TTIC 31110
Speech Technologies

April 14, 2020

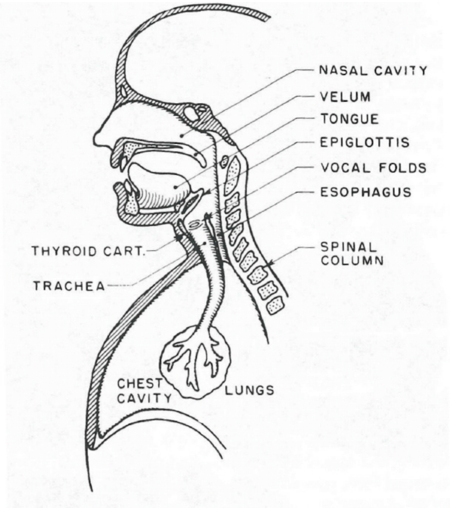
Announcements

- Survey responses available (see Canvas announcement), please respond if you haven't already
- Tutorial 1 slides are posted
- HW1 due Friday 4/17 7pm

Questions?

Outline

Recap: Speech production

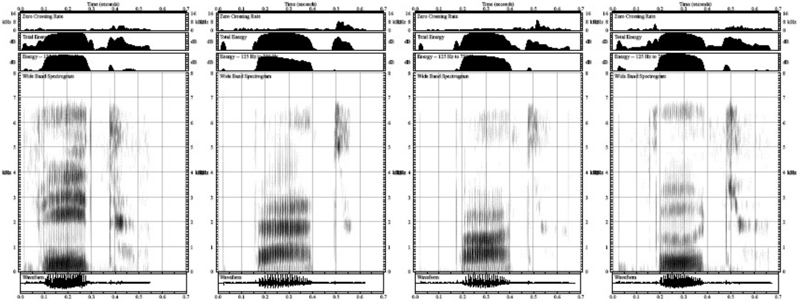


(fig. from [Flanagan 1972])

- *Source-filter model*: Vocal tract acts as a filter, modulating the spectrum of the source signal
- Source is air pressure waveform either from glottis (e.g. vowels) or from another constriction (most consonants)

Demo!

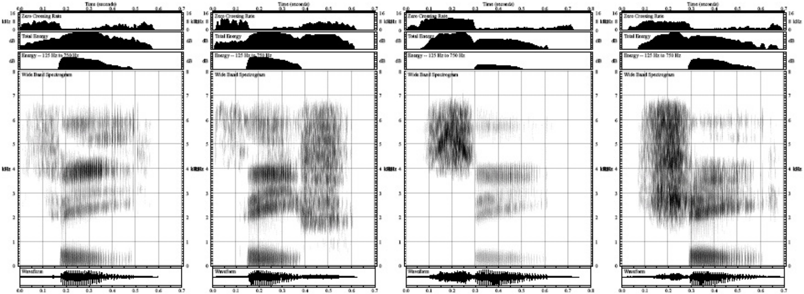
Recap: Spectrograms of the cardinal vowels



beet bat bott boot

(figs. from MIT 6.345 Spring '03, OpenCourseWare `http://ocw.mit.edu`)

Recap: Spectrograms of English voiceless fricatives

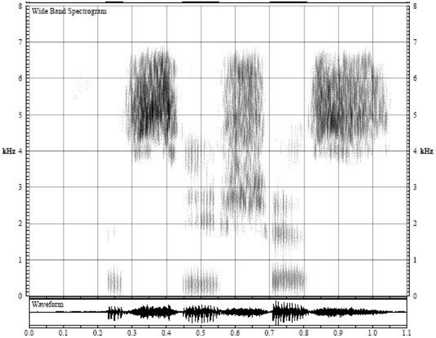


fee thief see she

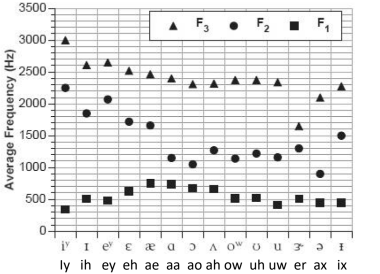
(figs. from MIT 6.345 Spring '03, OpenCourseWare `http://ocw.mit.edu`)

