Cognition and its Computation
Prof. Rajlakshmi Guha
Prof. Sharba Bandyopadhyay
Biotechnology and Bioengineering
Indian Institute of Technology, Kharagpur

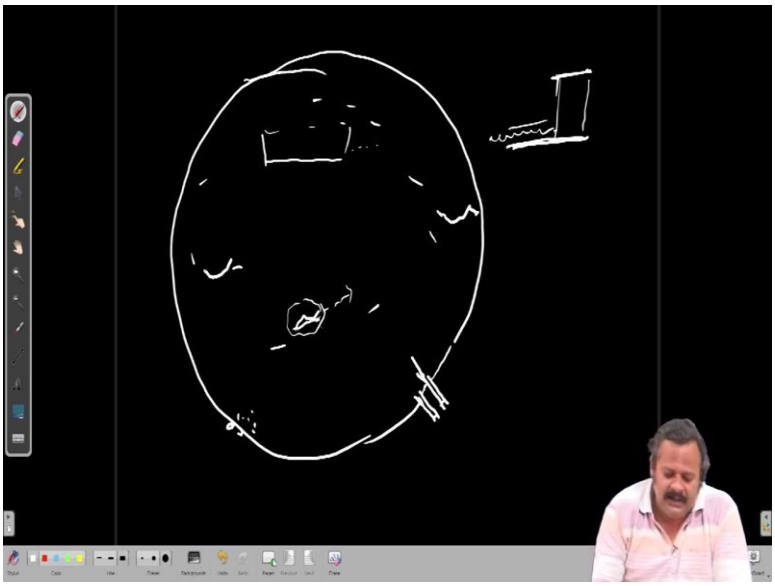
Lecture - 35
Auditory Scene Analysis, McGurk Effect

Welcome. We have been discussing about visual object recognition. So, we ended the discussion by saying that we will briefly touch upon similar aspects in auditory object recognition and what usually we call that area is auditory scene analysis and the problems there are distinct from the ones that we saw in the visual system in the sense that there is a temporal component that is continuously involved in auditory objects.

So, actually truly speaking it is difficult to even define an auditory object; however, if we think of a single source of sound that is that has a particular characteristics of it and that is continuing over time that may be called an auditory object and the auditory system has is being continuously bombarded with many such sources of sounds which gives rise to the idea of auditory streams over time.

And so, the auditory system has to identify and recognize and follow through each particular stream when whenever which one is whenever whichever is required and. So, to do that it has to rely on many many factors bottom up as well as top down.

(Refer Slide Time: 02:29)



And so, let us try to give an example of understanding what this whole idea of auditory scene analysis is. This is how actually the father of auditory scene analysis Bregman introduces auditory scene analysis and that is if you think of lake let us say a large lake and let us say we cut it out and there are two narrow channels that are cut out here from that lake.

So, this is full of water this is full of water and there are two narrow channels that are cut out and let us say there are two handkerchiefs that are hanging on top of the water. So, its as if for from the side view of the of one channel we will have a particular let us say handkerchief that is hanging across the flow of the water.

This is the water flowing along the channel. So, water is flowing into this channel and let us say this handkerchief is just hanging on top of the water surface and there are two of those and let us say we are making a video of the way the handkerchief keeps moving or the two of them keep moving based on the water flow in there.

And in the meantime in this lake let us say there is a large steamer large ship flowing in one direction there is a small boat going in one direction there is another boat that is let us say anchored here and may be waving in the water there are children playing on the beach jumping around on this on the shores where the water is touching the shores and there is someone who is swimming here on the on this lake and so, on and so, forth and

so, what the auditory system is doing is basically taking those movements of those handkerchiefs.

And trying to identify each of those each of those events that are going on or each of those phenomena that are happening in the lake. So, I hope if you recollect our original discussions of the auditory system, you would realize the connection that is being made here. So, those handkerchiefs are essentially like the ear drums and the waves are basically the sound pressure waveforms and there are many sources of those waves coming in.

And there are two ears separated by our head and those vibrations are impinging on those on those handkerchiefs or ear drums and from those vibrations from those fluctuations in the or the movement of the handkerchiefs, let us say the video of the movement of the handkerchiefs one has to tell what is present in the water where and so, on. Let us say I mean this is you can imagine how difficult this task is.

