Digital Signal Analysis (CS7.303)

Spring 2022, IIIT Hyderabad
19 Feb, Saturday (Lecture 12)

Taught by Prof. Anil Kumar Vuppala

Speech Production

Speech is produced by expelling air from the oral and nasal cavities. The air is supplied by the lungs, and passes through the vocal tract; here the vocal folds vibrate at a frequency, called the *pitch* of the voice.

Sounds for which the vocal folds vibrate are called *voiced* (like vowels and certain consonants) and those for which there is no vibration are called *unvoiced.* Voiced sounds tend to have periodic waveforms with more energy, while unvoiced sounds are irregular and low-energy.

Speech production can be mathematically modelled as an LTI system (linear prediction) whose input is a periodic signal for voiced sounds and noise for unvoiced sounds.

If we consider the waveform within a small window, we can assume it to be periodic, and predict the waveform from its previous p samples:

$$\hat{s}(n) = -\sum k = 1^p a_k s(n-k),$$

where $\hat{s}$ is the predicted waveform. Then we call the error (or *excitation*)

$$e(n) = s(n) - \hat{s}(n),$$

or

$$e(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k).$$

The values of a_k represent the vocal tract.

This will give us

$$S(z) = E(z) \cdot \left(\frac{1}{1 + \sum_{k=1}^{p} a_k z^{-k}} \right).$$

The term in brackets is the ZT $H(z)$ of the impulse response of the system.