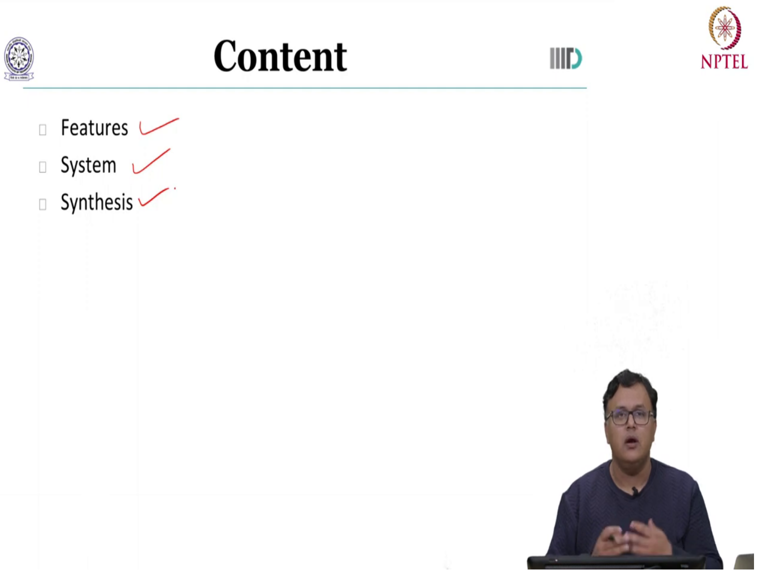
Affective Computing

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

Week - 05
Lecture - 02
Automatic Speech Analysis based Affect Recognition

Hello and welcome. I am Abhinav Dhall from the Indian Institute of Technology Ropar. Friends, today we are going to discuss about the aspects of Automatic Speech Analysis based Affect Recognition. This is part of the Affective Computing course.

(Refer Slide Time: 00:45)



Content

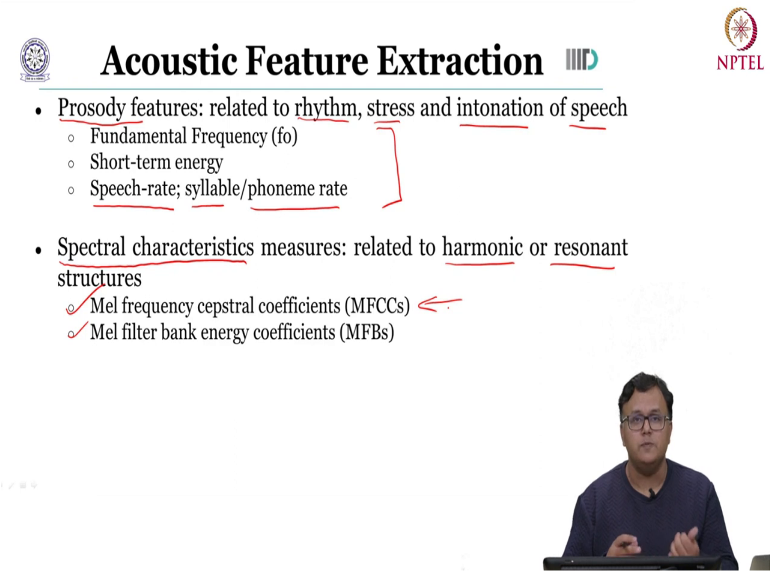
- Features ✓
- System ✓
- Synthesis ✓

So, in the last lecture, we discussed about why speech and voice-based emotion recognition is a useful modality for understanding the affective state of the user. We discussed about the application and challenges. Then we looked at some of the commonly used speech-based affect recognition databases.

Today, we are going to discuss about the feature analysis aspect. So, I will mention to you about some of the commonly used hand engineered features which are used in speech analysis. Then we will look at a system for normalization of the speech feature and later on we will see an example of affect induced speech synthesis.

So, we are not only interested in understanding the emotion of the user using the speech modality, but we are also interested in that if the feedback can be in the form of speech which is generated by a machine, then how can appropriate emotions be added to the generated speech itself for a more engaging and productive interaction with the user, alright.