However, amazingly the auditory system indeed actually solves this problem and of course, not based on the videos, but by transducing those vibrations and breaking them down into frequency first as you remember through the auditory nerve then creating parallel pathways and gradually along the hierarchy, there is integration of different features to finally, form the percept of each auditory so, to speak object in the acoustic scene around us.

So, the one of the biggest problems or one of the one of the classical problems that we talk about in auditory scene analysis is the cocktail party phenomena and that is basically you are at a party where you have a large number of people around you. And let us say there is someone playing the piano, there is bell ringing from the waiter calling others there is someone there is a large loud debate going on between two groups of people and then there is also one person in front of you who is speaking to you and you are trying to listen to this person.

We do amazingly in terms of simply picking out this person conversation the persons speech and understand what this person is doing saying and we can continue the conversation in spite of all the other sounds going on. So, to speak here based on these vibrations coming in we can keep on saying which way this swimmer is going in this lake based on the vibrations that are being picked up.

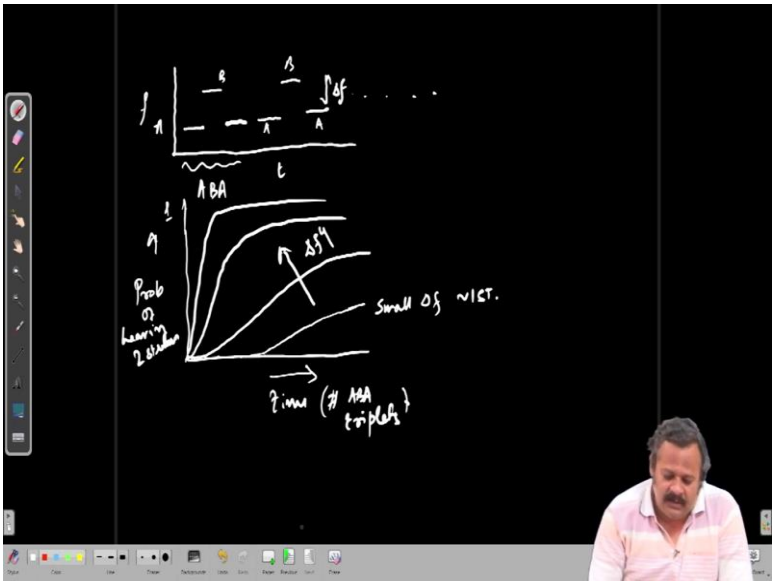
And from and remove the vibrations that are created on the water the waves that are created on the water from all the other movements going on all the other phenomena that are happening. So, again in that cocktail party effect we also know that if we want we can simply shut out the speech of the person in front of us and immediately turn our attention on to the debate that is going on at the corner of the room between two groups.

And we can we are picking up the conversation from that corner of the room and not listening to the person right in front of us trying to tell us something. I mean the of course, it is rude, but that is something that we often tend to do if there is something that is very interesting that catches our attention.

But so, it is this brings us to the point that the auditory scene analysis is not just bottom-up kind of phenomena where the sensory inputs decide on what we are listening to, but also a big role is being played by top down inputs or attention letting us selectively attend to a particular conversation or particular sounds or particular streams of sounds coming to us.

And so, this brings us to the idea of how the auditory system has been or rather the auditory scene analysis has been studied is with the ideas of streaming and grouping where a very although artificial, but one can understand how the percept of streams using simple pure tones.

(Refer Slide Time: 10:08)



So, let us say this axis is frequency, this axis is time and let us say we have this we will be representing a spectrogram that is how the energy in the sound is varying with time as at different frequencies as we have referred earlier in our auditory circuits lectures.

So, let us say there is a particular frequency being played that is frequency A or the sound A and then it is followed by a different frequency sound that is B followed by another A. So, these triplets ABA triplets then keep repeating. So, this is ABA and this goes on. So, depending on this separation delta f as time progresses one can listen to a single stream or it could also be two different streams.

One of the frequency A and the other of the frequency B. So, this is so, the it is a little probabilistic in the sense that the listeners if they are asked to say whether they are hearing two streams or one stream then the probability of hearing two streams gradually increases with time. So, essentially if the two streams this is the probability of hearing two streams.

