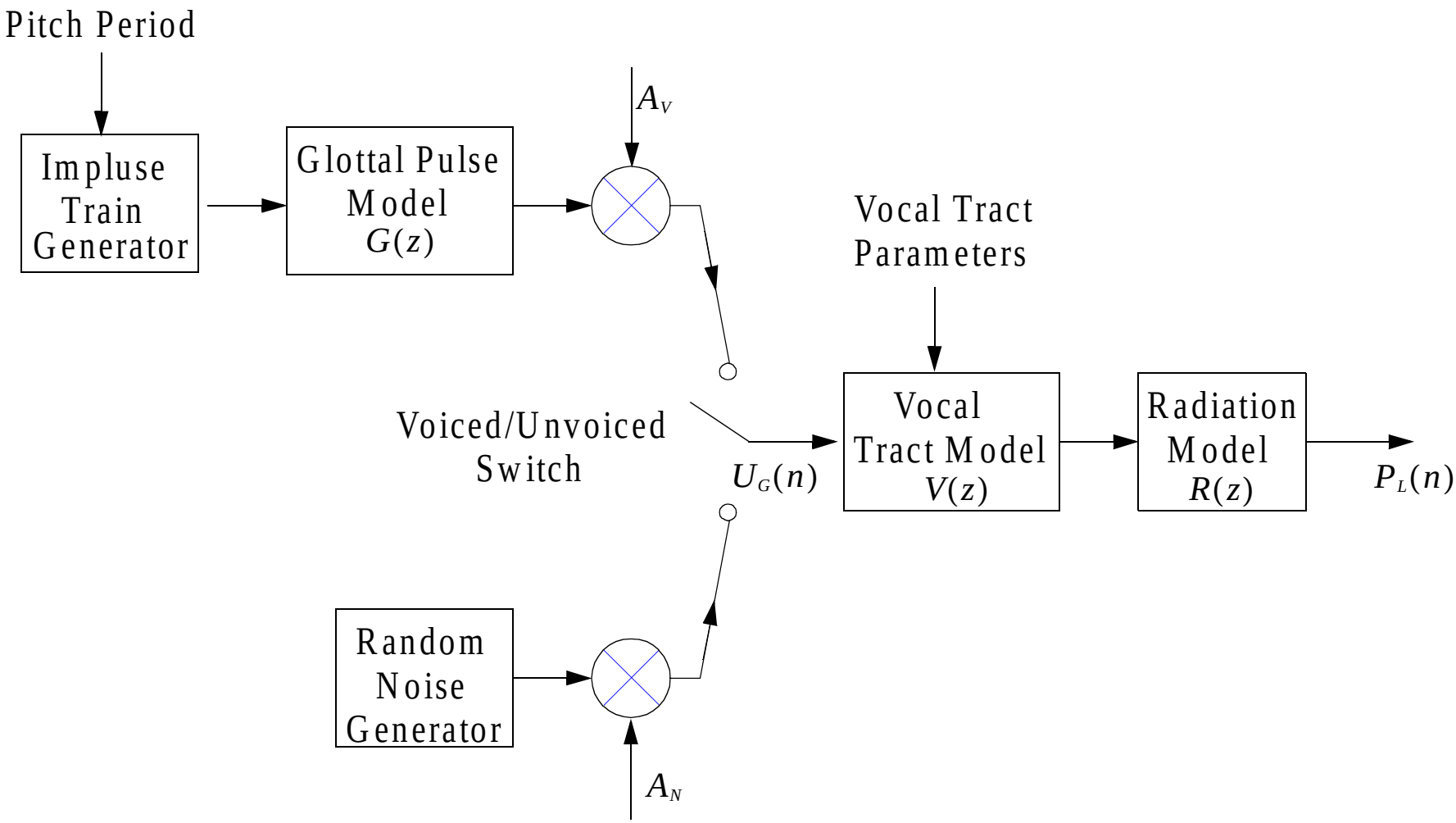
Develop a 2 tube model to derive a frequency response that approximates some vowels.

By solving a complicated wave equation, the frequency response can be found.

Look up equation in R & S.

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial}{\partial t}(u/A)$$

$$-\frac{\partial u}{\partial t} = \frac{1}{\rho c^2}\frac{\partial}{\partial t}(pA) + \frac{\partial A}{\partial t}$$




Viewgraph 2 : General Discrete-Time Model for Speech Production.

Assumption in this Model :

Vocal Tract Model - Time varying

Radiation Model - May be time varying

Glottal Pulse Model - Usually considered independent of vocal tract model, but later we'll examine this wave closely

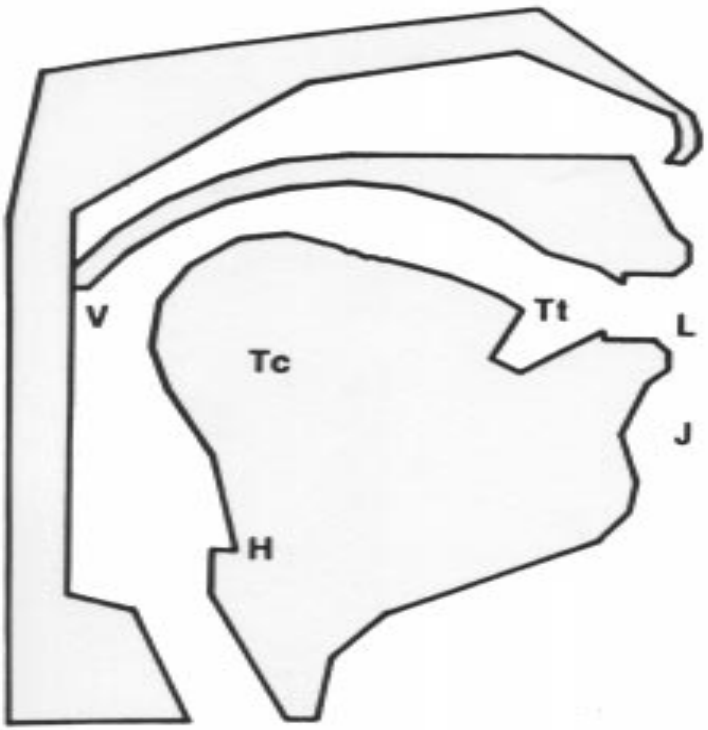
$$u(x,t) = u^{+}\left(t - \frac{x}{C}\right) - u^{-}\left(t + \frac{x}{C}\right)$$

$$p(x,t) = Z_o\left[u^{+}\left(t - \frac{x}{C}\right) + u^{-}\left(t + \frac{x}{C}\right)\right]$$

