Simple Model for Speech Generation

Key aspects of speech (review from previous module)

1. Speech is generated by the passage of air from the lungs through the glottis into the throat and oral and nasal cavities. This entire system starting with the glottis is called the vocal tract

Physiology of the vocal tract

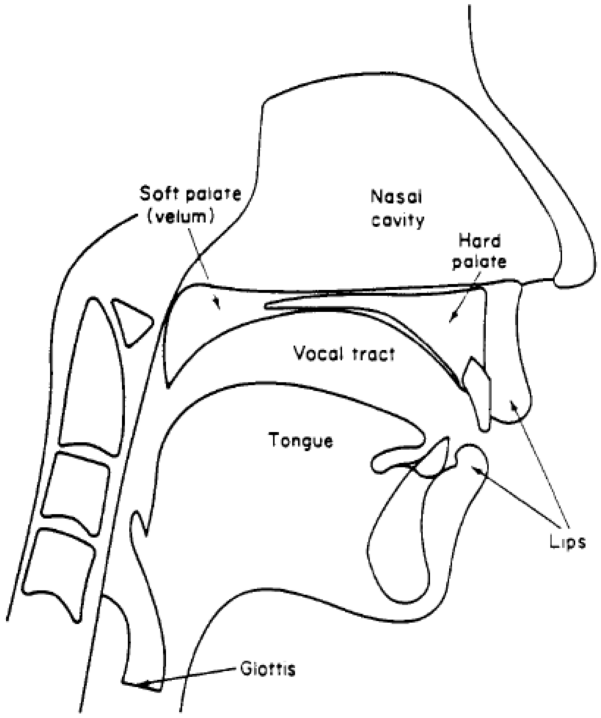


FIGURE 3-1. Cross-sectional view of the vocal mechanism, (after Markel and Gray).

2. Speech is categorized by phonemes that represent distinct sounds that last for a short interval of time called an epoch
3. There are two primary modes for speech

   a. Voiced – glottis opens and closes periodically, emitting short pulses of air with interval equal to the pitch period
   b. Unvoiced – glottis opens and stays open permitting a continuous flow of turbulent air to pass into the throat
4. The vocal tract functions as a resonant cavity
   a. The resonant frequencies are called formants – the first two or three formants are the most important for determining the speech characteristics
   b. The formants are nearly the same for a large population of speakers of the English language
   c. The vocal tract characteristics are independent of whether the speech is voiced or unvoiced
   d. The vocal tract response is determined by the shape of the vocal tract
      i. If the shape remains fixed throughout the epoch of the phoneme, the phoneme is continuant.
      ii. If the shape varies during the epoch of the phoneme, the phoneme is non-continuant
5. The two most important classes of phonemes in the English language are vowels and consonants
   a. Either type of phoneme can be voiced or unvoiced

Vowels: IPA Symbols, Typical Words, and Formants

Table 3.2 Average Formant Frequencies for the Vowels. (After Peterson and Barney [11].)

FORMANT FREQUENCIES FOR THE VOWELS					
Typewritten Symbol for Vowel	IPA Symbol	Typical Word	F_1	F_2	F_3
IY	i	(beet)	270	2290	3010
I	ɪ	(bit)	390	1990	2550
E	ɛ	(bet)	530	1840	2480
AE	æ	(bat)	660	1720	2410
UH	ʌ	(but)	520	1190	2390
A	ɑ	(hot)	730	1090	2440
OW	ɔ	(bought)	570	840	2410
U	ʊ	(foot)	440	1020	2240
OO	u	(boot)	300	870	2240
ER	ɜ	(bird)	490	1350	1690

