■ 2.2 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. We will 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 in speech processing are:

- Speech synthesis.
- Speech recognition.
- Speech 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 natural-sounding as if spoken by a person, enabling, for example, your computer to talk to you.

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, which human accuracy, speech from any speaker of a given language. One application is being able to dictate to a computer instead of typing. Another common application is automated telephone answering systems, which recognize vocal commands to determine the next action.

The goal of speech coding or compression is to reliably represent speech signals with as few bits as possible. This is very important for storage and transmission. When you store a lot of data, you would like to conserve space as much as possible; and 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 very redundant.

■ 2.2.1 Voiced and Unvoiced Speech

Voiced sounds, e.g., 'a', 'b', are essentially due to vibrations of the vocal cords, and are oscillatory. Therefore, over short periods of time, they are well modeled by sums of sinusoids. This makes short-time Fourier transform—to be discussed later—a useful tool for speech processing. Unvoiced sounds such as 's', 'sh', are more noise-like, as shown in Fig. 2.39.

For many speech applications, it is important to distinguish between voiced and unvoiced speech. There are two simple but effective methods for doing it:

- 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, $P_{av,voiced} > P_{av,unvoiced}$.

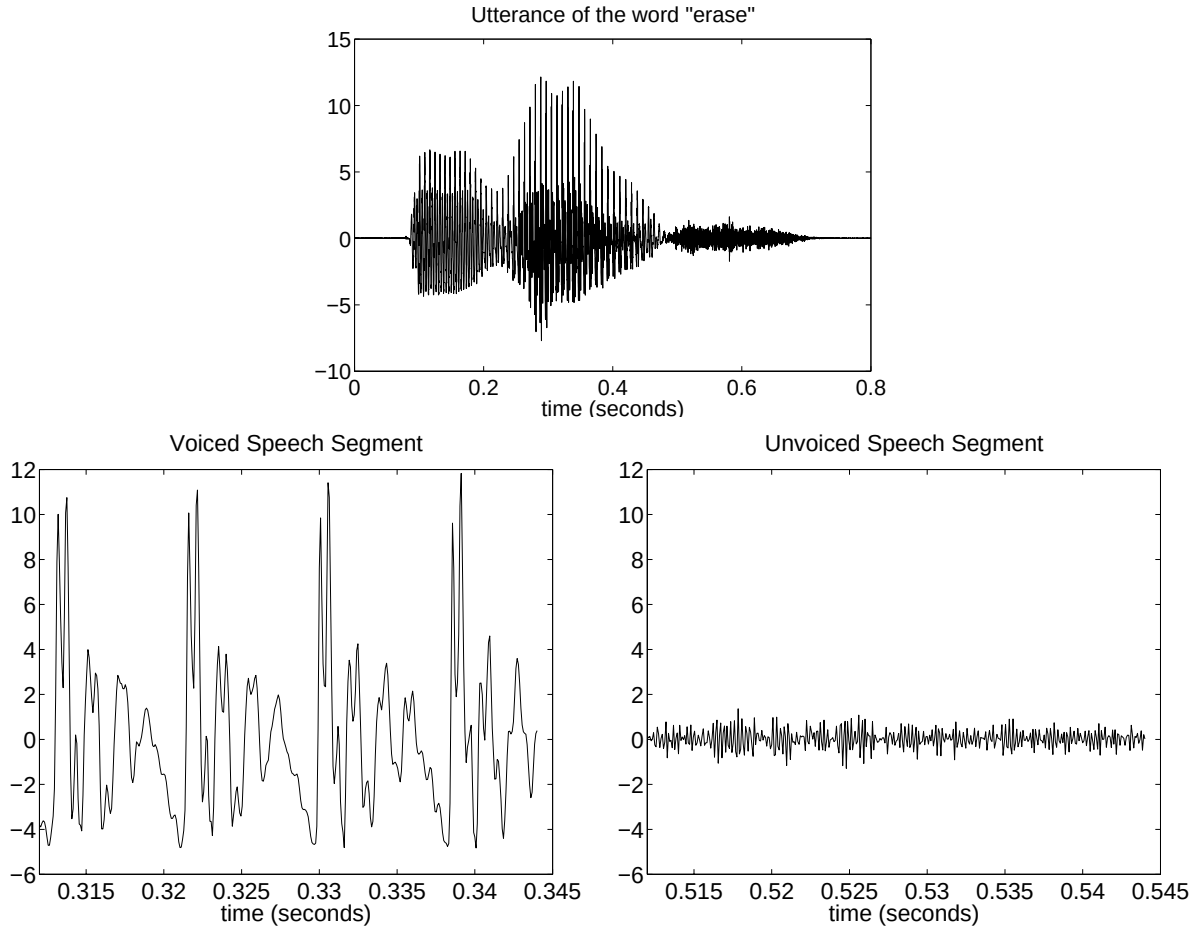


Figure 2.39. Distinction between voiced and unvoiced speech.

