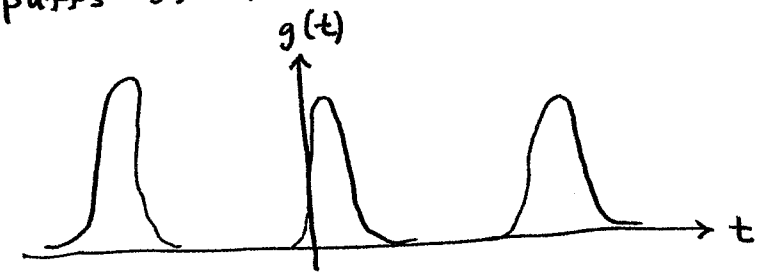
Speech and Fourier Series

Speech waveform (vowel sound)

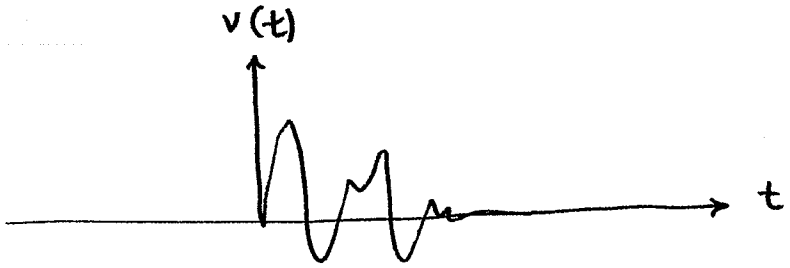
vocal cords slap (few hundred times/sec)

⇒ puffs of air



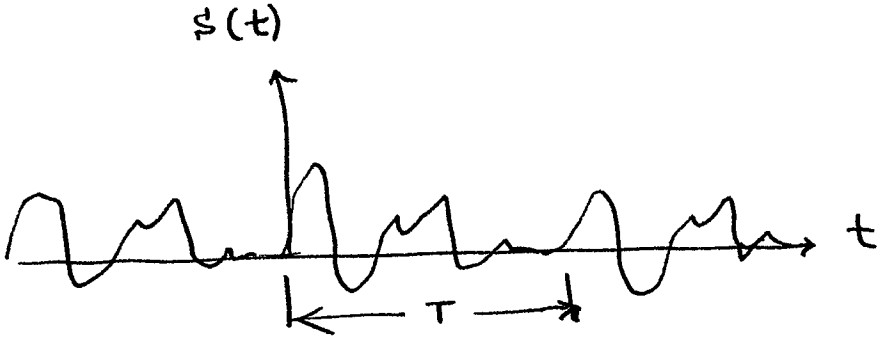
vocal tract (throat, mouth, lips, etc)

impulse response has resonant peaks in freq domain, sounds like chord



Speech $s(t) = g(t) \otimes v(t)$ convolution

repetitive waveform (from glottal pulses)



Can use Fourier series for $s(t)$.

$$s(t) = a_0 + \sum_{n=1}^{\infty} a_n \cos(2\pi n f_0 t) + b_n \sin(2\pi n f_0 t)$$

$$a_n = \frac{2}{T} \int_0^T s(t) \cos(2\pi n f_0 t)\, dt$$

$$b_n = \frac{2}{T} \int_0^T s(t) \sin(2\pi n f_0 t)\, dt$$

Problem: Need analog computer or filter bank with infinite number of filters to compute a_n, b_n.

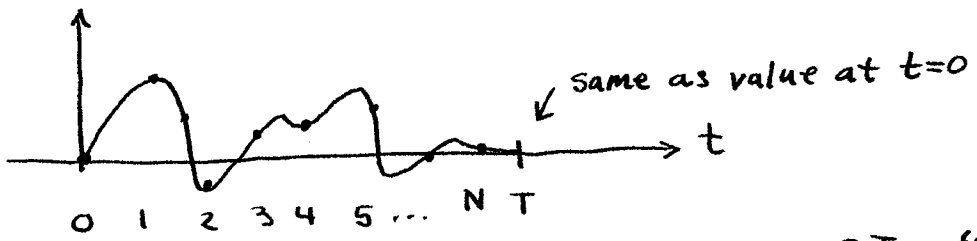
Observe: It turns out a_n, b_n decay as n gets larger. We can truncate series at some $n = N$. Gives $\hat{s}(t)$.

Q. Effect on accuracy of $\hat{s}(t)$?

A. $\hat{s}(t)$ smoother than $s(t)$. High freq's give fine details. Truncated series is like smoothed signal or *low-passed* signal.

Problem: Still need filters. Expensive, awkward.

