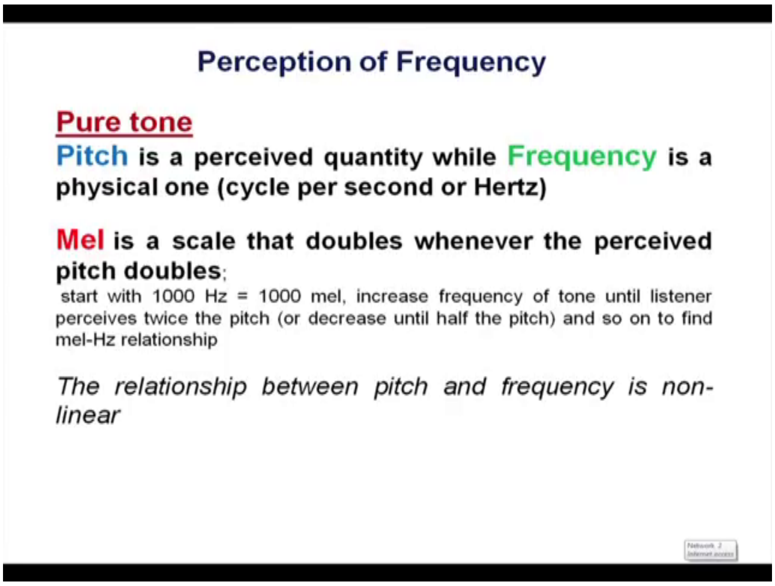
Digital Speech Processing
Prof. S. K. Das Mandal
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

Lecture – 18
Speech Perception – Part III

So, last class we have discussed about the perception of amplitude or we can say the perception of intensity that is loudness and how the loudness is converted into dB, what is 0 db? What is equate loudness curve? What is scone curve? That we have discussed. Today, we want to discuss about, how human being perceive the frequency.

(Refer Slide Time: 00:39)



Perception of Frequency

Pure tone

Pitch is a perceived quantity while Frequency is a physical one (cycle per second or Hertz)

Mel is a scale that doubles whenever the perceived pitch doubles;
start with 1000 Hz = 1000 mel, increase frequency of tone until listener perceives twice the pitch (or decrease until half the pitch) and so on to find mel-Hz relationship

The relationship between pitch and frequency is non-linear

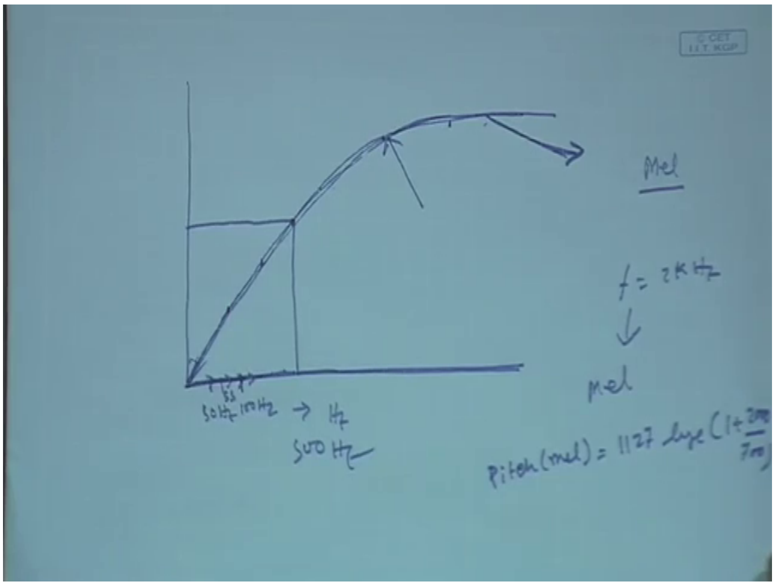
If you see that we have the pitch, we generally call the pitch of this sound is very high. Pitch of this sound is very low. Your pitch is very high pitch.

So, what is pitch? So, pitch is an perceptual parameters, where as the frequency which is related to the pitch is the physical parameter frequency can be measurable, but pitch is a physiological parameter which is perceptually measurable. So, this is a perceptual parameter. So, pitch and frequency are not same. If I say pitch is a perceptual of the perception of the particular frequency. If the frequency not in a pure tone, then the pitch may be different from the frequency like that, if I give an example you have heard it that suppose there is a harmonium and there is a guitar, there is a like sitar, there is a string different kind of string instruments are there every instruments are playing the let us the base [FL] is a particular frequency, let us 250 Hertz frequency.

But, if you perceive the sound, you can easily understand which one which sound is coming from guitar, which sound coming for sitar, which sound is coming from other string instrument so; that means, the perception of frequency is not as that the physical frequency. So, pitch is mainly a perceptual dimension of the frequency and pitch mainly correspondent to the fundamental frequency, but not it is one to one corresponds ok.

Let us we have to interest that human being as, suppose human being perceive the frequency yes we perceive the frequency. How good we perceive the frequency; that means, in which scale in which how we perceive the frequency.

(Refer Slide Time: 02:43)



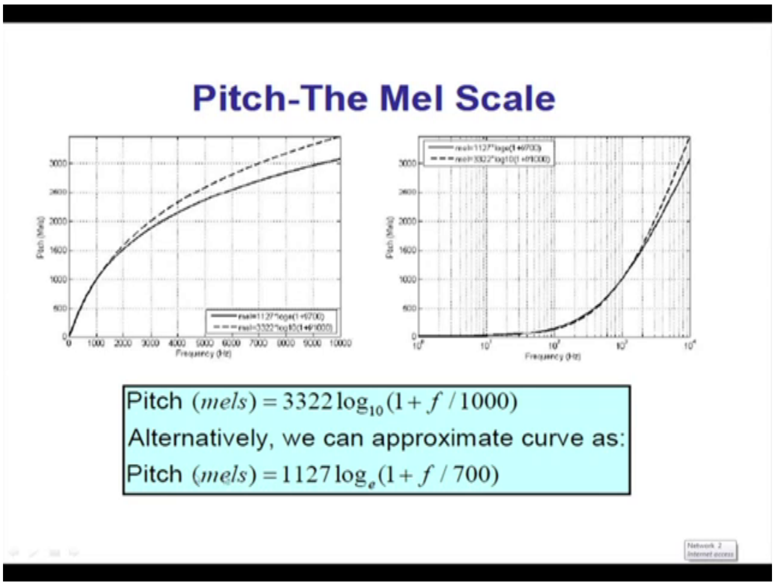
So, suppose I have a x axis, there is a scale physical frequency in Hertz. Now, I want to know if we perceive the frequency what is my resolution power? In which scale a human being normal shearing human being perceive the frequency. So, that is scale is called Mel scale. What is the logarithmic scale? Mel scale. How it is derived.

Now, suppose I do a perception text, that I varying the frequency of a pure tone signal and playing in the signal in a normal audio or micro phone in a loud speaker sorry loud speaker and I told the listeners when you perceive that the frequency is double then raise your hand like that amplitude. So, I can develop a scale on which human being perceive the frequency. If you see it will come look like this. So, this scale is called Mel scale, which is a logarithmic scale ok.

So, I can say the human perception of frequency the, I can derive by experiment these axis the frequency. So, I generate a pure tone of different frequency and I generate a pure tone, for suppose I generate a pure tone of 50 Hertz, then I said that you raise your hand when you perceive the frequency is double. So, let us it is come around 100 Hertz, then I just 55 Hertz, all kind of variation I play and he raise in here; raise your hand in here raise hand in here like here then may be here then may be here. So, from that point I can draw a curve and find out that the perception of the human frequency perception of the human is not linear.