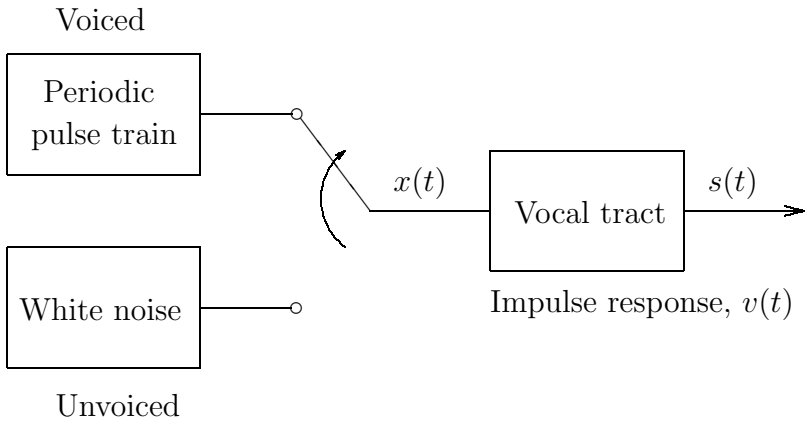


Figure 2.40. Human speech production system.

- Zero-crossing rate: "the signal $x(n)$ has a zero-crossing at n_0" means that

$$x(n_0)x(n_0+1) < 0\,.$$

  Unvoiced signals oscillate much faster, so they will have a much higher rate of zero-crossings.

■ 2.2.2 Source-Filter Model of Speech Production

Sound is variations in air pressure. The creation of sound is the process of setting the air in rapid vibration. Our model of speech production will have two major components:

- Excitation: How air is set in motion.

  Voiced sounds: Periodic air pulses (such as in Fig. 2.41(a)) pass through vibrating vocal chords.

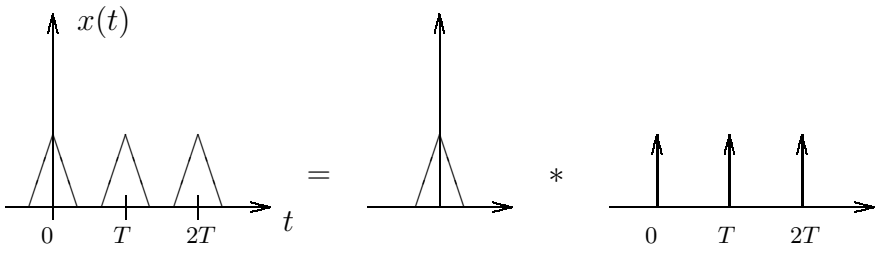
  Unvoiced sounds: Force air through a constriction in vocal tract, producing turbulence.

- Vocal tract: Guides air.

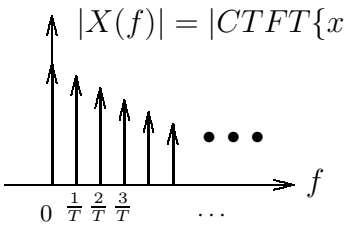
This system is illustrated in Fig. 2.40. Its upper part (the production of voiced sounds) is very much akin to playing a guitar (Fig. 2.40): you produce a sequence of impulsive excitations by plucking the strings, and then the guitar converts it into music. The strings are sort of like the vocal cords, and the guitar's cavity plays the same role as the cavity of the vocal tract.

A periodic pulse train excitation is illustrated in Fig. 2.41(a). The period T is called the *pitch period*, and $1/T$ is called the *pitch frequency*.

On average,



(a) Periodic pulse train.



$\frac{1}{T}$ = pitch frequency

(b) Frequency response.

Figure 2.41. Time domain and frequency domain perspectives of voiced sounds.

male: $T \approx 8ms \Rightarrow$ pitch $\approx 125Hz$;

female: $T \approx 4ms \Rightarrow$ pitch $\approx 250Hz$.

Vocal tract:

- Different voiced sounds are produced by changing the shape of the vocal tract $\Rightarrow$ this system is time-varying.
- However, it is slowly varying. As we saw in Fig. 2.39, changes occur slowly compared to the pitch period.

In other words, each sound is approximately periodic, but different sounds are different periodic signals. This implies that we can model the vocal tract as an LTI filter over short time intervals. Moreover, since the vocal tract is a cavity, it resonates. In other words, 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.

Therefore, the magnitude response of the vocal tract for one voiced sound (phoneme) can be modeled as in Fig. 2.42.