Sol'n: Use sampled signal. Use N samples for N coeff's:



Sample spacing $= \frac{T}{N}$ so $t = 0, \frac{T}{N}, \frac{2T}{N}, \frac{3T}{N}, \ldots \frac{(N-1)T}{N}$

$$t = m\frac{T}{N} \quad \text{for } m = 0, \dots, N-1$$

We have ↙ truncated series

$$s\left(m\frac{T}{N}\right) = a_0 + \sum_{n=1}^{N} a_n \cos\left(2\pi n f_0 m\frac{T}{N}\right) + b_n \sin\left(2\pi n f_0 m\frac{T}{N}\right)$$

($f_0 = \frac{1}{T}$)

problem: We have $1 + 2N$ coeff's to be determined from N samples. Need to truncated series sooner.

$$s\left(m\frac{T}{N}\right) = a_0 + \sum_{n=1}^{(N-1)/2} a_n \cos(2\pi nm) + b_n \sin(2\pi nm)$$

We have N eq'ns in N unknowns. Can solve exactly.

Matrix form:

$$
\begin{array}{ll}
m=1 & s(0) \\
m=1 & s\left(\frac{T}{N}\right) = \\
m=2 & s\left(2\frac{T}{N}\right) =
\end{array}
\begin{bmatrix}
1 & 1 & 0 & 1 & \cdots \\
1 & \cos\frac{2\pi}{N} & \sin\frac{2\pi}{N} & \cos\frac{4\pi}{N} & \cdots \\
1 & \cos\frac{4\pi}{N} & \sin\frac{4\pi}{N} & \cos\frac{8\pi}{N} & \cdots \\
\vdots & & & \ddots & \\
1 & \cos\left(2\pi \frac{N-1}{2}\frac{N-1}{N}\right) & \cdots & &
\end{bmatrix}
\begin{bmatrix}
a_0 \\ a_1 \\ b_1 \\ a_2 \\ b_2 \\ \vdots \\ a_{\frac{N-1}{2}} \\ b_{\frac{N-1}{2}}
\end{bmatrix}
$$

(column labels: $n=1$, $n=2$ …)

or

$$\vec{s} = A \vec{a}$$

Fourier coeff's: $\vec{a} = A^{-1}\vec{s}$

FFT ≡ Fast Fourier Transform = Efficient calc of A^{-1}

Note: A^{-1} can be written down and looks very much like A.

FFT exploits redundancy such as

$$\cos\left(\frac{2\pi}{N}\right) = \cos\left(2\pi - \frac{2\pi}{N}\right)$$

$$= \cos\left(2\pi\,\frac{N-1}{N}\right)$$

Compute only once. Can recursively exploit redundancy.

problem: High freq's are in signal. When sampled, high freq (above samp rate /2) looks exactly like low freq (below samp rate /2).

This is *aliasing*.

$$\cos\left(2\pi\, n\,\frac{mt}{N}\right) = \cos\left(\left[2\pi N - 2\pi\frac{n}{N}\right]t\right)$$

$$(\cos 2\pi Nt)(\cos 2\pi\frac{n}{N}t)$$

$$+ (\sin 2\pi Nt)(\sin 2\pi\frac{n}{N}t)$$

for $t = \frac{m}{N}$ we get equality

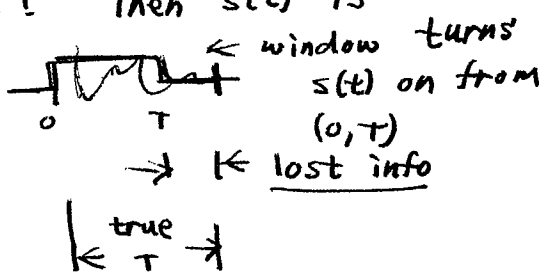
soln: low-pass filter signal before sampling

prob: loss of detail.

But speech < 20kHz (very little energy above 20k)

so sampling not too bad.

concern: What if window $(0,T)$ ~~not~~ ≠ one period of waveform? Then $s(t)$ is



like waveform *

soln: Try to track pitch period.

prob: glottal pulse rate wobbles, (croaky sound).
Difficult to track pitch period.

concern: If glottal pulse wobbles, T changes.
Then a_n for one window ≠ a_n for next window.

Conclusion: Speech is difficult prob to solve with Fourier techniques.

Note: Typical approach is to use fixed-length windows (e.g. 30 msec) and overlap them.

prob: Many speech sounds are <u>not</u> repetitive.
p, t, d are popping sounds.
Better to analyze p, t, d in time domain, although info in FFT = info in sampled signal.
IFFT = sampled signal * invertible matrix