$p(l,t) = 0$ : open tube

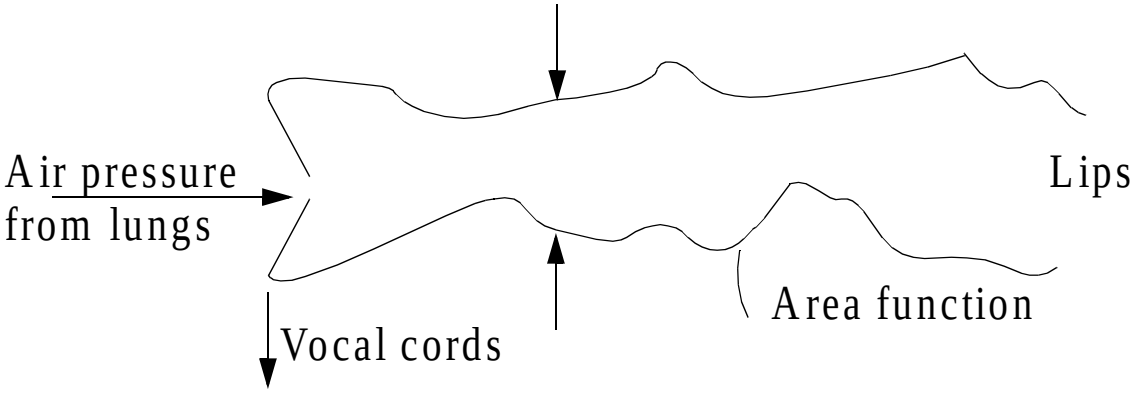
$$u^{+}\left(t - \frac{l}{C}\right) = -u^{-}\left(t + \frac{l}{C}\right)$$

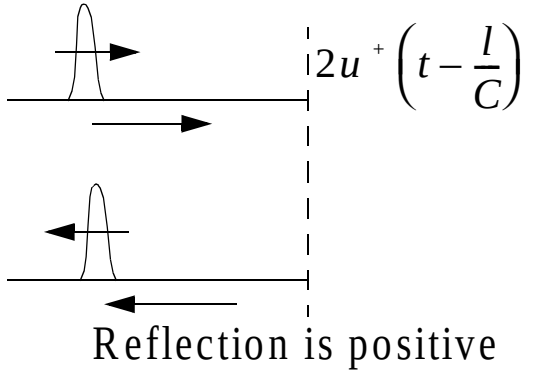
$$u(l,t) = 2u^{+}\left(t - \frac{l}{C}\right)$$




H = HYOID POSITION
J = ANGLE OF JAW OPENING
L = LIP PROTRUSION AND ELEVATION
Tc = TONGUE CENTER
Tt = POSITION OF TONGUE TIP
V = VELUM OPENING

