COMP 546

Lecture 19

Sound 2: frequency analysis

Tues. March 27, 2018
Speed of Sound

Sound travels at about 340 m/s, or 34 cm/ ms.

(This depends on temperature and other factors)
Wave equation

\[ \text{Pressure} = I_{\text{atm}} + I(X,Y,Z,t) \]

\( I(X,Y,Z,t) \) is not an arbitrary function.

Rather:

\[ \left( \frac{\partial^2}{\partial X^2} + \frac{\partial^2}{\partial Y^2} + \frac{\partial^2}{\partial Z^2} \right) I(X,Y,Z,t) = \frac{1}{v^2} \frac{\partial^2}{\partial t^2} I(X,Y,Z,t) \]

\[ \nu = 340 \text{ m/s} \]
The wave equation + boundary conditions give complicated shadow and reflection effects.

What happens when sound enters the ear?

plane wave + single slit

sea waves + islands
Musical sounds

(brief introduction)
Modes are \( \sin(\frac{\pi}{L} jx) \) where \( L \) is the length of the string, \( j \) is an integer.

Write one string displacement at \( t = 0 \) as sum of sines.

Example: guitar

\[
\begin{align*}
\text{\textbullet} & \quad (\quad) + (\quad) + (\quad) + (\quad) + e^{+c}
\end{align*}
\]
Physics says:

$$\omega = \frac{c}{L}$$

where constant $c$ depends on physical properties of string (mass density, tension)
Modes of a vibrating string each have fixed points which reduce the effective length.

Physics says:

$$\omega = \frac{c}{L} \quad \frac{2c}{L} \quad \frac{3c}{L} \quad \frac{4c}{L}$$
The temporal frequency $m \omega_0$ is called the $m$-th harmonic.
For stringed instruments, most of the sound is produced by vibrations of the instrument body (neck, front and back plates). http://www.acs.psu.edu/drussell/guitars/hummingbird.html

The lines in the sketches below are the nodal points. They don't move.

These are vibration modes, not harmonics. The guitar sound is a sum of these modes.
Difference of two frequencies $\omega_1$ and $\omega_2$:

$$\log_2 \frac{\omega_2}{\omega_1} \text{ octaves.}$$

e.g. 1 octave is a doubling of frequency.
(Western) Musical Notes

Each “octave” ABCDEFGA is divided into 12 “semitones”, separated into 1/12 octave.

C-D, D-E, F-G, G-A, A-B are two semitones each
E-F, B-C are one semitone each.
Q: How many semi-tones are there from $\omega_0$ to $\omega$?
Q: How many semi-tones are there from $\omega_0$ to $\omega$?

A: $12 \log_2 \frac{\omega}{\omega_0}$

\[ \omega \]

Fundamental frequency of note
The fundamental frequencies of successive notes define a *geometric progression*.

This is different from the harmonics of a vibrating string which define an *arithmetic progression*. 
Speech Sounds

- alveolar ridge
- hard palate
- oral cavity
- nasal cavity
- teeth
- lips
- tongue
- pharynx
- epiglottis
- esophagus
- larynx (vocal cords)
What determines speech sounds?

- voiced vs. unvoiced

  ‘zzzz’ vs. ‘ssss’, ‘vvvv’ vs. ‘ffff’

- articulators (jaw, tongue, lips)

Voiced sounds are produced by “glottal pulses”.

\[ n_{\text{glottal}} \sum_{j=0}^{\infty} g(t - j T_{\text{glottal}}) \]
Exercise 16 Q7.

\[ g(t - t_0) = g(t) * \delta(t - t_0) \]
Voiced sounds are produced by “glottal pulses”.

\[
\sum_{j=0}^{n_{\text{glottal}}} g(t - j T_{\text{glottal}}) = g(t) \ast \sum_{j=0}^{n_{\text{glottal}}} \delta(t - j T_{\text{glottal}})
\]
\[ n_{glottal} \sum_{j=0}^{n_{glottal}} g(t - j T_{glottal}) \]

decrease \( T_{glottal} \) by increasing tension in vocal cords

\( \equiv \) increase frequency of pulses
Let $a(t)$ be the impulse response function of the articulators. 
(jaw, tongue, lips)

$$I(t) = a(t) \ast g(t) \ast \sum_{j=0}^{n_{glottal}} \delta(t - j T_{glottal})$$
Q: What is the Fourier transform of

\[ I(t) = a(t) * g(t) * \sum_{j=0}^{n_{\text{pulse}}-1} \delta(t - j T_j) \]
Q: What is the Fourier transform of

\[ I(t) = a(t) * g(t) * \sum_{j=0}^{\text{n_pulse}-1} \delta(t - jT_j) \] ?

