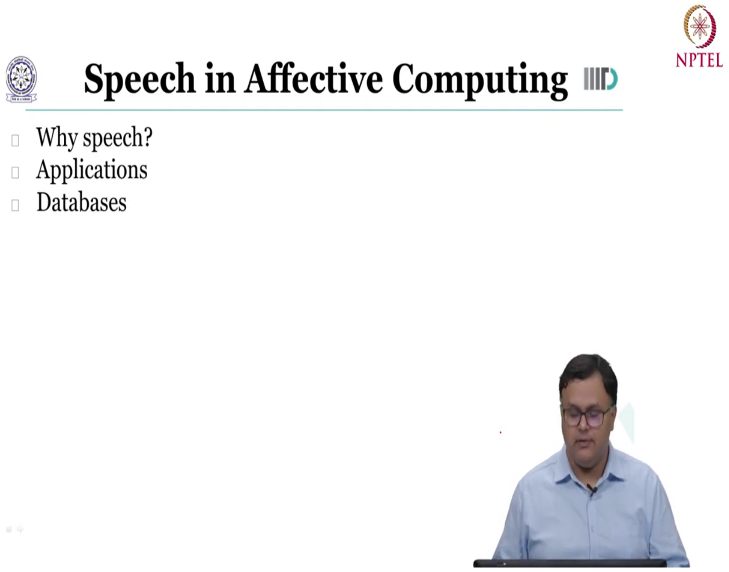
Affective Computing
Dr. Abhinav Dhall
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Week - 05
Lecture - 01
Speech Based Emotion Recognition

Hello and welcome. I am Abhinav Dhall from the Indian Institute of Technology, Ropar. Friends, this is the lecture in the series of Affective Computing. Today, we will be discussing about how we can recognize emotions of a user by analyzing the voice. So, we will fixating on the voice modality. So, the content which we will be covering in this lecture is as follows.

(Refer Slide Time: 00:53)



Speech in Affective Computing

- Why speech?
- Applications
- Databases

First, I will introduce to you, give you some examples about why speech is an extremely important cue for us to understand the emotion of a user. Then we will discuss several applications where speech and voice based affective computing are already being used. And then we will switch gears and we will talk about the first component which is required to create a voice based affective computing system, which is labeled datasets.

And in this pursue, we will also discuss the different attributes of the data and the conditions in which it has been recorded. Now, if you look at me, let us say I have to say a statement about how I am feeling today and I say, well today is a nice day and I am feeling contented.

Now, let me look down a bit and I say, today is a wonderful day and I am feeling great. Now, in the first case, you could hear me and you could see my face very clearly. And in the second case, my face was partially visible, but you could hear me clearly. And, I am sure you can make out that in the first case, I was showing neutral expression. And in the second case, even though my face was facing downwards, not directly looking into the camera, I was sounding to be more positive right.

So, there was a more happy emotion which could be heard from my speech. So, this is one of the reasons why we are using voice as one of the primary modalities in effective computing. You talk to a friend, you understand their facial expressions, you look at the person's face, but in the parallel, you are also listening to what that person is speaking. So, you can actually tell how that person is feeling from their speech. And that is why we would be looking at different aspects of how voice can be used in this environment.

(Refer Slide Time: 03:36)

Speech in Affective Computing

Please listen again carefully

Source: VGAF dataset

So, here is an example. So, this is a video which I am going to play from a audio visual group effect dataset. So, let me play the video.

Protects.

Protects. (Refer Time: 03:47) So, you protect yourself. No.

The other delivers protects.

(Refer Time: 03:52).

You protect yourself no.

(Refer Time: 03:55).

The other delivers.

So, if you notice in this case, the body language of the subjects here, that is trying to be a bit aggressive. So, this looks like a training video. But if you hear the voice over, the explanation voice in this video, you can tell that there is no fight going on, there is no aggressive behavior, it is simply a training going on. And how are we able to find that? By simply looking at the tonality of the voice.

If it was, let us say, actually a fight or some aggressive behavior shown by the subjects in the video and the voice was also from one of the subjects, we would also hear a similar pattern which would tell us that let us say the subjects could be angry. But in this case, even though the body language facial expression says that they are in an aggressive pose, but from the voice, we can tell that this is actually a training video. So, it is the environment is actually neutral. Now, let us look at and hear another video.

(Refer Time: 05:10) [FL].

Now, in this case, the video has been blacked out. You can hear the audio and you can tell that there are several subjects in the audio video sample and the subjects are happy right. How are we able to tell that? We can hear the laughter.

(Refer Slide Time: 05:34)

Speech in Affective Computing

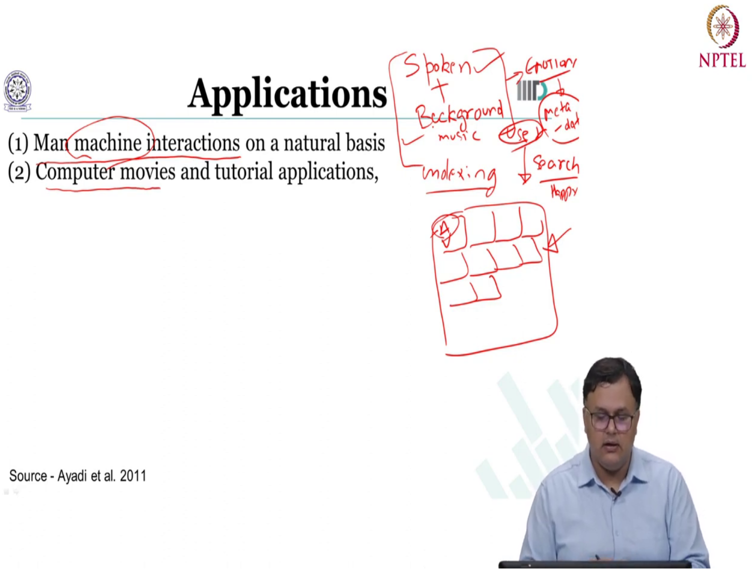
Source: VGAF dataset

Now, if I was to play the video, now this is the video which we had earlier blacked out. So, you can look at of course, the facial expressions, but even without looking at the facial expressions and the gestures just by hearing you can tell that the subject are happy. So, this gives us enough motivation to actually pursue voice as a primary modality in affective computing.