Initial period may be up to 500 Hertz to some 500 Hertz its scale is linear, but after that it becomes a non-linear scale. This scale is called Mel scale.

(Refer Slide Time: 05:00)



Pitch-The Mel Scale

Pitch $(mels) = 3322\log_{10}(1 + f/1000)$

Alternatively, we can approximate curve as:

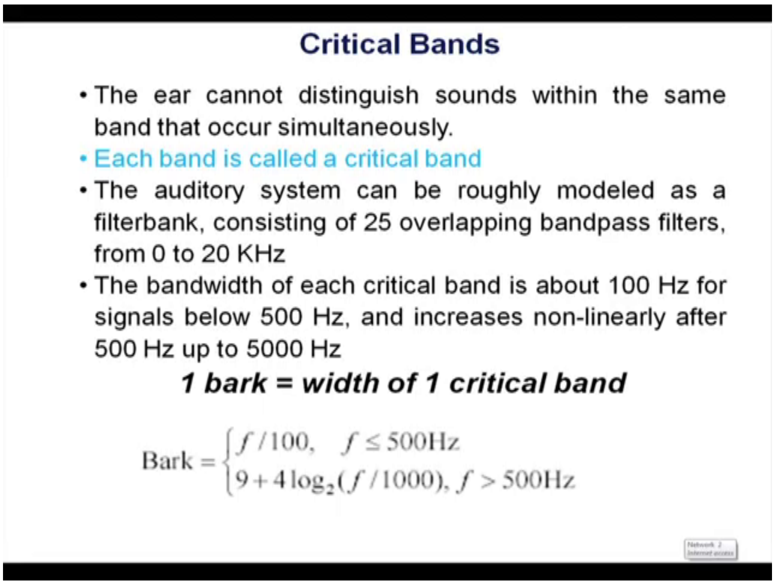
Pitch $(mels) = 1127\log_{e}(1 + f/700)$

So, in Mel scale I can derive the mathematical equation of this curve and the Mel scale even pitch in Mel, there is a two equation I can use any one of them like 3322 log 10, 1 plus f by 1000 or 1127 log e 1 plus f by 700. So, this equation is actually fitting this curve.

So, this equation is called picth Mel scale Mel equation, pitch in Mel scale. So, suppose I have f is equal to 2 Kilohertz, then I can find out what is the value of the f in Mel scale?

So, I can put the pitch in Mel scale, let us pitch in Mel scale Mel is equal to 1127 log let us e 1 plus 2 Kilohertz 2000 divided by 700. So, I can find out the Mel frequency. So, this is called Mel scale you can say Mel scale ah conversion of the hertz to linear perception of the frequency.

(Refer Slide Time: 06:14)



Critical Bands

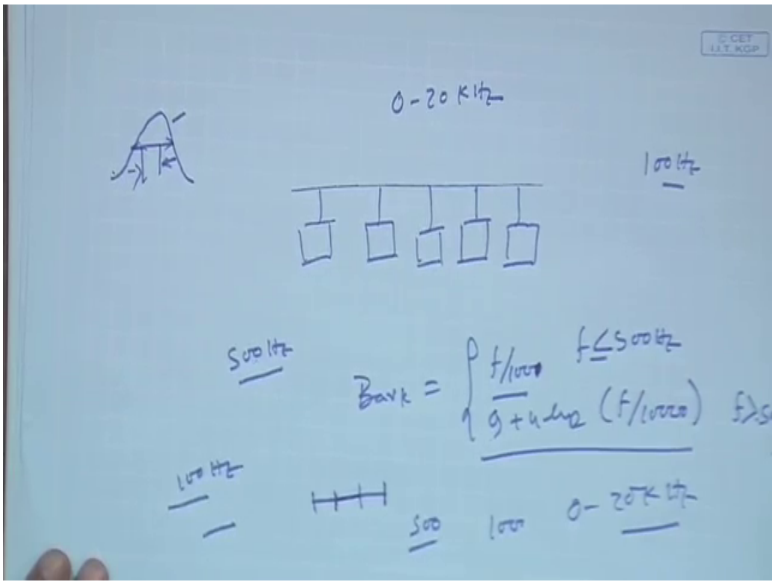
- The ear cannot distinguish sounds within the same band that occur simultaneously.
- Each band is called a critical band
- The auditory system can be roughly modeled as a filterbank, consisting of 25 overlapping bandpass filters, from 0 to 20 KHz
- The bandwidth of each critical band is about 100 Hz for signals below 500 Hz, and increases non-linearly after 500 Hz up to 5000 Hz

1 bark = width of 1 critical band

$$\text{Bark} = \begin{cases} f/100, & f \le 500\text{Hz} \\ 9 + 4\log_2(f/1000), & f > 500\text{Hz} \end{cases}$$

Next, if I heard of the I will say I have discussed it, during the basilar membrane that we perceive that each sensor along the basilar membrane is corresponding to a particular band of frequency so that means, the ear cannot distinguish sound within the same band that occur simultaneously.

(Refer Slide Time: 06:40)



$$\text{Bark} = \begin{cases} f/1000 & f \leq 500\text{ Hz} \\ 9 + 4\log_2 (f/1000) & f > 5 \end{cases}$$

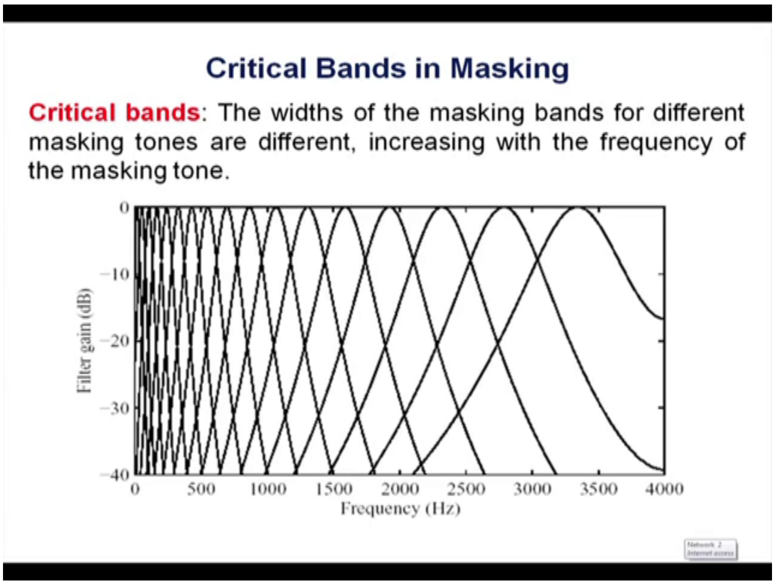
Suppose, a cochlear that along the basilar membrane particular band of frequency is perceive by a one sensor. So, if within this band if the frequency occur, then I cannot distinguish the difference between the two frequency.

So, this band is called critical band. So, the auditory system can be roughly modelled as a filterbank consist of 25 overlapping band pass filter, which is varies from 0 to 20 Kilohertz. So, that is band so I can say instead of human auditory systems I can think about engineering model of some bandpass filter, whose frequency bands are non-linear and 25 overlapping non-linear bandpass filter, non-linear band width bandpass filter can be completely model the human auditory system.

Now, bandwidth are each critical band is about 100 Hertz. So, signal below 500 Hertz it is linear and if it is increase above 500 Hertz it is becomes non-linear. So, that band width is define as a bark. So, one bark is equal to width of a critical bandwidth. So, bandwidth in bark scale. So, bark scale is equal to f by 1000 if f 500 Hertz 9 plus 4 log 2 f by 1000, sorry; 100, 1000 if it is f is greater than 500 Hertz.