A:

\[ \hat{I}(\omega) = \hat{a}(\omega) \cdot \hat{g}(\omega) \cdot \mathcal{F} \left\{ \sum_{j=0}^{\text{n_pulse}-1} \delta(t - jT_j) \right\} \]
Glottal Pulse shape

$g(t)$

$|\hat{g}(\omega)|$

Glottal pulse $g(t)$ has a shape roughly between a Gaussian and impulse.
Oral and nasal cavity have resonant modes of vibration, like air cavity in guitar does.
Time domain

\[ a(t) * g(t) \]

Temporal frequency domain

\[ |\hat{g}(\omega)| \]

\[ |\hat{a}(\omega)| \]

Peaks are called “formants”
\[ F \sum_{j=0}^{n_{glottal}} \delta(t - j T_{glottal}) = ? \]

\( T_g \) is the period of the glottal pulse train.

The pulse train has \( n_{glottal} \) pulses in \( T \) time steps, i.e.

\[ T_{glottal} \ n_{glottal} = T. \]

Assume that the Fourier transform is taken over \( T \) samples.
Assignment 3: Show

\[
\mathcal{F} \sum_{j=0}^{n_{glottal}-1} \delta(t - j T_{glottal}) = n_{glottal} \sum_{m=0}^{T_{glottal}-1} \delta(\omega - m n_{glottal})
\]
Units of temporal frequency $\omega$

$T_{glottal}$ is the period of the glottal pulse train.

$n_{glottal}$ pulses in $T$ time samples.

To convert ‘pulses per T time samples’ to ‘pulses per second’, we multiply by ‘time samples per second’.

High quality audio uses 44,100 samples per second.
$n_{glottal}$ is the fundamental frequency of the voiced sound. It determines the "pitch".

Adult males : 100-150 Hz
Adult females : 150-250 Hz
Children: over 250 Hz
\[ \omega_0 = 100 \text{ Hz} \]

- glottal pulse spectrum
- "formants"
- sound spectrum

\[ \omega_0 = 200 \text{ Hz} \]
Voiced vowel sounds

formants

[a]

[i]

[u]

after Fant
Unvoiced sounds

*noise instead of glottal pulses*

\[ I(t) = a(t) \ast n(t) \]
Unvoiced sounds

noise instead of glottal pulses

\[ I(t) = a(t) * n(t) \]

\[ \hat{I}(\omega) = \hat{a}(\omega) \hat{n}(\omega) \]

Flat amplitude spectrum on average (‘white noise’)
Consonants

Restrict flow of air by moving tongue, lips into contact with the teeth & palate.

Fricatives
- voiced    z, v, zh, th (the)
- unvoiced  ?

Stops
- voiced    b, d, g
- unvoiced  ?

Nasals (closed mouth)
- m, n, ng
Consonants

Restrict flow of air by moving tongue, lips into contact with the teeth & palate.

Fricatives
- voiced       z, v, zh, th (the)
- unvoiced     s, f, sh, th (theta)

Stops
- voiced       b, d, g
- unvoiced     p, t, k

Nasals (closed mouth)
- m, n, ng
The amount of each frequency component can vary over time. To analyze this, we partition time into $B$ blocks.

We then take the Fourier transform of each block.
Spectrogram

Partition a sound signal into $B$ blocks of $T$ samples each (i.e. the sound has $BT$ samples in total).

Take the Fourier transform of each block.
Spectrogram

Partition a sound signal into $B$ blocks of $T$ samples each (i.e. the sound has $BT$ samples in total).

Take the Fourier transform of each block.

Let $b$ be the block number, and $\omega$ units be cycles per block.

$$\hat{I}(b, \omega) = \sum_{t=0}^{T-1} I( b T + t) e^{-i \frac{2\pi}{T} \omega t}$$
\[ \omega_0 = \frac{44000}{T} \]
e.g. \( T = 512 \text{ samples (12 ms)}, \quad \omega_0 = 86 \text{ Hz} \)
You cannot simultaneously localize the frequency and the time. This is a fundamental tradeoff. We have seen it before (recall the Gaussian).

e.g. $T = 512$ samples (12 ms), $\omega_0 = 86$ Hz

$T = 2048$ samples (48 ms), $\omega_0 = 21$ Hz
Narrowband

(good frequency resolution, poor temporal resolution ... ~50ms)

Wideband

(poor frequency resolution, good temporal resolution)
Examples: Spectrograms of 10 vowel sounds