
SPEECH PROCESSING

Speech Production Modelling

Patrick A. Naylor
Spring Term 2008/9

Imperial College London

Aims of this module

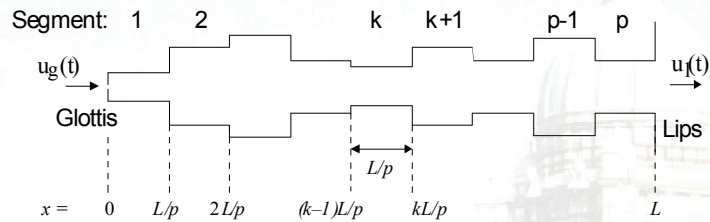
- Derive a theoretical model of how sound waves are affected by the vocal tract
- Describe a model for lip radiation
- Describe a model for the pulsating glottal waveform during voiced speech
- Assemble the components of a simple speech synthesiser

Imperial College London

2

Tube model of the vocal tract

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

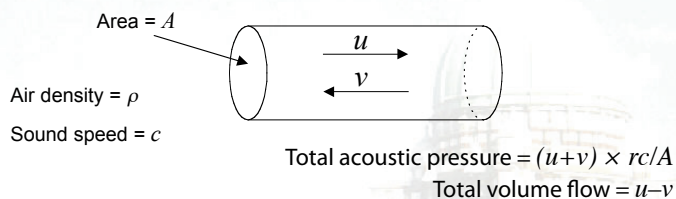


- u_g and u_l are the volume flows of air at the glottis and lips respectively (measured in litres per second).

- 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 = $1/f_{\text{samp}}$
- Number of tube segments needed = $2L/cT \approx 0.001 f_{\text{samp}}$

Sound Waves in a Tube

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



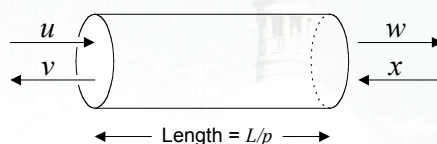
- Exactly analogous to transmission lines:
 - Volume flow \approx Current, Pressure \approx Voltage
 - Acoustic Impedance of tube = rc/A
- Assumptions:
 - Sound waves are 1-dimensional: true for frequencies < 3 kHz whose wavelengths are long compared to the tube width
 - No frictional or wall-vibration energy losses

Imperial College London

5

Segment Delays

- Time taken for sound to travel along one segment = L/cp



- Hence:

$$v(t) = x\left(t - \frac{L}{cp}\right) \quad \text{and} \quad u(t) = w\left(t + \frac{L}{cp}\right)$$
- Segment length chosen to correspond to half a sample period. If we take z-transforms, this time delay corresponds to multiplying by $z^{-1/2}$:

- In matrix form: $V(z) = z^{-1/2}X(z)$ and $U(z) = z^{+1/2}W(z)$

$$\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} z^{+1/2} & 0 \\ 0 & z^{-1/2} \end{pmatrix} \begin{pmatrix} W \\ X \end{pmatrix} = z^{+1/2} \begin{pmatrix} 1 & 0 \\ 0 & z^{-1} \end{pmatrix} \begin{pmatrix} W \\ X \end{pmatrix}$$

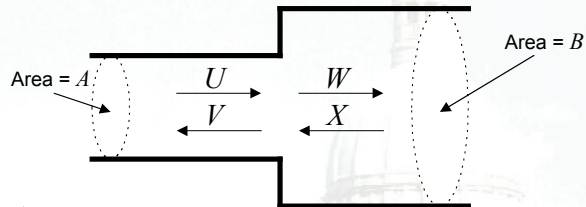
Imperial College London

6

Segment Junctions

- Flow continuity:

$$(U - V) = (W - X)$$



- Pressure continuity:

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

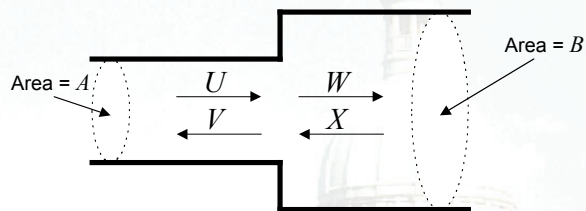
- Combine in matrix form:

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

Imperial College London

7

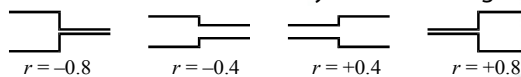
Reflection Coefficients



- Reflection coefficient $r = \frac{B - A}{B + A}$

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

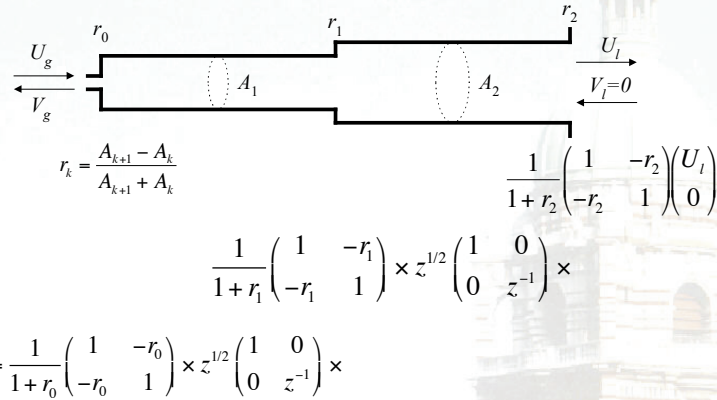
- Reflection coefficients always lie in the range ± 1 :



