Sound perception:
Understanding how we hear sounds and how we perceive speech leads to better design and implementation of robust and efficient systems for analyzing and representing speech.
The better we understand signal processing in the human auditory system, the better we can (at least in theory) design practical speech processing systems (Speech coding, speech recognition and etc.)

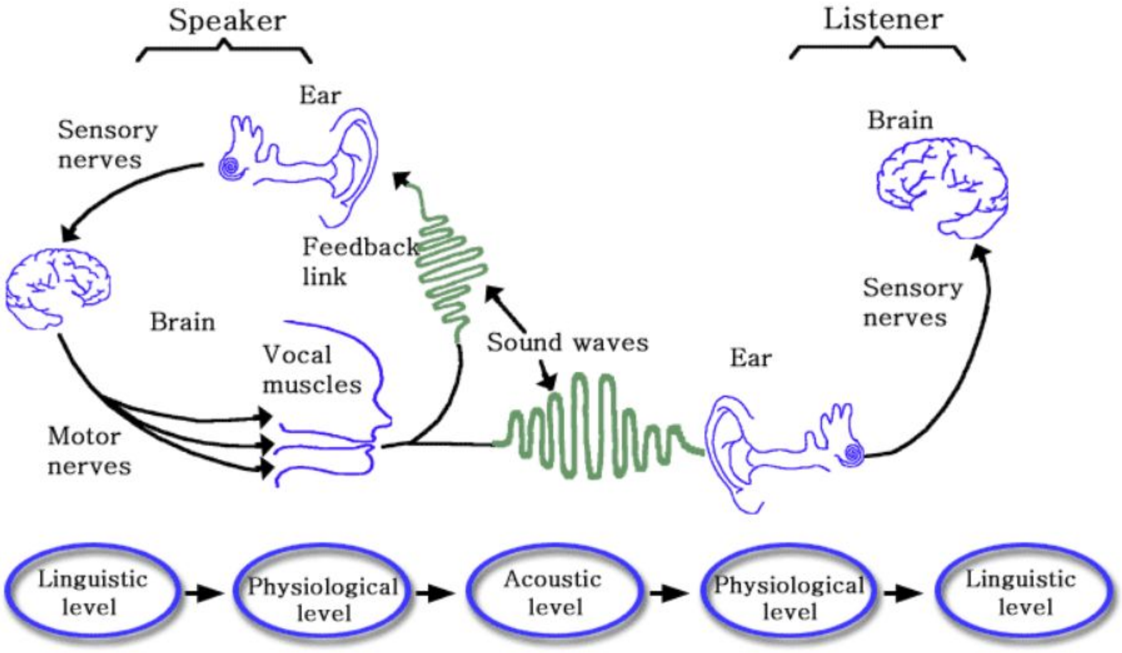


Figure 1. Speech production chain

In speech chain, the message to be conveyed by speech goes through five levels of representation between the speaker and the listener, namely:

– the **linguistic level** (where the basic sounds of the communication are chosen to express some thought of idea)
– the **physiological level** (where the vocal tract components produce the sounds associated with the linguistic units of the utterance)
– the **acoustic level** (where sound is released from the lips and nostrils and transmitted to both the speaker (sound feedback) and to the listener)
– the **physiological level** (where the sound is analyzed by the ear and the auditory nerves), and finally
– the **linguistic level** (where the speech is perceived as a sequence of linguistic units and understood in terms of the ideas being communicated)

Auditory system:

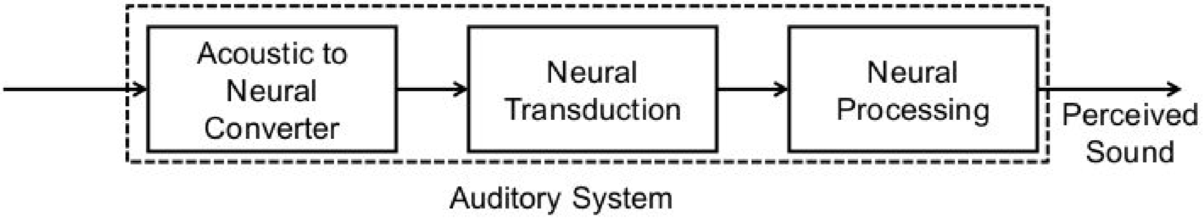


Figure 1. Auditory system block diagram

Acoustic to neural converter: The acoustic signal first converted to a neural representation by processing in the ear is depicted in above block diagram. The conversion takes place in stages at the outer, middle and inner ear.