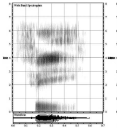
What is this word? (Hint: It contains only vowels and voiceless fricatives)

Spectrogram of target word

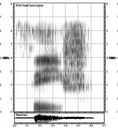


For reference: Vowel formants, fricative spectrograms

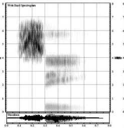




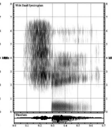
fee



thief



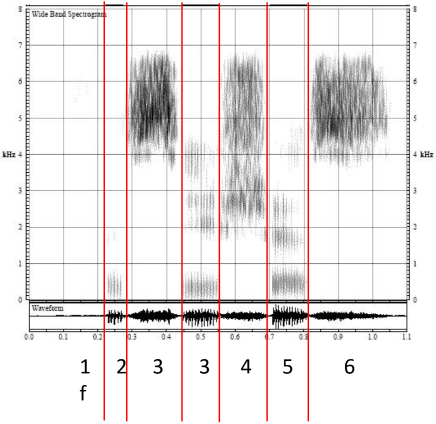
see



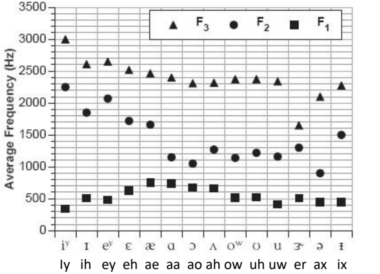
she

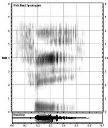
What is this word? (Hint: It contains only vowels and voiceless fricatives)

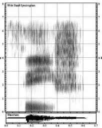
Spectrogram of target word

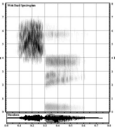


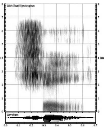
For reference: Vowel formants, fricative spectrograms




fee

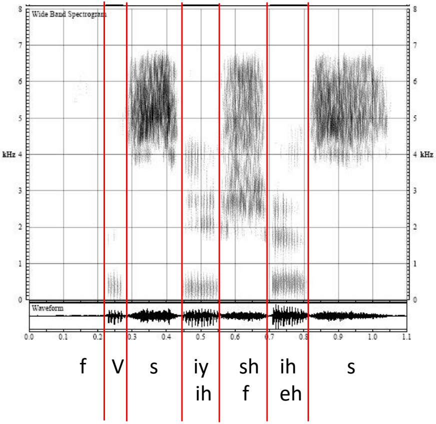

thief


see

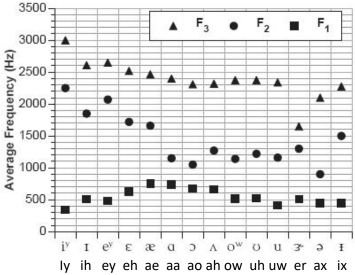

she

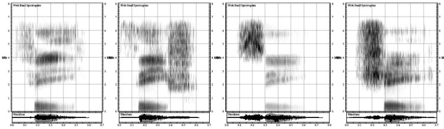
What is this word? (Hint: It contains only vowels and voiceless fricatives)

Spectrogram of target word



For reference: Vowel formants, fricative spectrograms





fee thief see she

What do speech production/perception tell us about acoustic features?

