7. The Speech Signal

Take care of the sense and the sounds will take care of themselves.

Lewis Carroll (1832–1898)

The speech signal, as it emerges from a speaker's mouth, nose and cheeks, is a one-dimensional function (air pressure) of time. Microphones convert the fluctuating air pressure into electrical signals, voltages or currents, in which form we usually deal with speech signals in speech processing. Digital-to-analog converters change the analog voltages into binary (or n-ary) digital signals. Bandlimited speech signals (bandlimited by a telephone system, for example) of less than 4000 Hz bandwidth can be represented, according to the sampling theorem, by 8000 samples per second. Each sample can be quantized to 256 levels (8 bits) with little audible degradation if the levels properly cover the voltage range of the signal. (One or two bits per sample can be saved by a judicious, non-uniform choice of levels at the cost of only minor audible distortion.) Thus the total information rate required for a high-quality representation of a speech signal bandlimited to 4 kHz is 8 bits/sample times 8000 samples/second or 64 kbits per second. (For comparison, the bit rate on a stereo compact disc (CD) exceeds 1.4 Mbits/second.) The aim of speech compression is to reduce this bit rate as much as possible for more efficient storage and transmission.

Although the speech signal is a one-dimensional function (air pressure) of a one-dimensional variable (time), it is generated by a plethora of parallel nerve commands from the brain, controlling the muscles of the various organs participating in the articulatory process – vocal cords, tongue body, tongue tip, lips, soft palate (velum), etc. These nerve commands do not only occur in parallel, they are noticeably desynchronized to compensate for different delays on different nerve fibers and to promote the numerous "coarticulatory" effects observed in speech in which the articulation of one speech sound is substantially influenced by its neighbors. These phenomena are documented in great detail for one widely understood language in the book *Acoustics of*

M. R. Schroeder, *Computer Speech*

American English Speech: A Dynamic Approach by J. P. Olive, A. Greenwood, and [7.1].

In spite of the great complexity of the speech production process – from thought and intent in the brain to the acoustic signal – a more simplistic view of speech signals, disregarding most of these complexities, suffices for simple speech signal compression. But some of these complexities cannot be safely ignored in speech recognition and particularly in speech synthesis from written material. Not paying proper attention to the human production process is precisely the reason why machine speech, to this day, has a flavor of, well, machine speech. Exorcising the "electronic accent" from synthetic speech is a continuing challenge.

7.1 Spectral Envelope and Fine Structure

Of the many distinctive features of speech that even speech compression cannot ignore is the dichotomy between *voiced* and *unvoiced* sounds. For voiced sounds, like the vowels and voiced consonants in non-whispered speech, the vocal cords vibrate more or less periodically, chopping up the air stream from the lungs into individual puffs of air at a fundamental frequency f_0 ranging from roughly 50 Hz (low male) to 300 Hz (high female). The resulting "quasiperiodicity" of the speech signal is manifest both in the waveform and the short-time spectrum: the waveform shows a repetitive pattern at the rate f_0 and the spectrum has equidistant peaks ("lines") at integer multiples of the fundamental frequency f_0 ("harmonics"), see Fig. 4.2.

7.2 Unvoiced Sounds

During unvoiced sounds, the vocal cords do not vibrate and they do not undulate the air flow from the lungs. (But they may be nearly closed causing audible friction as for the /h/ sound.) The acoustic energy is produced by turbulence at one or several narrow air passages in the mouth (tongue tip against or between the teeth, tongue against the palate etc.). This turbulent energy has a "smooth" spectrum like that of noise, without a line structure, see Fig. 4.3.

7.3 The Voiced–Unvoiced Classification

But not all speech sounds are either purely voiced or turbulent. The voiced fricatives /v/ as in *veal*, /z/ as in *zeal*, and /ʒ/ as in *pleasure* are voiced, because the vocal cords vibrate, and they are turbulent because of noise generated at narrow constrictions in the vocal tract. The speech signal therefore