Neural transduction: This step takes place between the output of the inner ear and the neural pathways to the brain. It consists of a statistical process of nerve firings at the hair cells of the inner ear which are transmitted along the auditory nerve to the brain.

Neural processing: The nerve firing signals along the auditory nerve are processed by the brain to create the perceived sound corresponding to the spoken utterance.

For better understanding of auditory perception, it is necessary to understand the functioning of different parts of a ear. The ear is composed of three sections: the **outer ear**, **middle ear**, and **inner ear** Figure 2. The outer ear directs speech pressure variations toward the eardrum, where the middle ear transforms these variations into mechanical motion. The inner ear converts these vibrations into electrical firings in the auditory neurons, which lead to the brain.

Outer ear:

The pinna (a cartilaginous flap of skin) helps in sound localization [13], and by its asymmetric shape makes the ear more sensitive to sounds coming from in front of the listener than to those coming from behind. The meatus, an air-filled cavity open at one end (pinna) and closed at the other (eardrum), acts as a **quarter-wavelength resonator**. The first resonance is near 3 kHz. This resonance amplifies energy in the 3-5 kHz range by up to 15 dB, which likely aids perception of sounds having significant information at these high frequencies.

Middle ear:

The middle ear accomplishes an impedance transformation between the air medium of the outer ear and the liquid medium of the inner ear. Spectrally, the middle ear acts as a **lowpass filter with attenuation of about -15 dB/oct above 1 kHz**. The middle ear also protects the delicate inner ear against very strong sounds. As sound intensity increases, the stapes motion changes from a pumping action to one of rotation, so that inner ear oscillations do not increase proportionally with sound levels.