Production:

- Separate the source from the filter information
- Only use the filter information for speech recognition (for non-tonal languages, like English; we'll get back to other languages later)

Perception:

- Work in the frequency domain
- Warp the frequency scale as humans do

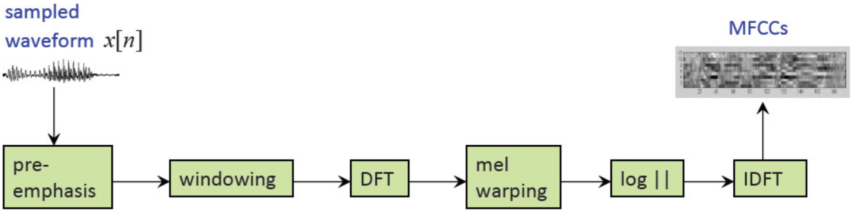
Aside: Acoustic features in the neural networks era

In recent neural network-based approaches, 3 types of acoustic features are common:

- No signal processing, just use the raw signal: Works best with very large amounts of data
- (Lightly post-processed) spectrogram: Works in many typical settings
- Traditional signal processing-based features: Work best in low-data settings

Example acoustic features: MFCCs

- MFCCs = mel-frequency cepstral coefficients [Davis & Mermelstein 1980]
- Most popular type of signal processing-based features
- Many of the signal processing steps are also used in computing spectrograms or other features



Sampling

All digital signal processing starts with sampling

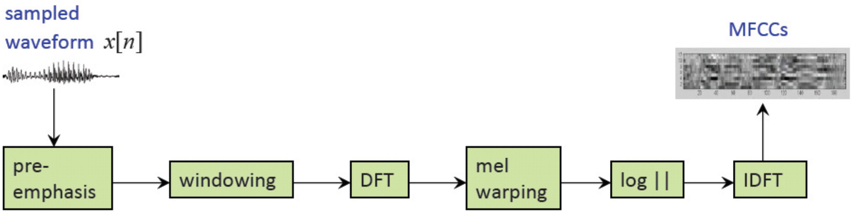
- Sampling $=$ measuring a continuous-time signal at some fixed sampling rate $f = 1/T$
- Rate f measured in Hz, sampling period T in seconds
- Sampled signal is $x[n] = x(nT)$

Sampling

What sampling rate f should we use?

- Ideally, should be higher than the *Nyquist rate* = twice the highest frequency in the signal
- (Roughly, this is so that we sample each period of each frequency component at least twice)
- Most useful speech information is in [0,8000 Hz] $\rightarrow$ sample at $>$ 16 kHz
- Don't always have a choice, e.g. land-line phone lines transmit only up to $\sim$4 kHz ($f = 8$ kHz)
- If the signal has any higher frequencies than $1/2$ the sampling rate, then they are filtered out before sampling

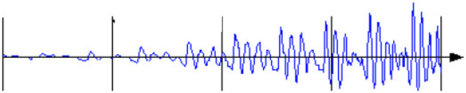
Pre-emphasis



- For most speech sounds, the higher frequencies have much lower amplitudes than the low frequencies
- Pre-emphasis equalizes, to some extent, the amplitudes at the low and high ends: $x_{pre}[n] = x[n] - ax[n-1]$ for positive a

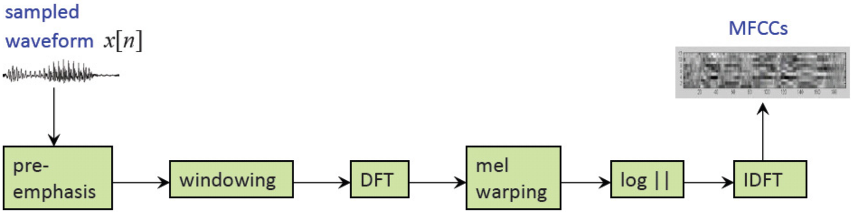
Windowing

- Speech is non-stationary, so we extract spectrum over short windows (*frames*) rather than entire signal



- What length window?
  - Short enough to assume stationarity within the window
  - Long enough for good frequency resolution ("Uncertainty Principle")

MFCCs


sampled
waveform $x[n]$
pre-
emphasis
windowing
DFT
mel
warping
log | |
IDFT
MFCCs