Now, as I mentioned earlier, there are a large number of applications where speech of the user that is being analyzed to understand the effect. And friends this is similar to how when we discussed in the last lecture about facial expression analysis, the several applications were there in health and in education.

We find the similar use cases for voice based effect, but which are applicable in different circumstances in circumstances, in scenarios where it could be non-trivial to have a camera look at the user. Of course, there is a privacy concern which comes with the camera as well. So, instead we can use microphones and we can analyze the spoken speech and the information which is there in the background.

(Refer Slide Time: 07:03)



Applications

(1) Man machine interactions on a natural basis

(2) Computer movies and tutorial applications,

Source - Ayadi et al. 2011

Now, the first and quite obvious application of voice in affective computing is understanding the man machine interaction on a natural basis. What does that mean? Let us say there is a social robo. Now, the robo is greeting the user and the user has just entered into the room. The robo greets the user and the user replies back.

Now, based on the voice of the user and the expression which is being conveyed by the user, the machine which is the robo in this case is able to understand the emotion of the user and then the feedback to the user can be based on the emotional state. Let us say if the user is not so cheerful.

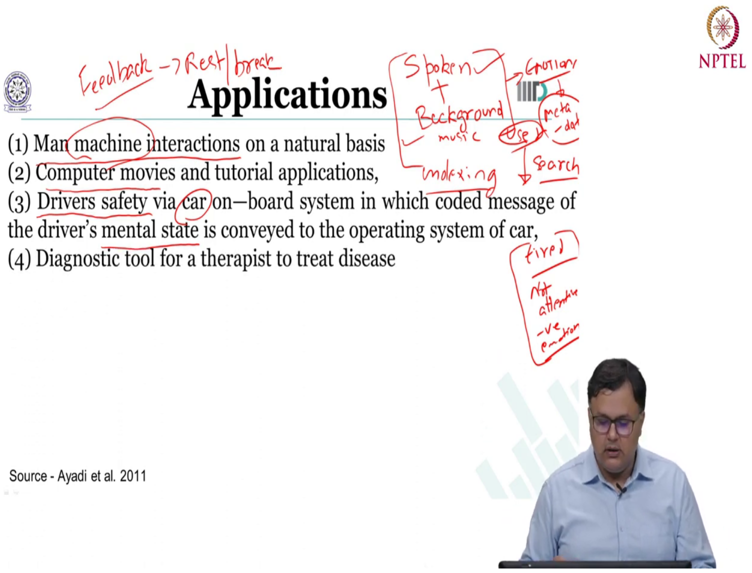
So, the robo reacts accordingly and then tries to understand with a question which could better make the either the user more comfortable, relaxed or the robo tries to investigate a bit. So, that it can have a conversation which is appropriate with respect to the emotion of the user.

The 2nd we see is in entertainment particularly looking at computer movies. So, friends in this case we are talking about the aspect of indexing. Let us say you have a large repository of movies ok. So, these are let us say a large repository. Now, the user wants to search, let us say the user wants to search all those videos which are belonging to a happy event ok.

You can think of it as a set of videos in your phone's gallery and you want to fetch those videos which let us say are from events such as birthdays which are generally cheerful right. So, from this audio visual sample we can analyze the audio which would mean the spoken content by the subject and the background voices could be music. So, we can analyze and get the emotion.

Now, this emotion information it can be stored as a metadata into this gallery. So, let us say the user searches for all the happy videos. We look through the metadata which tells us that when we analyze the audio these are the particular audio video samples which based on their spoken content and the background voice or music sound cheerful. So, the same is then shown to the user.

(Refer Slide Time: 10:42)



Feedback -> Rest | break

Applications

(1) Man machine interactions on a natural basis
(2) Computer movies and tutorial applications,
(3) Drivers safety via car on—board system in which coded message of the driver's mental state is conveyed to the operating system of car,
(4) Diagnostic tool for a therapist to treat disease

Source - Ayadi et al. 2011

Now, moving on to another very important application here let me first clear the screen a bit. So, looking at the aspects of operator safety let us say there is a driver and the driver is operating a complex heavy machinery. You can think of an environment for example, in mining where a driver is handling a big machinery which has several controls. What does that apply? Well, harsh working environment, a large number of controls of the machine and the high cost of error.

So, the driver would be required to be attentive right. Now, you can clearly understand the state of the driver by listening to what they are speaking and how they are speaking. From the voice pattern one could easily figure out things such as if the person is sounding tired, is not attentive, has negative emotion. So, if these attributes can be figured out the machine let us say the car or the mining machine it can give a feedback to the user.

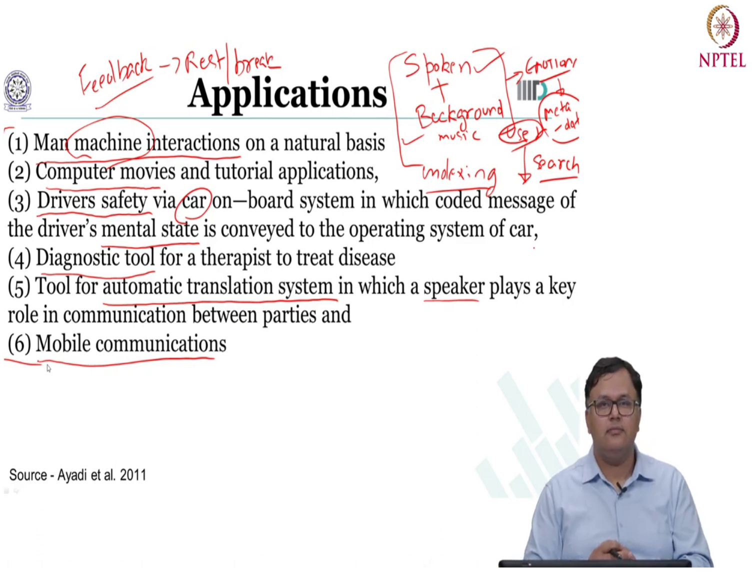
An example feedback can be please take a break right before any accident happens please take a break. Because when I analyzed your voice I could figure out that you sounded tired, distracted or an indication of some negative emotion, which can hamper the productivity and affect the safety of the user and of the people who are in this environment around the user.

