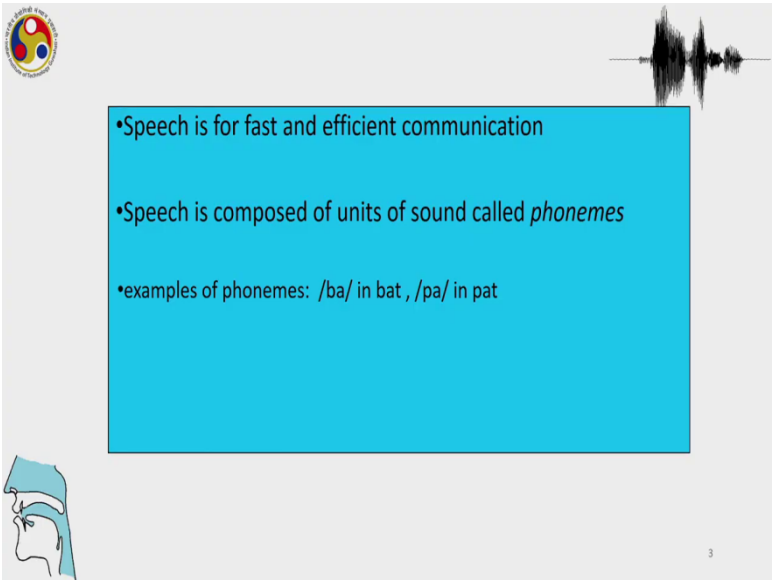
Phonetics and Phonology: A Broad Overview
Professor Shakuntala Mahanta
Department of Humanities and Social Sciences
Indian Institute of Technology, Guwathi
Lecture 12
Unit 4
Speech Perception

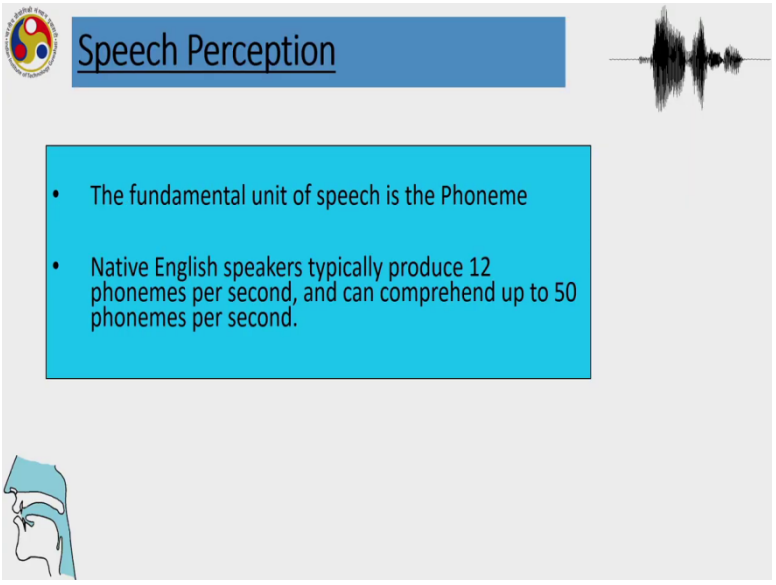
Welcome to the fourth unit of the course, Phonetics and Phonology, A Broad Overview. This unit is on speech perception. And in this unit we will cover how we hear how we make sense of the sounds that we hear.

(Refer Slide Time: 00:54)



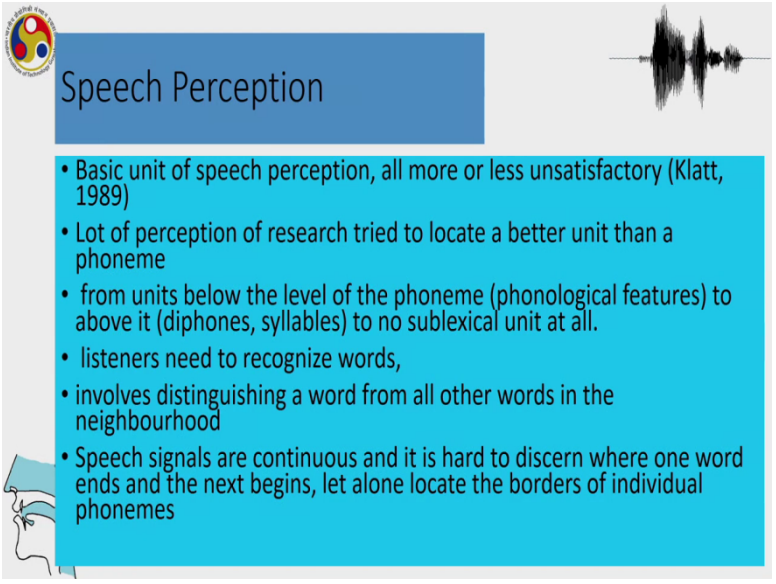
So, as we have seen in the three units before this, speech is for fast and efficient communication and speech is composed of units called phonemes. So, ba, pa are two phonemes in suppose any language or bat and pat in English.

(Refer Slide Time: 01:15)



So, the fundamental unit is generally considered to be the phoneme. Native English speakers typically produce 12 phonemes per second and can comprehend up to 50 phonemes per second. So, this is a very fast rate of recognition as well as production.

(Refer Slide Time: 01:39)

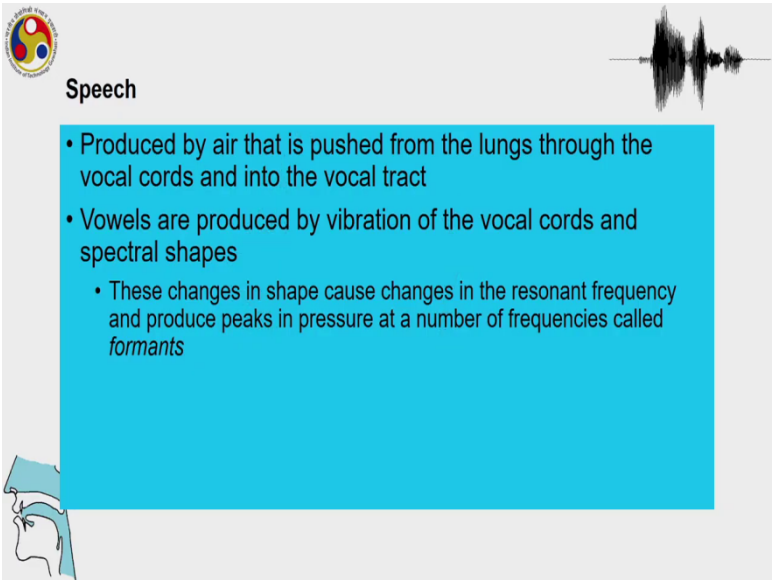


When we talk about speech perception, is it true that even in speech perception phoneme is considered the basic unit? Yes, it is. Even in speech perception phoneme is considered to be the basic unit for understanding speech. And there have been some perception research which tried to locate a better unit than a phoneme for understanding how humans perceive speech. But the basic unit of speech perception in all these studies were not satisfactory, the ones that were proposed were not satisfactory.

So, the units that you proposed are below the level of the phoneme of knowledge your features to above it to diphones syllables to know lexical unit or sub lexical units at all. All these approaches were not finally good enough to capture the processes involved in understanding speech or to perceive speech. Listeners need to recognize words units of chunks of speech.

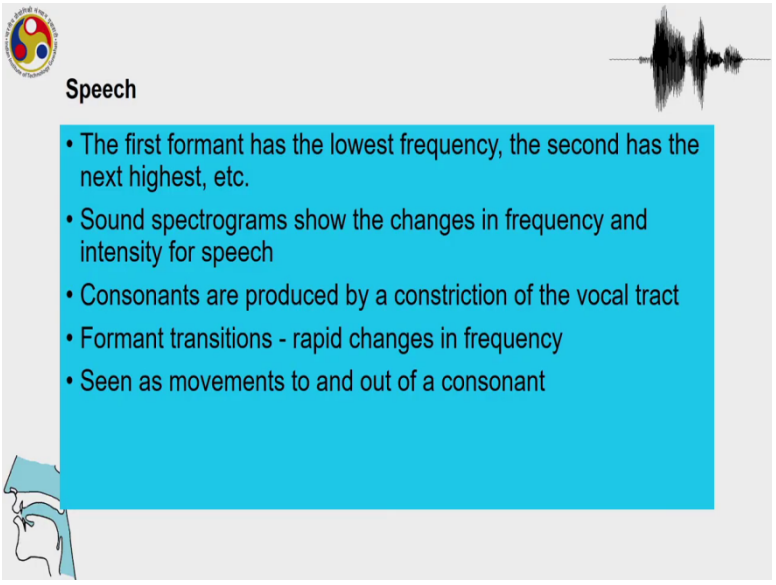
And it involves distinguishing a word, one word from all other words in a context in the neighbourhood of a sentence. So, speech signals are continuous, and it is hard to discern where one word ends and another begins, let alone locate the boundaries of individual phonemes. So, it is not only difficult to determine the units of speech, but it is also difficult to locate the boundaries of these units of speech. So, speech perception is indeed a difficult process that is performed by humans all the time, while speaking and listening and understanding speech.

(Refer Slide Time: 03:41)



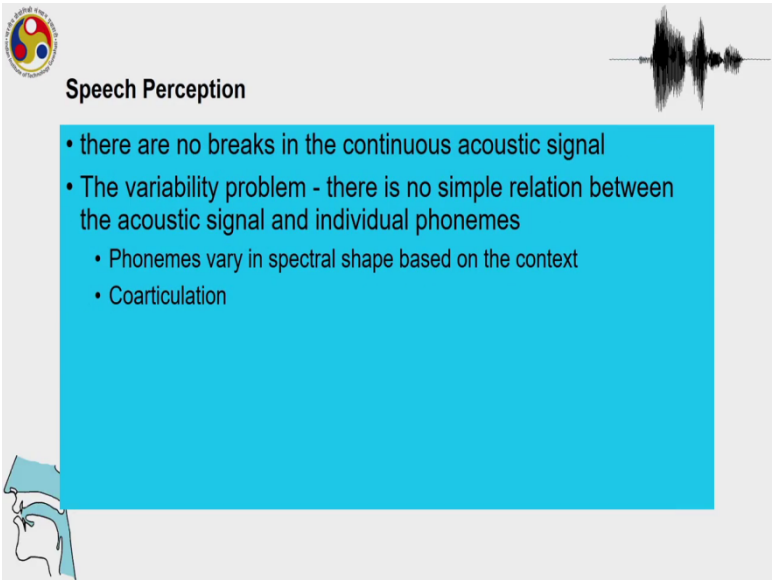
So, what happens in this process? Speech is the result of air that is pushed from the lungs, through the vocal cords and into the vocal tract. Now vowels are produced by vibration of the vocal cords and the spectral shape given to it by the vocal tract. These changes the spectral shape changes, causes changes in the resonances there resonant in the frequency and produce peaks in energy and peaks in pressure at a number of different frequencies, and these are called formants.