Imperial College London

8

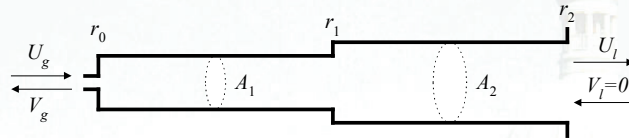
2-segment Vocal Tract Tube Model



Computing the 2-segement Model

- 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:

$$\begin{pmatrix} U_g \\ V_g \end{pmatrix} = \frac{z^{+1}}{\prod_{k=0}^2 (1+r_k)} \begin{pmatrix} 1 + (r_0 r_1 + r_1 r_2) z^{-1} + r_0 r_2 z^{-2} \\ -r_0 - (r_1 + r_0 r_1 r_2) z^{-1} - r_2 z^{-2} \end{pmatrix} U_l$$

We can ignore V_g as it gets absorbed in the lungs.

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

$$\begin{aligned} \frac{U_l}{U_g} &= \frac{\prod_{k=0}^2 (1+r_k) \times z^{-1}}{1 + (r_0 r_1 + r_1 r_2) z^{-1} + r_0 r_2 z^{-2}} \\ &= \frac{G z^{-1}}{1 + (r_0 r_1 + r_1 r_2) z^{-1} + r_0 r_2 z^{-2}} \\ &= \frac{G z^{-1}}{1 - a_1 z^{-1} - a_2 z^{-2}} \end{aligned}$$

Imperial College London

11

Multi-segment Vocal Tract Model

- Using: $\frac{1}{1+r} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix} \times z^{1/2} \begin{pmatrix} 1 & 0 \\ 0 & z^{-1} \end{pmatrix} = \frac{z^{1/2}}{1+r} \begin{pmatrix} 1 & -r z^{-1} \\ -r & z^{-1} \end{pmatrix}$

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

$$\begin{pmatrix} U_g \\ V_g \end{pmatrix} = \frac{z^{1/2 p}}{\prod_{k=0}^p (1+r_k)} \prod_{k=0}^{p-1} \begin{pmatrix} 1 & -r_k z^{-1} \\ -r_k & z^{-1} \end{pmatrix} \times \begin{pmatrix} 1 \\ -r_p \end{pmatrix} U_l$$

- This results in a transfer function of the form:

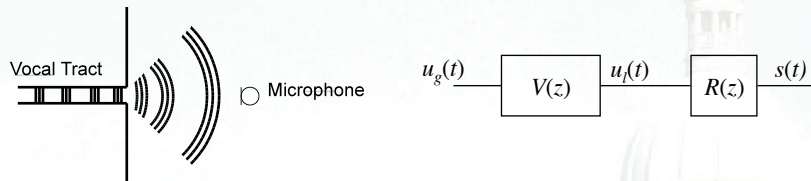
$$V(z) = \frac{U_l}{U_g} = \frac{G z^{-1/2 p}}{1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_p z^{-p}}$$

- G is a gain term
- $z^{-1/2 p}$ is the acoustic time delay along the vocal tract
- The denominator represents a p^{th} order all-pole filter

Imperial College London

12

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

$$R(z) = \frac{S(z)}{U_l(z)} = 1 - z^{-1} \Rightarrow |R(z)| = 2 \sin\left(\frac{\omega T}{2}\right)$$

Imperial College London

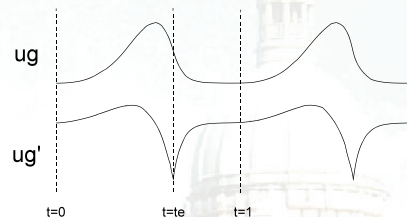
13

Spectrum of the Glottal Waveform

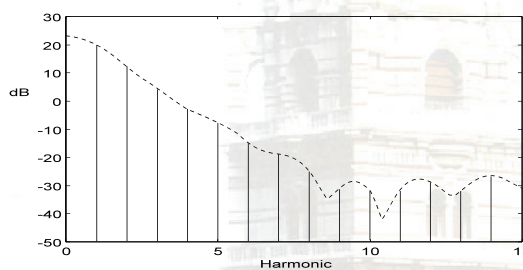
- "LF Model" (Liljencrants & Fant)

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

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



- Line Spectrum of u_g
(approx -12 dB/octave):

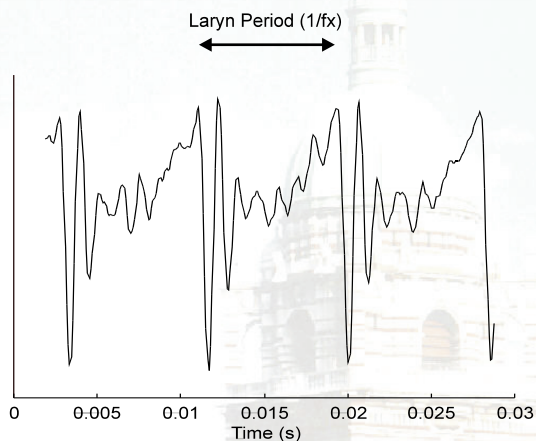


Imperial College London

14

Vowel Waveform

- Vowel /A/ from "part"
- Larynx frequency ~ 130 Hz
- First 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: /A/ vowel ("part")