Example speech waveform

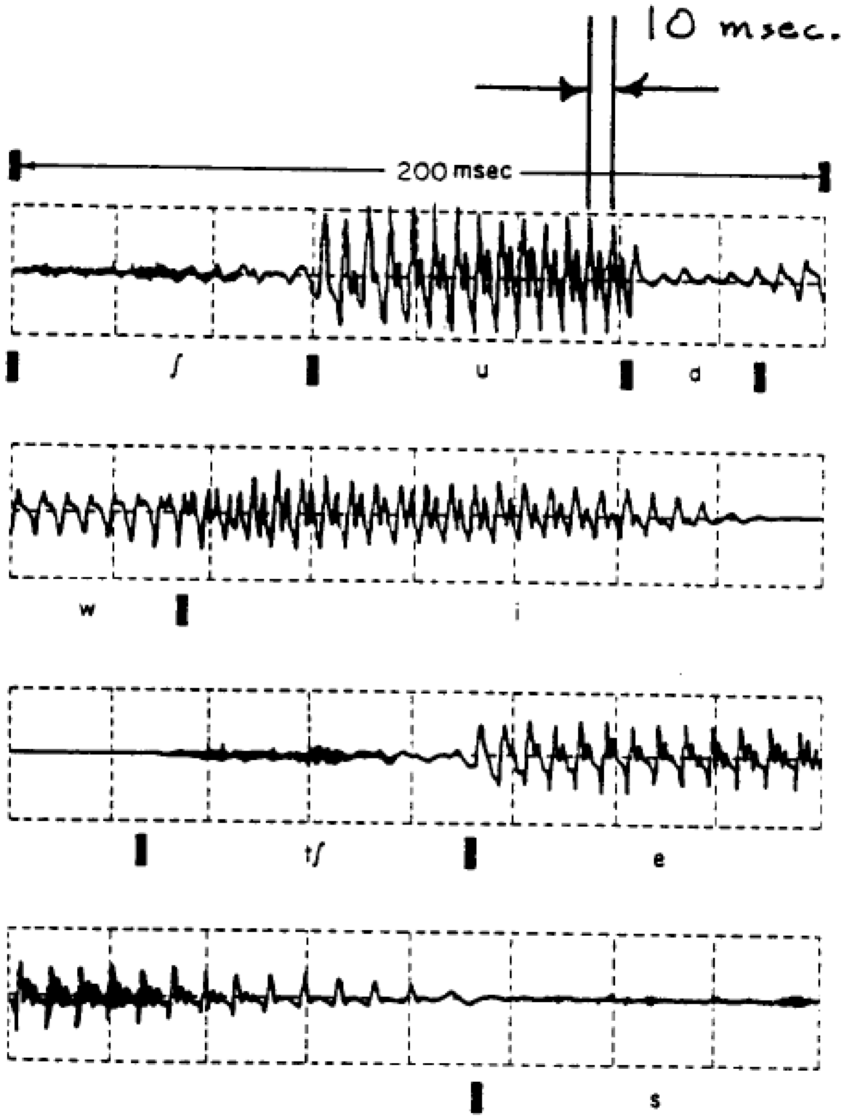
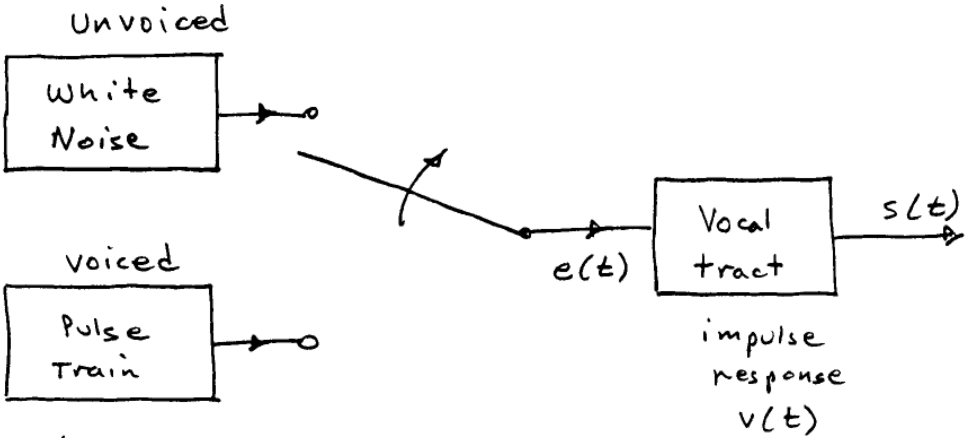


