

Lec 28, 10/31/01 Wed

Speech Processing.

We will now address another very important application, namely, processing of speech signals. This is a very rich area, which has many interesting problems. During the next couple of weeks, we'll be applying some of the things we studied about systems, frequency analysis, and random processes, to some of these problems.

Three main areas of interest:

- speech synthesis
- speech recognition
- speech ~~compression~~ coding

The goal of speech synthesis is to develop a machine that accepts a piece of text as input and converts it to speech, which would be as intelligible and as natural-sounding as if spoken by a person. For example, you may want your computer to talk to you, ~~and probably~~

Speech recognition is important because you may want to talk to your computer. The ultimate goal here is to produce a system which can recognize, with human accuracy, speech from any speaker of a given language. ~~Some other~~ ~~applications are~~ One application is that you dictate to your computer instead of typing. Another one is automated telephone answering systems, which recognize vocal commands to determine the next action.

The goal of speech ~~compression~~ ^{coding or compression} is to ~~represent~~ reliably represent speech signals with as few bits as possible. This is very important for storage and transmission. When you store ~~many signals~~ a lot of data, you would like to conserve space as much as possible; when you transmit, you want to use as few bits as you can to transmit as much information as you can. Efficient coding of speech turns out to be possible, because speech is ~~by no means a white noise process~~ redundant. The simplest example is that if you are saying "aa" for a second, you don't need 8000 samples to represent that.

The physiology of speech production is rather intricate and interesting, but for us the most important distinction will be voiced vs unvoiced speech.

Voiced and Unvoiced Speech

Voiced sounds, ^{like 'a', 'e', 'o'} are essentially due to vibrations of the vocal cords, ~~are~~ and are oscillatory. Therefore, over short periods of time, they are well modelled by sums of sinusoids. And next week, as a matter of fact, we'll ~~look at~~ ^{look at} that time Fourier transform for speech processing.

- voiced: periodic over short ~~time~~ ^{time intervals}
- unvoiced: ~~other~~ noise, e.g. "sh" ^{e.g. "bear"}

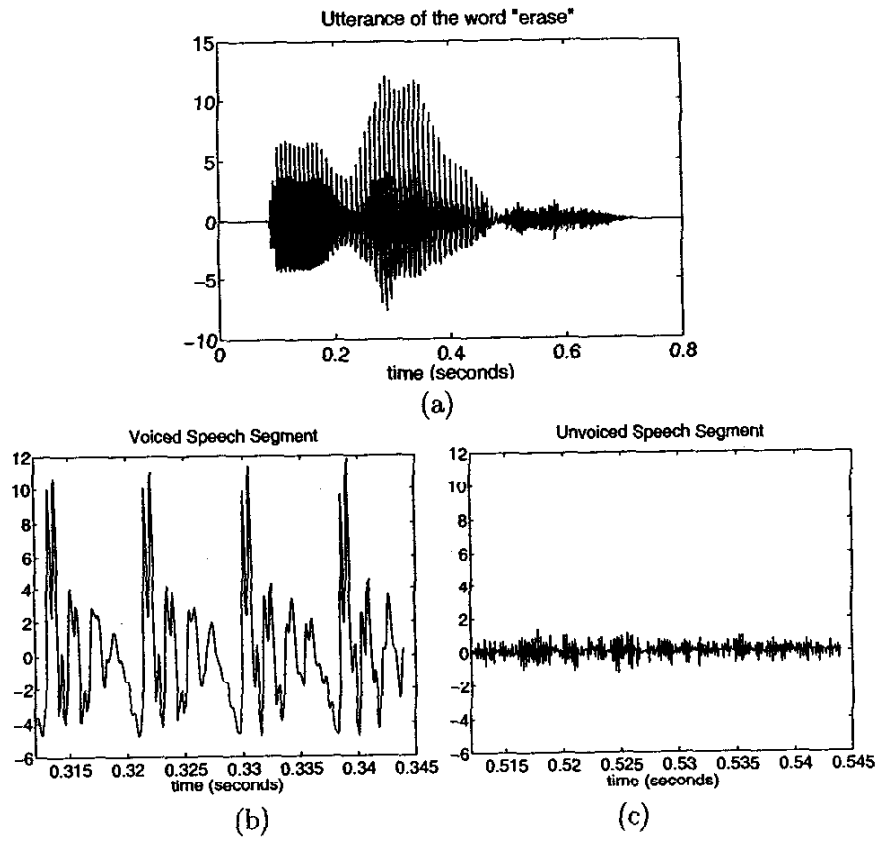


Figure 2: (a) Utterance of the word "erase". (b) Voiced segment. (c) Unvoiced segment.