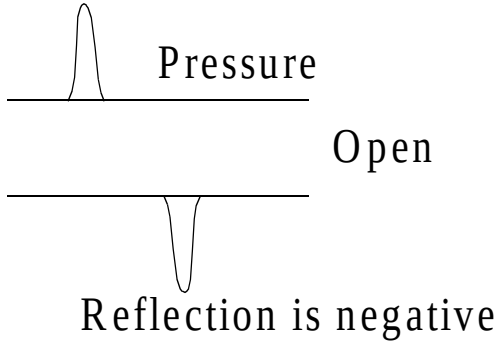
# Model of Vocal Tract




Closed Tube $u(l,t) = 0$

so $u^+\left(t-\frac{l}{C}\right) = -u^-\left(t+\frac{l}{C}\right)$




- Given Area Function we can compute Spectrum

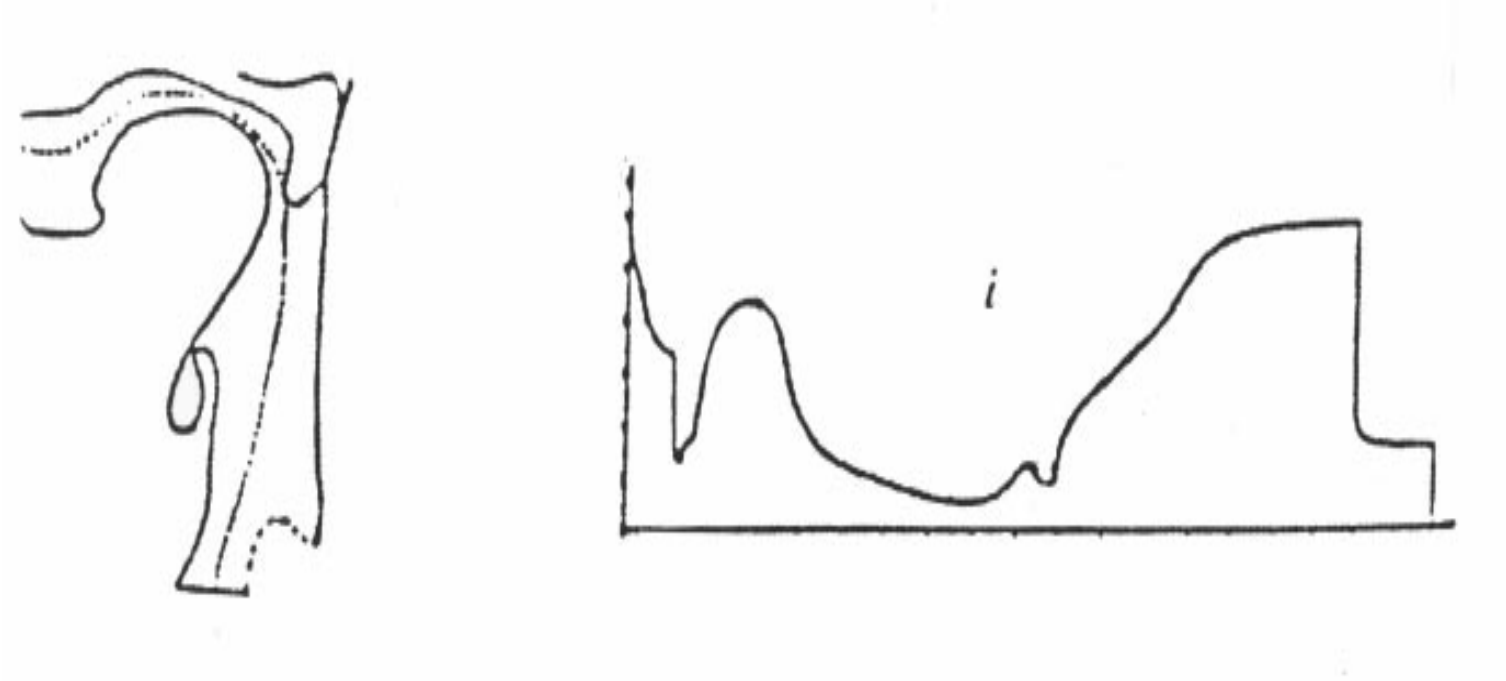



Figure 11.1: X-ray tracing and area function for phoneme /i/

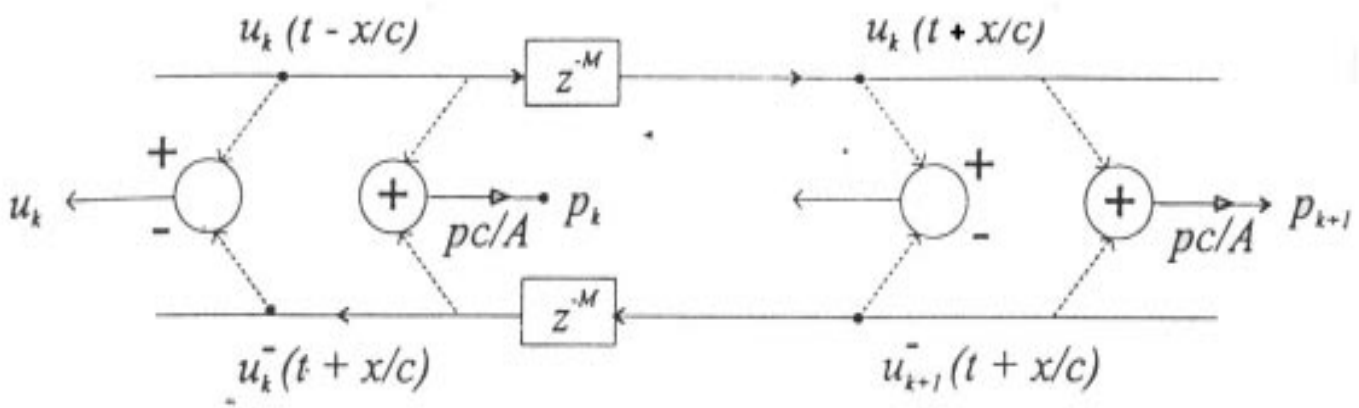