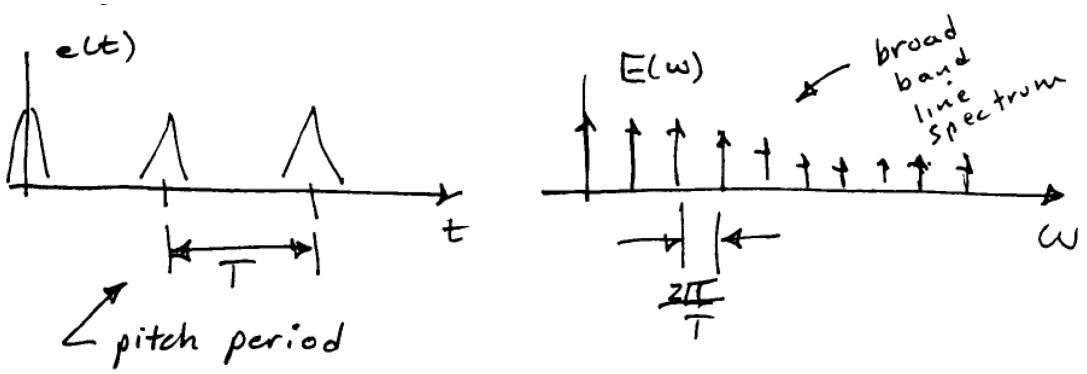
FIGURE 3-2. Example of a speech waveform illustrating different classes of sounds. The utterance is "should we chase . . .".

A simple model for speech generation



Excitation $e(t)$

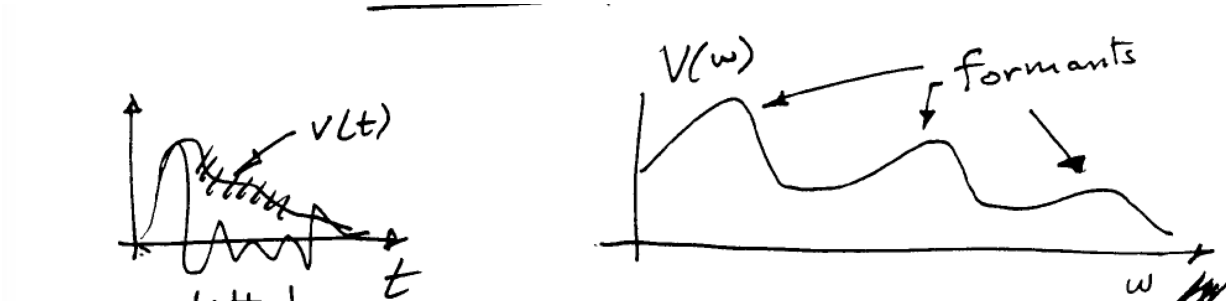
1. Voiced – vocal tract is driven by pulses of air at an interval given by the pitch period



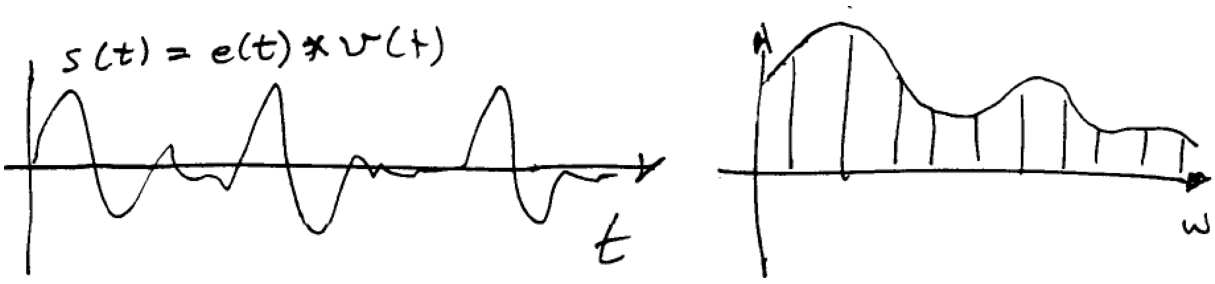
   a. We assume that the pulses are very narrow compared to the duration of the vocal tract response. So relative to the vocal tract response, they look like impulses.
2. Unvoiced – vocal tract is driven by continuous flow of turbulent air
   a. Model $e(t)$ as a white noise stochastic process
   b. Flat power spectrum, i.e. equal energy at all frequencies

Vocal tract

1. The configuration of the vocal tract changes with each different phoneme
    a. For *continuant* phonemes, it is constant throughout the epoch of the phoneme
    b. For *non-continuant* phonemes, it varies continuously throughout the epoch of the phoneme
2. During a short time interval, we model the vocal as a *linear time-invariant system* with vocal tract (impulse response) $v(t)$
    a. In fact, the vocal tract is time-varying.
    b. There exists theories to deal with *linear, time-varying linear systems*. However, we will not formally try to represent the time-varying nature of this system.
3. As discussed previously, the vocal tract will exhibit resonances that correspond to the formants