(Refer Slide Time: 02:11)



Acoustic Feature Extraction

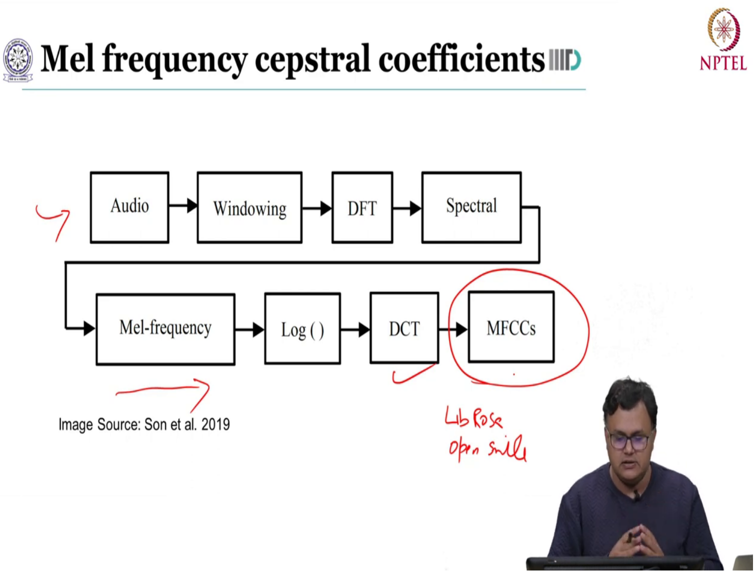
- Prosody features: related to rhythm, stress and intonation of speech
  - Fundamental Frequency (fo)
  - Short-term energy
  - Speech-rate; syllable/phoneme rate
- Spectral characteristics measures: related to harmonic or resonant structures
  - Mel frequency cepstral coefficients (MFCCs)
  - Mel filter bank energy coefficients (MFBs)

So, let us dive in. So, first we will talk about the automatic feature extraction for speech. So, friends, the very commonly used features are referred to as a prosody features. These relate to the rhythm stress and the intonation of the speech. These are generally computed in the form of the fundamental frequency, the short-term energy of the input signal and simple statistics such as the speech rate, syllable and the phoneme rate.

Now, along with this, we also have features for speech analysis based on the measurement of the spectral characteristics of the input speech. Now, these are related to the harmonic or resonant structures. And typically, commonly these are based on your Mel frequency cepstral coefficients, the MFCCs and the Mel filter bank energy coefficients.

In fact, MFCCs is one of the most commonly used speech analysis feature for not only just emotion recognition, but also for other speech related applications such as automatic speech recognition.

(Refer Slide Time: 03:43)



Mel frequency cepstral coefficients

Audio → Windowing → DFT → Spectral → Mel-frequency → Log ( ) → DCT → MFCCs

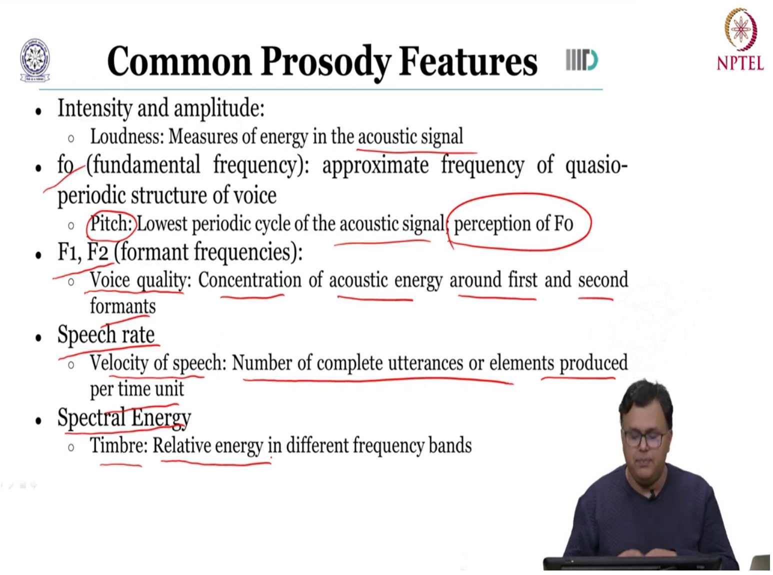
Image Source: Son et al. 2019

LibRosa
open smile

Now, the MFCCs are as I mentioned the most commonly used feature. So, from a practitioner’s perspective, you can extract the MFCCs with library such as librosa or openSMILE. Now, the steps are as follows, here you have continuous audio which is coming in or since it is a continuous audio signal, we divide it into chunks. So, we are doing, we are doing here. So, we would then apply the discrete Fourier transform where we move in the incoming audio signal into the Fourier domain.

Then we are going to apply the Mel filter banks to these power spectra which we are getting from spectral analysis of the DFT outputs and then we are going to sum the energy in each filter. Later on, we compute the logarithm of these filter bank energies and take the discrete cosine transform of the input. Further, this gives us the MFCCs where we will keep the certain DCT coefficients and discard others.