Figure 11.2: Single section of digital wave guide

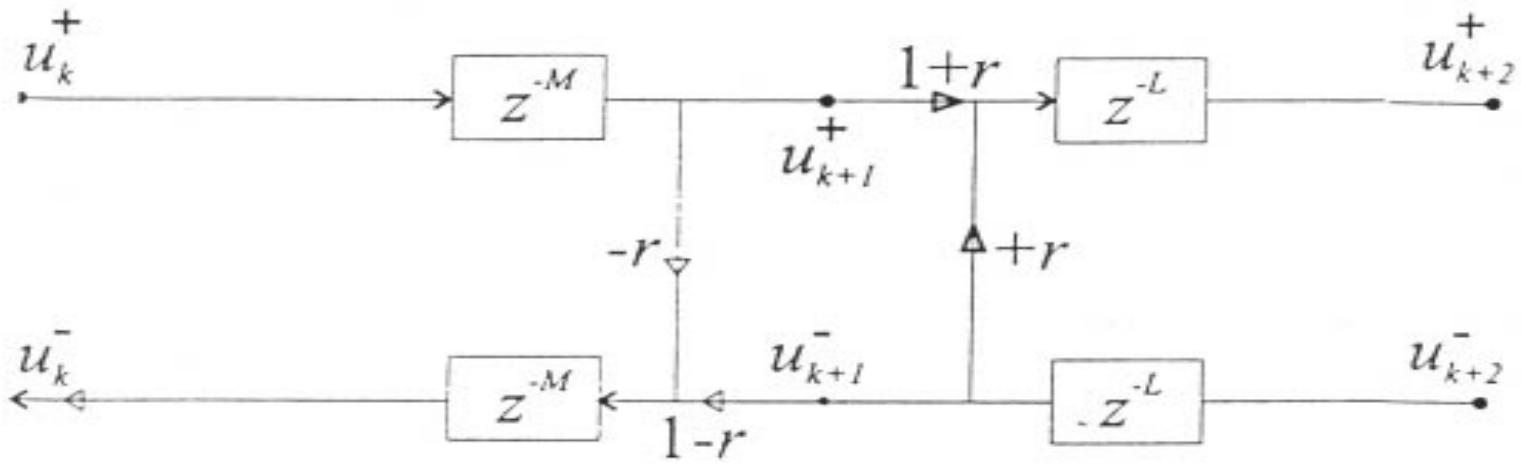



Figure 11.3: Two section digital wave guide

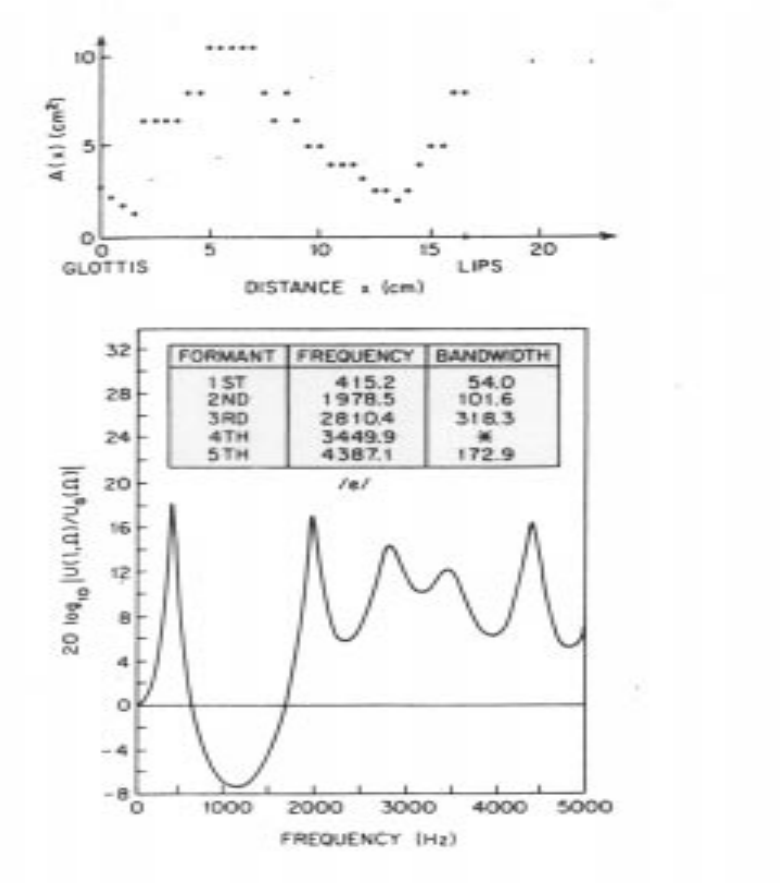



Figure 3.24 Area function and frequency response for the Russian vowel /*e*/

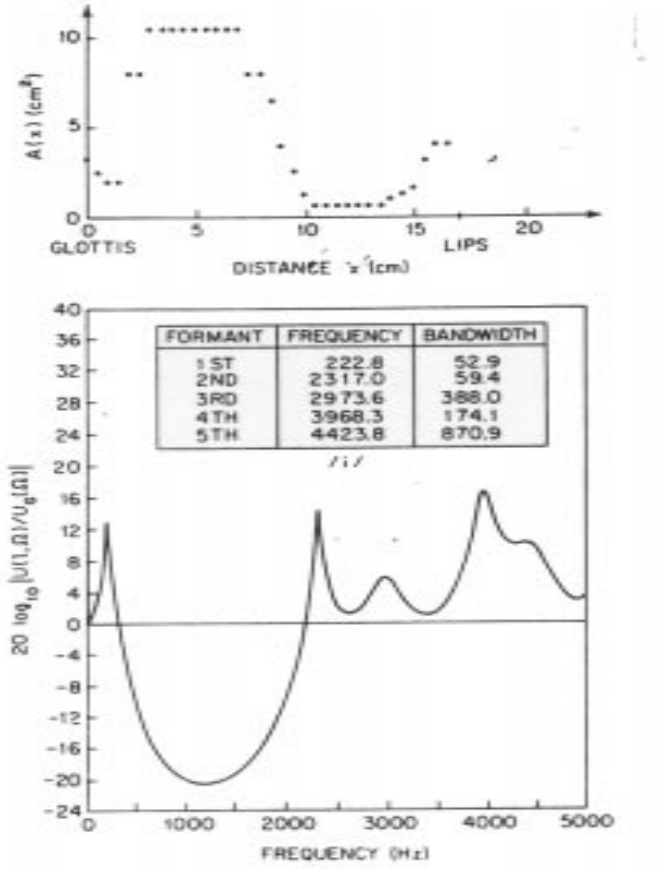
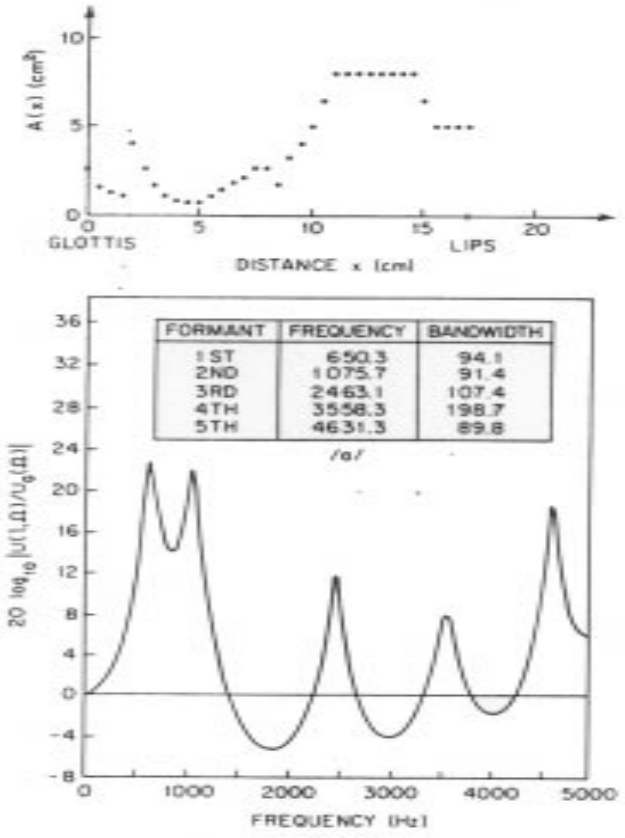
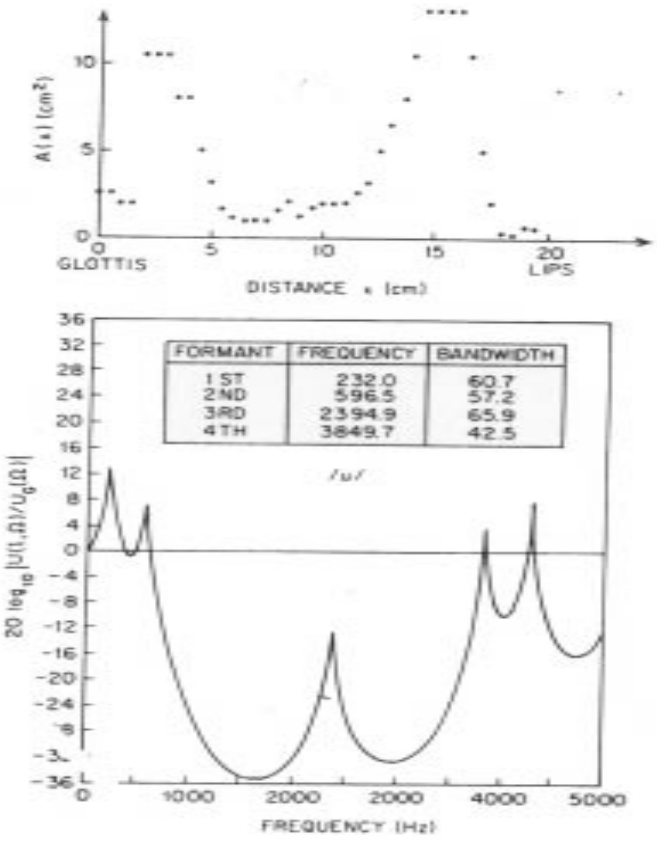