The waveform for this particular phoneme will then be the convolution of the driving periodic pulse train $x(t)$ with the impulse response $v(t)$, as illustrated in Fig. 2.43(b), and the magnitude of its spectrum $|S(f)$ will be the product of $X(f)$ and the magnitude

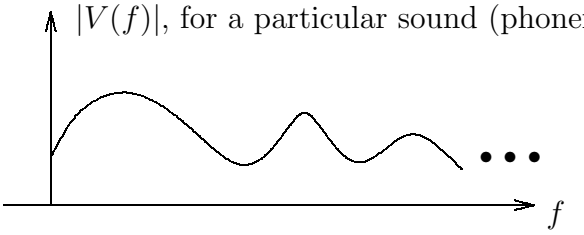
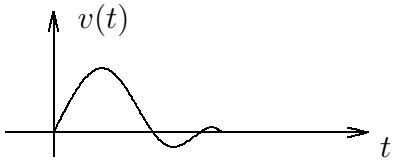
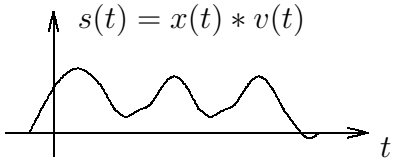


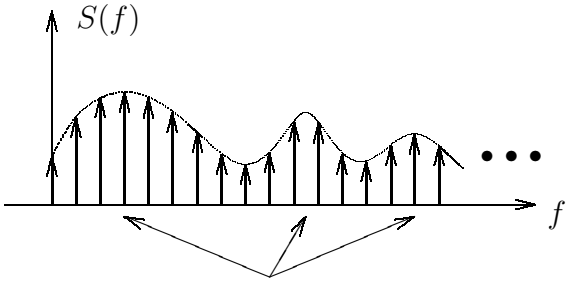
Figure 2.42. Magnitude response of the vocal tract.



(a) Voice tract transfer function.



(b) Resulting sound wave.



(c) Magnitude response of the resulting sound wave.

Figure 2.43. The vocal tract as an LTI filter.

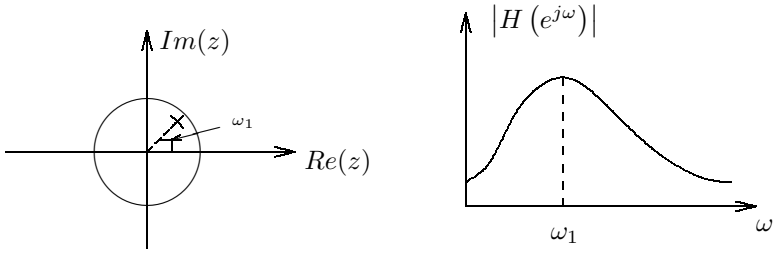
response $|V(f)|$, as illustrated in Fig. 2.43(c). The maxima of $|S(f)|$ are called the *formant frequencies* of the phoneme.

- Typically, one formant per 1 kHz.
- Locations are dictated by the poles of the transfer function.

- Roll-off is such that the first 3 – 4 formants (range: up to 3.5 kHz) are enough for reasonable reconstruction. Thus, sampling at 3.5·2 kHz = 7 kHz is typically enough. Depending on the application, the sampling rate is usually 7 – 20 kHz.

Suppose we discretized speech, and want to model the vocal tract as a digital filter. The following gives a very rough idea of how to this.

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



(a) A pole near the unit circle. (b) Magnitude response.

Figure 2.44. Poles near the unit circle correspond to large values of $H(e^{j\omega})$.

Poles of $H(z)$ near the unit circle correspond to large values of $H\left(e^{j\omega}\right)$ (Fig. 2.44). So, we can design an all-pole filter, with poles which are close to the unit circle, corresponding to formant frequencies. The larger the magnitude response at the formant frequency, the closer the corresponding pole(s) to the unit circle.

Example 2.21. *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.

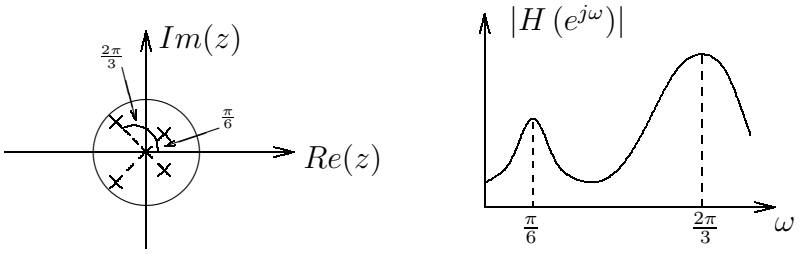
Sketch the approximate pole locations of $H(z)$.

Solution: DT frequency $\omega = 2\pi$ corresponds to 6 kHz. Therefore,

100 Hz corresponds to $\frac{2\pi}{6000} \cdot 100 = \frac{2\pi}{60} \Rightarrow D = 60$;

500 Hz corresponds to $\frac{2\pi}{6000} \cdot 500 = \frac{\pi}{6}$;

2000 Hz corresponds to $\frac{2\pi}{6000} \cdot 2000 = \frac{2\pi}{3}$.



(a) Complex plot of Example 2.21. (b) Magnitude response.

Figure 2.45. Poles near the unit circle correspond to large values of $H(e^{j\omega})$.

Remarks:

- *A pair of complex conjugate poles for each formant, to make the frequency response real.*
- *The ones at* $\pm\frac{2\pi}{3}$ *are closer to the unit circle, since the corresponding peak of* $\left|H\left(e^{j\omega}\right)\right|$ *is larger.*