(Refer Slide Time: 05:08)



Common Prosody Features

- Intensity and amplitude:
  - Loudness: Measures of energy in the acoustic signal
- fo (fundamental frequency): approximate frequency of quasio-periodic structure of voice
  - Pitch: Lowest periodic cycle of the acoustic signal, perception of Fo
- F1, F2 (formant frequencies):
  - Voice quality: Concentration of acoustic energy around first and second formants
- Speech rate
  - Velocity of speech: Number of complete utterances or elements produced per time unit
- Spectral Energy
  - Timbre: Relative energy in different frequency bands

Now, if you look at the prosody-based features. So, we are interested in analysing the intensity and the amplitude of the input signal wherein you can use loudness which is a measure of the energy in your input acoustic signal. We also very commonly use the fundamental frequency. Now, the fundamental frequency, this is based on the approximation of the frequency of your quasio periodic structure of voice.

And based on the fundamental frequency from the perception perspective, we have the feature attribute called pitch, which is essentially the lowest periodic cycle of your input acoustic signal. It is commonly referred to as the perception of your fundamental frequency.

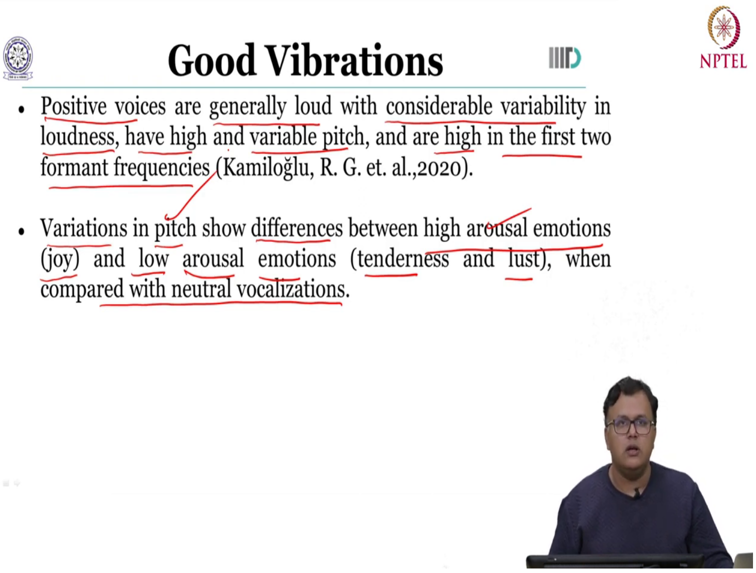
Typically, this means you will require listeners to be able to quantify the pitch of an input signal, but from a compute perspective wherein we have an automatic system, we are going to compute the fundamental frequency. Then friends moving on, we have the formant frequencies F1 and F2.

Now, you can compute the quality of a voice through this wherein we are looking at the concentration of the acoustic energy around the first and the second formants. From this another commonly used prosody feature is the speech rate as the name suggests. We are interested in let us say computing the velocity of the speech, which is basically number of complete utterances or elements produced per time unit, ok. The other one which is very commonly used is your spectral energy.

Now, through this you can compute the timbre which essentially the relative energy in the different frequency bands of your input signal. Now, a point to note here is these features are very fundamental basic features have been extensively used in speech analysis for emotion understanding of a user and in other applications. But again, these features such as MFCC are hand engineered features.

So, Friends similar to how we discussed for automatic facial expression recognition wherein early on histogram of gradients, scale in within feature transform features these were used. But the community in academia and industry it moved to representation learning through deep learning and we have these pre-trained networks through which we are extracting the features.

(Refer Slide Time: 08:22)



Good Vibrations

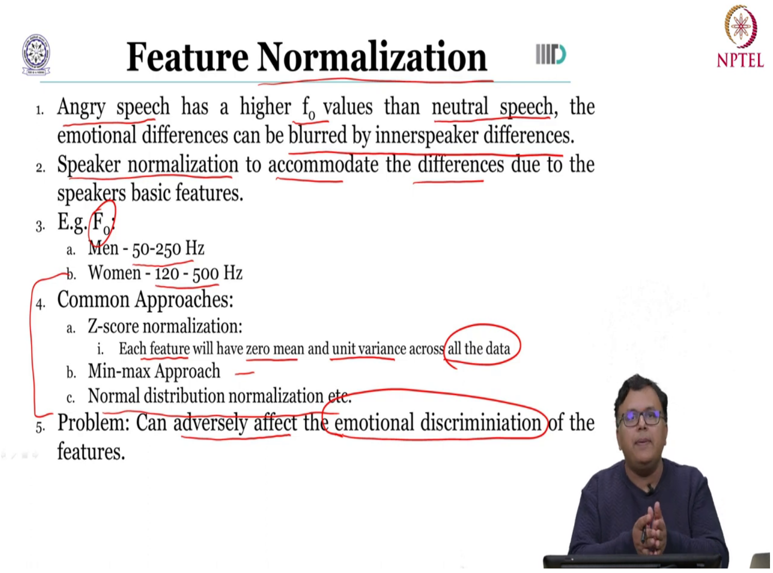
- Positive voices are generally loud with considerable variability in loudness, have high and variable pitch, and are high in the first two formant frequencies (Kamiloğlu, R. G. et. al.,2020).
- Variations in pitch show differences between high arousal emotions (joy) and low arousal emotions (tenderness and lust), when compared with neutral vocalizations.