Now, friends the other extremely important aspect of where we use this voice based affect is in the case of health and well being. Now, an example of that which is right now being experimented in a large number of academic and industrial labs is looking at the mental health through the voice patterns. So, an example of that is let us say we want to analyze data of patients and healthy controls who are in a study where the patients have been clinically diagnosed with unipolar depression.

So, when we would observe the psychomotor retardation which I briefly mentioned in the facial expression recognition based lecture as well. The changes in the speech in terms of that is the frequency of words which are spoken, the intensity, the pitch you could learn a machine learning model which can predict the intensity of depression. Similarly, from the same perspective of objective diagnostic tools, which can assist clinicians let us say there is a patient with ADHD.

So, when a clinician or an expert is interacting with the patient we can record the speech of the interaction, we can record the voices and then we can analyze how the patient was responding to the expert to the clinician and what was the emotion which was elicited when a particular question was asked. That can give very vital useful information to the clinician.

(Refer Slide Time: 15:36)



Applications

(1) Man machine interactions on a natural basis
(2) Computer movies and tutorial applications,
(3) Drivers safety via car on—board system in which coded message of the driver's mental state is conveyed to the operating system of car,
(4) Diagnostic tool for a therapist to treat disease
(5) Tool for automatic translation system in which a speaker plays a key role in communication between parties and
(6) Mobile communications

Source - Ayadi et al. 2011

Now, another aspect where voice based effective computing is being used is for automatic translation systems. Now, in this case a speaker would be let us say communicating between a party or translating right. So, let me give you an example to understand this. Let us say we have speakers of language one you know group of people who are in a negotiation deal trying to negotiate a deal with group of people who speak language too and both parties do not really understand each other's language.

Now, here comes a translator could be a machine, could be a real person who is listening to group 1 translating to group 2 and vice versa. Now, along with the task of translation from language 1 to language 2 and vice versa there is a very subtle yet extremely important information which the translator needs to convey.

Since the scenario is about negotiation, let us say a deal is being cracked. The emotional aspect of what is the emotion which is conveyed when the speakers of language 1 are trying to make a point to the other team that also needs to be conveyed. And based on this simply by understanding the emotional the and the behavioral part 1 could indicate one could understand if let us say the communication is clear.

And if the two parties are going in the direction as intended you can think of it as an interrogation scenario as well. Let us say interrogator speaks another language and the person who is being interrogated speaks another language right. So, how do we understand that in what is the direction of communication are they actually able to understand each other and when the context of the communication has changed all of a sudden a person let us say who was cooperating is not cooperating, but speaks another language.

So, that is where we analyze this voice when you analyze the voice you can understand the emotion and that is a very extremely useful cue in this kind of a diagnostic conversation or multiparty interaction. And of course, in this case the same is applicable to the human machine interaction as well across different languages. Friends, also another use case is mobile communication. So, let us say you are talking over a device could be on a using a mobile phone.

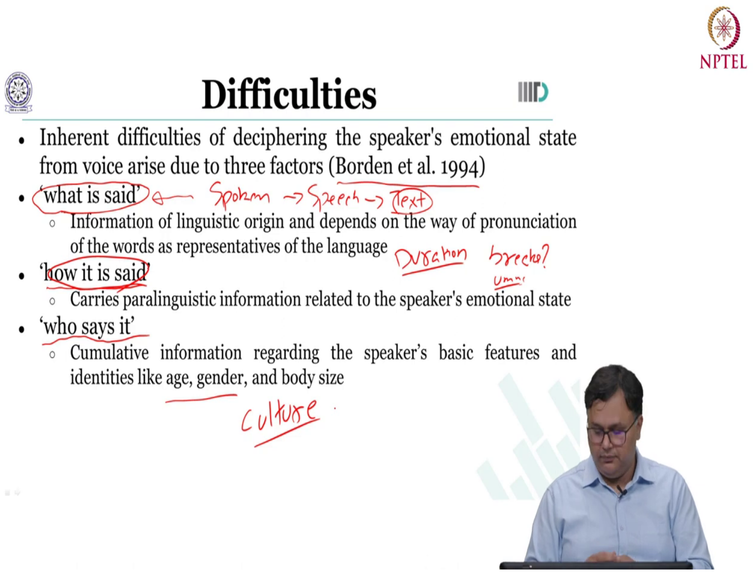
Now, from strictly privacy aware health and well being prospective can the device compute the emotional state of the user and then let us say after the call or communication is over maybe in a subtle way suggest some feedback to the user to perhaps let us say calm down or simply indicate that you have been using the device for n number of hours this is actually quite long you may like to take a break right.

Now, of course, you know in all these kind of passive analysis of the emotion of the user the privacy aspect is extremely important. So, either that information is analyzed used as is on the device and the user is also aware that there is a feature like this on the device or it could be something which is prescribed suggested to the user by an expert. So, the confidentiality and privacy that need to be taken care of.

Now, this is a very interesting aspect friends on one end we were saying well when you use a camera to understand the facial expression of a person there is a major concern with the privacy. Therefore, microphone could be a better sensor. So, analysis or voice could be a better medium.

However, the same applies to your voice based on analysis as well because we can analyze the identity of the subject through the voice and also when you speak there could be personal information. So, where the processing has to be done to understand the affect through voice is it on device of the user where is it stored. So, these are all very extremely important applications which come into the picture when we are talking about these applications.