So; that means, within 500 Hertz bandwidths are 100 Hertz bandwidth, with overlapping if it is 50 percent overlap. So, that is a band 100 Hertz bandwidth, then from 50 Hertz, another 100 Hertz bandwidth. So, up to 500 Hertz bandwidth are linear 100 Hertz. After 500 Hertz the bandwidth is non-linear. So, I can find out how many critical band is required to cover the 0 to 20 Kilohertz. So, that bandwidth is called bark scale.

(Refer Slide Time: 09:17)

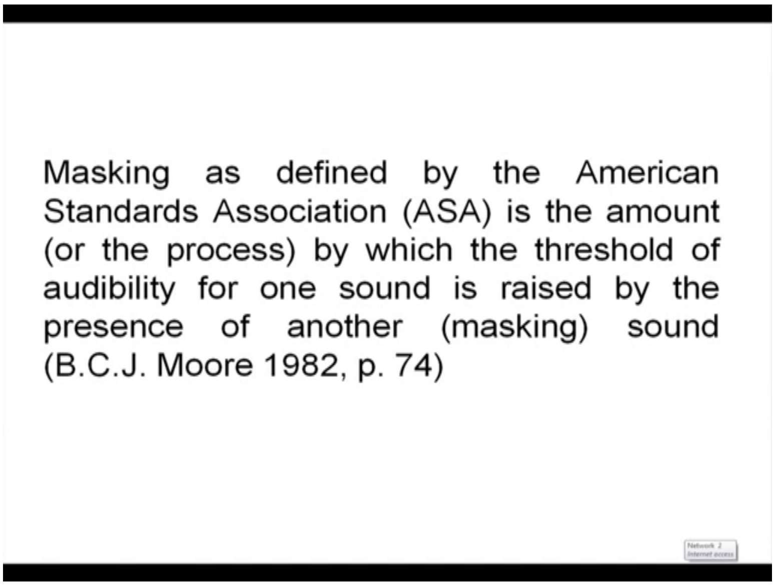


Critical Bands in Masking

Critical bands: The widths of the masking bands for different masking tones are different, increasing with the frequency of the masking tone.

So, that is the critical band pictures of critical band, and I will again I will discuss with whenever do think about the (Refer Time: 09:20) kind of things ok.

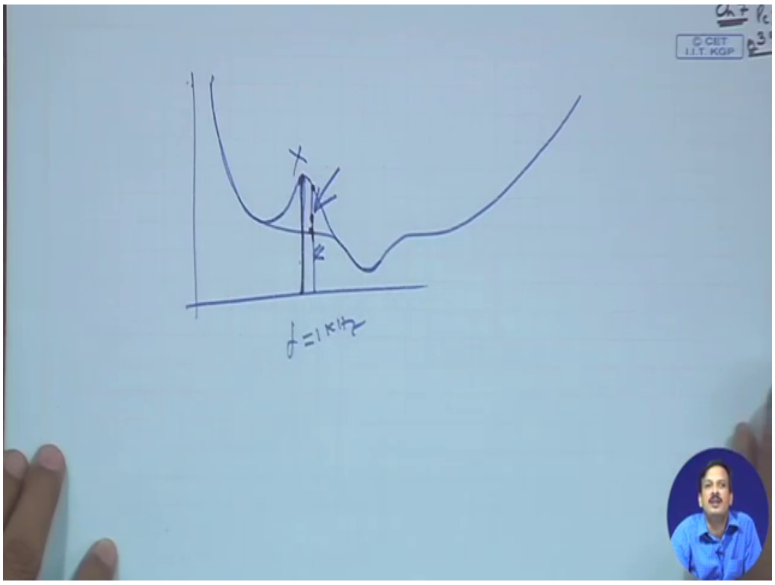
(Refer Slide Time: 09:24)



Masking as defined by the American Standards Association (ASA) is the amount (or the process) by which the threshold of audibility for one sound is raised by the presence of another (masking) sound (B.C.J. Moore 1982, p. 74)

So, this is the bark scale. Next another phenomenon is called frequency masking. So, perception of frequency and another is called masking. So, what I said the human being has a threshold of earing. So, threshold of earing curve, I can say this is the frequency this is the threshold of earing ok.

(Refer Slide Time: 09:42)

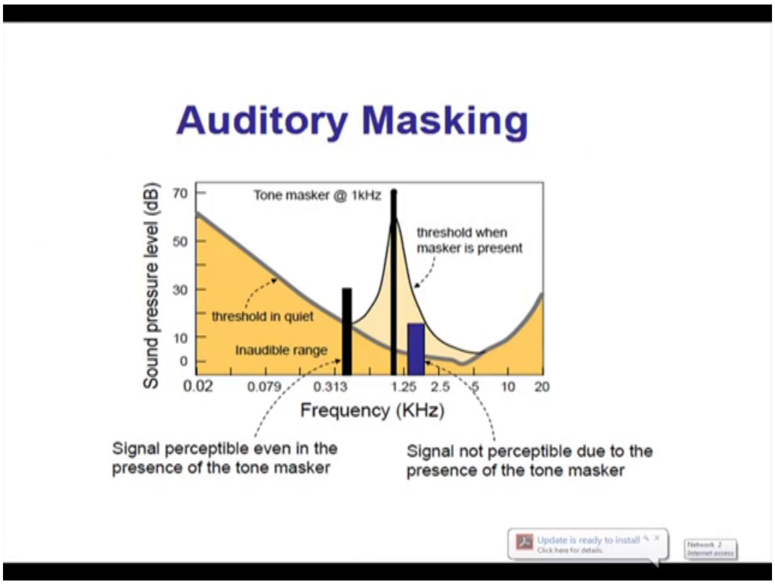


Now, it is said that presence of a particular tone, if some tone is present in here; pure tone suppose some tone is high tone is present in here then it is said that nearby bandwidth threshold of bandwidth threshold of earing bandwidth will change; that means, if there is a strong particular tone is present less f is equal to 1 Kilohertz, 1 Kilohertz; let us 1 Kilohertz then nearby; frequency threshold of earing is shifted upwards.

So, I cannot perceive if it is this tone is not present, I can perceive this frequency, but since this tone is present and if I have frequency amplitude of like this I cannot perceive it required a amplitude to cross this limit. So, that then; that means, I can hide this frequency because of presence of this tone. So, this phenomenon is called frequency masking. And this is utilize in speech coding to hide the coding hide the oh noise. So, suppose the coding generate a noise. If the noise is within this limit, then I can say this noise is not persuadable.

So, to hide this noise frequency masking is much more useful. I am not going details about the frequency masking there is a lot of details on frequency masking.

(Refer Slide Time: 11:29)



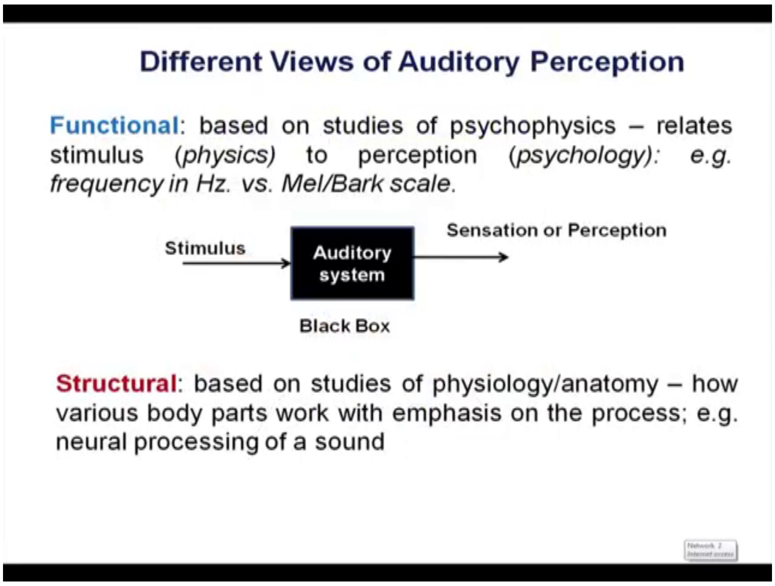
Auditory Masking

Signal perceptible even in the presence of the tone masker

Signal not perceptible due to the presence of the tone masker

So, if you are studying the speech coding, then auditory masking is very important. Next one is different view of auditory perception. So, there is a I can say the auditory perception has a 2 view: one is called functional view, which is means which means.