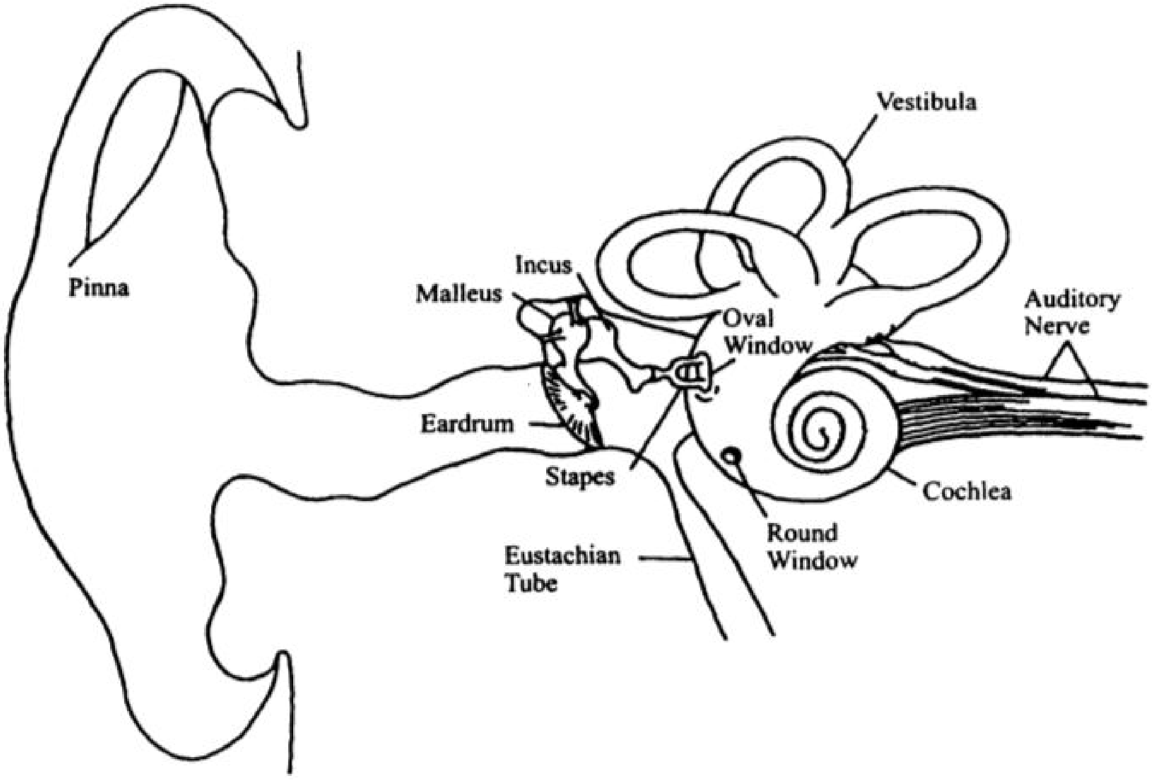


Figure 1. The structure of the peripheral auditory system

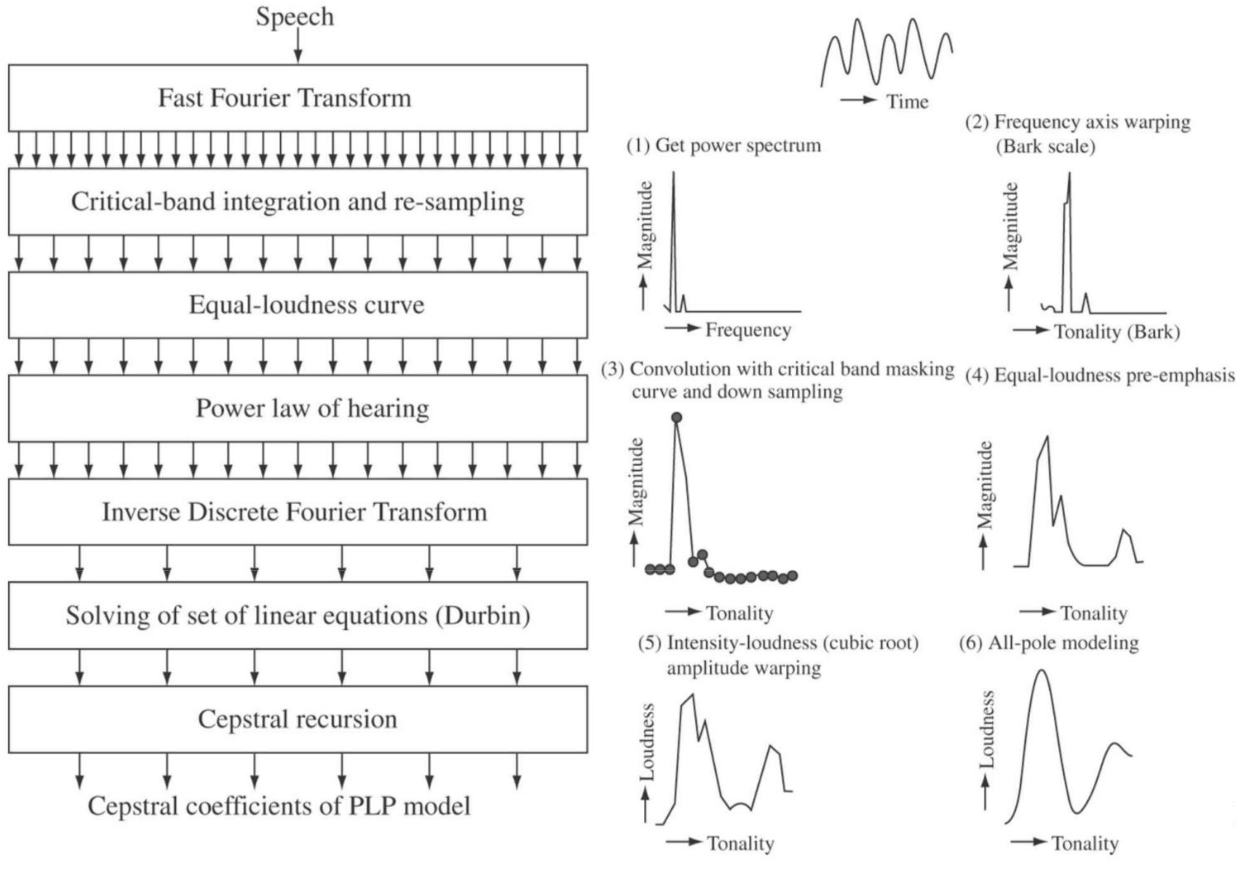
Inner ear:

Mechanical input from middle ear starts traveling wave moving down **Basilar membrane**. Each location along the BM has a characteristic frequency (CF), at which it vibrates maximally for a given input sound. Varying stiffness and mass of BM results in continuous variation of resonant frequency. For every input frequency, there is a point on the BM of maximal vibration. I.e Basilar membrane decomposes audio signal on non-linear scale (log scale). BM acts as a collection of damped resonators. In decomposing the audio signal it will not provide any phase delay. At resonance, traveling wave energy is dissipated in BM vibration. **Hair cells** on BM convert motion into nerve impulses (firings). **Inner Hair Cells** detect motion. IHCs convert BM vibration into **nerve firings**. **Auditory nerve** collects all responses of IHC neural firings and passed to brain.