(Refer Slide Time: 21:04)



Difficulties

- Inherent difficulties of deciphering the speaker's emotional state from voice arise due to three factors (Borden et al. 1994)
- 'what is said' ← Spoken → Speech → Text
  - Information of linguistic origin and depends on the way of pronunciation of the words as representatives of the language Duration breaks?
- 'how it is said'
  - umm
  - Carries paralinguistic information related to the speaker's emotional state
- 'who says it'
  - Cumulative information regarding the speaker's basic features and identities like age, gender, and body size

culture

Now, let us discuss about some difficulties in understanding of the emotional state through voice. So, according to Borden and others there are three major factors which are the challenges in understanding of the emotion through voice. The first is what is said. Now, this is about the information of the linguistic origin and depends on the way of pronunciation of words as represented tips of the language. What did the person actually say right? The content for example, I am feeling happy today right.

So, the content what is being spoken the interpretation of this based on the pronunciation of the speaker that could vary if that varies if there is any noise in the understanding of this content which is being spoken then that can lead to noisy interpretation of the emotion as well. The second part the second challenge is how it is said you know how is a particular statement said.

Now, this carries again paralinguistic information which is related to the speaker's emotional state. And example is let us say you were in discussion with a friend and you asked ok do you agree to what I am saying? The person replies in scenario 1. Yes. Yes, I agree to what you are saying. In scenario 2, the person says hm, yes. Hm, I agree. Now, in these two examples, there is a difference right. The difference in which how the same words were said the difference was the emotion.

Let us say, the confidence in this particular example of how the person agreed to the other if the person agreed or not or was the bit you know hesitant. So, we have to understand how the content is being spoken which would indicate the emotion of the speaker. Now, looking at the third challenge third difficulty in understanding emotion from voice which is who says it ok. So, this means you know the cumulative information regarding the speaker's basic attributes and features for example, the age, gender and even body size.

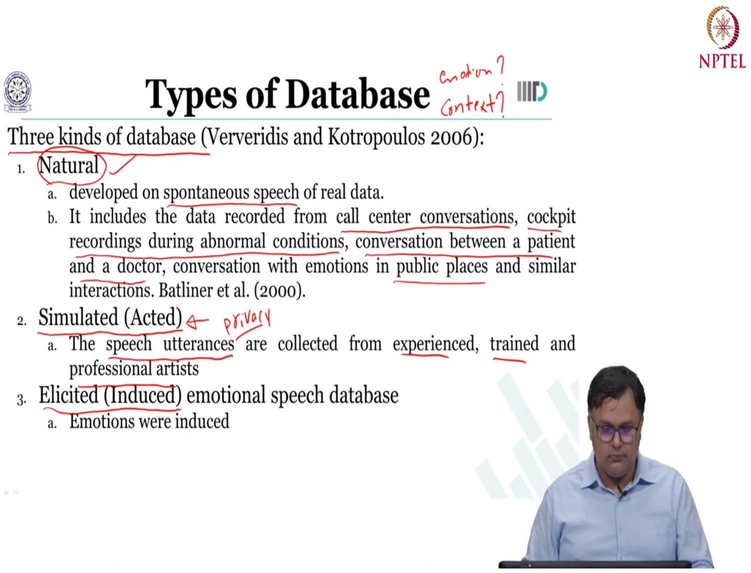
So, in this case let us say a young individual saying I am not feeling any pain you know as an example versus an individual you know adult saying I am not feeling any pain right a young individual versus an adult speaking the same content I am not feeling any pain. Maybe the young individual is a bit hesitant maybe the adult who is speaking this is too cautious. So, what; that means, is the attributes the characteristics of the speaker which not only is based on just their age, gender and body type, but also their cultural context.

So, in some cultures it could be a bit frowned upon to express certain type of emotion in a particular context right. So, that means, if we want to understand the emotion through voice of a user from a particular culture or a particular age range we need to equip our system our affective computing system with this Meta information. So, that the machine learning model then could be made aware during the training itself that there could be differences in the emotional state of the user based on their background their cultures.

So, this means to understand emotion we need to be able to understand what is spoken ok. So, you can think of it as speech to text conversion then how it is said a very trivial way to explain will be you got the text what was the duration in which the same was said were there any breaks were there any umms and repetition of the same words you know.

So, that would indicate how it is being said and then the attributes of the speaker. So, we would require all this information when we would be designing this voice based affect computing system.

(Refer Slide Time: 27:11)



Types of Database

Emotion?
Context?

Three kinds of database (Ververidis and Kotropoulos 2006):

1. Natural
   a. developed on spontaneous speech of real data.
   b. It includes the data recorded from call center conversations, cockpit recordings during abnormal conditions, conversation between a patient and a doctor, conversation with emotions in public places and similar interactions. Batliner et al. (2000).
2. Simulated (Acted) ← privacy
   a. The speech utterances are collected from experienced, trained and professional artists
3. Elicited (Induced) emotional speech database
   a. Emotions were induced

Now, as we have discussed earlier when we were talking about facial expression analysis through cameras the extremely important requirement for creating a system is access to data which has these examples which you could use to train a system right. Now, when we are talking about voice based affect then there are three kind of databases you know the three broad categories of databases which are existing, which have been proposed in the community.

Now, the attributes of these databases is essentially based on how the emotion has been illustrated. So, we will see what does that mean and what is the context in which the participant of the database have been recorded ok. So, let us look at the category. So, the first is natural data ok. Now, this you can very easily link to facial expressions again. We are talking about spontaneous speech in this case.

Spontaneous speech is what you are let us say creating a group discussion you give a topic to the participants and then they start discussing on that. Let us say they are not provided with much constraints it is supposed to be a free form discussion and during that discussion within the group participants you record the data ok. So, that would be spontaneous replies spontaneous questions and within that we will have the emotion which is represented by a particular speaker.