(all speech applications)
For coding, it is important to distinguish between voiced and unvoiced speech. Lab 9 will consider two simple but effective methods for doing that. (Lab 9 wk 1) ④

- Short-time power function
split the speech signal $x(n)$ into blocks of 10-20 ms, and calculate the power within each block:

$$P_{av} = \frac{1}{L} \sum_{n=1}^L x^2(n)$$

Typically, $P_{av, \text{voiced}} > P_{av, \text{unvoiced}}$

• Zero crossing rate:

"the signal $x(n)$ has a zero crossing at n_0 "
means

$$x(n_0)x(n_0+1) < 0$$

Unvoiced signals oscillate much faster

⇒ have a much higher rate of zero crossings.

Source-Filter Model of Speech Production.

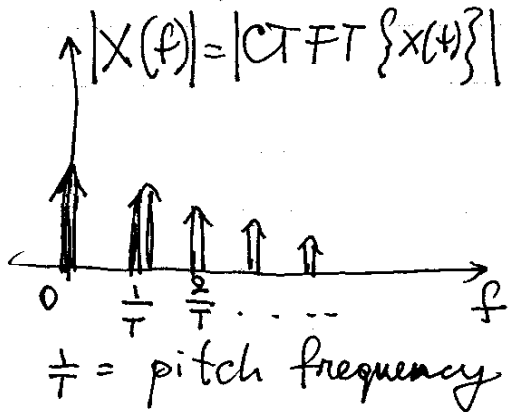
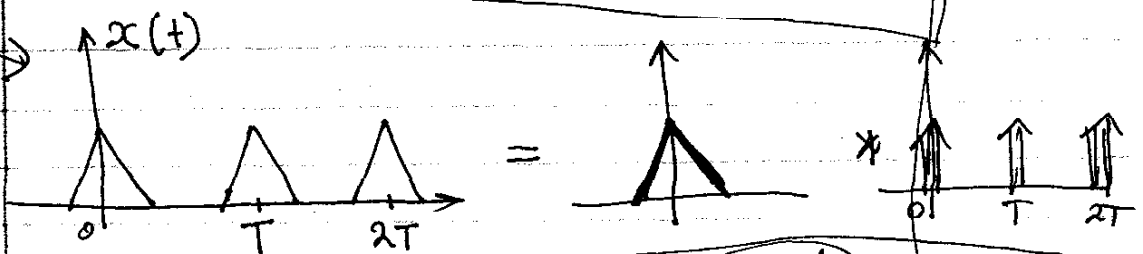
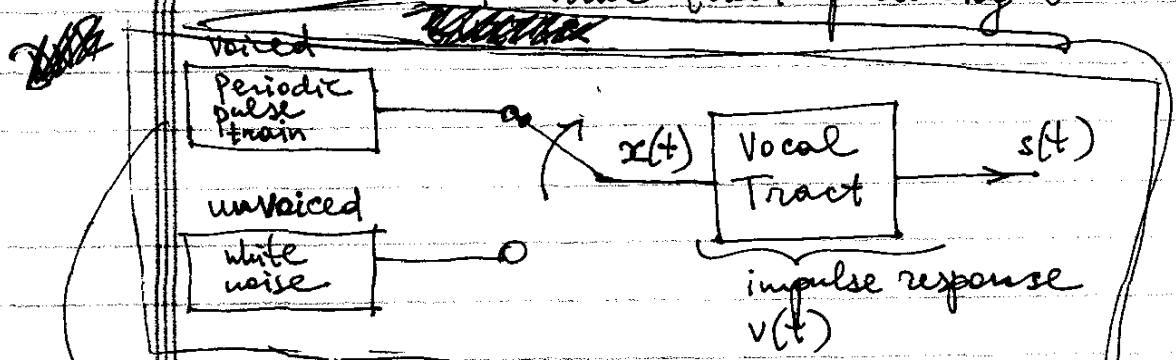
Sound: variations in air pressure.
 Creation of sound: set air in rapid vibration.

The system has two main components:

- Excitation: how air is set in motion
- Vocal Tract: guides air

Voiced Sounds: periodic air pulses pass through vibrating vocal chords

Unvoiced Sounds: force air through a constriction in vocal tract producing turbulence



^{upper part} This is like playing a guitar: you produce a sequence of impulsive excitations by plucking the strings, and then the guitar converts it into music. The strings are sort of like the vocal cords, and the guitar's cavity is like the cavity of the vocal tract.