Speech waveform $s(t)$ and its Fourier transform $S(f)$



$$s(t) = v(t) * e(t)$$
$$e(t) = \text{rep}_P\left[\delta(t)\right]$$
$$s(t) = \text{rep}_P\left[v(t)\right]$$

Therefore

$$S(f) = \frac{1}{P}\text{comb}_{\frac{1}{P}}\left[V(f)\right]$$

Note that this analysis only applies to a continuant vowel that is spoken for many pitch periods, i.e. a relatively long epoch

Fourier spectrum of a real speech waveform

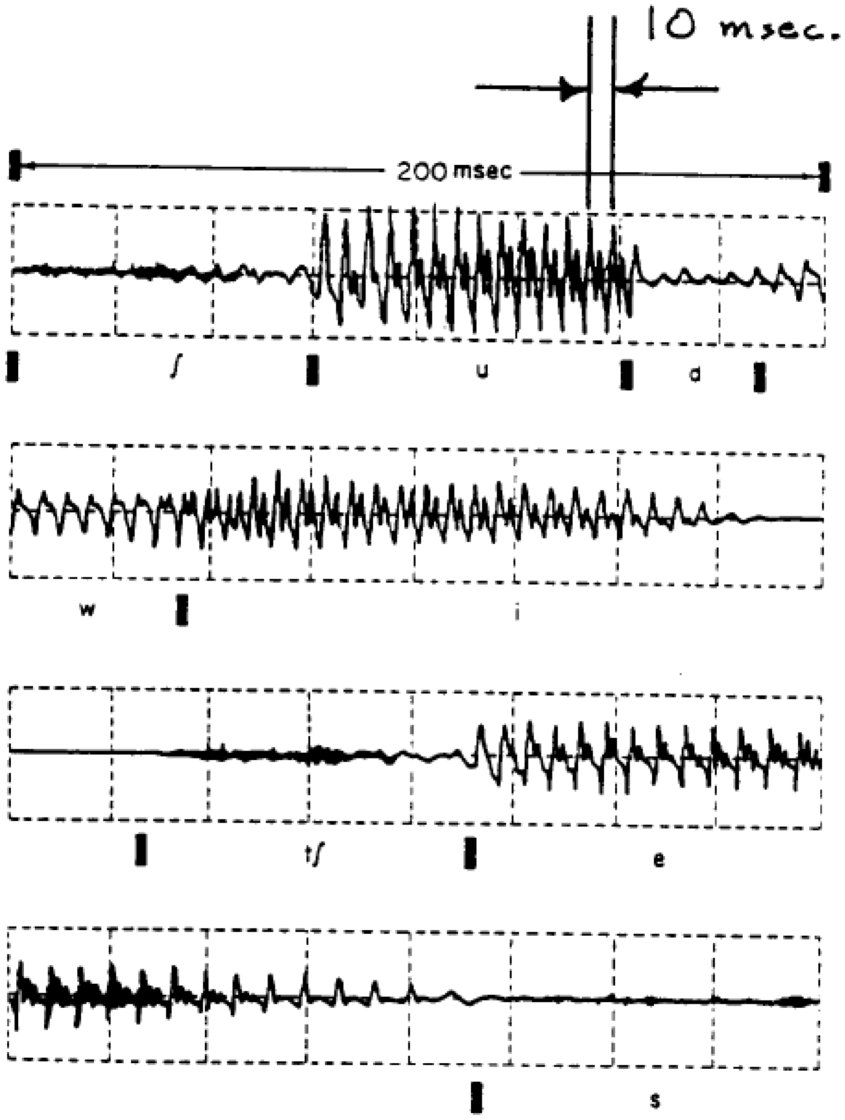


FIGURE 3-2. Example of a speech waveform illustrating different classes of sounds. The utterance is "should we chase . . .".

$$S(f) = \int_{-\infty}^{\infty} s(t) e^{-j2\pi ft}\, dt$$

$$= \; ???$$

1. The Fourier integral includes many different phonemes with different spectral characteristics
2. The result will be mush
3. We need a new Fourier analysis method that only looks at a short interval of the speech waveform at any particular time
4. This is called *short-time Fourier analysis*
5. It can be based on the CTFT, the DTFT, or the DFT
6. We will use the DTFT