Discrete Fourier transform

- For each frame of length N samples starting at sample t, the spectrum is computed using the discrete Fourier transform (DFT):

$$X[k] = \sum_{n=t}^{t+N-1} x[n] e^{-j2\pi kn/N}, \quad k = 0, \ldots, N-1$$

- $X[k]$ is the value of the spectrum at the k^{th} frequency
- Reminder (?): $e^{ja} = \cos(a) + j\sin(a)$
- Equivalently, we can consider $k = -N/2, \ldots, 0, \ldots, N/2$
- $X[k]$ is in general complex-valued, but we will only use its magnitude

Discrete Fourier transform

- For each frame of length N samples starting at sample t, the spectrum is computed using the discrete Fourier transform (DFT):
$$X[k] = \sum_{n=t}^{t+N-1} x[n]e^{-j2\pi kn/N}, \quad k = 0, \dots, N-1$$
- $X[k]$ is the value of the spectrum at the k^{th} frequency
- In practice the *fast Fourier transform* (FFT) algorithm is used with a window length $M = 2^m$ for some m
- After doing this for all frames, the result is a *spectrogram*
- Typically, the frequency axis is then warped to the mel scale, giving a "mel spectrogram"

Discrete Fourier transform

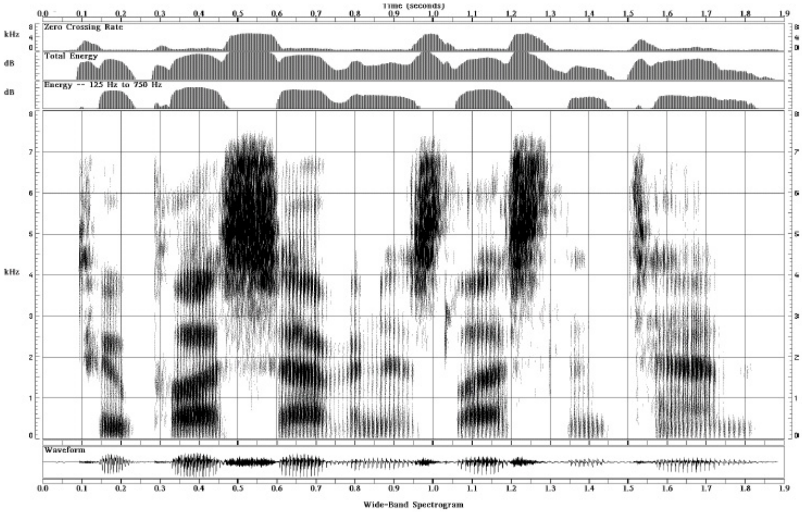
- Note on units:
  - Each time sample n corresponds to a time in seconds
  - Each frequency sample k corresponds to a frequency $f(k) = \frac{k}{N}R$ in Hz, where R is the sampling rate
- Important properties:
  - Spectra of real signals have magnitude that is symmetric about $k = 0$
  - Spectra of periodic signals with period λ seconds consist of delta functions at multiples of the fundamental frequency $\frac{1}{\lambda}$ Hz

Periodic signals and their spectra

Vowels and many other **voiced** phones are examples of periodic signals

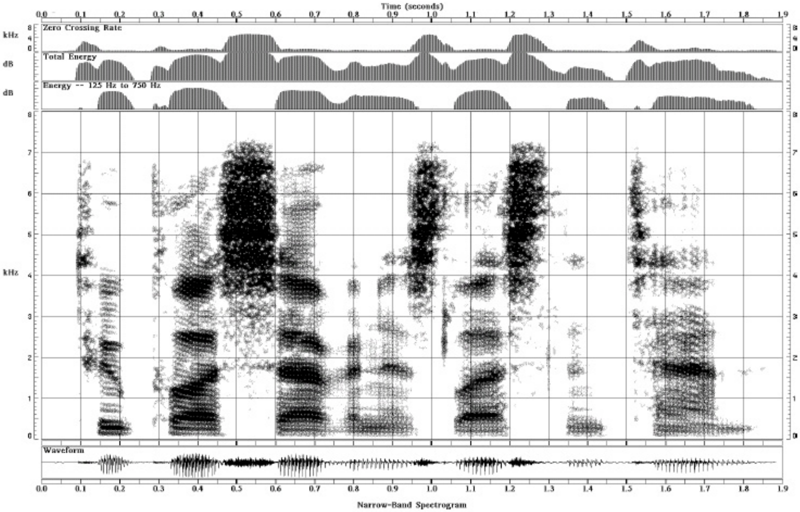
- Spectra of periodic signals with period λ seconds consist of delta functions at multiples of the fundamental frequency $\frac{1}{\lambda}$ Hz
- The fundamental frequency is often denoted $F0, F_0$, or f_0
- Some phones, like voiced fricatives ([z, v, dh]), include a periodic component (with some fundamental frequency) and an aperiodic component
- **Pitch** is the perceptual correlate of the fundamental frequency
- We will use the terms **pitch** and **fundamental frequency** interchangeably