(Refer Slide Time: 11:36)



Different Views of Auditory Perception

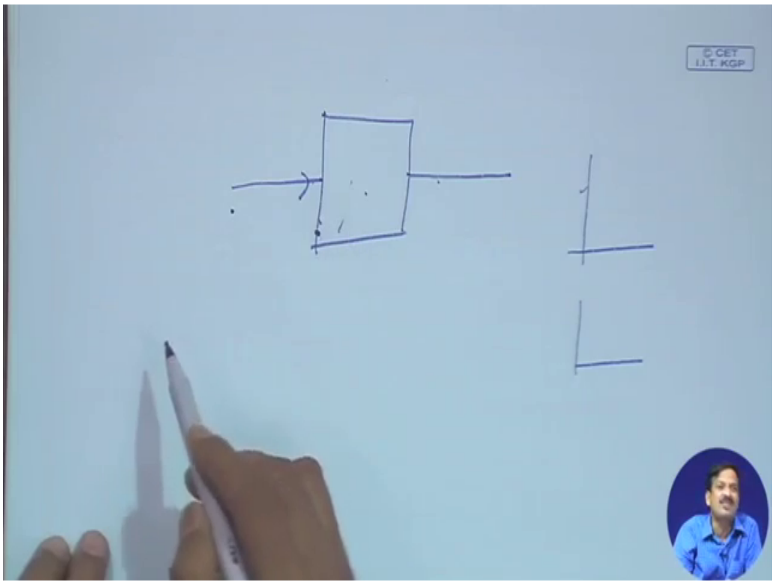
Functional: based on studies of psychophysics – relates stimulus (*physics*) to perception (*psychology*): *e.g. frequency in Hz. vs. Mel/Bark scale.*

Structural: based on studies of physiology/anatomy – how various body parts work with emphasis on the process; e.g. neural processing of a sound

Suppose, I cannot know what is happening in here, I cannot do the anatomy and I cannot biological system I cannot measure.

But, if I consider the human hair human auditory system perception is nothing, but a black box.

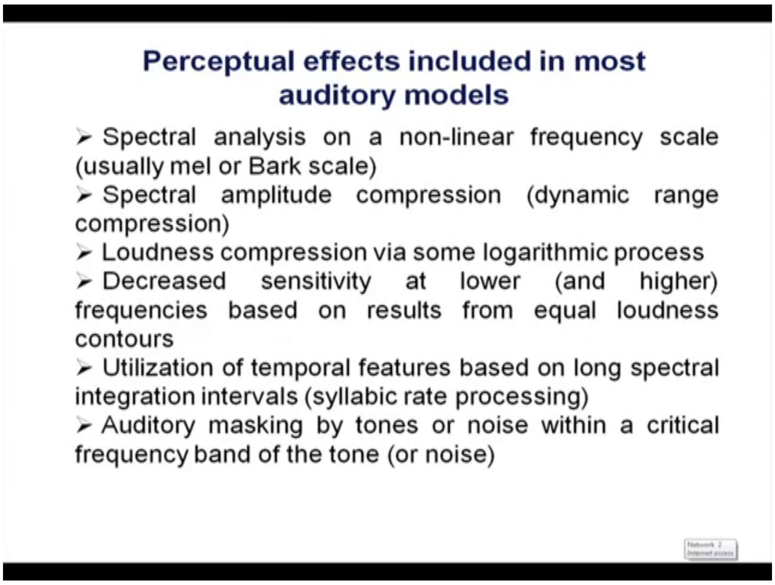
(Refer Slide Time: 12:02)



Which, I discuss in the first class, then I can say I can stimuli the systems. So, I can excited the system by a known stimulus. And then I can measure the physiological behaviour coming out from the system like that; clear the development of shown scale of mel scale is the example of functional modelling. Another one is called structural modelling, based on the study of physiology or anatomy. How various body parts work with emphasis on the process or emphasis on the process neural processing of sound? So, this is another kind of study now structural study I have to analyse or human anatomy excited the signal measure the nervous signal all kind of things can be done which is called structural analysis.

So, I can say the functional analysis like that, how human being perceive the frequency is the functional modelling? I play a different sound and human being listen the sound. So, I excited the human perception by external stimulus which is known stimulus and observe the output. Then can develop try to develop the how human being perceive the frequency that is scale; that is Mel scale. So, there is a functional kind of auditory perception ok.

(Refer Slide Time: 13:31)



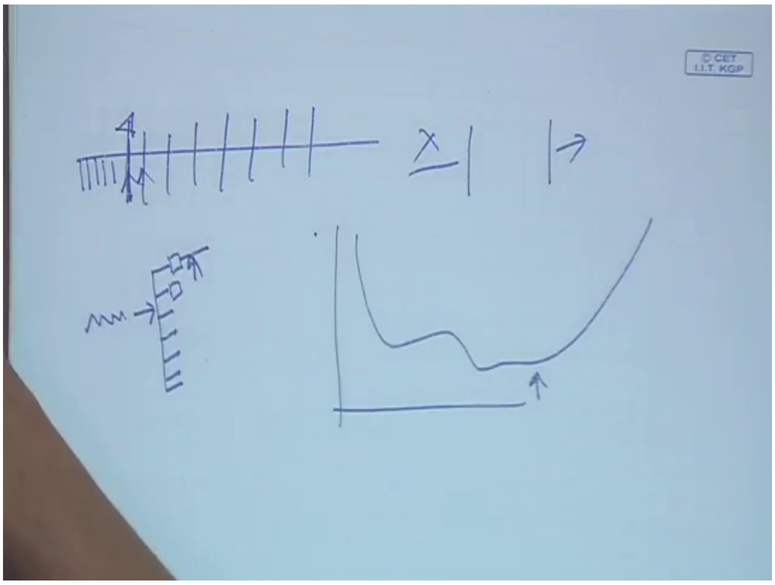
Perceptual effects included in most auditory models

- ➢ Spectral analysis on a non-linear frequency scale (usually mel or Bark scale)
- ➢ Spectral amplitude compression (dynamic range compression)
- ➢ Loudness compression via some logarithmic process
- ➢ Decreased sensitivity at lower (and higher) frequencies based on results from equal loudness contours
- ➢ Utilization of temporal features based on long spectral integration intervals (syllabic rate processing)
- ➢ Auditory masking by tones or noise within a critical frequency band of the tone (or noise)

Second one is the why we have to know the perceptual modelling or how human being perceive the frequency. So, perceptual effect include the most auditory model: spectral analysis on a non-linear frequency scale, spectral amplitude compression, loudness compression via logarithmic scale. So, we are not we are not we can say that physical loudness is not physical intensity is not equivalent to the perceptual loudness**.** So, there is a logarithmic compression is there.