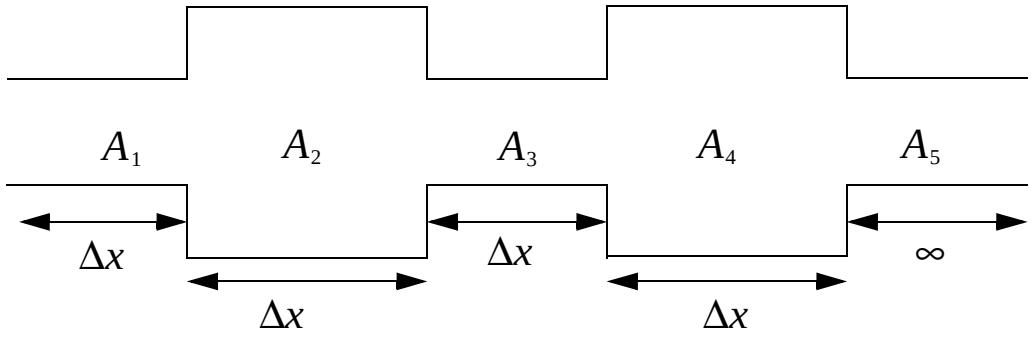

Figure 3.25 Area function and frequency response for the Russian vowel /*i*/


A(x) (cm²)
GLOTTIS
LIPS
DISTANCE x (cm)
FORMANT FREQUENCY BANDWIDTH
1 ST 650.3 94.1
2ND 1075.7 91.4
3RD 2463.1 107.4
4TH 3558.3 198.7
5TH 4631.3 89.8
/a/
20 log10 |U(1,Ω)/UG(Ω)|
FREQUENCY (Hz)

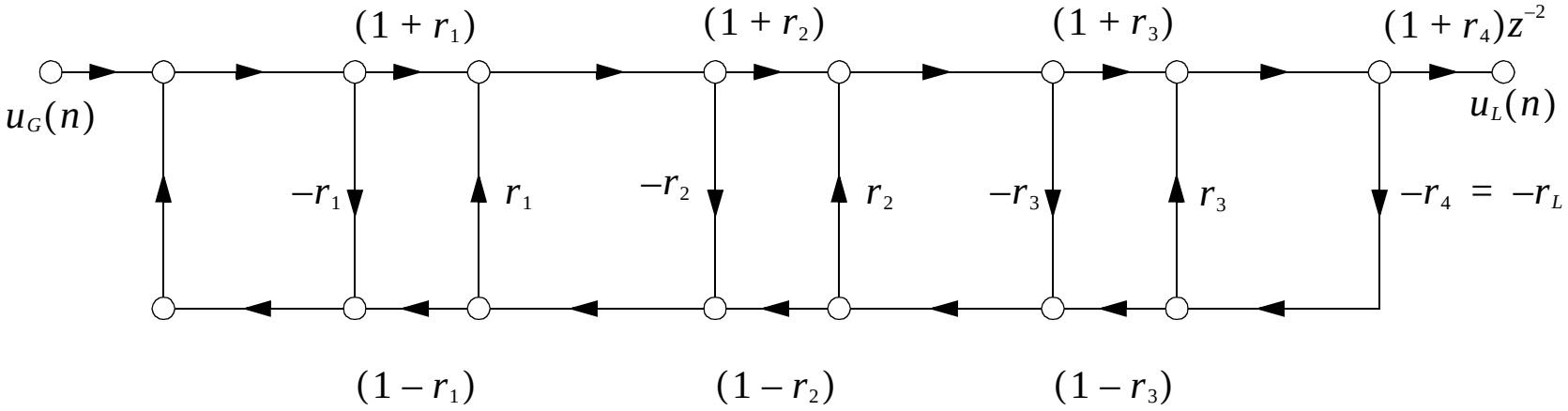

A(x) (cm²)
GLOTTIS
LIPS
DISTANCE x (cm)
FORMANT FREQUENCY BANDWIDTH
1 ST 232.0 60.7
2ND 596.5 57.2
3RD 2394.9 65.9
4TH 3849.7 42.5
/u/
20 log10 |U(1,Ω)/UG(Ω)|
FREQUENCY (Hz)

# More Complex Tube Structure

a)




b)