(Refer Slide Time: 04:24)



The first formant has the lowest frequency and then the second and so on there, there are many formants whenever there is some fundamental frequency. Sounds spectrograms show the changes in frequency and intensity for speech. And consonants are produced by construction at any part of the vocal tract and forming changes or rapid changes in frequency and seen as movements to and out of continent. So, this is the part of speech the acoustic aspects which we studied in unit two, and you have a good knowledge of the acoustic aspects of speech which we are repeating now.

(Refer Slide Time: 05:04)



When it comes to speech perception, the part which is important in the acoustic signal, the part that there are no breaks in the continuous acoustic signal. Whenever we are saying a

sentence we are not particularly paying attention to giving breaks between the words we speak everything in one go, unless there are other prosodic aspects to be taken care of, we are not really paying attention to the boundaries between the words.

This creates what is known as the variability problem, there is no simple relation between the acoustic signal and individual phonemes. So, this means that the phonemes the acoustically the phonemes are subject to a lot of changes, some of which you had seen in the lectures on acoustic phonetics. Phonemes vary in spectral shape based on the context based on the in the context in which in which they appear.

So, depending on the following vowel cart we will have the K, they will have a different spectral different form and transitions, unlike kit. So, this is what we talked about in the first class, that they are different in the way the acoustic shape of the properties will show and part of the problem there is also that of co-articulation. So, none of the sounds that we produce in the words are actually completely separate.

So, while producing a vowel, it carries information about the following consonant and so on and so forth. So, they cannot be really separated one cannot be separated from the other and as a result of co-articulation. So, as we already studied, these are gestures these are movements from one position to from one target to another, and before the completion of one target, the movement for the second target already starts and which results in issues of variability and co-articulation etc.

(Refer Slide Time: 07:35)

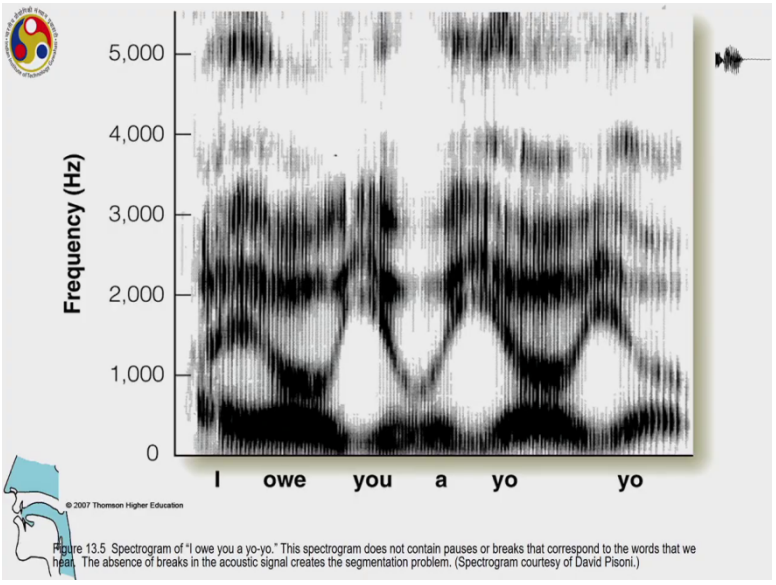
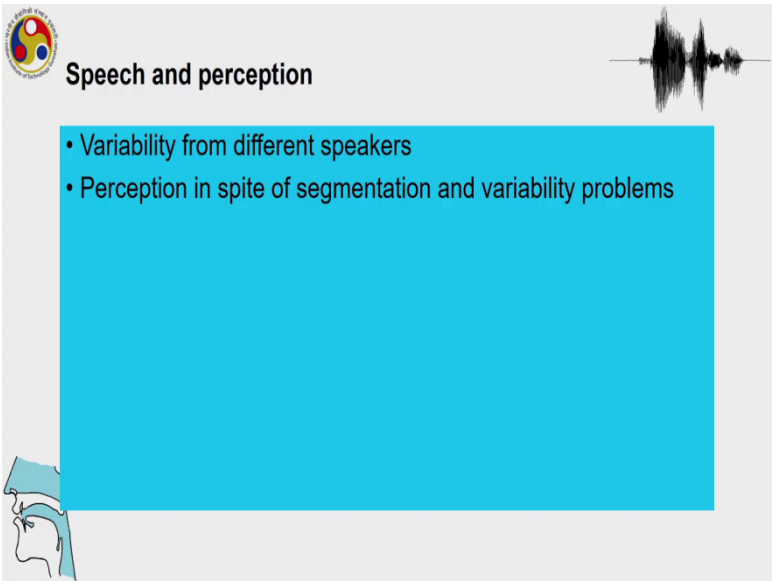


Figure 13.5 Spectrogram of "I owe you a yo-yo." This spectrogram does not contain pauses or breaks that correspond to the words that we hear. The absence of breaks in the acoustic signal creates the segmentation problem. (Spectrogram courtesy of David Pisoni.)

So, this is spectrogram showing the spectrogram of I owe you a yo-yo. So, there are no pauses or breaks between the words and the absence of breaks in the acoustic signal creates the segmentation problem.

(Refer Slide Time: 07:53)

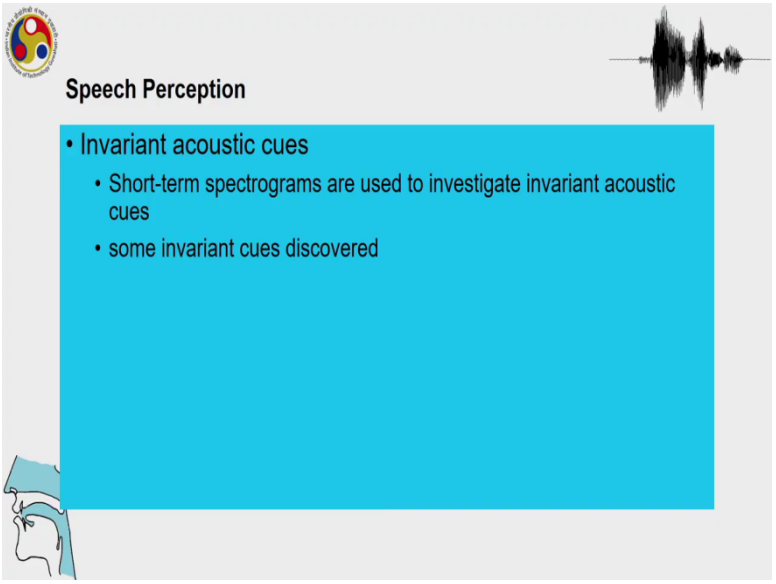


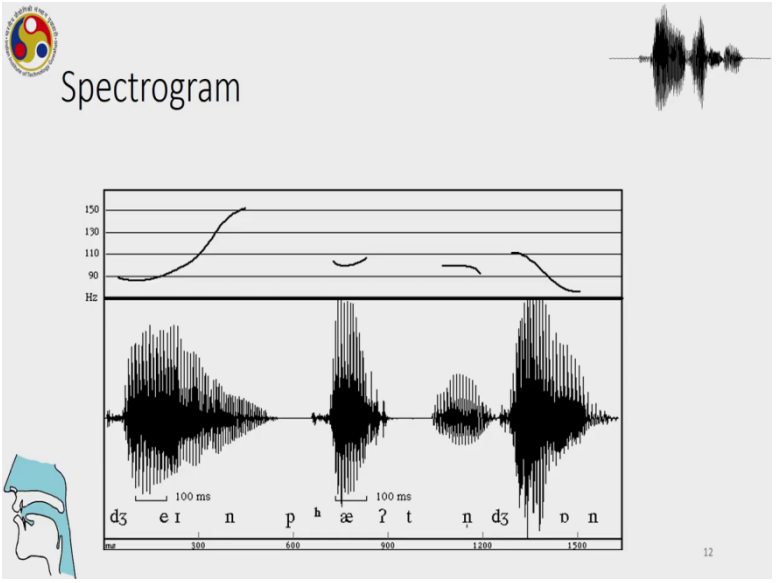
So, now, what are the other issues related to variability? Different speakers, so, speakers speak differently each speaker different speakers will have different ways of non-saying production of sounds and there could be mannerisms, etc, which would lead to differences in the way sounds are produced. Also, there are other issues which can be related to a person's geographical origins.

So, that is to say dialects. So, different regions speak different produce different dialectal variations, there could be also variations based on gender based on the people have shown that there could be different groups of society can produce can have their own unique way of producing sounds depending on various issues like social strata, etc. So, those are also factors affecting variability of the sound.

So, perception in spite of segmentation and variability problems. Now we perceive human beings perceive speech quite well despite all these variability in the signal. So, there are many many variables which occur in the speech of any individual and the listeners are able to decipher or able to understand whatever is conveyed by the speaker.