And this is time that is the number of or rather the number of ABA triplets ABA triplets. So, for a small delta f initially there is 0 probability. In fact, of hearing two streams and then gradually the probability of hearing two streams increases and as the delta f. So, this is a smallest delta f or small delta f maybe about one semitone and as the delta f increases the this probability actually increases higher and higher to 1.

So, this is as delta f is increasing as delta f is increasing. So, this is basically when someone is passively listening and they are asked whether they are hearing one stream or two streams and this is the general kind of observation that we have and this gives us an idea that given simultaneous presentation of two maybe let us say two different sounds although not spatially segregated. In this case from the same source, we based on the input stimulus itself we may be able to differentiate two different streams based on the characteristics of the sound itself.

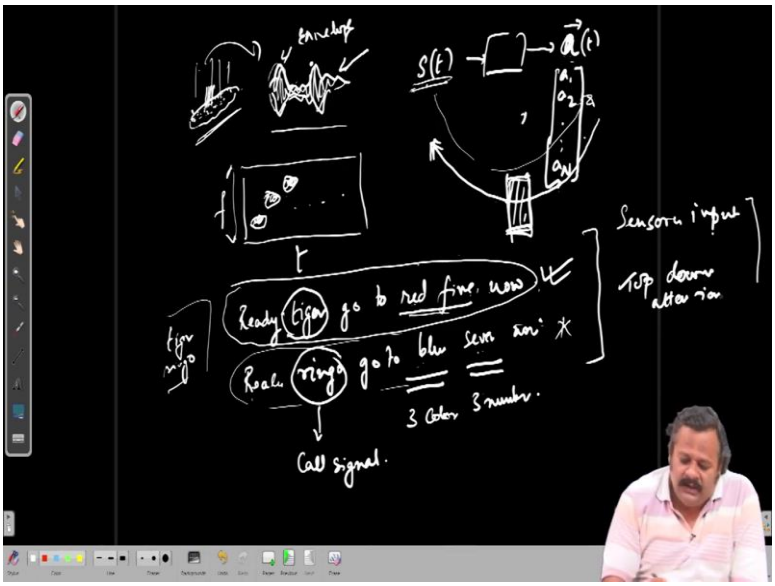
As we are saying that as the delta as the A and B come closer and closer to each other it appears that it is a single stream of sound we do not we cannot perceive two different streams, but as the delta f increases it actually clearly we hear two distinct streams of sound coming. So, this is kind of the bottom up way of separating different streams or different sources of sounds around us.

And here we have taken a small example of a toy example of single frequency tones, but extension of these ideas probably can explain streaming of different kinds of more complex features; however, as we have already noted and we know that in our course on cognition and its computation, we know that attention plays a huge role in terms of our what we perceive and in fact, what we our sensory cognition.

So, in that case the best example that or study where we know how or rather we get an inkling of how the brain is trying to do this segregation comes from studies where people have looked at simultaneous speakers and recorded from brain regions and identified and shown that actually when someone is paying attention to one particular sound that representation is enhanced and the other is reduced.

So, the way that experiment was done the most recent one was from about a decade ago 10 years ago that showed clearly from human listeners.

(Refer Slide Time: 16:24)



So, there are people with epilepsy who need surgery and require electrode placement on their on the on that in the temporal regions and actually in the temporal gyrus superior temporal gyrus which is involved in speech perception.

And so, in that particular region they have an array of electrodes implanted it just happens to be that those patients had an array of electrodes implanted in those particular regions and from these electrodes basically signals like local field potentials or at least

the low frequency not really spiking signals that we have talked about, but much more fine and more elaborate than eeg signals.

Because we are directly collecting the data from the neurons or being near the neurons because it is invasive and. So, when speech is presented so, this is over time this is the envelope of speech that is the amplitude envelope and inside there are different frequency contents that define the actual speech and we know that these low frequency temporal components play a big role in speech perception.

So, it is we will see that with the responses in the population of electrodes we have information about this speech envelope in them. So, like we had studied earlier in the auditory circuits part, where we talked about spectro temporal receptive fields similarly here one can look at the spectro temporal receptive fields of these particular population of electrodes based on the spectrogram of the envelope and that is the low frequency part.

So, we based on so, that is the forward aspect that is we have the speech envelope over time S t that is ultimately being represented in the activity of a population of neuronal responses or a that is a vector many different many different electrode positions and that is representing the activity at the different position. So, this is a let us say this is a 1, a 2 up to let us say there are n electrode electrodes and this is varying as a function of time.