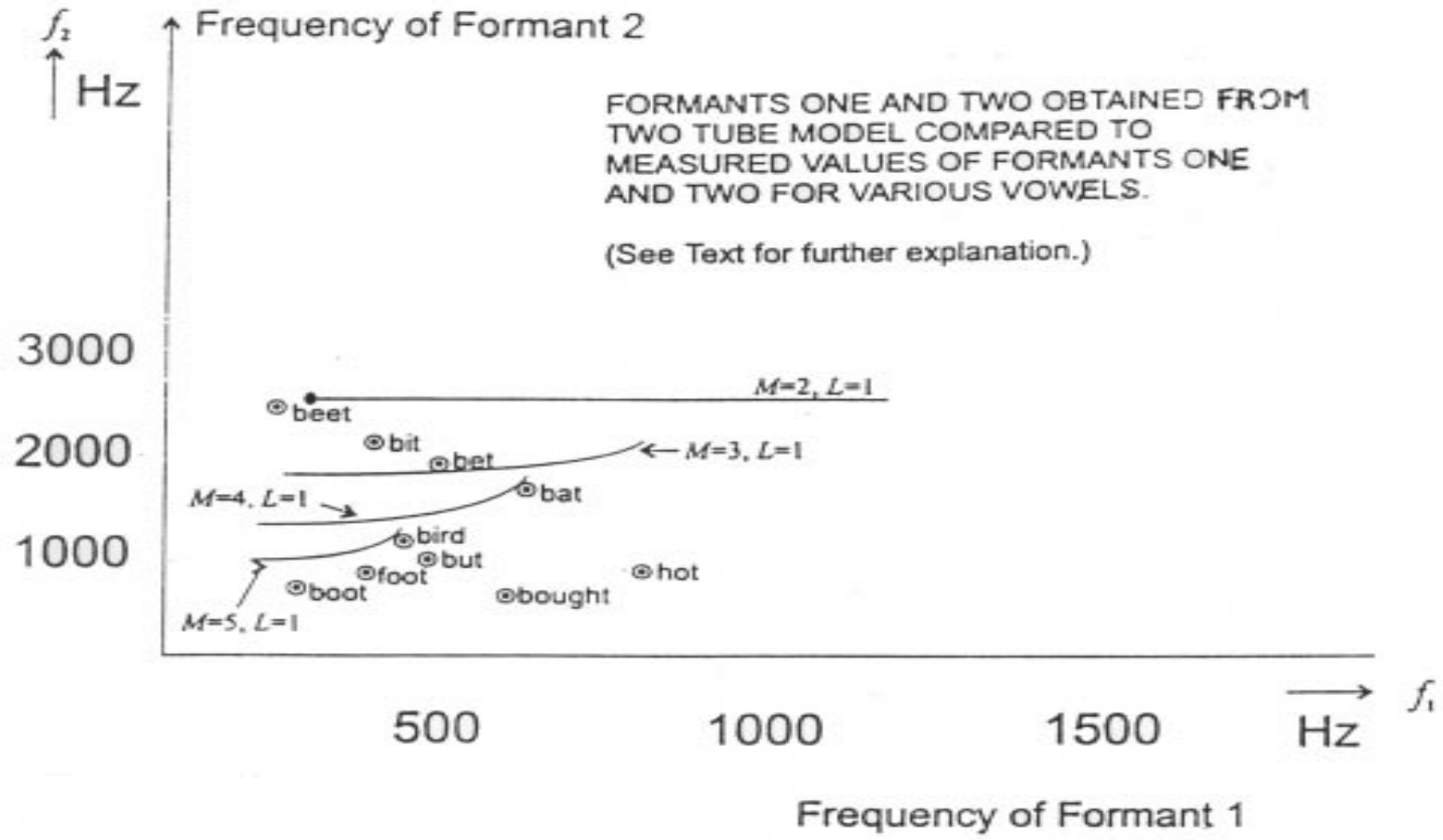



Figure 11.4 Formants 1 and 2 obtained from two tube model

## Physics!

# Physics in Speech

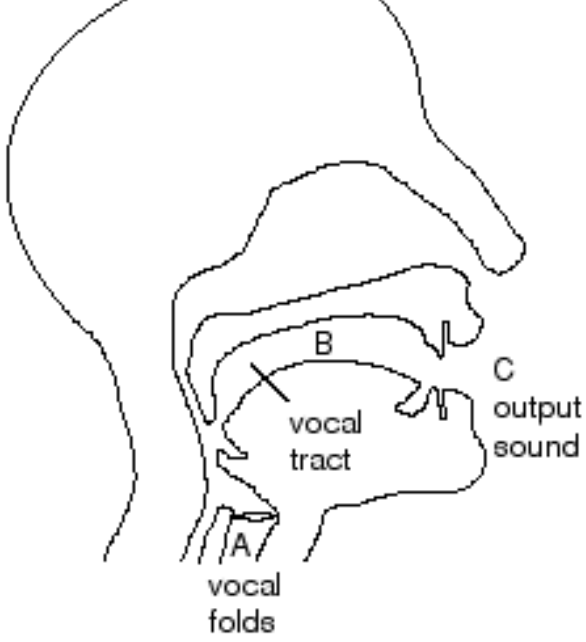
**An introduction to some of the physics in speech**
**(including some notes about helium speech)**

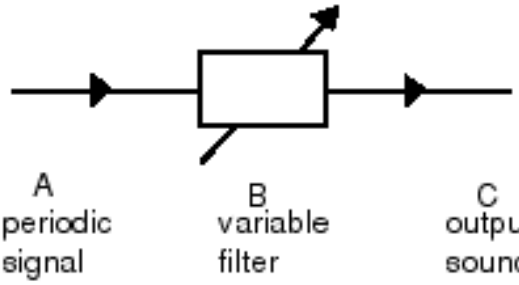
Content : Joe Wolfe

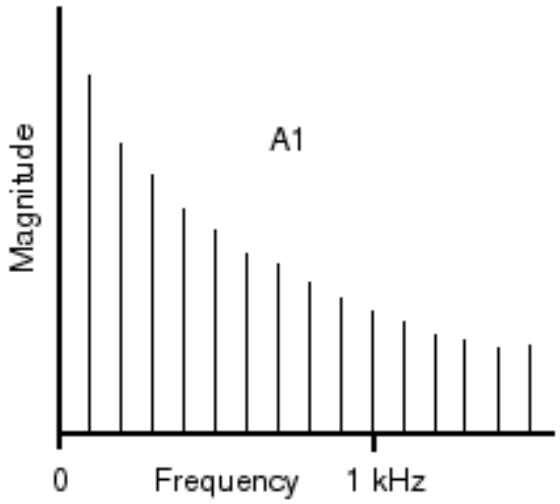
This short document describes a simple model of the vocal tract and the production of voiced speech used in the production of some sustained phonemes - especially the vowels. It also includes some brief notes about helium speech.