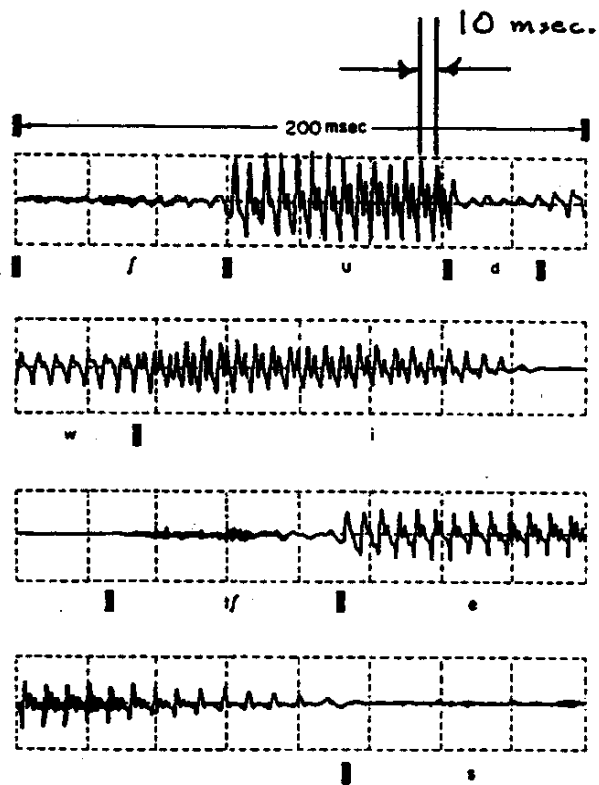


FIGURE 3-2. Example of a speech waveform illustrating different classes of sounds. The utterance is "should we chase ...".



On average, male $T \approx 8\text{ms} \Rightarrow \text{pitch} \approx 125\text{Hz}$
female $T \approx 4\text{ms} \Rightarrow \text{pitch} \approx 250\text{Hz}$

What can we say about the vocal tract?

~~periodic~~

Vocal tract:

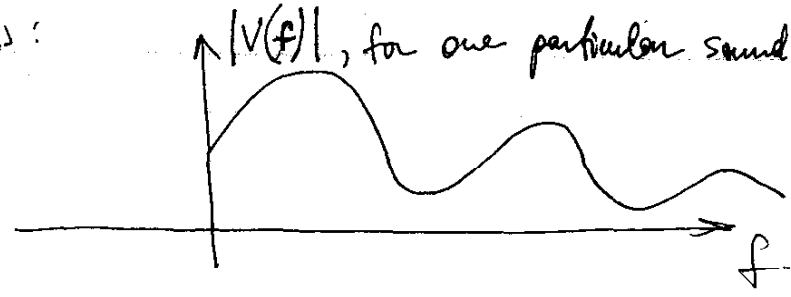
- different voiced sounds are produced by changing the shape of the vocal tract
 \Rightarrow system is time-varying
- BUT: slowly varying
(As we saw in the example, changes occur slowly compared to the pitch period)

(each sound is periodic, but different sounds are different periodic signals)

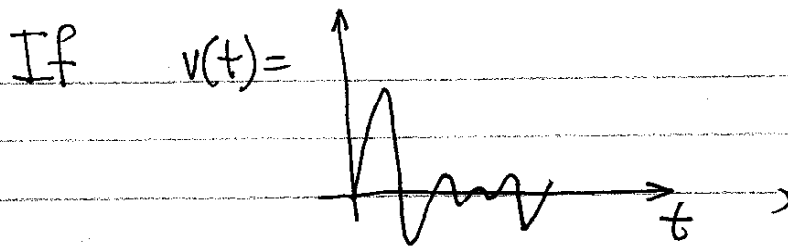
\Rightarrow We ~~can~~ can model the vocal tract as an LTI filter over short time intervals.

Since vocal tract is a cavity, it resonates. In other words, ~~there is~~ when a wave propagates in a cavity, there is a set of frequencies which get amplified. They are called natural frequencies of the resonator, and depend on the shape and size of the resonator.

So, the ~~response~~ magnitude response of the vocal tract can be modeled as something like this:

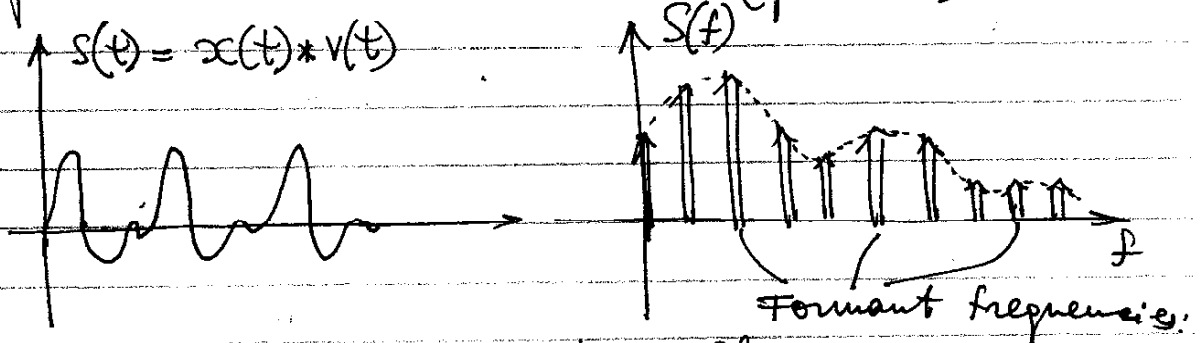


8



then we have the following waveform for this particular ~~some~~ voiced sound (phoneme):