Now, other environments scenarios where you could get this kind of spontaneous speech data which is reminiscent of representative of natural environment is for example, also in call center conversations ok. So, in this case you know let us say customer calls in there is a call center representative conversation goes on and if it is a real one then that could give you know spontaneous speech.

Similarly, you could have you know cockpit recordings during abnormal conditions. Now, in this case what happens right let us say there is an adverse condition there is an abnormal condition the pilot or the user they would be communicating based on you know how they would generally communicate when they are under stress. And in that whole exercise we would get the emotional speech right.

Then also conversation between a patient and a doctor and I already gave you an example right when we were talking about how voice could be used for affect computing in the case of health and well-being right. A patient asking questions to sorry a patient replying to questions to doctors a patient replying to the questions of the doctor and in that case you know we would have these conversations about emotions. Same goes for you know these communication which could be happening in public places as well.

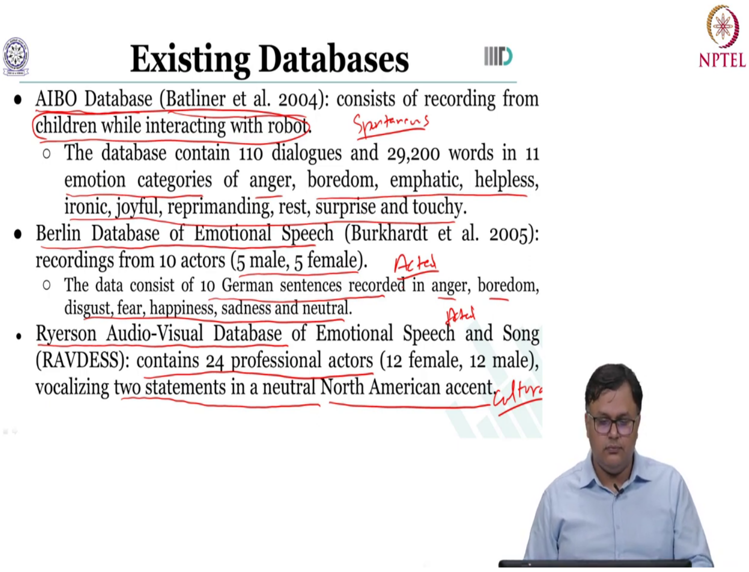
Now, the other category for the voice based data set friends is simulated or acted. Now, in this case the speech utterances the voice patterns they are collected from experienced trained professional artist. So, in this case you know you would have let us say actors who would be coming to a recording studio and then they could be given a script or a topic to speak about and you know that data would be recorded.

Now, there are several advantages when it comes to the simulated data. Advantages, well it is relatively easier to collect as compared to natural data. Since the speakers they are already you know informed about the content or the theme which they are supposed to speak they also have given an agreement. So, you know the cash compared to your natural data privacy could be better handled in this case, but I should say easier to handle.

Now, the issue of course, is when you are talking about simulated data, acted data, then not all examples which you are capturing in your data set could be the best examples of how the user behavior will be in the real world. Now, the third category friends, is elicited emotion which is induced.

So, in this case an example of course, is you know let us say you show a stimuli, a video which contains positive or negative effect and after the user has seen the video, you could ask them to answer certain questions about that video and the assumption is that the stimuli would have elicited some emotion into the user right. And that would be affected represented shown when the speaker the user in the study is answering questions.

(Refer Slide Time: 33:46)



Existing Databases

- AIBO Database (Batliner et al. 2004): consists of recording from children while interacting with robot. Spontaneous
  - The database contain 110 dialogues and 29,200 words in 11 emotion categories of anger, boredom, emphatic, helpless, ironic, joyful, reprimanding, rest, surprise and touchy.
- Berlin Database of Emotional Speech (Burkhardt et al. 2005): recordings from 10 actors (5 male, 5 female). Acted
  - The data consist of 10 German sentences recorded in anger, boredom, disgust, fear, happiness, sadness and neutral.
- Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): contains 24 professional actors (12 female, 12 male), vocalizing two statements in a neutral North American accent. Culture

Acted

Now, let us look at the databases which are very actively used in the community. The first is the AIBO Database by Batliners and others. Now, this contains the interaction between children and the AIBO robo contains 110 dialogues and the emotion categories, the labels are anger, boredom, empathetic, helpless, ironic and so forth.

So, the children are interacting with this robo, the robo is a cute you know Sony AIBO dog robo. So, the assumption here is that the participant would get a bit comfortable with the robo and then emotion would be elicited within the participant. And we can have these labels these emotion categories you know labeled afterwards into the data which has been recorded in during the interaction between the robo and the children.

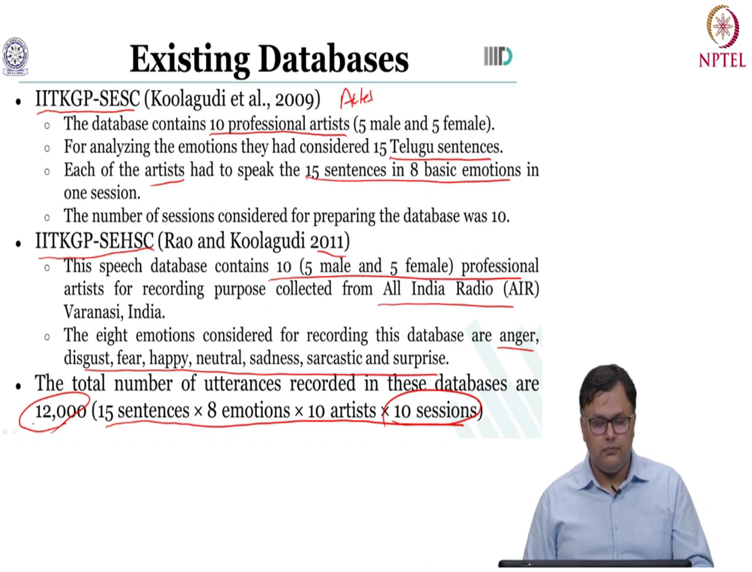
The other dataset which is very commonly used is the Berlin Database of Emotional speech which was proposed by Burkhardt and others in 2005. Now, this contains 10 subjects and you know the data consist of 10 German sentences which are now recorded in different emotions. Notice this one is the acted one ok. So, this is the acted type of dataset where in the content was already provided to the participants, actors and you know they try to speak it in different emotions.