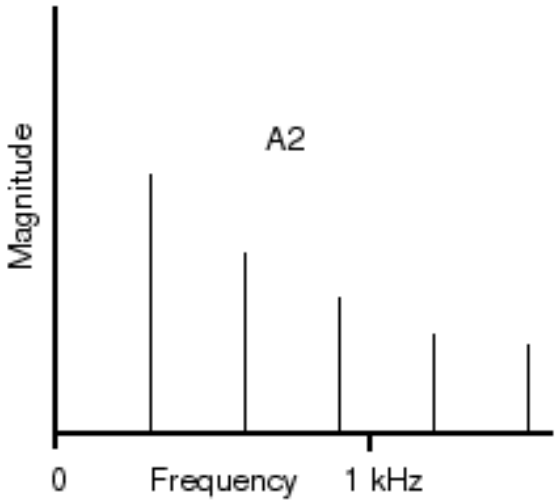
### The source-filter model of the vocal tract

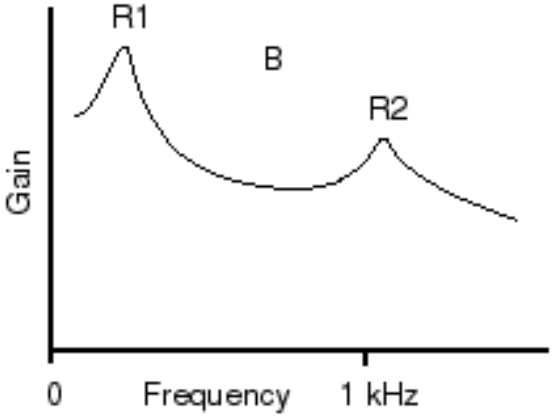
The vibration of the vocal folds produces a varying air flow which may be treated as a periodic source (A). (A periodic signal is cyclic: its motion is reproduced after a time interval called its period. A consequence is that its spectrum is made up of harmonics. Go to 'What is a sound spectrum?' for an introduction.) This source signal is input to the vocal tract. The tract behaves like a variable filter (B) in that its response is different for different frequencies. It is variable because, by changing the position of your tongue, jaw etc you can change that frequency response. The input signal and the vocal tract, together with the radiation properties of the mouth, face and external field, produce a sound output (C). Because the source is harmonic, we can say that the gain of the tract (B) is sampled at multiples of the pitch frequency F0. In the case at left, the resonances R1 and R2 can be determined approximately from the peaks in the envelope of the sound spectrum. These peaks are called the formants (F1 and F2).

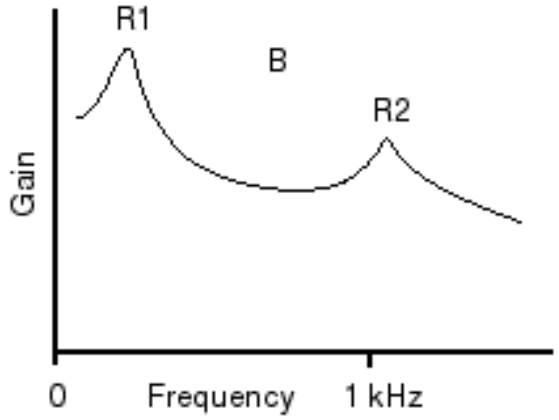

B
C
output
sound
vocal
tract
A
vocal
folds

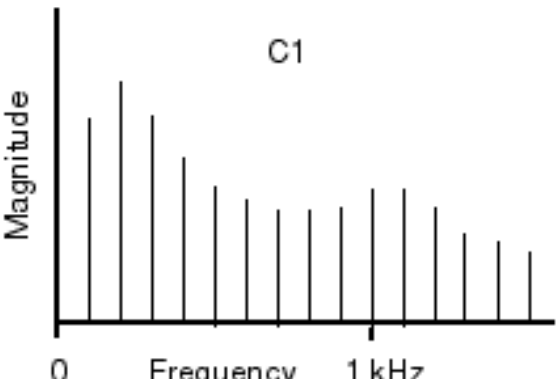

A
periodic
signal
B
variable
filter
C
output
sound

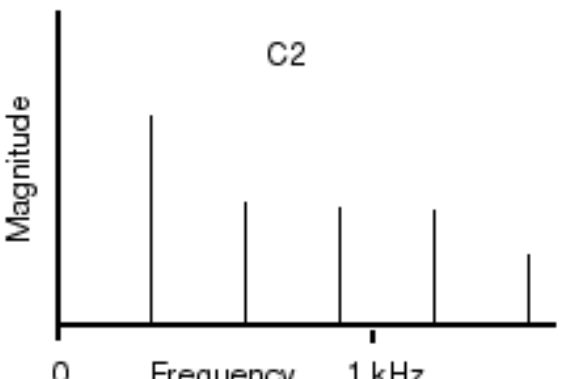

Magnitude
A1
0
Frequency
1 kHz

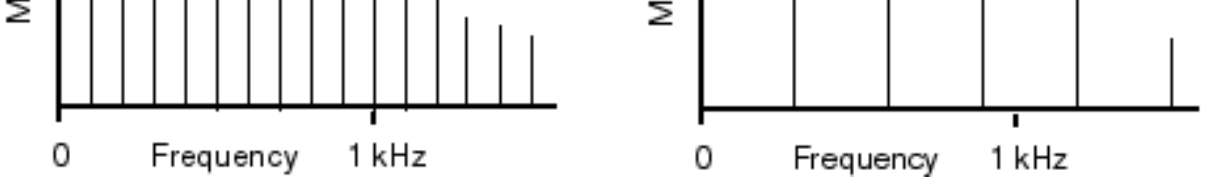

Magnitude
A2
0
Frequency
1 kHz


R1
B
R2
Gain
0
Frequency
1 kHz


R1
B
R2
Gain
0
Frequency
1 kHz


C1
Magnitude
0
Frequency
1 kHz


C2
Magnitude
0
Frequency
1 kHz




Note that the detail in the spectrum is easier to see if F0 is low, e.g. for a low pitched man's voice (diagram at left), than it is for a child's or woman's voice - shown at right.

The lowest resonance is determined to a considerable extent by the end effect of your mouth: if you lower your jaw, R1 rises. R2 is affected by the jaw position too, but it is primarily affected by the position of the constriction inside your mouth. Moving your tongue forwards and backwards changes R2 (and also R1, but to a lesser extent). A map of (R1,R2) for Australian English is given on our speech research page.