So, from the speech input we are getting this activity and if you remember from these activity, we can predict the behavior or the receptive field of this box which could be a single neuron or the entire pathway in this case and similarly. So, that is to describe how this S t is being converted to a or activity.

Similarly, we can do mathematically with based on this a as a function of time we can actually go backwards and predict what S t was. So, that is the reverse problem. So, in this particular case what the thus the people the scientists did is that based on these activity they found out the particular transformation from this activity back on to this speech and so, essentially that is the transformation of this matrix of each vector over time back into the spectrogram of the envelope of the speech signal.

So, from here one is predicting what the envelopes spectrogram would be like. So, this is time and frequency this is going to be low frequencies because envelope is a low

frequency part of the overall amplitude variation and. So, the energies are different in different time points at different frequency regions and so, on. So, they have a model of how to go back given a set of activity of these electrodes.

They can pretty well predict what the stimulus spectrogram was or the speech envelope spectrograms. So, they have a model for this box here the backward solving problem. So, now, what they did is. So, they estimated this based on many different speech segments being presented and the average function was obtained. Now they did this cool experiment where the speaker or rather the listener the subject was presented with two competing speakers.

So, both of them said similar sentences like ready which was common in both cases tiger go to red 5 now and ready ringo go to blue let us say 7 now. They had three colors here 3 colors and 3 numbers and this tiger and ringo is the call signal or cue as to which sentence or which speaker the patient has to listen to and at the end of the sentence report the color number combination.

So, before the sentences the two competing sentences presentation is started the person is shown the picture of the tiger or ringo one of them and that tells the listener that they have to identify the color number pair of that particular speakers sentence. So, if it is skewed to listen pay attention to the tiger sentence, then it has it would have to report red five at the end of the sentence or if the cue is to listen to the sentence which contains ringo, then they have to report the color combination colored number combination in that particular sentence.

So, they have these competing sentences presented to the listeners or subjects and they report the color number combination. So, as you can see these and the speakers have variable and there are many many different such combinations possible and many different speakers are there and there then they had different pitch like male and female, they had different rates of speaking and so, on.

So, they could actually make a lot of errors in terms of understanding which particular number and color was being said by one particular person in this competing environment and so, what they did is that while the person was listening to these competing sentences presented simultaneously and the subject was paying attention to one particular stream of

sound that is the tiger. So, the person has to separate or segregate the tiger stream from the ringo stream the two speakers are different it they both have different characteristics.

And the person has to pay attention to one particular stream and ignore the other stream by and that is controlled by the fact that we know that they are hearing the correct stream and they are paying attention to that stream because they have to report this color number combination and it is not that they do this 100 percent of the time.

They do it 100 percent of the time when a single sentence is presented of course, but when two of them are presented they do not they are unable to predict they are unable to report the correct color number combination all the time 100 percent of the time and what they did is during this period they are similarly recording from the all the electrodes array the array of electrodes in the auditory region of the brain.

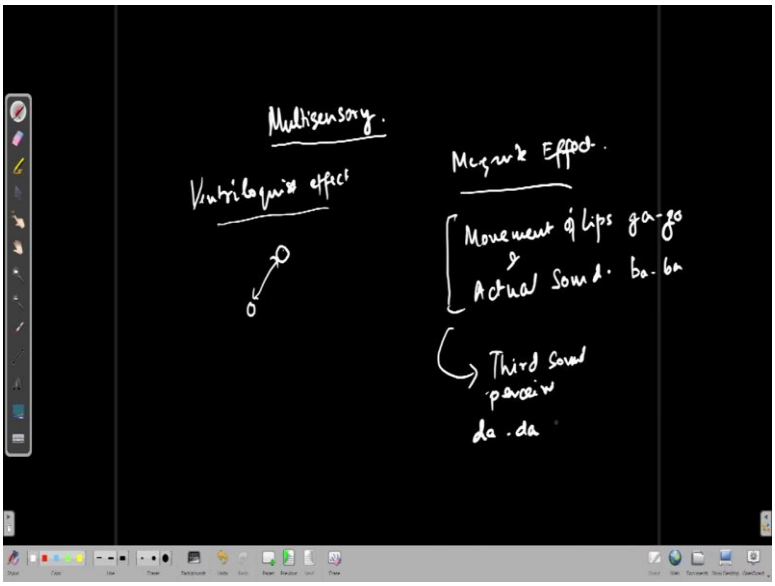
And when they look at that activity and reconstruct the signal from that activity they find that the activity closely matches with the attended speech and not and is not like the unattended speech and they also see that when they make errors in the reporting of the colored number then the correlation or rather the reconstruction of the speech envelope spectrogram does not match either of them or rather parts of it match with one parts of it match with other.