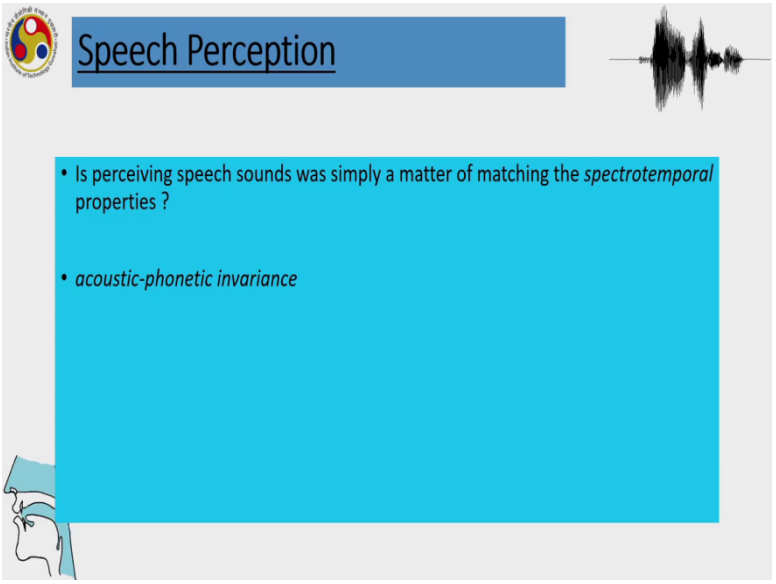
(Refer Slide Time: 09:35)

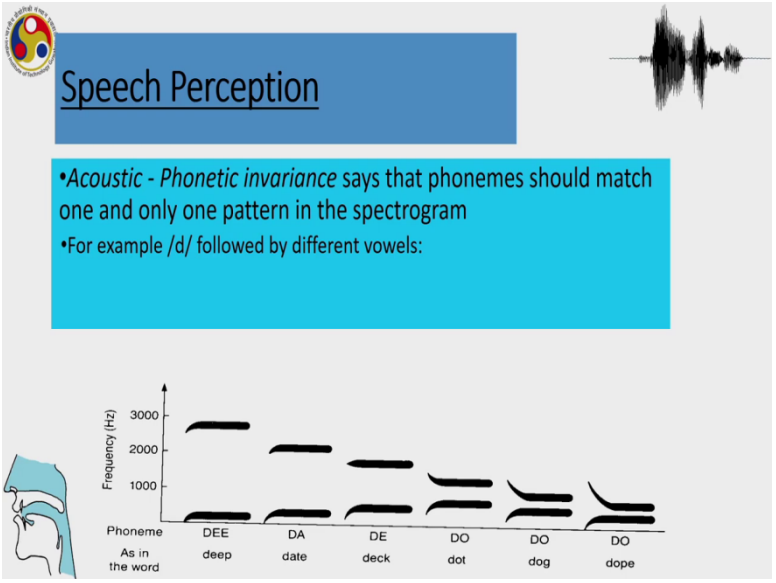




So, in acoustic research, there have been research on invariant acoustic cues. So, short terms spectrograms are used to investigate invariant acoustic cues, some invariant cues, what have been actually discovered for speech. So, this is a spectrogram and a wave form but short term spectrograms can give more information about the harmonics about the amplitude etc. And whereas, wideband spectrograms gives you more information about formants etc. So, the spectrograms are used for acoustic analysis and they have been shown to show invariant acoustic cues.

(Refer Slide Time: 10:27)





So, is perceiving speech, simply a matter of the spectral properties the timing properties and can acoustic phonetic invariants always determine perception. So, here we show a few diagrams, which show that actually the acoustic phonetic invariants cannot always tell us about perceptual cues. So, because of the variability that we already talked about. So, acoustic phonetic invariance says that phoneme should match one and only pattern in a spectrogram that is to say that, each consonant each vowel will have only one pattern.

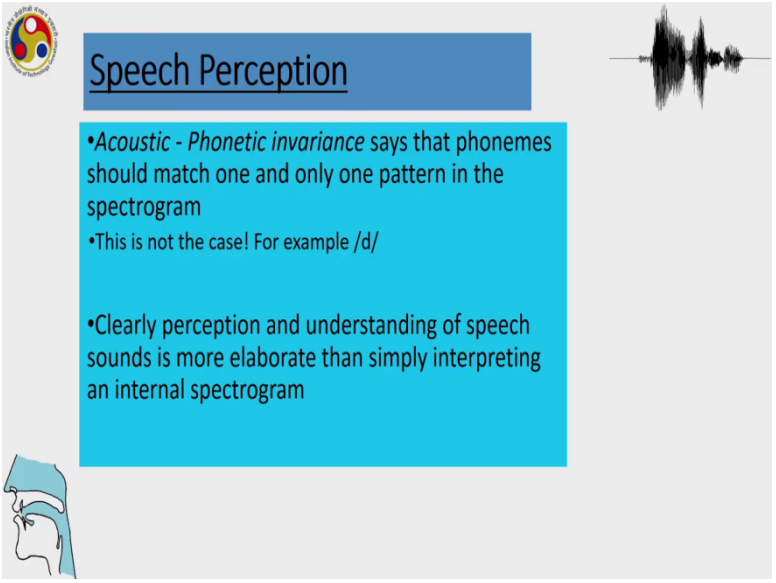
Now, this is shown in the diagram in front of you, you can see that these are formants transitions for one continent du and this is term transitions that you see, all along in this graph, it starts from D as in deep, and Day is in Date and Day as in Dec and Dot or, Doo as in

Dot, and Dog and Dope. So, in all these vowel, in the presence of these different vowels, we can see that the formants vary greatly.

So, we know the formants transitions are quite different. So, the formants themselves will be different. Of course, we know that because the vowels are determined. So, we know from our lectures on acoustic phonetics, that if it is a high vowel, then we have low f1. And the difference between f1 and f2 is quite big and that is because of deep the ease there, but what is important to note here is that formants transitions are very different are can be the movement in and out could be quite drastically different.

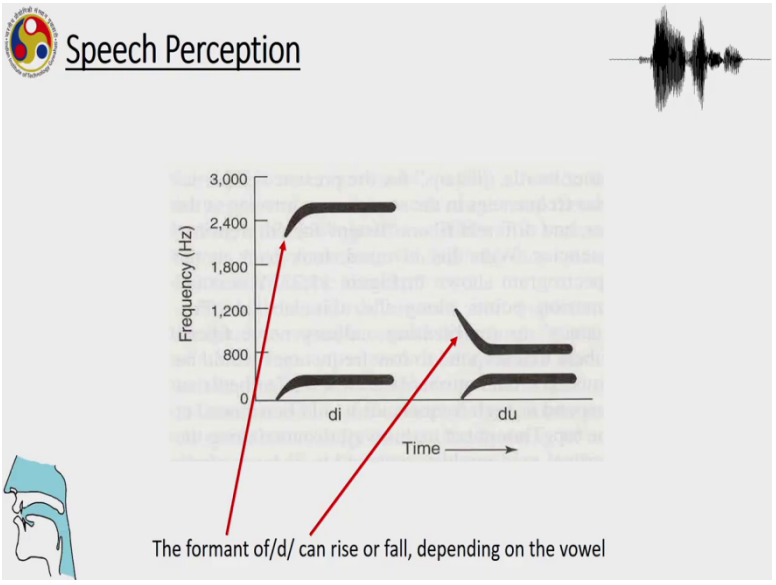
And you can see that for the front vowels in the back vowels, they are completely different while it is sort of rising. So, you can see have like a falling one second f2 for Do and for Dot and Dog and Dope quite different from D as in Deep and De as in Date and Days in Deck.

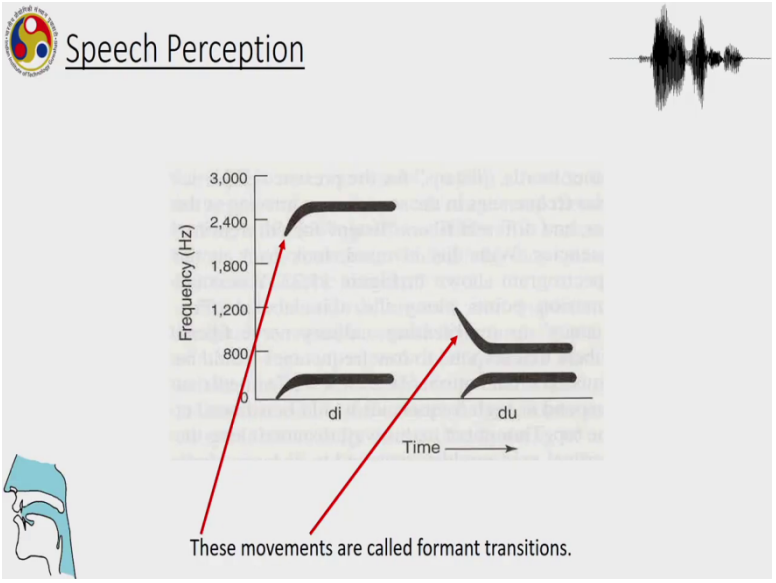
(Refer Slide Time: 13:09)

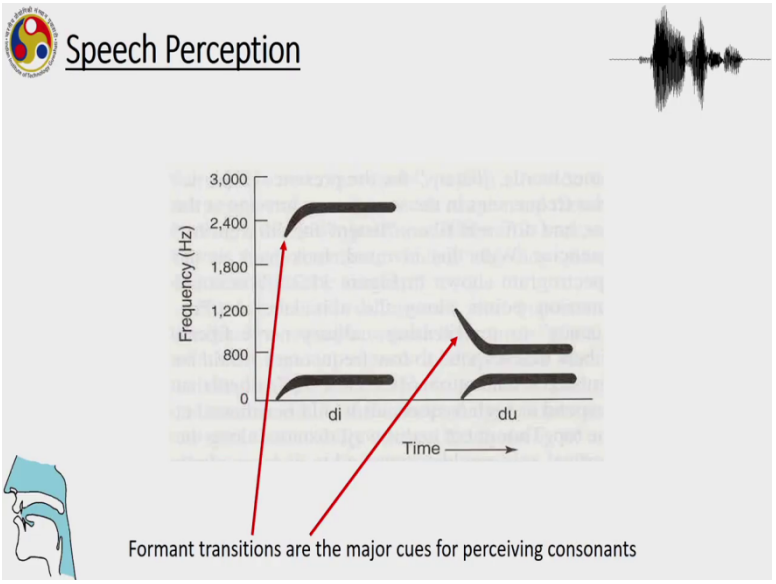


So, is it possible that phonemes will match only one and only one pattern in the spectrogram? So, we saw the example of du and that is not true.

