COMP 546

Lecture 20

Sound 2: frequency analysis

Tues. March 30, 2017
Plucking a guitar string

Write displacement as a sum of sines (similar to a Fourier transform). Each "mode" has its own vibration frequency (cycles/sec).

\[ \text{Displacement at } t = 0 \]

Modes are \( \sin\left(\frac{\pi}{L} jx\right) \) where \( L \) is the length of the string, \( j \) is an integer.
Physics says that the temporal frequency $\omega$ of vibration of string of length $L$ is

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

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

\[ \omega = \frac{c}{L} \quad \frac{2c}{L} \quad \frac{3c}{L} \quad \frac{4c}{L} \quad \text{etc.} \]
The temporal frequency $m \omega_0$ is called the $m$-th harmonic.
\( \omega \) - cycles per second

\( v = 340 \) metres/sec

\[ \lambda = \frac{v}{\omega} \text{ metres/cycle} \]

(wavelength)

(Don't confuse \( \lambda \) with \( L \) from slides above.)
For stringed instruments, most of the sound is produced by vibrations of the instrument body (neck, front and back plates). The body has its own vibration modes.

http://www.acs.psu.edu/drussell/guitars/hummingbird.html

The lines in the sketches below are the nodal points of the vibration (i.e. these points don't move).
Recall that two frequencies $\omega_1$ and $\omega_2$ are separated by

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

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

Each “octave” ABCDEFGA... is divided into 12 semitones e.g. piano keyboard

Semitones are spaced every \(\frac{1}{12}\) octave.

\[\cdots C \ D \ E \ F \ G \ A \ B \ C \ D \ E \ \cdots\]

C-D, D-E, F-G, G-A, A-B is 2 semitones (\(\frac{2}{12}\) octave)

E-F, B-C is 1 semitone (\(\frac{1}{12}\) octave)
Q: How many notes to go from $w_0$ to $w$?

A: 

$$n = 12 \log_2 \frac{w}{w_0}$$
\( \omega_0 \) is the fundamental frequency of some 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.*
fundamental
Speech Sounds
What determines speech sounds?

- **Voiced vs. Unvoiced**
  Tensing vs. relaxing the vocal chords
  e.g. ‘zzz’, ‘vvv’ vs. ‘sss’, ‘fff’
  (whispering is unvoiced)

- **Articulators** (jaw, tongue, lips)
  Their positions determine the shape of the oral cavity which amplifies some frequencies and attenuates others.
  e.g. vowels “aaaa”, “eeeee”, “oooo”
Voiced sounds produced by glottal pulses.

$$\sum_{j=0}^{n-1} g(t - jT_g) = g(t) \ast \sum_{j=0}^{n-1} \delta(t - jT_g)$$
decreasing $T_g$ (increasing tension in vocal cords)

$\Rightarrow$ higher frequency of pulses

$\Rightarrow$ ? (described below)
Let $a(t)$ be the impulse response function of the articulators, so a glottal pulse is transformed to $a(t) * g(t)$ and a glottal pulse train becomes:

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

\[ I(t) = a(t) \ast g(t) \ast \sum_{j=0}^{n_{\text{pulse}}} \delta(t - j T_j) \]

A:

\[ \hat{I}(\omega) = \hat{a}(\omega) \cdot \hat{g}(\omega) \cdot F \left\{ \sum_{j=0}^{n_{\text{pulse}}} \delta(t - j T_j) \right\} \]
Glottal pulse $g(t)$ has a shape roughly between a Gaussian and impulse.
What about $a(t)$?

Like the cavity of a guitar, the oral and nasal cavity have resonant modes of vibration (of the air).
\[ g(t) \]

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

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

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

peaks are called **formants**
$T_g$ is the period of the glottal pulse train.

The pulse train has $n_{\text{pulse}}$ pulses in $T$ time steps, i.e.

$$T_g \ n_{\text{pulse}} = T.$$

Assume that the Fourier transform is taken over $T$ samples.
Exercise: Show

\[ F \sum_{j=1}^{n_{\text{pulse}}} \delta(t - jT_g) = n_{\text{pulse}} \sum_{m=0}^{T_g-1} \delta(\omega - mn_{\text{pulse}}) \]
\( T_g \) is the period of the glottal pulse train. The pulse train has \( n_{\text{pulse}} \) pulses in \( T \) time steps.

To convert ‘pulses per \( T \) time samples’ to ‘pulses per second’, we multiply by ‘time samples per second’. High quality audio uses 44,000 samples per second.
$n_{pulse}$ is the fundamental frequency of the voiced sound. It determines the "pitch".

Very roughly....

Children: over 250 Hz  
Adult females: 150-250 Hz  
Adult males: 100-150
\[ \hat{g}(\omega) = \sum_{n} \delta(\omega - \frac{n}{T}) \cdot G_n(\omega) \]

Where:
- \( \hat{g}(\omega) \) represents the Fourier transform of the impulse response.
- \( G_n(\omega) \) is the source spectrum at frequency \( \omega \).
- \( n/T \) are the frequencies where impulses occur.

\[ \hat{g}(\omega) \times \hat{h}(\omega) = \hat{y}(\omega) \]

Where:
- \( \hat{h}(\omega) \) is the filter function.
- \( \hat{y}(\omega) \) is the output energy spectrum.

\( F_\Omega = 100 \text{ Hz} \)
\( F_\Omega = 200 \text{ Hz} \)
Vowel sounds

\begin{itemize}
  \item a
  \item i
  \item u
\end{itemize}

formants

\begin{itemize}
  \item [AH] as in "FATHER"
  \item [EE] as in "HEED"
  \item [OO] as in "POOL"
\end{itemize}
Unvoiced sounds
noise instead of glottal pulses

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

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

Flat on average
(recall ‘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
Speech production:
switch between voiced and unvoiced sounds (within words)
Spectrograms

The amount of each frequency component can vary over time. To analyze this, we partition time into $B$ blocks.

$0 \ 1 \ 2 \ 3 \ 4 \ 5 \ 6 \ \ldots$

We then take the Fourier transform of each block.
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.

\[
\hat{I}(b, \omega) = \sum_{t=0}^{T-1} I(bT + t) e^{-i \frac{2\pi}{T} \omega t}
\]
$\omega_0 = \frac{44000}{T}$

Cycles per second (Hz)

Time (samples)
e.g. \( T = 512 \) samples (12 ms) \( \omega_0 = 86 \) Hz \hspace{1cm} T = 2048 samples (48 ms) \( \omega_0 = 21 \) Hz

Note that you cannot simultaneously localize the frequency and the time. This is a fundamental tradeoff. We have seen it before (recall the Gaussian).

FYI, this is essentially the Heisenberg Uncertainty principle from physics.
Narrowband
(good frequency resolution, poor temporal resolution)

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