A wideband (short-window) spectrogram



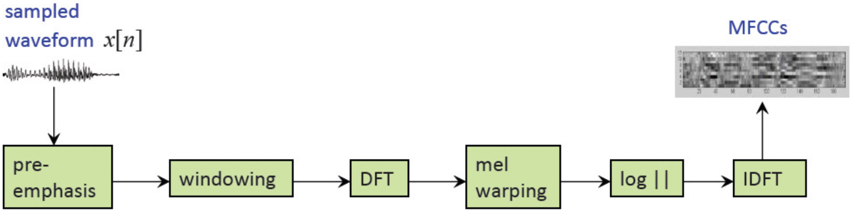
Two plus seven is less than ten

A narrowband (long-window) spectrogram



Two plus seven is less than ten

Log spectra



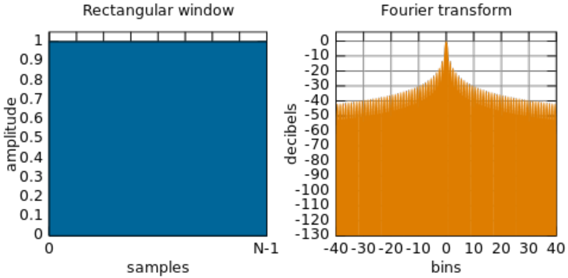
- The source-filter theory says that $X[k] = S[k]F[k]$, where $S[k]$ = source spectrum and $F[k]$ = vocal tract filter
- We are mostly interested in the filter (vocal tract shape)
- We take the *log* magnitude: compresses the dynamic range, will allow us to separate the source from the filter
- We typically convert to a decibel (dB) scale, $X_{dB}[k] = 20 \log_{10} |X[k]|$

More on windowing

- Windowing is equivalent to multiplying the signal by a function $w[n]$, i.e. $x_w[n] = x[n]w[n]$
- E.g. rectangular window with length N and starting at time t:
$$w_r[n] = \begin{cases} 1 & t \leq n \leq t + N - 1 \\ 0 & \text{otherwise} \end{cases}$$
- The spectrum of two multiplied signals is the *convolution* of their spectra: $X_w[k] = X[k] * W[k] = \frac{1}{N} \sum_l X[l]W[k-l]$
- Properties of convolution (can use to derive any spectrum of a windowed signal):
  - If $x[n]$ contains a single frequency F, $X[k] = \delta[k-F]$, then $X_w[k] = W[k-F]$
  - Convolution is a *linear* operation:
    $X[k] = A[k] + B[k] \implies$
    $X[k] * W[k] = A[k] * W[k] + B[k] * W[k]$
  - How does spectrum of a windowed periodic signal look?