(Refer Slide Time: 13:22)

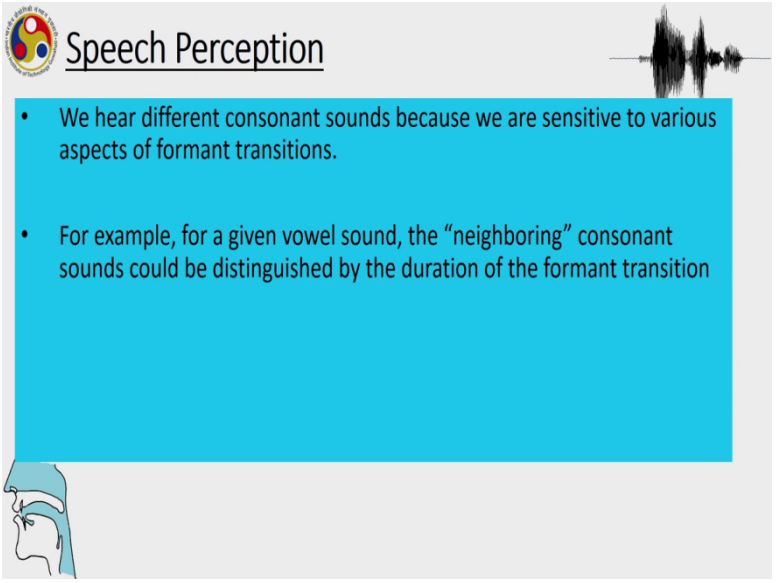






So, again here, the format of the can rise or fall depending on the vowel as you can see. And again here Du, Di, Do they are quite different for both in one these movements are called formant transitions, which you already know from our acoustic analysis class, lecture and formant transitions are the major cues for perceiving consonants and we will have a look at that again.

(Refer Slide Time: 13:57)

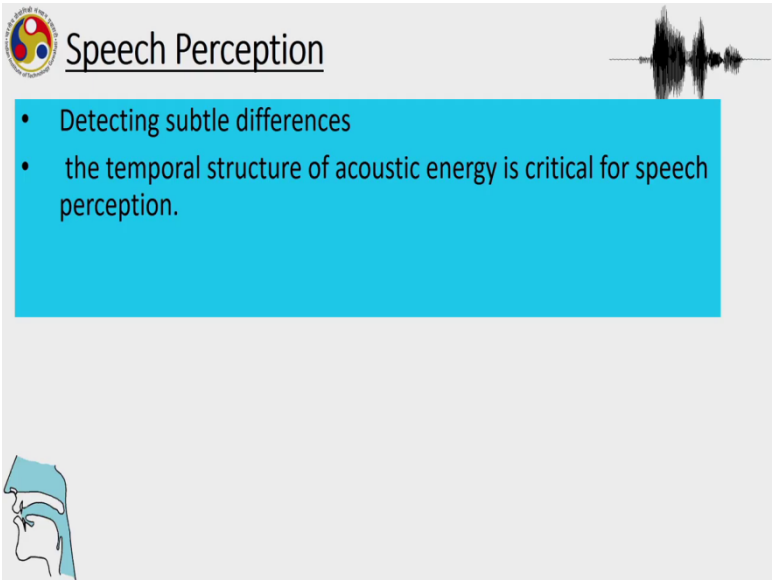


So, we hear different consonant sounds, because, we are sensitive to various aspects of formant transitions. But even for one sound, there are different formant transitions and we are sensitive to minute differences in formant transitions for the same consonant. So, for example, for a given vowel sound, a neighboring consonant sounds could be distinguished by

the duration of form and transition and apart from the movement of rising and falling, it is also duration which plays a role.

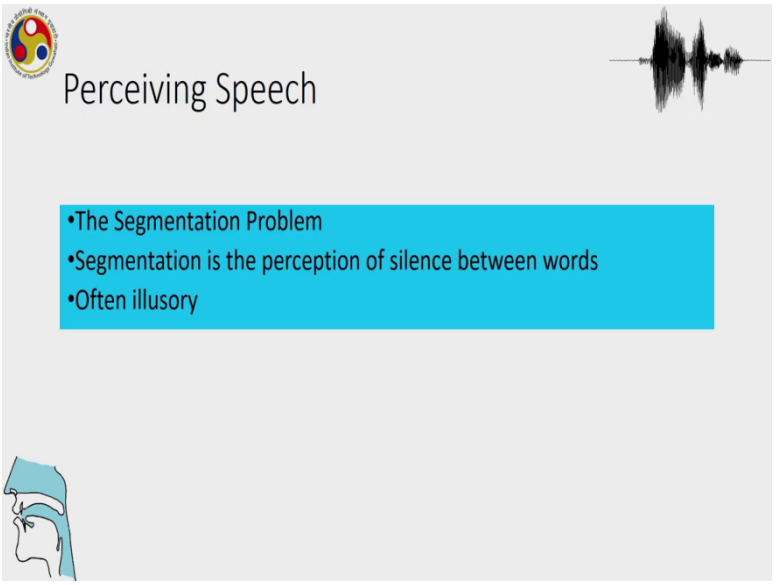
So, we saw that where the acoustic phonetic invariants issue that one phoneme will always have specific acoustic cues associated to it is cannot be true because we saw with the example of Du that, depending on the vowel, the formant transitions are very different.

(Refer Slide Time: 14:59)



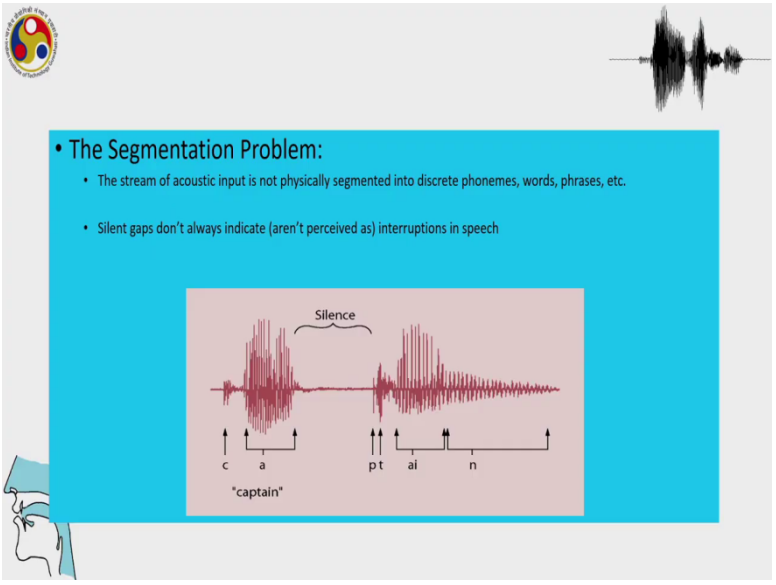
The temporal structure of acoustic energy is critical for speech perception. So, apart from the spectral shape, the temporal structure is also important for speech perception.

(Refer Slide Time: 15:13)



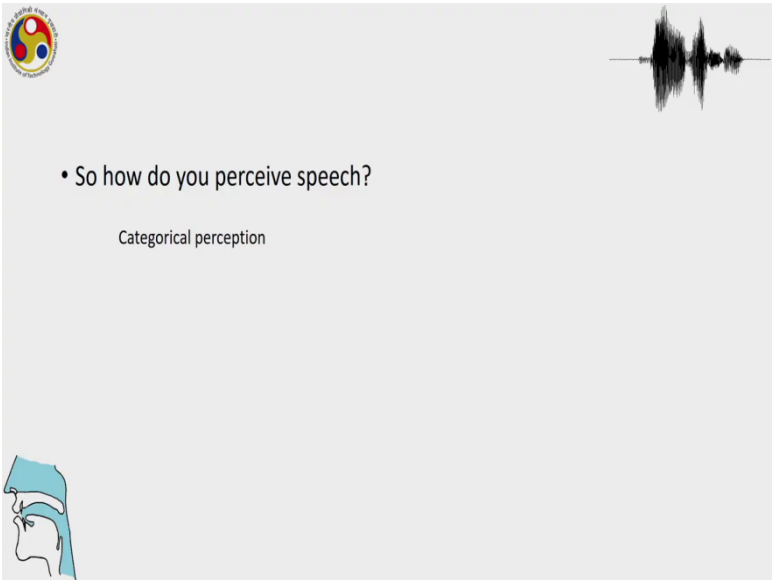
So, now, we come to a big issue in speech perception, which is called the segmentation problem. So, segmentation problem arises because speech is not always speech, a sentence or continuous speech is not always marked by periods of long silences between words.

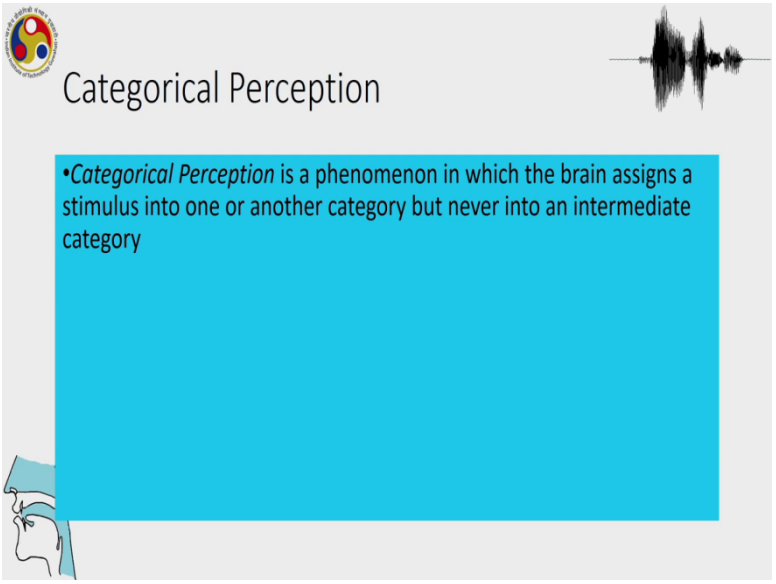
(Refer Slide Time: 15:38)

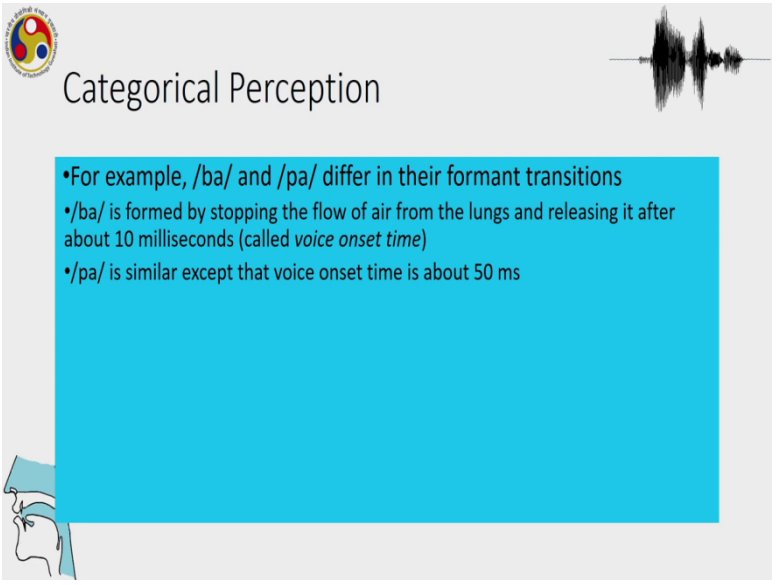


So, this is another silence that we need to talk about the silence between two stops. So, the silence between the presence of a stop. So, this one, which we know that is the period of silence because of stops, because the two stops here needs a period of closure and then release and that silence is always compensated for and here in the word captain, so, the silence that was preceded before the release of the two stops, seems easily compensated for.