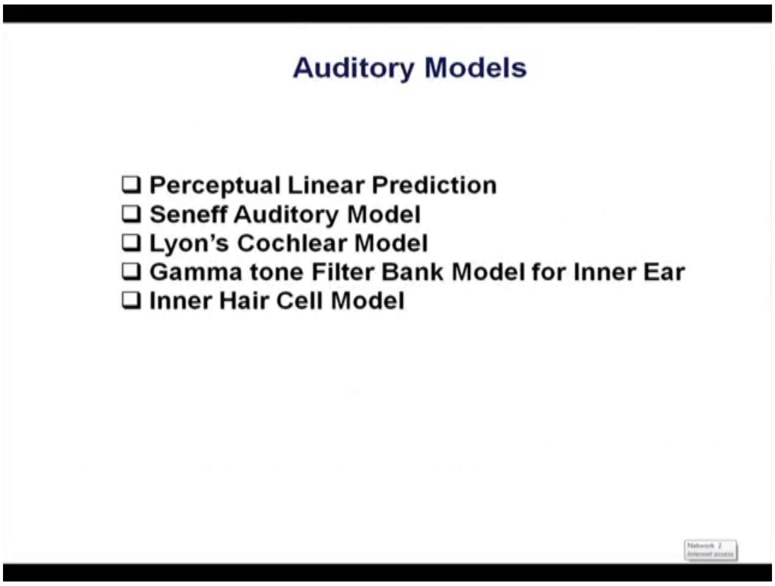
Decreased sensitivity at lower frequency; you know that lower frequency are very we have sensitive to the lower frequency, but higher frequency we average there is a bandwidth critical bandwidth critical band is very big. So, average. So, resolution human resolution of the higher frequency is very rough. I can say roughly approximate. So, I can say decreased sensitivity at lower frequency and increase sensitivity at; decreased sensitivity at lower frequency or you can say the change of perception of the frequency and lower than do it is much linear and the upper bandwidth is non-linear.

(Refer Slide Time: 14:35)



Then, utilization of temporal features and auditory masking of tones, so those phenomenon can be used in auditory modelling or when we extract the speech parameter from the speech, we have to include this perceptual variation. So, that the parameter directly represent the human speech perceive what we have done in the inside the ear ok.

(Refer Slide Time: 15:07)



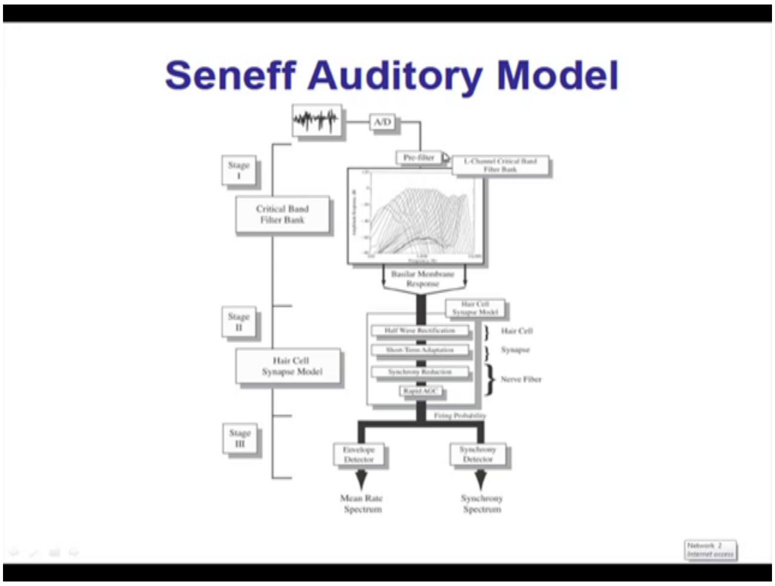
Auditory Models

- ❑ Perceptual Linear Prediction
- ❑ Seneff Auditory Model
- ❑ Lyon's Cochlear Model
- ❑ Gamma tone Filter Bank Model for Inner Ear
- ❑ Inner Hair Cell Model

So, that is called perceptual modelling. So, the different auditory models are available. Perceptual linear prediction this is called PLP only known as PLP. So, this details I will cover during the linear prediction analysis this is called perceptual scale linear prediction.

Then, Seneff auditory model, Lyon cochlear model, Gamma tone filter bank model or inner hair or inner hair cell model. So, all are called auditory modelling of human speech processing. So, let us I just discuss one or two model ad then you can study it.

(Refer Slide Time: 15:40)



Seneff Auditory Model

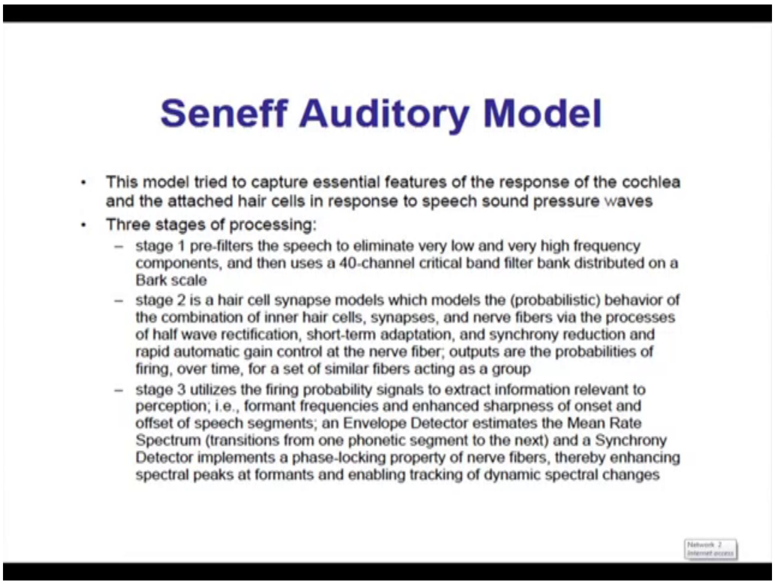
So, one model is that Seneff auditory model. If you see, what is auditory model? So, how human being perceive the speech sound, I have to implement in auditory model if you see there is a. So, stage 1, stage 2 and stage 3. So, what is stage 1? What we have done? We have a series of basilar membrane filter.

So, I can say I can develop a basilar membrane filter, by a filter bank; non-linear filter bank. So, there is a pre filtering; per filtering means since it is a digitized signal. So, what I can do I can develop a pre filtering after a d c. So, that high frequency and low frequency signal are corrupt. And then I pass the signal through a bandpass or equation of chunk of bandpass filter which is called basilar critical band filter.

So, critical band filter bank. So, I have a speech signal I pass this speech signal let us forget the pre filtering, pass this pre signal through a various frequency band filter. So, each filter output or you can say the critical band filter output give me the response of a basilar membrane filter bank each sensor. Now I have to know which sensor is firing I have to define which; so this is called the hair cell firing. So, this hair cell firing I have to find out which filter is firing based on the collected energy at the output of the each filter band. So, what the stage? So, second stage is called you can say the modelling of hair cell. So, this is called the hair cell modelling.

So, half wave rectifier, rectification for find out the energy, then short term, adaptation, synchrony residence, now the rapid a g c automatic gain control, then I can get the envelope detector mean rate spectrum and synchrony detector synchrony spectrum. So, I get a spectrum how human being perceives frequency at the output of this stage. So, details are here stage 1, stage 2 stage 3 you can read.