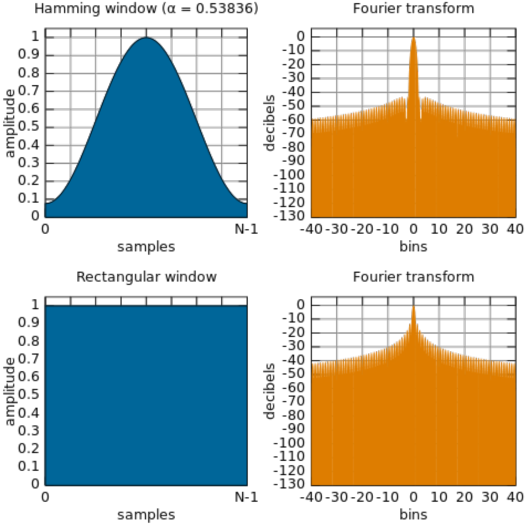
Windowing (cont'd)

The spectrum of the rectangular window $W_r[k]$:

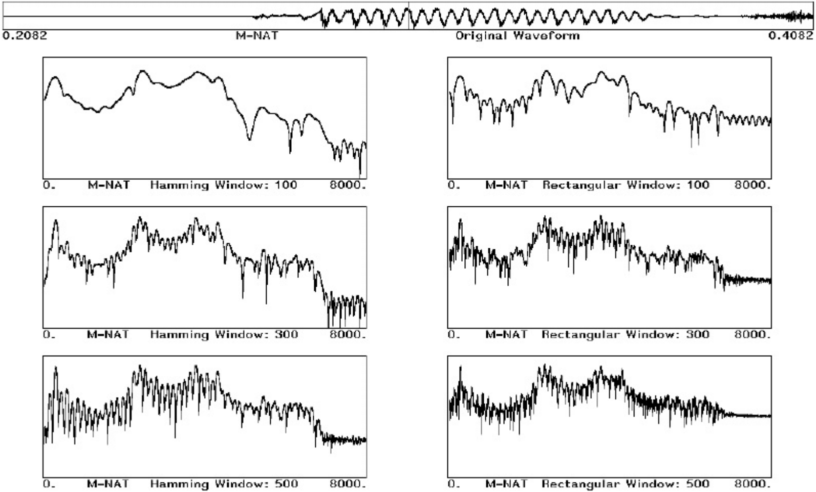


A better window?

Hamming window: $w_h[n] = .54 - .46cos(2\pi n/N)$



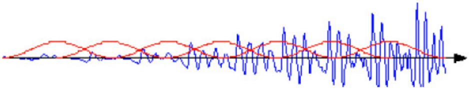
Hamming vs. rectangular windows



Note: Periodic signal with period $\lambda \Longrightarrow$ spectrum with peaks at multiples of the pitch (fundamental frequency) $f = 1/\lambda$.

Spectrograms: Additional notes

Overlapping windows: Typically used to smooth out edge effects

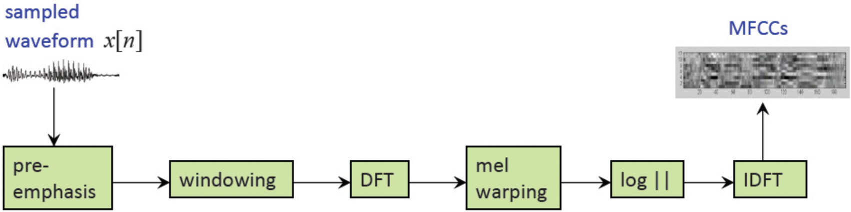


Some typical settings:

- Frame size (window length) = 25ms
- Frame shift (time between successive windows) = 10ms

Depending on details of the window, frame size, and frame shift, it **may be possible** to invert a spectrogram to reproduce the waveform

MFCCs


sampled
waveform $x[n]$
MFCCs
pre-
emphasis
windowing
DFT
mel
warping
log | |
IDFT

Inverse DFT

- Finally, we take the inverse DFT to obtain the *cepstrum*:
$$c[n] = \frac{1}{N} \sum_{k=0}^{N-1} X_{dB}[k] e^{j2\pi kn/N}, \quad n = 0, \ldots, N-1$$
- n is called "quefrency"
- The log spectrum was symmetric, so this is equivalent to taking another DFT
- The cepstrum is the "spectrum of the (log magnitude) spectrum"!