(Refer Slide Time: 16:19)







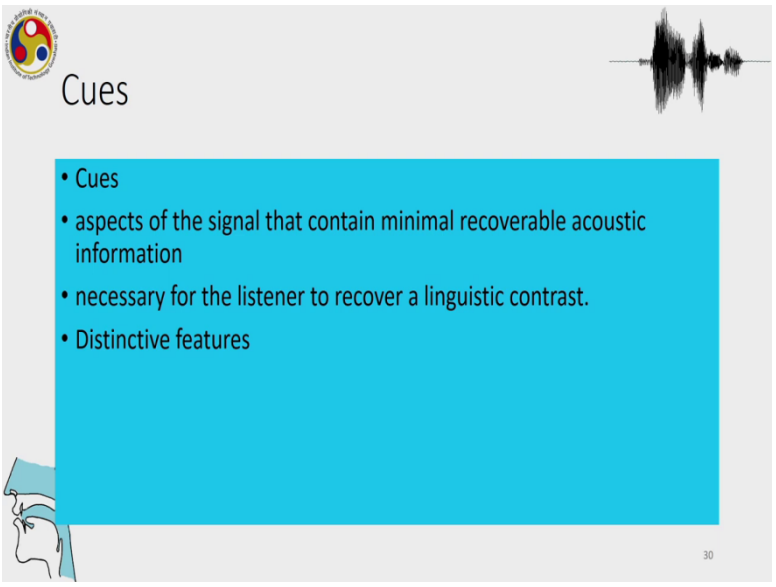
So, how do we perceive speech? One of the very important things that speech research has shown over the years is categorical perception. And what is categorical perception categorical perception is a phenomenon in which the brain assigns a stimulus into one category, but never into the intermediate category. So, basically, the brain divides sounds into categories and nothing in between.

So, there are no categories which are intermediate. So, they are either this category or that category. For example, let us take two sounds two syllables ba and pa and ba formed by stopping the flow of air from the lungs. So, the voice onset time is shorter. And unlike pa, the voice onset time for pa is longer because it is a voiceless sound.

(Refer Slide Time: 17:24)

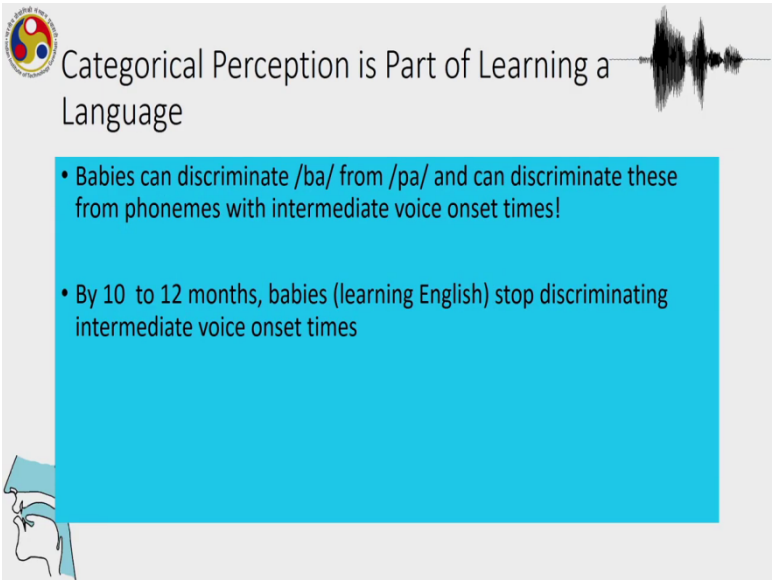
Also, again, what is categorical perception? Categorical perception is the ability to hear categories, and never anything intermediate. So, English speakers will hear ba or pa, but will not hear anything in between.

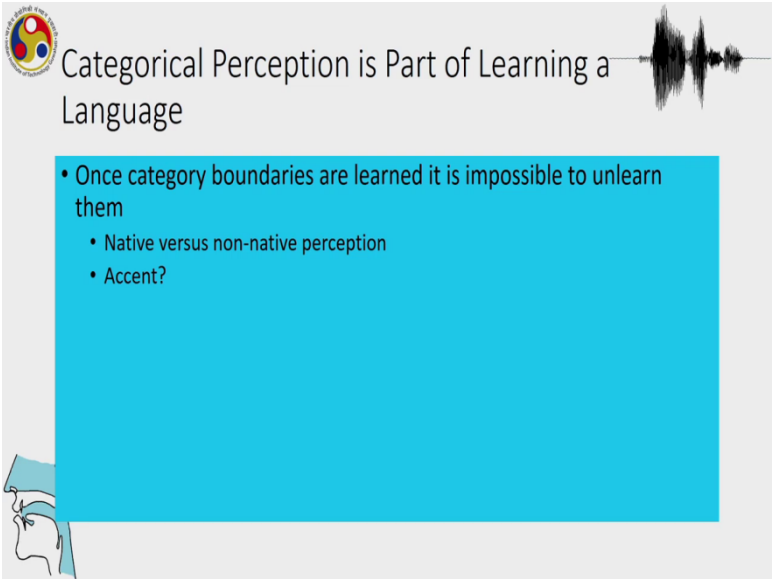
(Refer Slide Time: 17:40)



So, let us also talk about cues we have already mentioned, invariant cues. So, what are queues, queues are aspects of the signal that contains minimum recoverable acoustic information. And they are necessary for the speaker to recover a contrast, and they are related to developments in distinctive features.

(Refer Slide Time: 18:03)

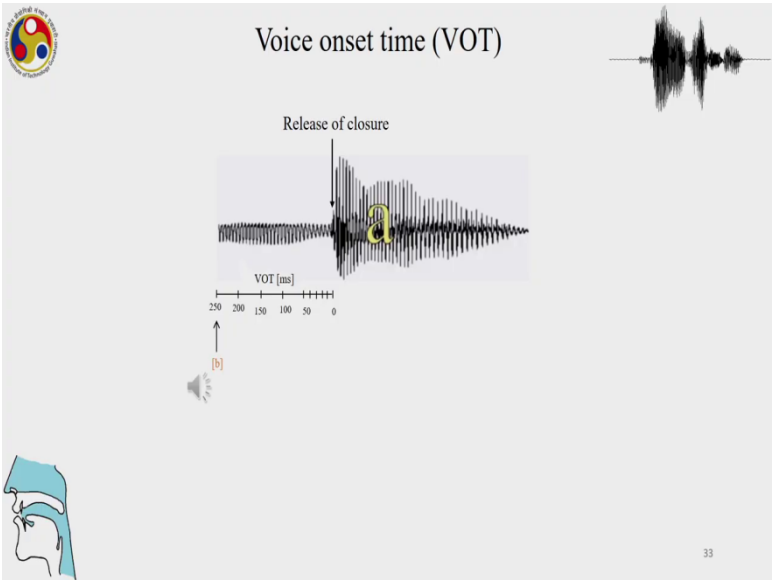


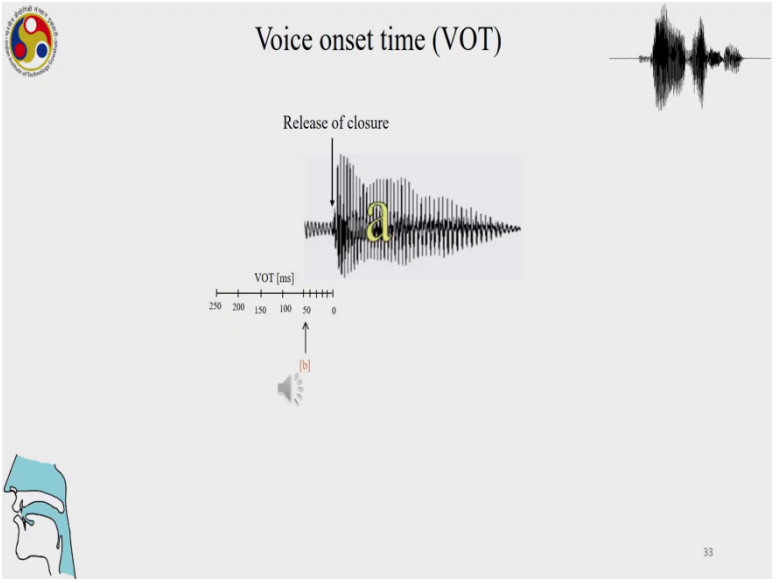


So, categorical perception is considered to be part of learning a language babies can discriminate ba from pa and can discriminate these phonemes with intermediate voice ones at times. So, by ten to twelve months baby stop discriminating intermediate voice on set times. Now, after learning the category boundaries, it is not possible for infants to or unlearn them or and as they grow, they are more set in their ways such that the boundaries are fixed.

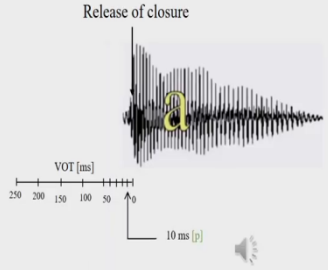
So, this results in what we know as native perception versus non-native perception. And once the category boundaries become fixed, it becomes very difficult to unlearn those boundaries and learn other sounds and also is supposed to lead to accent in speakers, because of the early perception of category boundaries.