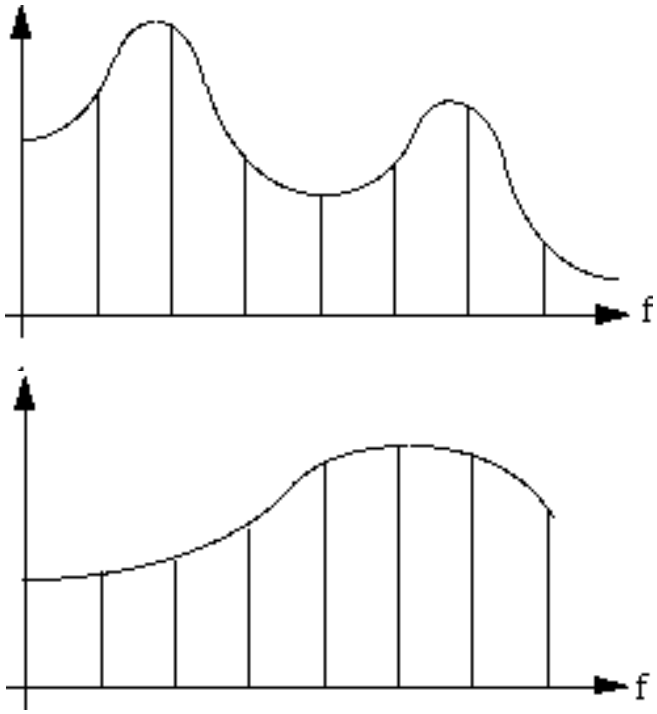
Nearly all information in speech is in the range 200 Hz-8 kHz. (The telephone carries only 300 Hz - 3 kHz but speech is reasonably intelligible and the telephone company's hold music still sounds okay.) The pitch is determined by the spacing of harmonics as much as or more than by the fundamental. Thus you can tell the pitch of a man's voice on the phone even though the fundamental of that signal is not present. Note the size of the vocal tract (~170 mm long) gives resonances > ~ 500 Hz. In fact a closed tube of this length is a functional approximation of the tract for the vowel "er" as in "herd". For this 'neutral' vowel, the first five resonances of the author's vocal tract are indeed at values of about 500, 1500, 2500, 3500 and 4500 Hz.

You can investigate this model by changing the speed of sound. Inhaling helium changes the frequencies of the resonances. As you would expect, it does not change the pitch, which is determined by the tension, mass and geometry of vocal folds, and some other effects. It does however change the timbre. In speech, you may have the illusion that the pitch has changed because one doesn't think much about pitch when listening to speech. To make it clear, you can sing with and without a lung full of He and listen.

***Warnings:***

- He is suffocating and conducts heat well.
- After one inhalation of He, breathe air normally for a few minutes.
- In a gas cylinder, He is under high pressure.
- Do not inhale directly from a gas cylinder.
- Fill a toy balloon and inhale from that.

**What helium does to speech**




The first diagram shows a schematic picture of the spectrum for a particular configuration of the vocal tract *filled with air*. The solid line is the spectral envelope; the vertical lines are the harmonics of the vibration of the vocal folds. The second diagram shows the effect of replacing air with helium, but keeping the tract configuration the same (i.e. trying to pronounce the same vowel as before, but with a throat full of helium). The speed of sound is greater, so the resonances occur at higher frequencies: the second resonance has been shifted right off scale in this diagram. The flesh in your vocal folds still vibrates at the same* frequency, so the harmonics occur at the same frequency.

What does this sound like? Well if you listen for the pitch, you will hear that it is the same note as previously (it is easier to hear the pitch if you sing rather than speak, because in speech we are much less conscious of the pitch). If you do the experiment with someone who has a bit of experience with singing, (and if s/he doesn't laugh too much on hearing helium voice) then the pitch will be the same in the two cases. The pitch is determined by the frequencies of the harmonics and these have not changed*. The speech does however sound 'like Donald Duck'. There is less power at low frequencies so the sound is thin and squeaky. This alteration to the timbre changes vowels in a spectacular way. Although we can understand whole sentences (using contextual clues) we find that individual vowels are very difficult to identify. (By the way, an articulate but otherwise standard duck would have a shorter vocal tract than ours so, even while breathing air, Donald would have resonances at rather higher frequencies than ours.)

* If you keep the muscle tensions the same, that is, the frequencies will not change much. There could be a small change because the less dense He loads the vocal folds a bit less than the air, but this effect is slight. The effect on the resonances is large, however. Its size depends on how pure the He in your vocal tract is.)

| Audio File | File Format | |
|---|---|---|
| Ordinary Speech | wav | au |
| Helium Speech | wav | au |
| Pitch in Air | wav | au |
| Pitch in Helium | wav | au |

**Some other phoneme classes (very briefly)**

Fricatives (f, sh, ss etc) are produced by turbulence at a small constriction. This produces broad band sound with characteristic frequencies. Initial plosives (b, d, k etc) have a short burst of broad band sound then a characteristic transient (as the constriction opens) in the following vowel. Final plosives have a transient (as the constriction shuts) followed by short silence and then the broad band sound. The relative timing of voicing (vocal fold vibration) is important. The presence of voicing distinguishes v from f, zz from ss, b from p etc.

**Gear for further investigations:**
**A microphone and oscilloscope** with a sensitive input range (~ mV) or else a pre- amplifier. Appropriate connectors. To start, try 100 ms/div on the time base, then look more closely. If the CRO is digital (or a virtual one running on your PC), the storage mode is very useful.
**A PC with a sound card** and analysis/edit software is useful. The sampling feature is effectively a storage CRO, and the analysis feature is effectively a spectrum analyser.
You can put your fingers on your throat to determine whether vocal fold vibrate or not ('voiced' or not).

**Some explanatory notes**

- What is a decibel?
- What is a sound spectrum?
- What is acoustic impedance and why is it important?

**Related pages**

- "Vocal tract acoustics" (A web resource about our work in this area)
- "Musical acoustics" (A web resource with both introductory and research material)

- "French vowels"

---

## Further Information


### Email

- Joe Wolfe : J.Wolfe@unsw.edu.au
- Musical Acoustics Group

### Phone Number

- 61 2 9385 4954 (UT +10, +11 Oct-Mar)

### Facsimile Number

- 61 2 9385 6060