$s(t) = x(t) * v(t)$



- typically, one formant per 1 kHz
- locations are dictated by the poles of ~~the~~ the transfer function
- roll-off is such that

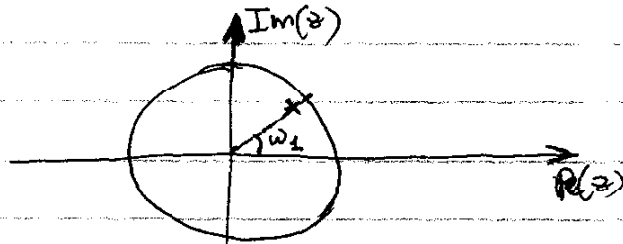
the first 3-4 formants (range: up to 3.5 kHz) are enough for reasonable reconstruction

~~Depending on~~
⇒ sampling at $3.5 \cdot 2 \text{ kHz} = 7 \text{ kHz}$ is OK.
Depending on the application, the sampling rate is normally 7-20 kHz.

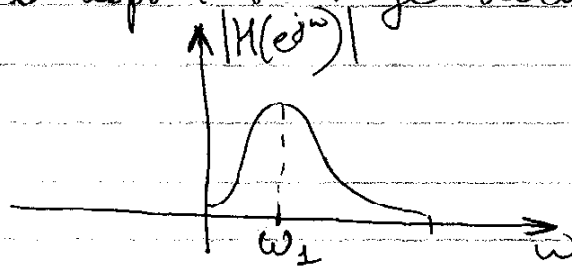
Suppose we discretized speech, and want to model the vocal tract as a digital filter. Here is a very rough idea of how to do this.

9

If we knew the formant frequencies, we could use what we learned about designing frequency selective filters



Poles ~~are~~ of $H(z)$ near the unit circle correspond to large values of $H(e^{j\omega})$:



⇒ Design an all-pole filter, with poles ^(which are close to the unit circle) corresponding to formant frequencies. The ~~closer the pole~~ larger the magnitude resp. at a formant frequency, the closer the corresponding pole(s) to the unit circle.

Example ~~Exercise 10.10, 10.11, 10.12, 10.13~~ (10). A phoneme whose pitch is 100 Hz, is sampled at 6 kHz. It has two formants: a weak one at 500 Hz and a stronger one at 2 kHz.

Find D , the DT pitch period.

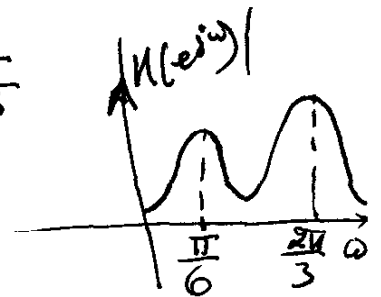
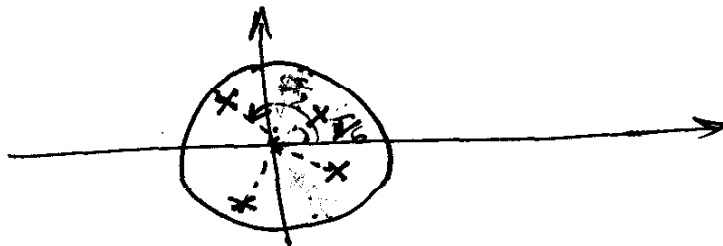
Sketch the approx. pole locations of $H(z)$.

Solution. DT freq. $\omega = 2\pi$ corresponds to 6 kHz.
Therefore, 100 Hz corresp. to $\frac{2\pi}{6000} \cdot 100 = \frac{2\pi}{60}$

$$\Rightarrow D = 60$$

$$500 \text{ Hz corresp. to } \frac{2\pi}{6000} \cdot 500 = \frac{\pi}{6}$$

$$2000 \text{ Hz " " } \frac{2\pi}{6000} \cdot 2000 = \frac{2\pi}{3}$$



Remarks:

- a pair of complex conjugate poles for each formant, to make the freq. resp. real
- the ones at $\pm \frac{2\pi}{3}$ are closer to the unit circle, since the corresp. peak ~~is lower~~ of $|H(e^{j\omega})|$ is larger.