(Refer Slide Time: 18:12)

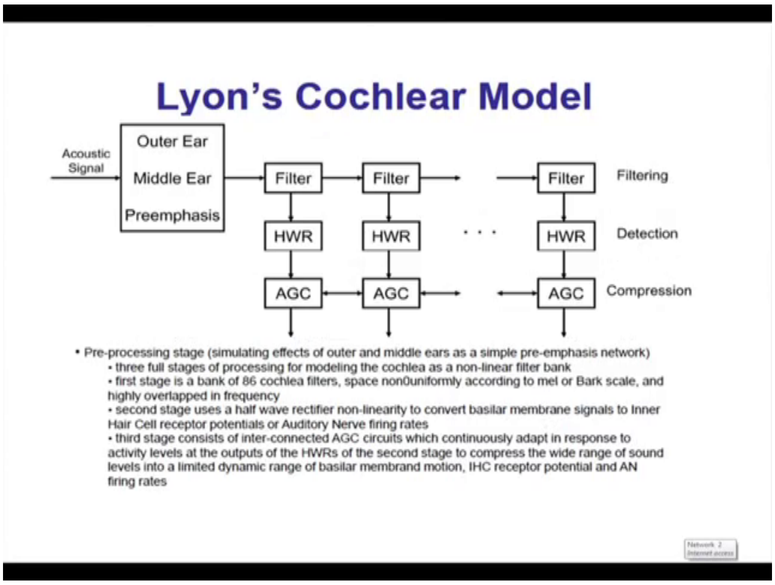


Seneff Auditory Model

- 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
- 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

So, this is nothing, but a idea is that I have a input speech I pass through that as a chunk of bandpass filter, which is actually critical band filter and each filter output is nothing, but a sensor response. So, I calculate the response and I have to adjust that firing throughout that you can say the amplitude compression that part as to be done. So, that pass I have done and I the estimate the firing of the hair cell to perceive the frequency that is called Seneff auditory model.

(Refer Slide Time: 18:51)



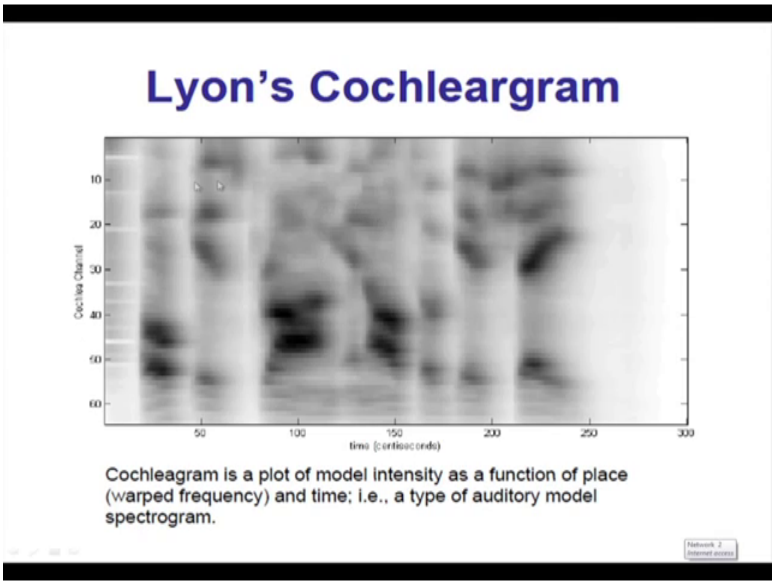
Lyon's Cochlear Model

- Pre-processing stage (simulating effects of outer and middle ears as a simple pre-emphasis network)
  - three full stages of processing for modeling the cochlea as a non-linear filter bank
  - first stage is a bank of 86 cochlea filters, space non0uniformly according to mel or Bark scale, and highly overlapped in frequency
  - second stage uses a half wave rectifier non-linearity to convert basilar membrane signals to Inner Hair Cell receptor potentials or Auditory Nerve firing rates
  - third stage consists of inter-connected AGC circuits which continuously adapt in response to activity levels at the outputs of the HWRs of the second stage to compress the wide range of sound levels into a limited dynamic range of basilar membrand motion, IHC receptor potential and AN firing rates

Then, Lyon's cochlear model same things. So, acoustic signal outer ear, middle ear, pre emphasis. So, acoustic signal is if we pass through a you can see we have a this kind of a frequency response of the middle ear and outer ear. So, these can be pre emphasised signal can be pre emphasized using inverse response. After the pre emphasis I can pass the signal through a chunk of filter. So, so that is 86 cochlear filter banks; here it is designed 86 cochlear filter bank in mel or bark scale.

Then it is pass to the half wave rectifier to detect the amplitude. Then a g c automatic gain compression to find out the frequency response of this acoustic signal, as per the auditory response of human being.

(Refer Slide Time: 19:54)

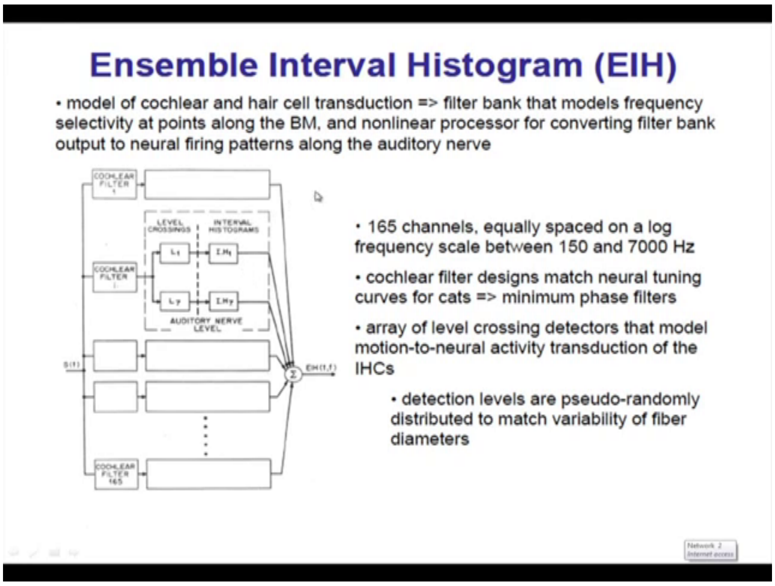


Lyon’s Cochleargram

Cochleagram is a plot of model intensity as a function of place (warped frequency) and time; i.e., a type of auditory model spectrogram.

Then there is another model. So, this is the Lyon’s cochlear gram this I will come later on spectrogram.

(Refer Slide Time: 19:58)



Ensemble Interval Histogram (EIH)

- model of cochlear and hair cell transduction => filter bank that models frequency selectivity at points along the BM, and nonlinear processor for converting filter bank output to neural firing patterns along the auditory nerve

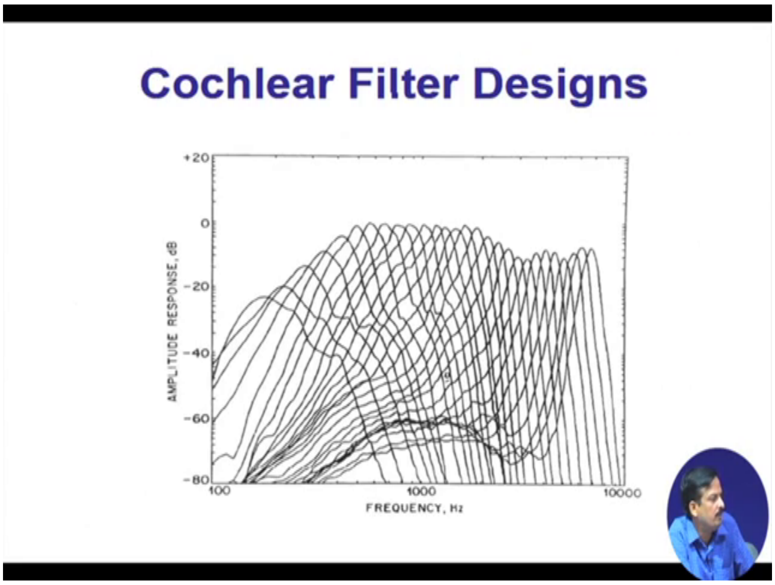
- 165 channels, equally spaced on a log frequency scale between 150 and 7000 Hz
- cochlear filter designs match neural tuning curves for cats => minimum phase filters
- array of level crossing detectors that model motion-to-neural activity transduction of the IHCs
    - detection levels are pseudo-randomly distributed to match variability of fiber diameters

Then Ensemble interval histogram EIH model. So, the model of cochlear hair cell transduction. So, I can say this we along the basilar membrane; there is a lot of there is a cochlear hair cell sensors each sensor consist of ten fibres.