(Refer Slide Time: 19:12)

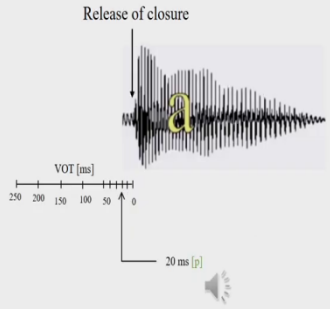




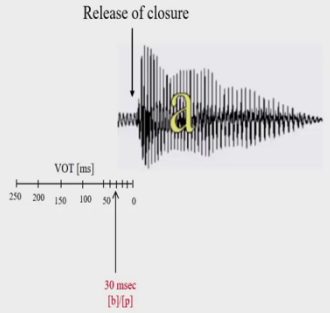
Voice onset time (VOT)



Voice onset time (VOT)



Voice onset time (VOT)



So, we just talked about voice onset time. So, this is what we want to show here is voice onset time. So, you will hear three sounds. Now in this diagram, let us explain this diagram a bit more. In this diagram, you will see that there is something noted as VOT, which you know is voice onset time, as when voicing starts. And then release of closure. So, it is a stop consonant if it is a stop consonant the release of closure.

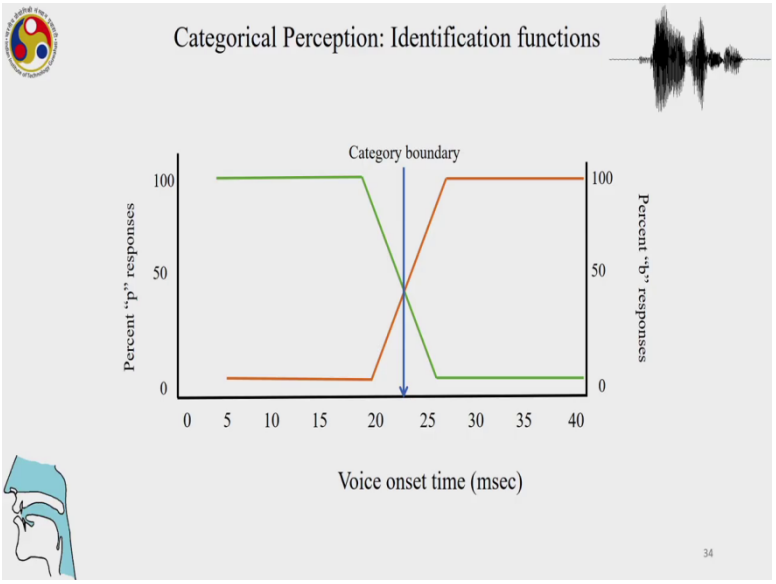
So, depending on the VOT, the category boundaries will be different. So, let us hear the three sounds here and see how the category boundaries could be different for these three sounds. So, we see that the voice onset time there is pretty, it takes a lot of time. So, 250 milliseconds. So the vowel is aa you heard one sound. So, now you hear another sound around 50 milliseconds. And now you hear a third sound around pa, pa 10 milliseconds.

So, what categorical perception with relation to VOT says is that you do not hear anything intermediate, you hear only two categories, which you must have heard, when we played the first two sounds, you probably heard the same sound, when we played this 10 millisecond pa sound, then you only heard another category. So, that is the ability of categorical perception that we have as human beings.

So, again let us play with 20 milliseconds pa. So, this is the category boundary that we are talking about around 30 milliseconds. So, at that category boundary, you can on either side of it, you can hear either on one side, you will hear ba on the other side, you will hear pa.

So, let us play all the four sounds again one by one, so that the point is clear, baa, ba, pa, paa. So, at 10 milliseconds and 20 milliseconds, what you heard must be the same sound and what you heard at 50 and 250, must be the same sound. So, that is what is called categorical perception of sounds and that is VOT voice onset time is supposed to be a very significant cue for voicing across languages.

(Refer Slide Time: 22:43)



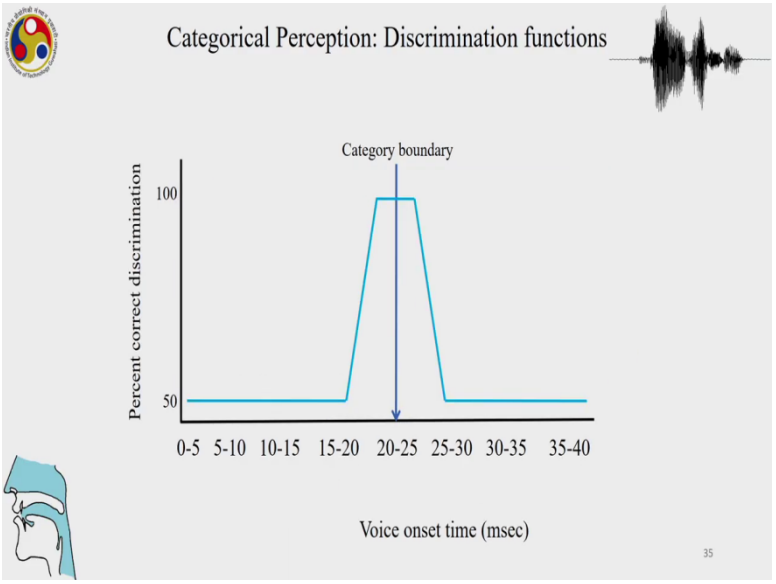
So, what you just heard, can be expressed like this in terms of results of experiments. So, these are these are called VOT experiments, which involve identification functions. And so, where participants are forced identification experiments where participants are asked whether they heard pa or paa. So, around this place between 20 to 25 milliseconds, there was a lot of change.

So, you can see that the at this category boundary percentage of power responses go down significantly and at this point, again, the perception of ba responses go up to 100 percent. So, this is the boundary that intersects between these two sounds and helps us to identify two categories and this is this is called category boundary or a phonetic boundary. And this boundary is pretty much stable.

And speakers learn such boundaries for learning languages for learning the consonants in languages and this helps you to distinguish between two categories. So, here you see that what this experiment means is that in the x-axis, you have all these tokens, which has speakers heard at 5 second, 10 second, 15, second, 20 second, 25 seconds, 30, 35 and 40 seconds, and these are the responses.

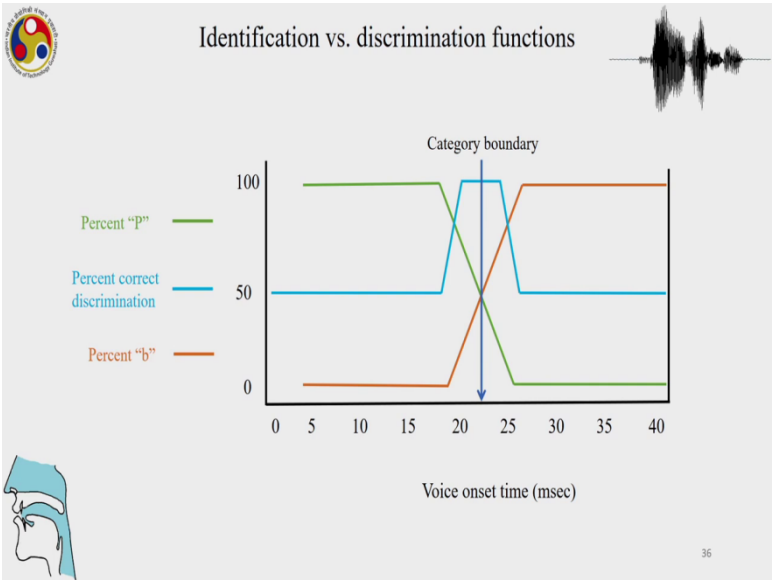
So, percentage responses, percentage responses go up to 100 percent. When voice onset is that 25 milliseconds, percentage bar responses go up to 100 percent and percentage pa go up to 100 percent when VOT is around 20 milliseconds. So, between these five seconds, we have proper category boundary which will decide whether a sound is this category or that category, but not nothing in between.

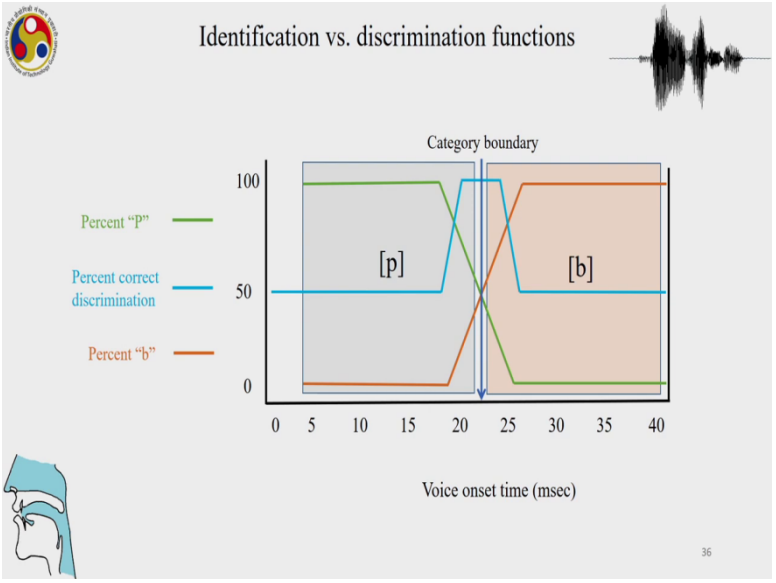
(Refer Slide Time: 25:04)



There are other types of experiments also called discrimination, where speakers are asked to discriminate whether they heard ba or pa, or they are asked to compare ba or pa with an sound and even there, where they asked to discriminate between the voicing that they have VOT that they have heard, even their the category boundary was the same between 20 to 25 milliseconds and percentage correct discrimination that is whether it is ba or pa goes up in around this region of 20 to 25 milliseconds.

