developing theories of speech perception. For example, the observation of large articulatory variability in the phoneme /r/, using magnetic resonance imaging, suggests that speakers and listeners rely on an acoustic representation in production and perception of speech, rather than an articulatory representation [13]. Similar acoustic features for the phoneme /r/ can be generated with very different vocal tract shapes involving different tongue positions and constrictions, as we described in Section 3.4.6. Such large articulatory variability can depend on phonetic context and has been observed in other phonemes as well.

In the motor theory of perception, prior to making an articulatory mapping, we detect (or estimate) the acoustic features that serve as perceptual cues. There is evidence that feature detection is carried out by the auditory system in the ear and auditory centers of the brain [5],[29]. Experiments have been carried out to study the possibility of special auditory detectors for the perception of formant energy, bursts, voice onset times, formant transitions, and the presence (or lack) of voicing. Different features were found to correspond to different neural cell firing patterns in the early and higher level auditory centers. There are two primary neural cell firing characteristics used in feature detection: increase in the firing rate with a sudden change in energy (within specific frequency bands) and the synchronizing of a firing pattern to frequency. With plosive consonants, for example, some cells in the higher auditory levels, called *onset cells*, appear to fire at both the burst release and the vowel onset and thus may directly encode voice onset time in their response pattern. There is also evidence that in the detection of plosive consonants, detectors may exist that are tuned to sense the burst and the formant transition into the following vowel. With vowels, formant energy can be identified in two ways. First, there is an increase in the firing rate of nerve fibers tuned to specific frequencies and, second, nerve fibers synchronize with the formant frequency (firing at the peaks in each cycle of these frequencies), a characteristic known as *phase synchrony*. Similar neural firing mechanisms are hypothesized to exist for identifying the pitch associated with harmonic spectra.

In the design of speech analysis and synthesis systems, we will need to consider how changes in the temporal and spectral properties of a speech waveform affect its perception. For example, changes in the short-time Fourier transform phase characteristics of speech may affect the phase synchrony of frequency components in firing patterns of auditory nerves. We will address such issues in Chapter 8 and other places throughout the text.

3.7 Summary

This chapter described qualitatively the main functions of the speech production mechanism and the associated anatomy. Articulatory and acoustic descriptors of speech sounds were given and, based on these features, the study of phonemics and phonetics was introduced. Some implications of sound production and perception mechanisms for signal processing algorithms were discussed. For example, the source mechanism can result in a variety of voice types, such as the *hoarse*, *breathy*, or *diplophonic* voice, that can influence speech analysis and synthesis strategies. Glottal secondary pulses that occur in the diplophonic voice, for instance, have in part motivated *multi-pulse* speech analysis/synthesis, and aspiration that occurs in the breathy voice has in part motivated *residual*-based speech analysis/synthesis, both of which we will study in Chapter 12.