So, that EIH-Ensemble interval histogram is nothing, but a model of cochlear hair cell transduction. This transduction this motion how it is transduced this motion to the transduction is done place take place that is model here. How it is model? 165 channel equal space on a log frequency scale between 150 to 700 Hertz. 150 channel filters sorry 165 channel filter. So, each cochlear filter design match neural tuning curve of for cats minimum phase filter.

Array of level crossing detector that model motion-of-neural activity transduction of the inner hair cell. And then we sum it and get the response; details you can also study if it is required.

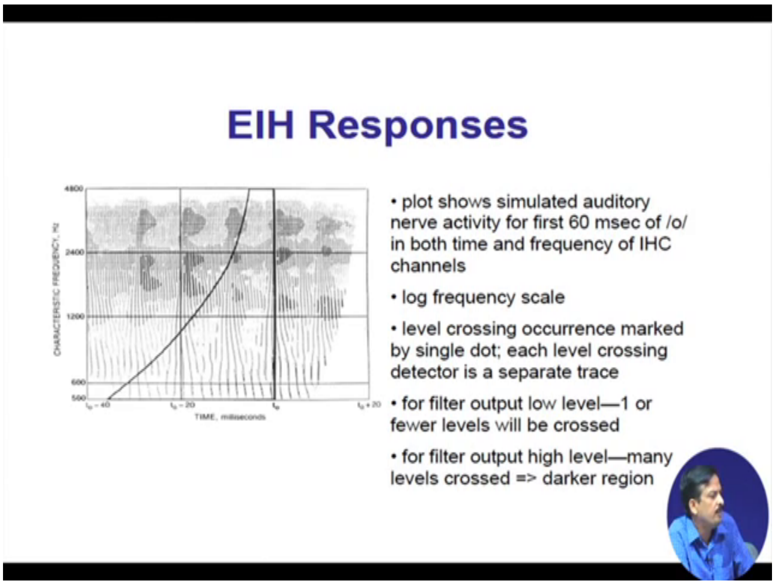
(Refer Slide Time: 21:11)



Cochlear Filter Designs

Then cochlear filter design how it is design.

(Refer Slide Time: 21:13)

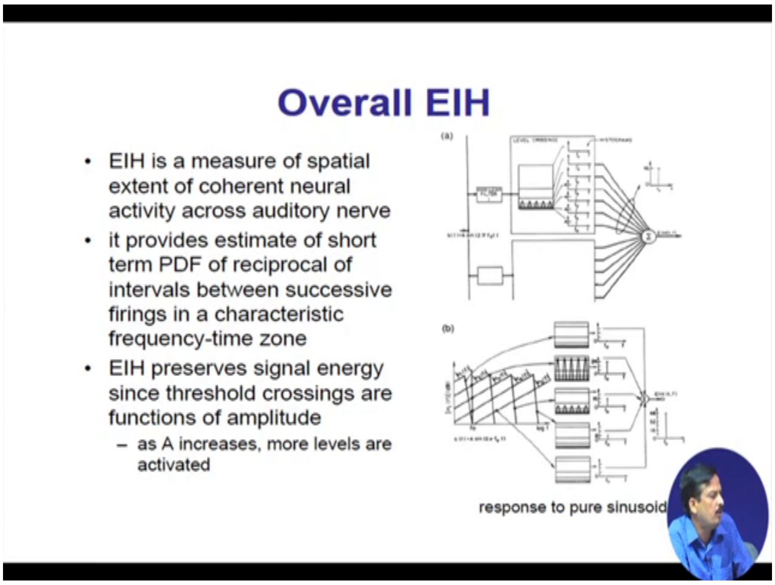


EIH Responses

- plot shows simulated auditory nerve activity for first 60 msec of /o/ in both time and frequency of IHC channels
- log frequency scale
- level crossing occurrence marked by single dot; each level crossing detector is a separate trace
- for filter output low level—1 or fewer levels will be crossed
- for filter output high level—many levels crossed => darker region

Then EIH response overall EIH response.

(Refer Slide Time: 21:16)



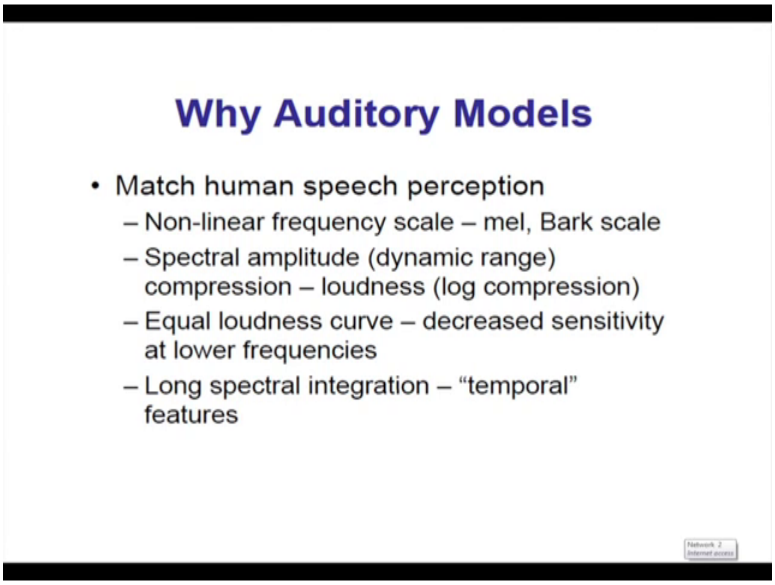
Overall EIH

- EIH is a measure of spatial extent of coherent neural activity across auditory nerve
- it provides estimate of short term PDF of reciprocal of intervals between successive firings in a characteristic frequency-time zone
- EIH preserves signal energy since threshold crossings are functions of amplitude
  - as A increases, more levels are activated

response to pure sinusoid

That is EIH measure the spatial extend of coherent neural activity across auditory nerve ok.

(Refer Slide Time: 21:30)



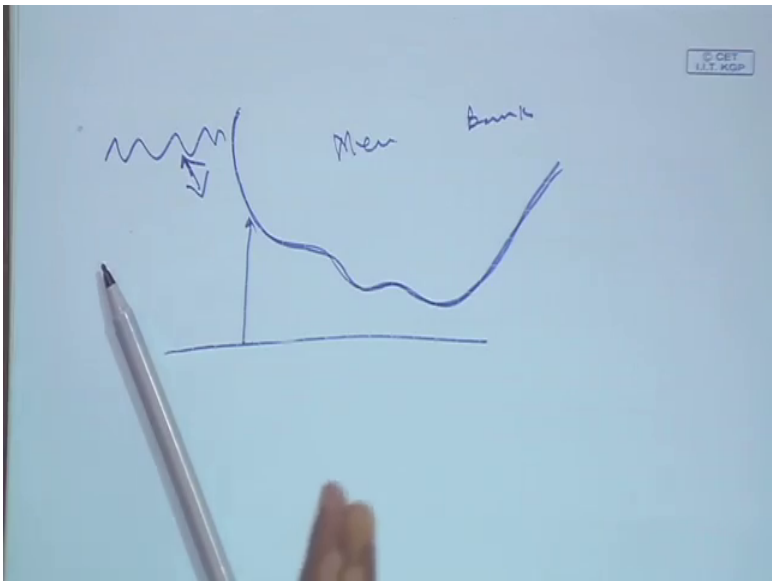
Why Auditory Models

- Match human speech perception
  - Non-linear frequency scale – mel, Bark scale
  - Spectral amplitude (dynamic range) compression – loudness (log compression)
  - Equal loudness curve – decreased sensitivity at lower frequencies
  - Long spectral integration – "temporal" features

So, this kind of auditory models are mainly we used that PLP that we will discuss during the linear perceptual coding LP.