So, what does this mean is now the quality of emotions which are reflected are based on the content and the quality of acting by the participant. Now, friends the third dataset is the Ryerson Audio Visual Database of emotional speech and song. Now, again this is an active dataset contains professional actors and these actors were given the task of vocalizing two statements in North American accent.

Now, of course, you know again the cultural context is coming into the picture as well you know this is also acted and if you compare that with the first dataset of the AIBO dataset then this was more spontaneous ok. Of course, you know you would understand that getting this type of interaction is non-trivial.

So, extremely important to be careful about the privacy and all the ethics approvals which are required to be taken. Now, in these kind of databases where you have actors it is relatively easier to scale the database because you know you could hire actors and you can have multiple recording sessions and you can give different content as well.

(Refer Slide Time: 37:15)



Existing Databases

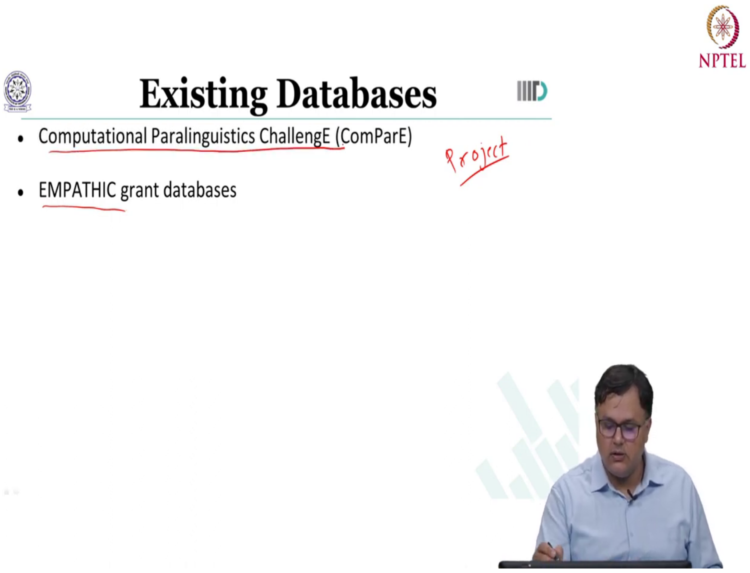
- IITKGP-SESC (Koolagudi et al., 2009) Acted
  - The database contains 10 professional artists (5 male and 5 female).
  - For analyzing the emotions they had considered 15 Telugu sentences.
  - Each of the artists had to speak the 15 sentences in 8 basic emotions in one session.
  - The number of sessions considered for preparing the database was 10.
- IITKGP-SEHSC (Rao and Koolagudi 2011)
  - This speech database contains 10 (5 male and 5 female) professional artists for recording purpose collected from All India Radio (AIR) Varanasi, India.
  - The eight emotions considered for recording this database are anger, disgust, fear, happy, neutral, sadness, sarcastic and surprise.
- The total number of utterances recorded in these databases are 12,000 (15 sentences × 8 emotions × 10 artists × 10 sessions)

Now, moving on to other databases friends the next is the IITKGP SESC dataset which was proposed in 2009 by Koolagudi and others. Again, an active dataset 10 professional artist. Now, this is a non-English dataset it is in an Indian language Telugu. Now, each artist participant here they spoke 15 sentences trying to represent 8 basic emotions in 1 session. Another dataset is again from the same lab called the IITKGP SEHSC again by Rao and Koolagudi and others 2011.

Now, in this case you again have 10 professional actors, but the recording is coming from radio jockeys ok so from all India radio. So, these are extremely good speakers and the emotions are 8 categories again acted. But if you have high quality actors then the assumption is that we would be able to get emotional speech as directed during the creation of the dataset.

Now, moving forward friends the number of utterances here you know this is a fairly good sized dataset 15 sentences, 8 emotions and then 10 artist, they were record in 10 sessions. So, we have 12000 samples which are available for the learning of an emotional speech analysis in system.

(Refer Slide Time: 39:01)

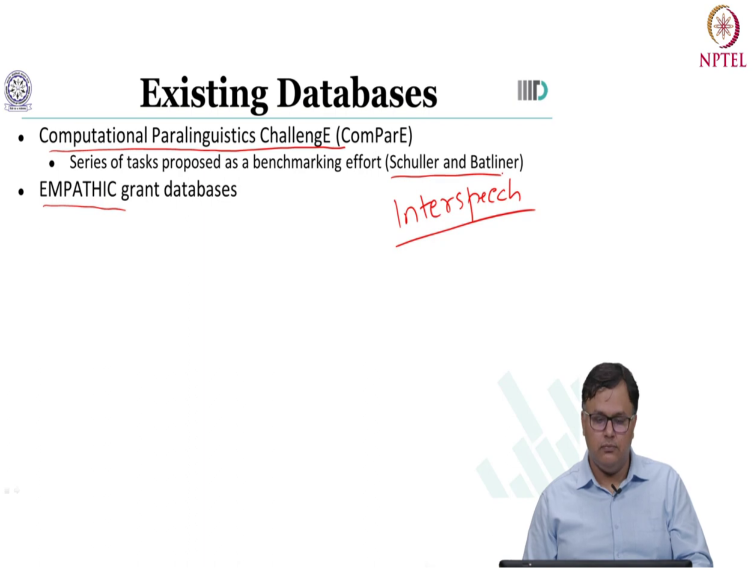


Existing Databases

- Computational Paralinguistics ChallengE (ComParE)
- EMPATHIC grant databases

Now, in the community there are several projects going on. Now, they are looking at different aspects of affect and behavior. So, one such is the empathetic grant in the EU. So, there as well you know there are these dataset resources which are used for analysis of emotions and speech. Another extremely useful very commonly used platform is the computational paralinguistic challenge platform by Schuller and Batliner.