(Refer Slide Time: 25:52)

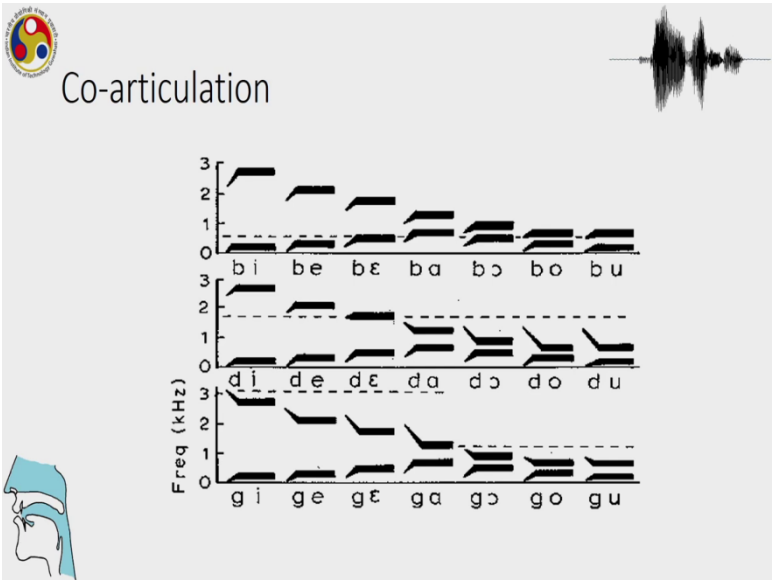




So, if we take both of these together, then we see that both identification and discrimination type of experiments, we see that between 20 to 25 milliseconds, we have our category boundary irrespective of the kind of experiment which is continuous, whether it is false identification or discrimination. That is whether the speakers are asked to discriminate, when asked is this ba or is this pa or whether they are asked to discriminate ba or pa or whether they are asked to match a sound with a ba or pa.

That is identification versus discrimination, they have similar, the experiments have produced similar results. So, this is a ba and from five to 20 milliseconds, it is always the result is always at a pa from between 20 to 25 milliseconds all the way to 40. And beyond, it is always it is always ba.

(Refer Slide Time: 27:05).

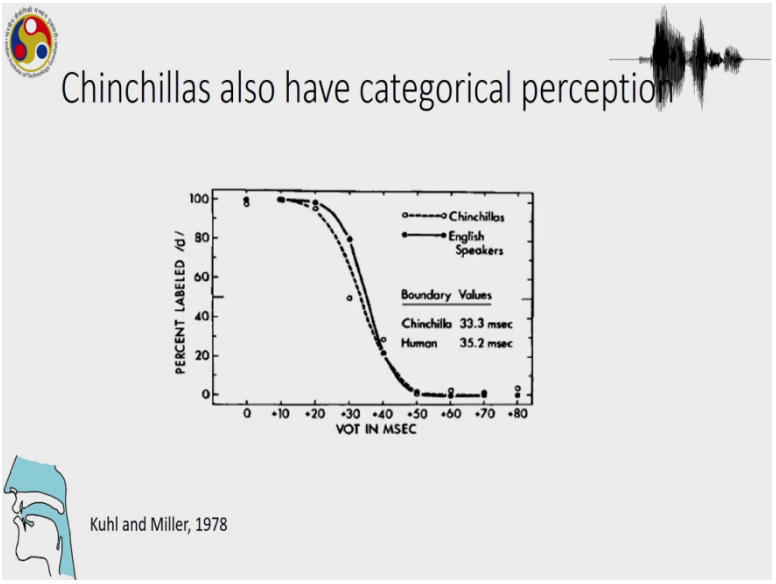


So, now what do we call co-articulation. So, the same phoneme in different contexts will be acoustically different. So, that is something we have seen in our discussions till now. So, these are schematic speech spectrograms of three voice consonants. Earlier, we had taken only a one D. Now you can see that this is for ba, da and ga. And then there are seven vowels across seven vowels we see the, the formant transitions for three different continents across seven vowels.

The formant transitions that designate the consonant where the form and changes over the consonant to the vowel can be also longer or shorter, or move in different directions. Depending on what the following vowel is. So, not just a direction, which is different for D, as we can see, it is also length. So, sometimes it is shorter, sometimes it is longer. So, all these very minute changes help us to distinguish these continents, but they are not invariant, unlike what we saw for voicing which is constant for ba and pa.

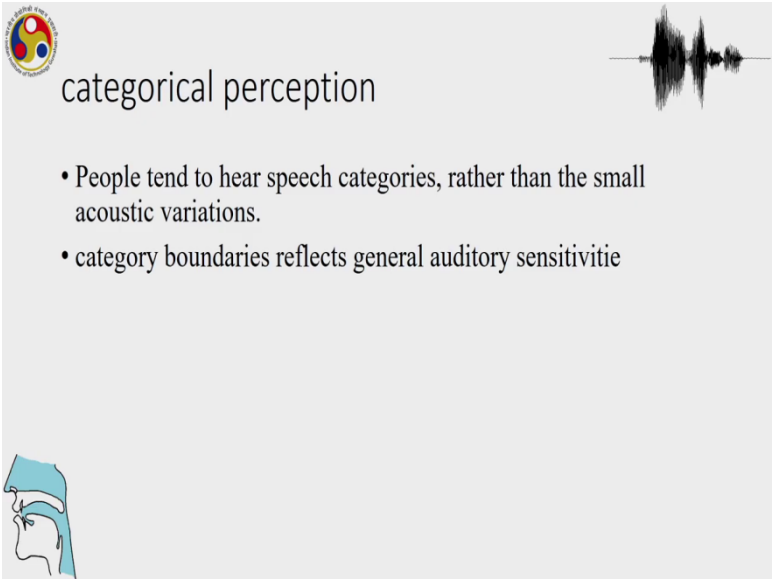
It is not dependent on a following vowel. Here depending on the vowel we have a whole varied range of formant transitions and length to contend with. So, just to compare the difference between the categorical perception and co-articulation which we have been talking that it may not be always invariant. So, we saw an example of first speech perception can have invariant acoustic cue, and we can have we can have variant ones like this one.

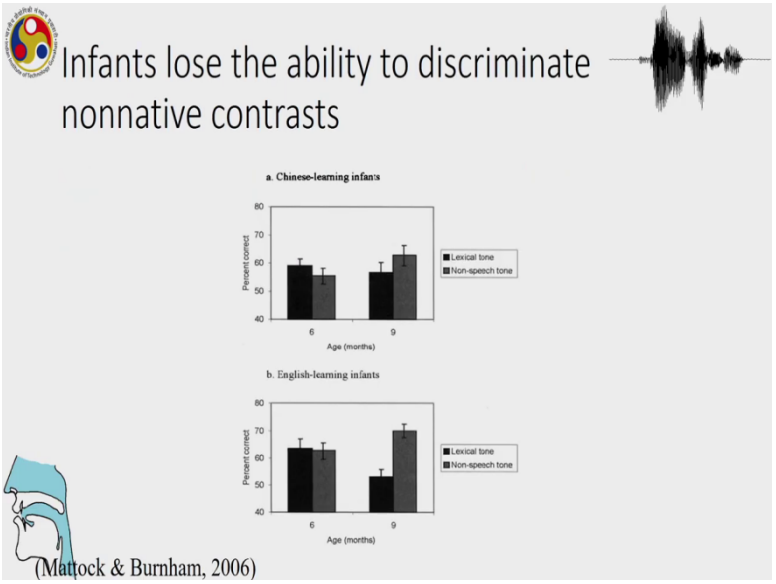
(Refer Slide Time: 29:17)



Now, going back to categorical perception. So, this is an experiment done by Kuhl and Miller 1978, they shown that chinchillas who learn to respond when they heard duh, but not when they heard guh. So, they produced an identification function that looks just like that of humans. So, it is hard to believe chinchillas which are, which are rodents how this has a special speech mechanism. This suggests that our ability to categories consonants can be shared with other species also.

(Refer Slide Time: 30:03)

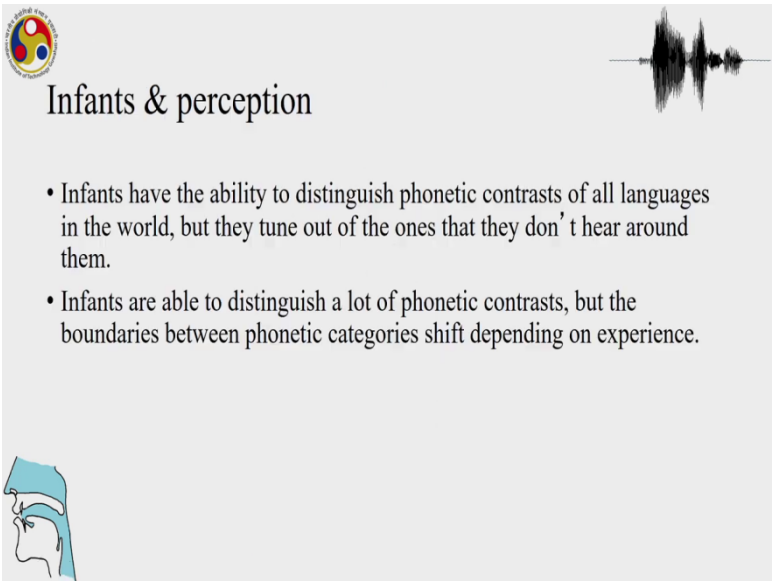




So, let us also look at infants. And these are their similar results of experiments looking at the perception of lexical tone by infants who are learning Chinese, and infants who are learning English. The infants were tested on lexical tone, and the un-discrimination of those tones that had the same difference. That is, they were either rising or falling, etc. So, both the groups of children whether they were Chinese or English speaking, learning infants growing in Chinese or English learning environments, could discriminate lexical tones, as well as the discriminated non speech tones.

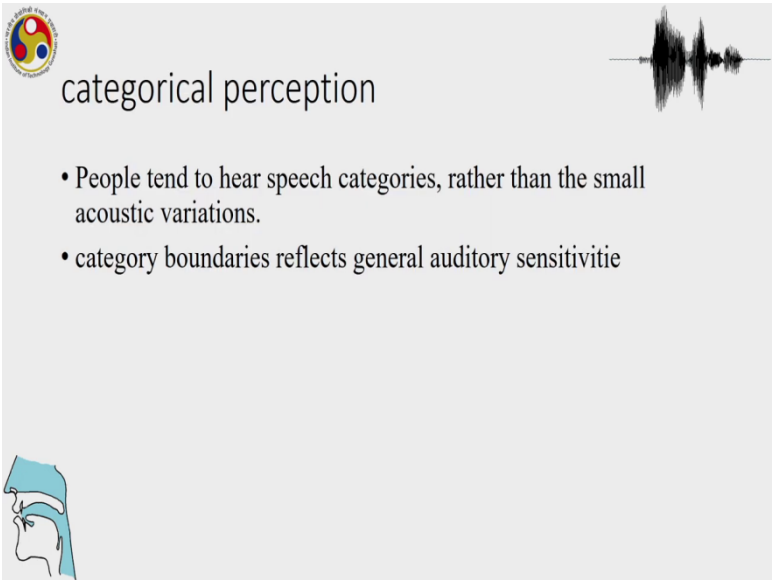
So, but they seem to have lost that ability and about nine months. And so while the Chinese learning infants maintain the ability to discriminate lexical tones, the English children seem to have lost it at nine months.