When we discuss the LP; that time we discuss the PLP details. Now, why these auditory models are important in human speech perception? So, non-linear frequency scales.

(Refer Slide Time: 21:57)



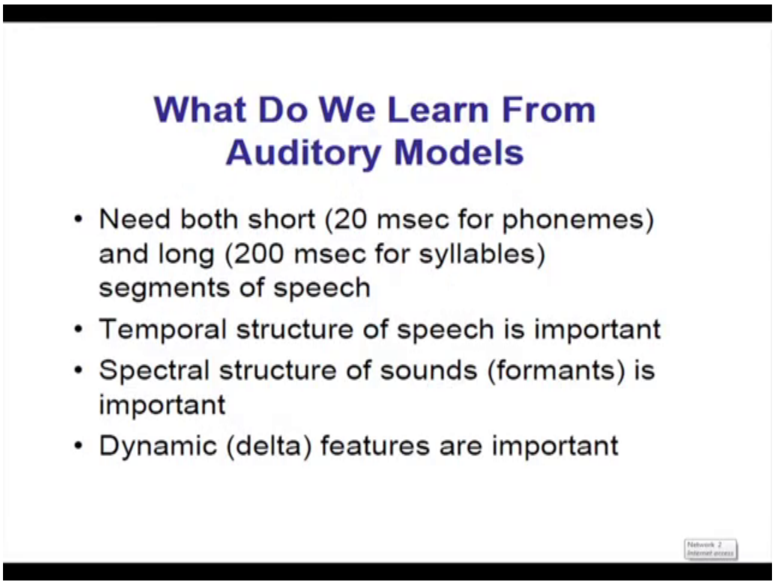
So, suppose I have a speech signal, I extract the parameter in linear frequency scale then physically I am doing is that I can extract the physical frequency, but how human being perceive, that is important also do we incorporate. So, if I want to incorporate that then all this frequency scale up the signal must be non-linear which is logarithmic scale either in mel scale or bark scale ok.

Then spectral amplitude dynamic range timber is in important parameter. So, spectral amplitude dynamic range or compression or loudness log. So, if I see each tone has an amplitude. Now perception of that tone, that particular frequency depends on the threshold of earing. So, perception of amplitude for all frequency is not same I have to go through the equal loudness curve. So, I can say spectral amplitude compression has for the equal loudness curve I have to do. So, I can dynamic range or amplitude of the spectrum as to be compressed as per the threshold of earing.

Those earing sensitivity of the human ear to the amplitude is not same or not equal to all frequency. So, as per the sensitivity is changing along the frequency so that sensitivity changing as to be model in spectral envelope also or equal loudness curve, log spectrogram integration and temporal features. So, I can say the temporal feature is very important in speech signal. How the spectral dynamics is changing is also a part of the speech signal also a features of the speech signal. So, that is also very important for speech perception.

Timber is one of the example, so suppose I am producing a signal somebody else is producing a signal same tone or same thing suppose that is example I have already given that suppose I am singing a song which is followed the same notation same lyrics same you can say the duration is also same I exactly copy my guru, but if you see the guru sound my sound cannot be equal because of complexity of the speech. So, dynamic of the spectrum is very much important to know the complexity of the speech ok.

(Refer Slide Time: 24:32)

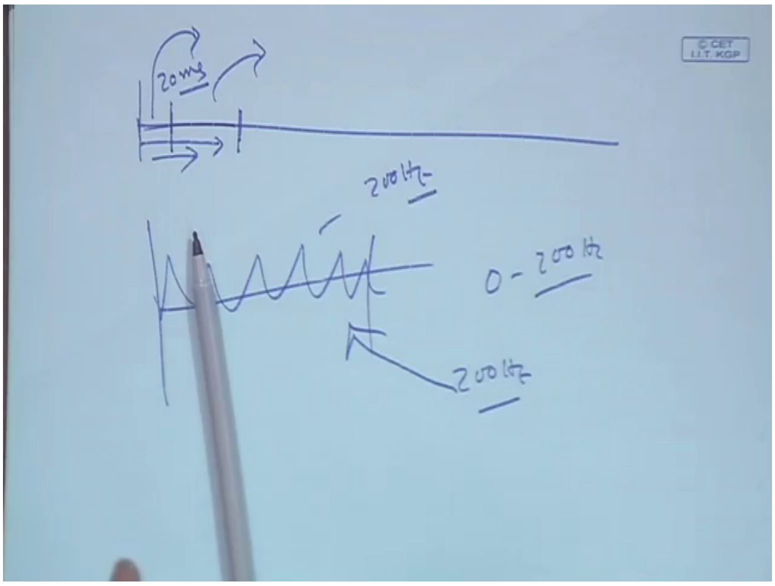


What Do We Learn From Auditory Models

- Need both short (20 msec for phonemes) and long (200 msec for syllables) segments of speech
- Temporal structure of speech is important
- Spectral structure of sounds (formants) is important
- Dynamic (delta) features are important

Then, what do learn from the auditory model? If I see speech; once I say the speech that if I say segmental and super I can say the speech is a not a stationary signal it is not a pure tone it is a stationary signal.

(Refer Slide Time: 24:51)



So, it change along the time; if I take a sort duration 20 millisecond for a let say for a phone and long duration for a speech segment. So, along the time property of the speech signal is change. So, dynamics of the speech signal is change. So, I can say 20 millisecond is also the parameter from the 20 millisecond is also in some kind of parameter we can get and from long interval also we can get some kind of parameter which is also important for speech processing.

So, temporal structure is very important for. So, I can say that speech contained, what kind of information not only the segmental information, suprasegmental information across the segment information is very much important. You can do an experiment; suppose you get a speech signal you record a speech signal and find out the fundamental frequency for this segment. Let us the fundamental frequency 200 Hertz. Design a filter or a you can say that I can let us 0 to 200 Hertz; design a 0 to 200 Hertz filter low pass filter cut the signal.

Again if you play the signal you still perceive the 200 Hertz is the fundamental frequency. How you got? That mean; spectral dynamics told you the 200 Hertz is the fundamental frequency. So, along the spectrum is also important and you can see the segmental and suprasegmental both information are necessary. So, dynamic features is also necessary the how the spectral dynamics is changed that also is necessary. So, dynamic features is change importance compression of loudness compression of the scaling of the frequency all are important which we learn from the speech perception.

So, in summary I can say how you perceive the intensity. Intensity is a physical parameter. So, how intensity related to the speech loudness is important. How the frequency which is a physical parameter is related to the perception of the speech of the signal? So, human perception of frequency, human perception of loudness and there mathematical model will be used in speech processing. So, that my parameter extraction when I do, I it should follow the human auditory systems. So, that is called speech perception ok.

Thank you.