(Refer Slide Time: 39:42)



Existing Databases

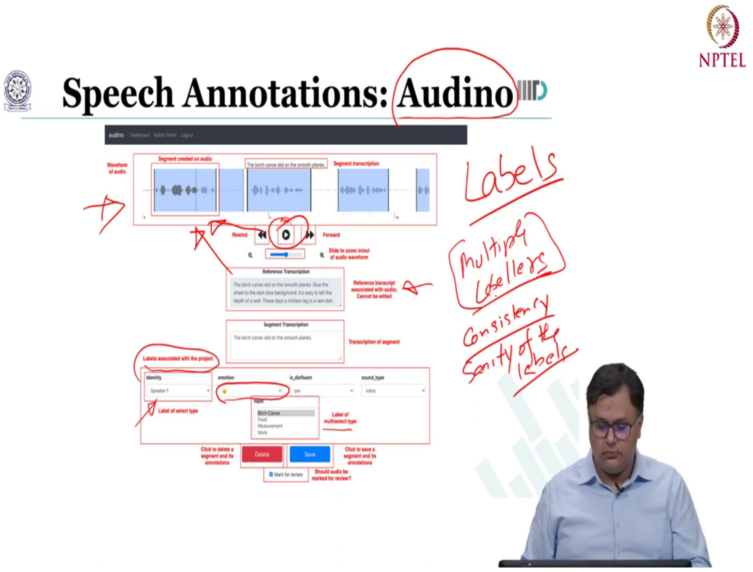
- Computational Paralinguistics ChallengE (ComParE)
  - Series of tasks proposed as a benchmarking effort (Schuller and Batliner)
- EMPATHIC grant databases

Interspeech

So, this is actually hosted as part of a conference called inter-speech. Now, this is a very reputed conference speech analysis. So, in the compare benchmarking challenge the organizers have been proposing every year different sub challenges which are related to speech analysis and a large number are related to emotion analysis and different task and different settings in which we would like to understand the emotion of a user or a group of users.

Now, there would be some acted and some spontaneous datasets which are available on this benchmarking platform.

(Refer Slide Time: 40:40)



Speech Annotations: Audino

The birch canoe slid on the smooth planks.

Reference Transcription

The birch canoe slid on the smooth planks. Glue the sheet to the dark blue background. It's easy to tell the depth of a well. These days a chicken leg is a rare dish.

Segment Transcription

The birch canoe slid on the smooth planks.

Identity: Speaker 1

is_disfluent: yes

sound_type: voice

topic: Birch Canoe

☑ Mark for review

NPTEL
audino
Dashboard
Admin Panel
Logout
Waveform of audio
Segment created on audio
Segment transcription
Rewind
Play
Forward
Slide to zoom in/out of audio waveform
Reference transcript associated with audio; Cannot be edited
Transcription of segment
Labels associated with the project
emotion
Label of select type
Food
Measurement
Work
Label of multiselect type
Click to delete a segment and its annotations
Delete
Save
Click to save a segment and its annotations
Should audio be marked for review?
Labels
Multiple Labellers
Consistency
Sanity of the labels

Now, moving on from the speech databases, let us see the databases have been collected could be acted could be spontaneous. The next task is to generate the annotations, the labels. Of course, before the recording is done the design of the experiment would already consider the type of emotions which are expected to be annotated generated from the data. So, one popularly used tool for annotation of speech is the audino tool.

Now, here is a quick go through of how the labeling is done. Let us say friends here is the waveform representation the labeler would listen to the particular chunk let us say this chunk can re-listen can move forward backwards. And they also could have access to the transcript what is being spoken during this time.

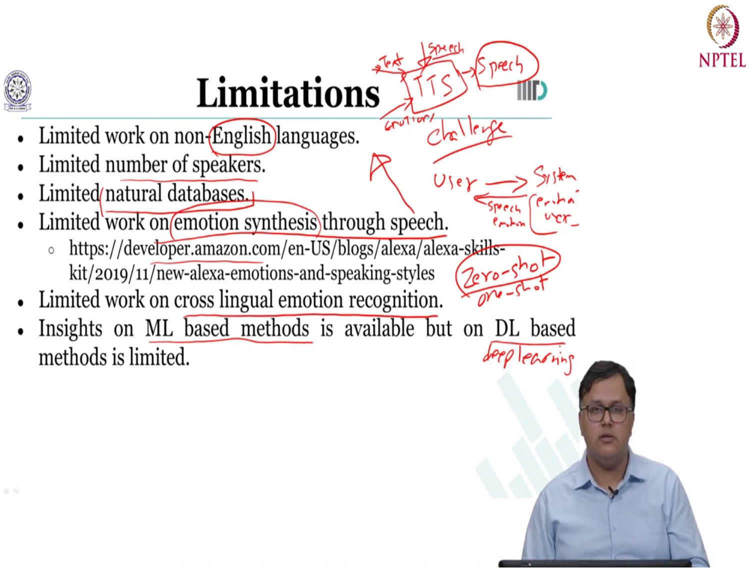
So, what they can do is you know they can then add the emotion which they interpret from this audio data point. And they can label you know different things such as the topic of the spoken text. And also things such as you know who is the speaker and the metadata. So, once they listen to the content they generate the labels, then they can save and then they can move forward.

Now, extremely important to have the right annotation tool because you may be planning to create a large dataset representing different settings. So, in that case you would also have multiple labelers right. So, if you have multiple labelers the tool needs to be scalable.

And as friends we have already discussed in the facial expression recognition lectures as well. If you would have multiple labelers you will have to look at things such as consistency for each labeler how consistent they are in the labeling process with respect to the sanity of the labels.

So, you would like the labels to be as less affected by thing such as confirmation bias. So, after you have the database collected annotated by multiple labelers you may like to do the statistical analysis of the labels for the same samples where the labels are generated from multiple labelers. So, that at the end we have one or multiple coherent labels for that data point.

