A brief introduction to speech

In Project 2B, you are asked to create a simple voice detector to decide whether a segment of speech was uttered by Homer or Lisa Simpson. The problem of voice recognition is complicated and involves many factors, most of which are outside the scope of the class. For the purposes of this project, you should apply the simple tools and reasoning from OS3 and lecture to write your voice detection function. The following is a brief introduction to speech production and speech processing and should provide you with enough understanding to complete the project.

Speech sounds may generally be classified as voiced or unvoiced. In voiced speech (for vowels and nasal consonants like /m/ or /n/), the vocal folds close and open periodically at a fundamental frequency $F_0$, which depends on the length and tension of the vocal folds. The perceived pitch of a person’s voice is largely determined by the frequency $F_0$. Adult males have larger vocal folds and thus their voices tend to have lower pitches around 125 Hz, while adult females average around 210 Hz. Pitch in children averages around 300 Hz.

Unvoiced speech is typically generated when the airflow through the vocal tract becomes constricted, resulting in noisy sounds (such as in fricatives like /f/ and /s/ or in unvoiced plosives like /p/, /t/, and /k/). During the production of these speech sounds, the vocal folds are not vibrating and thus do not produce a pitch percept. Another class of speech sounds combines voiced and unvoiced parts (such as in voiced fricatives like /z/ or /v/, or voiced plosives like /b/, /d/, or /g/).

In general, pitch information may be extracted during voiced sounds, particularly during the production of vowels, because voiced consonants are often too short to reliably estimate $F_0$. Sound waves traveling through the vocal tract generate resonances at natural frequencies which become the formants. The locations of these frequencies depends on the shape of the lips as well as the shape and position of the tongue and jaw, which are mostly the same across individuals. Therefore, the formant frequencies largely determine which vowel sounds are produced and do not provide pitch information. However, the presence of formants indicates that the sound is voiced and that the pitch may be extracted. Figure 1 depicts the time and frequency domain representations of a sound /a/ with formant frequencies at about 600 Hz, 1000 Hz, and 2500 Hz. The harmonic structure is evident in both the time domain (where there is clearly periodicity at about 10 ms) and the frequency domain (where there is obvious harmonic content at intervals of 100 Hz). The pitch of this utterance is around 100 Hz, indicating it was uttered by an adult male speaker.

Another way to describe a speech signal is with a spectrogram, which is a time-frequency-amplitude representation of a signal. Examples of a wideband and narrowband spectrogram of the same utterance are shown in Figure 10.22 of OS3. The formants of the vowels have mostly low frequency energy (formants 1-3 typically occur below 3 kHz). In contrast, unvoiced sounds like /s/ have lots of noisy high frequency content and give no information about the pitch of the speaker. Short unvoiced consonants like /t/ or /p/ look like short noise bursts and also reveal no pitch information. The harmonic structure reflecting the pitch is most evident during the production of vowel sounds. The pitch may be inferred from the wideband spectrogram by the period between the vertical lines, or from the narrowband spectrogram by the frequency difference between the horizontal lines.
Figure 1: The time- and frequency-domain presentation of vowel /a/. Taken from http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap3.html

Note that the data given are random segments of utterances spoken by Homer or Lisa. Therefore, your detector should be able to deal with combinations of voiced and unvoiced sounds. A couple ways to represent the information in speech signals have been presented here. You should explain in your project report what method you choose and why.