Auditory Models: Most auditory models includes the perceptual effects such as

1. spectral analysis on a **non-linear frequency scale** (usually mel or Bark scale)
2. **spectral amplitude compression** (dynamic range compression)
3. **loudness compression** via some logarithmic process
4. decreased sensitivity at lower (and higher) frequencies based on results from **equal loudness contours**
5. utilization of temporal features based on **long spectral integration** intervals (syllabic rate processing)
6. **auditory masking** by tones or noise within a critical frequency band of the tone (or noise)

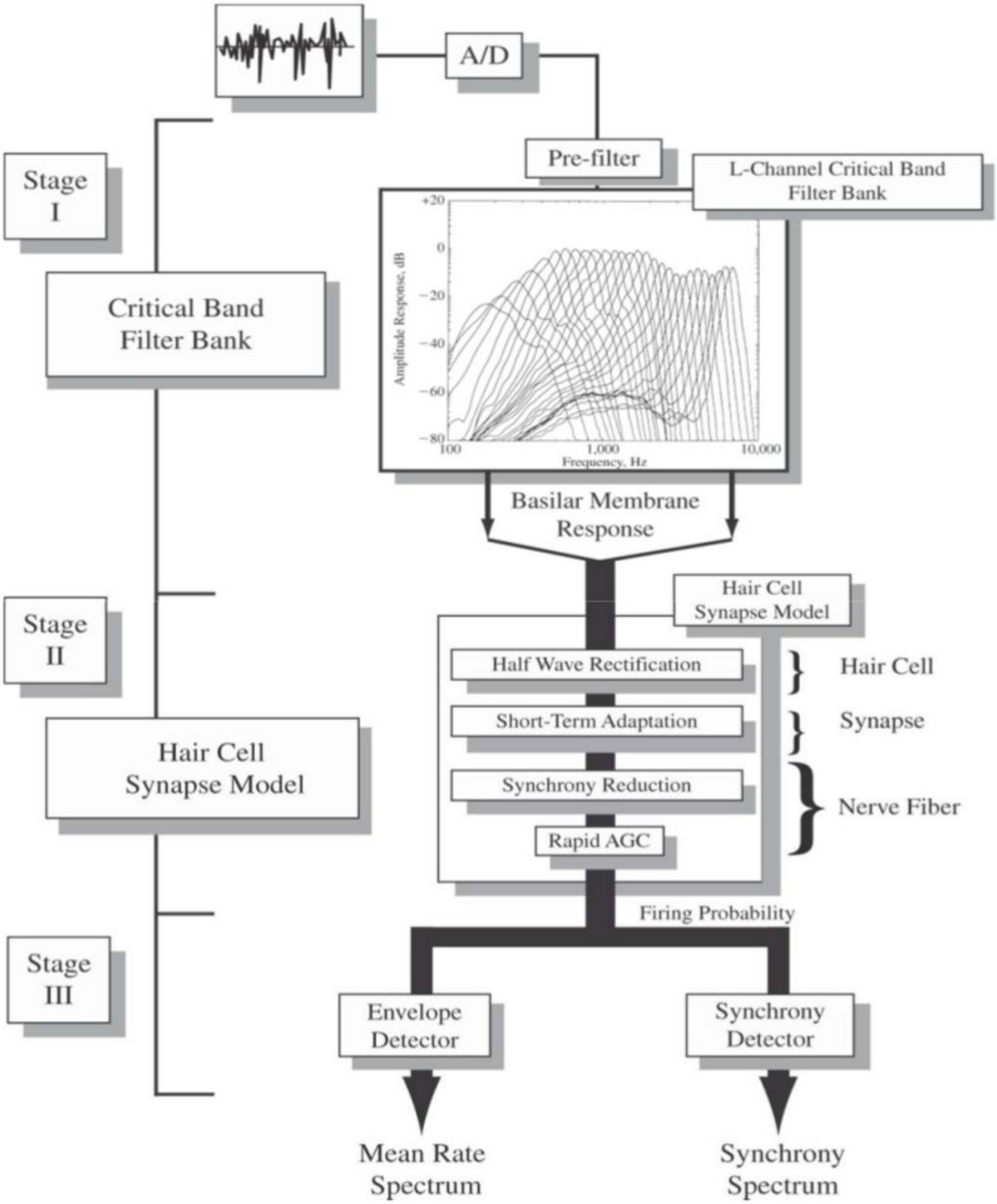
Perceptual Linear Prediction:


Speech
Fast Fourier Transform
Critical-band integration and re-sampling
Equal-loudness curve
Power law of hearing
Inverse Discrete Fourier Transform
Solving of set of linear equations (Durbin)
Cepstral recursion
Cepstral coefficients of PLP model
Time
(1) Get power spectrum
Magnitude
Frequency
(2) Frequency axis warping (Bark scale)
Magnitude
Tonality (Bark)
(3) Convolution with critical band masking curve and down sampling
Magnitude
Tonality
(4) Equal-loudness pre-emphasis
Magnitude
Tonality
(5) Intensity-loudness (cubic root) amplitude warping
Loudness
Tonality
(6) All-pole modeling
Loudness
Tonality

Perceptual effects included in PLP are:
– **critical band spectral analysis using a Bark frequency scale** with variable bandwidth trapezoidal shaped filters
– asymmetric auditory filters with a 25 dB/Bark slope at the high frequency cutoff and a 10 dB/Bark slope at the low frequency cutoff

– use of the **equal loudness contour** to approximate unequal sensitivity of human hearing to different frequency components of the signal
– use of the **non-linear relationship between sound intensity and perceived loudness** using a cubic root compression method on the spectral levels

Seneff Auditory Model:



This model tried to capture essential features of the response of the cochlea and the attached hair cells in response to speech sound pressure waves in three stages of processing:
– stage 1 pre-filters the speech to eliminate very low and very high frequency components, and then uses a 40-channel critical band filter bank distributed on a Bark scale.
– stage 2 is a hair cell synapse models which models the (probabilistic) behavior of the combination of inner hair cells, synapses, and nerve fibers via the processes of half wave rectification, short-term adaptation, and synchrony reduction and rapid automatic gain control at the nerve fiber; outputs are the probabilities of firing, over time, for a set of similar fibers acting as a group
– stage 3 utilizes the firing probability signals to extract information relevant to perception; i.e., formant frequencies and enhanced sharpness of onset and offset of speech segments; an Envelope Detector estimates the Mean Rate Spectrum (transitions from one phonetic segment to the next) and a Synchrony Detector implements a phase-locking property of nerve fibers, thereby enhancing spectral peaks at formants and enabling tracking of dynamic spectral changes.

Other auditory perception models: **Lyon's Cochlear model** and **Ensemble Interval Histogram model.**

Pitch of Sound:
This depends on the **fundamental frequency** of vibration of the waves. If the frequency of vibration is higher, we say that the sound is shrill and has a high pitch. On the other hand, if the sound is said to have lower pitch then it has a lower frequency of vibration. Ex: Voice of a woman has high pitch than that of a man.

Loudness of sound:
This phenomena of a sound depends on **amplitude or intensity** of the sound wave. If the intensity of the sound wave is large, then the sound is said to be loud. Loudness is directly **proportional to the square of the intensity of vibration**. If the amplitude of sound wave becomes double, then the loudness of the sound will be quadrupled. It is expressed in decibel (dB). Sounds above 80 dB becomes noise to human ears.

Intonation:
It is variation in pitch. Intonation, in phonetics represents the melodic pattern of an utterance. Intonation is primarily a matter of variation in the pitch. Stress and rhythm are also involved in intonation. Intonation conveys differences of expressive meaning (e.g., surprise, anger, wariness).

Self-check:
How the signal entering a listener's ears is converted into a linguistic message ?
How we discriminate gender, emotional state, health condition, and etc from speech?
What are **pitch, loudness, rhythm, and intonation** ?
How speech is perceived in noise?
What is **selective attention** (cocktail party example) mechanism ?
What are all different **audiotroy models**?