So, essentially one they were not paying attention correctly to them and hence the activity itself is telling them telling us that they were not paying attention and so, that is why they made the mistake. So, here as you can see the stream segregation is happening based on a top down input. Of course, the bottom up inputs are finally, providing a representation in the activity, but attention is shaping that final activity in the higher regions of the auditory pathway to make the activity such that it represents only the spoken or only the attended stream.

So, in terms of auditory scene analysis, these are the two things that we would like to convey to you as a being the basis for auditory scene analysis. So, the stimulus regularities or their predictive feature or their characteristics itself allows a stream of auditory information to be judged to be separate from the rest of the auditory scene or and or attention to a particular stream enhances the representation in the neural activity of that stream and actually suppresses the other stream.

So, in this case it is over time that the object is being formed. Unlike in the visual object recognition case we studied it was over space that integration is happening to form the percept of the visual object. So, here we have spoken the sensory input as one thing one aspect and top down attention as the other aspect.

(Refer Slide Time: 30:04)



Beyond this there is further another thing and that is the multi-sensory aspects that is not all objects are uni sensory. So, that sometimes or not sometimes most often actually objects are perceived as a whole from multiple different with multiple different modalities sensory modalities together. So, for straight away a very important example of that is any food.

So, it is not just the taste receptors but also the smell which we call flavor and the percept of that particular food is actually both of the taste and the flavor together and in fact, you can separate them out if you simply close your nose and stop breathing at that time then you will be able to separate the flavor from the taste and you should do this with some of the things that you eat which has a good flavor as well as a good taste or otherwise it is important to see how the percept totally changes if one of the cues is removed.

And similarly in the auditory somatosensory and visual system particularly in the auditory visual cross connections or multi-sensory integration what we find, we have many remarkable examples of percepts forming based on both auditory and visual inputs one of them is the ventriloquist effect and the other is the McGurk effect.

These are there are few others that are that have been studied strong intensely and so, here what happens is that we have a ventriloquist who is not moving their mouth and making the sound. So, the sound source is different, but the visual input of the sound source visual input that is biasing the sound source is the puppets mouth because the ventriloquist is moving the puppets lips differently. So, as you know we are very good at localizing sounds.

So, if we are in a show of a ventriloquist. So, let us say this is where the mouth of the ventriloquist is and this is where the mouth of the puppet is, this is separated by a few feet and if we are close enough it is easily separable in terms of visual in terms of sound sources and of course, visual sources, but the effect is that, we hear the sound as if it is coming from the visual coming from the puppets mouth as if the source of the sound is being captured by the visual input.

So, it is the; it is the integration of the two types of information two different sensory modalities information that is being combined together to produce this effect.

So, studies of the neural representation of this has shown that actually the independent estimates of the source from the two different modalities are made up to a very higher level until when a decision is made based on the two different modalities information and with some bias and almost like bias in decision making based on how much input is there from how much strong the or reliable the inputs are from one particular source the decision is made as to whether the percept is coming percept is together the source of the visual cues and the auditory cues are together.

Similarly, there is another effect which is the McGurk effect which is often studied and that is the movement of the lips and actual sound that is produced if these are in conflict there are examples where actually a third sound is perceived. So, for example, if the actual sound is ba ba and the movement of the lips is ga ga. So, if we have a video of a person who is saying ga ga and the audio is changed and the audio is made to play ba ba what we would hear is actually da da.

So, this is another classic example of how the auditory perception is influenced by other senses and. So, this is another phenomena that has to be kept in mind when we look at auditory scene analysis and so, with this we conclude our discussions on our visual object and auditory object recognition and so, we will continue later on with studies of

learning and examples of plasticity and learning in actual neural biological neural networks.

Thank you.