(Refer Slide Time: 44:01)



Limitations

- Limited work on non-English languages.
- Limited number of speakers.
- Limited natural databases.
- Limited work on emotion synthesis through speech.
  - https://developer.amazon.com/en-US/blogs/alexa/alexa-skills-kit/2019/11/new-alexa-emotions-and-speaking-styles
- Limited work on cross lingual emotion recognition.
- Insights on ML based methods is available but on DL based methods is limited.

Now, let us look at some of the limitations and this also is linked to the challenges in voice based effect analysis. What we have seen till now is that there is a limited work on non English language based effect analysis. You already saw the IIT KGP datasets which were around the Telugu language and then the Hindi language. There are a few Germans speaking datasets as well, Mandarin as well, but they are lesser in number, smaller in size as compared to English only based datasets.

So, what that typically would mean is let us say you have a system which is analyzing the emotion of a speaker speaking in English. You use that dataset and then you train a system on that dataset. Now, you would like to test it on other users who are speaking some other language.

Now, this cross dataset performance across different languages, that is a big challenge right now in the community. Why is it a challenge? Because you have already seen the challenges which are there are the three challenges you know what is being said, who said it and how it was said. So, these will vary across different languages.

The other is limited number of speakers. So, if you want to create emotion detection in a system based on voice which is supposed to work on a large number of users on a large scale, you would ideally like a dataset where you can get a large number of users in the dataset. So, that we learn the variability which is there when we speak right different people will speak differently, will have different styles of speaking and expressing emotions.

Now, with respect to the datasets of course, there is a limitation based on the number of speakers which you can have there is a practicality limit. Let us say you wanted to create a spontaneous dataset. So, if you try to increase the number of participants in the dataset, there could be challenges such as getting the approvals, getting the approval from the participant themselves and so forth.

Now, on the same lines friends issue is there are limited natural databases and I have already explained to you right creating spontaneous dataset is a challenge because if you the user is aware they are being recorded that could add a small bias. The other is the privacy concern needs to be taken into picture. So, the spontaneous conversations you know if the proper ethics and the permissions have been taken or not you know ethics based considerations are there or not so all that affects the number and size of the natural databases.

Now, this is fairly new, but extremely relevant as of today there is not a large amount of work on emotional speech synthesis. So, friends till now I have been talking about you have speech pattern, someone spoke machine analyzed we understood the emotion. But remember we have been saying right affect sensing is the first part of affective computing and then the feedback has to be there.

So, in the case of speech we can have emotional synthesis done. So, the user speaks to interacts with the system, system understands the emotion of the user and then the reply back let us say that is also through speech that can have emotion in it as well right. Now, with respect to the progress in the synthetic data generation of course, we have seen large strides in the visual domain data face generation, facial movement generation.

But comparatively there is a bit less progress in the case of emotional speech and that is due to of course, you know the challenges which I have just mentioned above. So, this of course, is being you know worked upon in several labs across the globe, but that is currently a challenge how to add emotion to the speech. Some examples you can check out for example, from this link from developer.amazon.com you know there are a few styles, few emotions which are added. But essentially the issue is as follows.

Let us say I want to create a text to speech system which is emotion aware. So, I could input into let us say this TTS text to speech system the text this is the text from which I want to generate the speech and let us say as a one hot vector the emotion as well. Now, this will give me the emotional speech. But how do you scale across large number of speakers? Typically high quality TTS systems are subject specific you will have one subjects text to speech model.

Of course, there are newer systems which are based on machine learning techniques such as zero short learning or one short you know where you would require lesser amount of data for training or in zero short what you are saying is well, I have the same text to speech system, which has been trained for large number of speakers along with the text and emotion I would also add the speech from the a speaker for which I want the new speech to be generated based on this text input, right.

So, that is the challenge, how do you scale your text to speech system across different speakers and have the emotion synthesized. The other friends, extremely important aspect which is the limitation currently is cross lingual emotion recognition. I have already given you an example when we are talking about the limited number of non English language based emotion recognition works.

So, you train on language 1 a system for detecting emotion test it on language 2 generally a large performance drop is observed. But one thing to understand is let us say for some languages it is far more difficult to collect data to create databases as compared to some other languages right, some languages are spoken more there are larger number of speakers other languages could be older languages are spoken by less number of people. So, obviously, the creating datasets would be a challenge.

Therefore, in the pursuit of cross lingual emotion recognition we would also like to have systems where, let us say you train a system on language 1, which is very widely spoken and the assumption is that you can actually create a large dataset. Then can be learn systems on that dataset and later borrow and do things such as domain adaptation adapt from that learn from that borrow information and fine tune on another language where the we have smaller datasets.

So, that you know now we can do emotion recognition on data from other smaller dataset. Now, another challenge limitation is this is applicable to not just voice or speech, but other modalities as well. When you are looking at the explanation part of why given a speech sample the system said the person is feeling happy.

If you use traditional machine learning systems for example, your decision trees or support vector machines it is a bit easier to understand why the system reached at a particular emotion why the system predicted a certain emotion. However, speech based emotion through deep learning.

So, this is deep learning friends, DL deep learning based methods. Even though it has the state of the art performance even then the explanation part of why you reached at a certain consensus based on the perceived emotion through the speech of a user that is still a very active area of research.

So, we would like to understand, why the system reached at a certain point with respect to the emotion of the user? Because if you are using this information about emotion state of the user in let us say a serious application such as health and well-being, we would like to understand how the system reached at that consensus.

So, friends with this we reach the end of lecture one for the voiced base emotion recognition and we have seen why speech analysis is important, why is it useful for emotion recognition, and then what are the challenges in understanding of emotion from speech from there, we moved on to the different characteristics of the databases the data which is available for learning voice based emotion recognition systems.

And then we concluded with the limitations which are currently there in voice based emotion recognition systems.

Thank you.