(Refer Slide Time: 31:19)



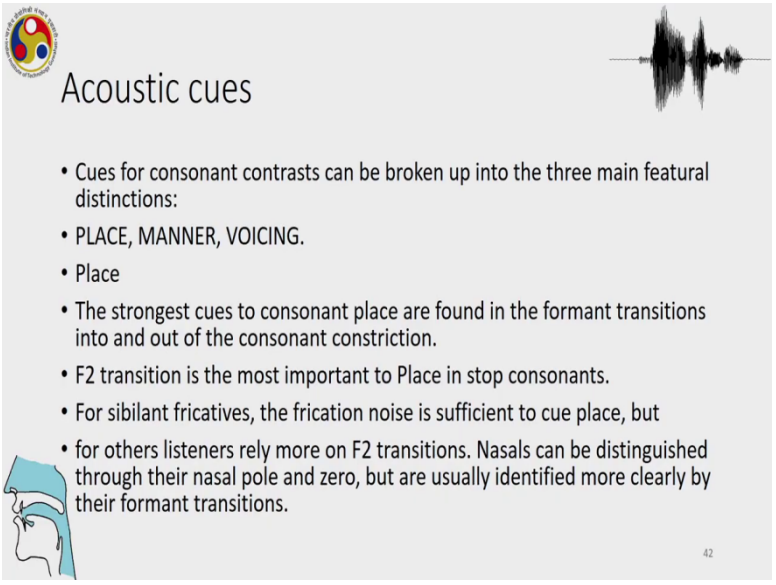
So, regarding perception, again, there are various ways in which this is seen theories, which theorize on perceptual abilities of children, it is said that infants have the ability to distinguish phonetic contrasts of all languages in the world, but the tune out of the ones that do not hear around them, and this is one approach and the other approach also says infants are able to distinguish a lot of phonetics contrast, but the boundaries shift depending on experience.

(Refer Slide Time: 32:00)



So, finally, we want to conclude this discussion on categorical perception by saying that people tend to hear speech categories rather than the small acoustic variations. So, that ability human seems to have so there are some abilities like speech categorical speech perception, which helps us to perceive speech and also that there are a certain auditory sensitivities that we have, and the category brown boundaries reflect our auditory abilities also.

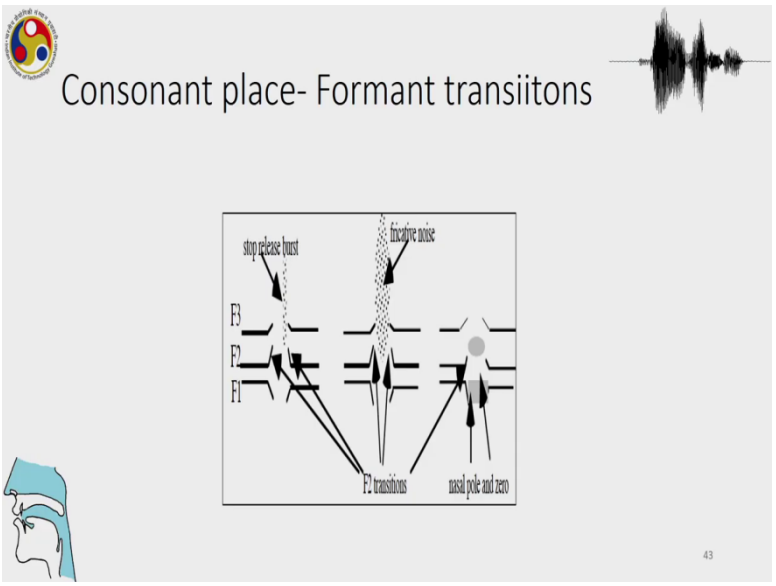
(Refer Slide Time: 32:52)

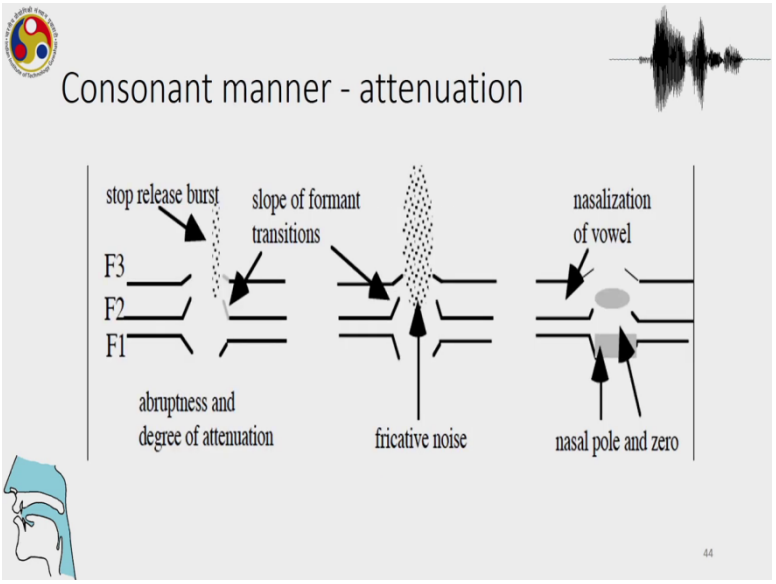


So, we had earlier talked about you know, acoustic cues. So, what are cues, if you recall, then, cues are the information which helps us to retrieve the contrast of information in the speech acoustic signal. So, cues for consonant contrast can be broken up into three main feature illustrations state that are called face, manner and voicing.

And the strongest cues to consonant plays are found in formant transitions into and out of the consonant constriction and F2 transition is the most important for place for sibling fricatives the friction noise is sufficient to cue place, but father listeners they rely more on F2 transitions. Nasals can be distinguished through their nasal pole and zero, but are usually identified more clearly by the formant transitions.

(Refer Slide Time: 33:53)



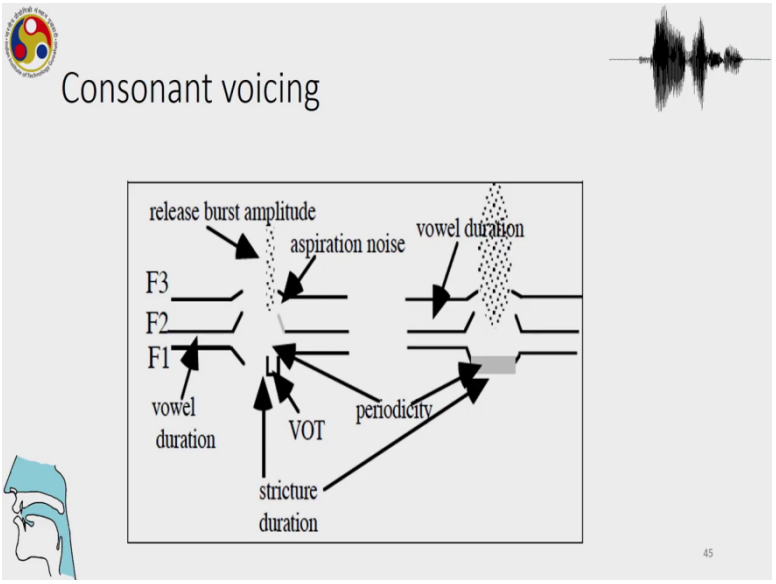


So, we had actually discussed these things when we discuss acoustic phonetics. So, so here are the F2 transitions, F2 transitions for stops, which and also for fricatives and for nasals. So, these three are important. However, for stop release burst is it is also a cue for fricatives the fricative noise is also a cue. And what you had seen in our classes on acoustic phonetics, nasal pole and zero the attenuated formants around F2 region, so, which we which we call the zero and the nasal pole, very low formant around F1.

So, and also attenuated low energy zones. So, these are very significant cues for nasals. And so, this consonant place, whether it is fricative or stop or nasal, the f2 is gives a lot of information, but from that when it comes to manner, the attenuation is very important. So, for stops, you would hear a stop release burst and the abruptness and degree of attenuation for fricative noise, it is rather slow the degree of attenuation is also a bit slow.

So, you can hear a lot of noise for fricatives. So, that is another big cue for fricatives and for nasals nasalization of the vowels and again the nasal pole, it is it is a significant cue and the other important cues are the slope of the formant transitions again for also manner, manner cues could also be expressed with the slope of formants transitions.

(Refer Slide Time: 35:59)



So, we have already seen how voicing is effectively cued by VOT the voice onset time that is important. But apart from that there are also other cues, not just the sole cue for voicing, we have a vowel duration which can be longer for voice consonants. Then for voiceless consonants, the stricture duration could be longer for voice then voiceless consonants. There could be aspiration noise for voiceless consonants, then voice consonants the release burst amplitude could be again, different for voice and voiceless.

And periodicity will be present in the consonant to cue, the voicing. So, these are the many aspects of perception with regard to consonants. And depending on whether it is a stop or a fricative, or nasals, the there will be cues related to their place of articulation, manner of articulation, and voicing. And these are not just one cue, as we saw that in the literature they are invariant acoustic cue has been is there is not a whole much water exactly because of the range of acoustic cues which are available for place, manner and voicing, and also for other aspects of consonants and vowels.

So, we come to the end of today's lecture on perception, and it shows that speech perception is related to the acoustic the acoustic phonetic component, the acoustic signal, but there are a variability of cues and not just one cue, which is associated with one sound and whether it is a vowel or a consonant, and there are very many variabilities associated with speech.

However, we seem to be able to compensate for those variabilities and have efficient communication, regardless of the issues of variability, co-articulation, regardless of absence

of invariance. In the cues, perception always seems to efficiently most of the time, present in speech communication.

And the search for cues will be an ongoing process in the literature in in speech, perception, research, and understanding how speech is perceived by humans and how to understand our perceptual abilities, our auditory sensitivities, which helped that perceptual ability, all the more and more knowledge in this area is going to be beneficial to understand speech in greater detail. Thank you for listening. And this is the first lecture on speech perception.