Now, later on I will show you an example of how the community in speech analysis has moved to the representation called spectrograms which is ideal for using convolution neural networks. Now, what is observed is based on these features which we discussed. If we have positive voices which are generally loud, they are with considerable variability in the loudness attribute.

And have high and variable pitch and they are having high in the first two format frequencies F1 and F2. Further it is observed that the variations in the pitch they tell us about the differences between high arousal emotions. So, again we are talking about the valence in arousal dimension where we are talking about the arousal emotion for example, representing joy.

And friend’s low arousal emotions such as tenderness and lust when compared with neutral vocalizations, ok so, pitch is an important feature for looking at the high and lower arousal emotion.

(Refer Slide Time: 09:44)



Feature Normalization

1. Angry speech has a higher f_0 values than neutral speech, the emotional differences can be blurred by innerspeaker differences.
2. Speaker normalization to accommodate the differences due to the speakers basic features.
3. E.g. F_0:
   a. Men - 50-250 Hz
   b. Women - 120 - 500 Hz
4. Common Approaches:
   a. Z-score normalization:
      i. Each feature will have zero mean and unit variance across all the data
   b. Min-max Approach
   c. Normal distribution normalization etc.
5. Problem: Can adversely affect the emotional discriminiation of the features.

Now, what typically happens is after let us say we have extracted a feature. We will have these features extracted from n number of different data points of those are speech feature-based data points which are coming from either different sources different speakers.

So, there is a very important step which is involved in the pipeline for speech-based emotion recognition which is the normalization of the feature. Now, an example of this which will set the motivation is as follows. When you have angry speech, it is observed that it has higher fundamental frequency values than compared with the frequency values fundamental frequency values for the neutral speech.

Now, this difference in the two emotions that is actually blurred by the inner speaker differences so, what that means, you have n speakers who let us say are speaking in turn both an angry speech and neutral speech. Though we observe differences between the fundamental frequency value of the angry and the neutral speech between one subject.

But if you have multiple subjects because of the difference in style of speaking of these subjects because of this intra-class variability the difference between the fundamental frequency of the angry speech and neutral speech would be varying, ok. So, you will observe a variation across different speakers. So, this observation that you know angry speech will have a higher fundamental frequency as compared to neutral speech would be blurred in some cases, but would be very evident in other speaker's cases.

Therefore, speaker normalization is used to accommodate these kind of differences which are introduced in my data set due to the differences in the speakers. Now, here is an example based on gender friends. For the fundamental frequency men would typically have signal in between 50 to 250 hertz and if you look at women subjects it would be between 120 and 500 hertz.

So, there is a difference in the range here. Now, for feature normalization there is a very simple approach let us talk about that one commonly used which is the Z-score normalization. What do we do in this Z-score normalization? We say well, I am going to take each feature and I am going to transpose it so that it will have zero mean and unit variance across all the data.

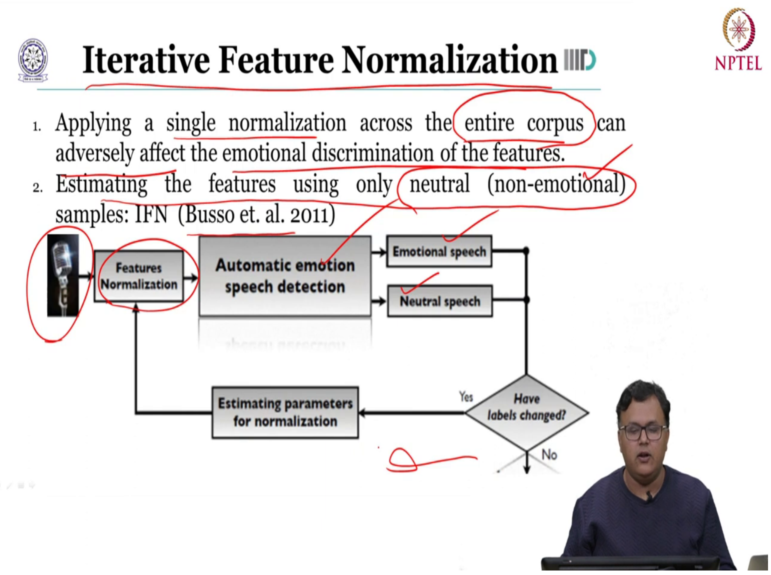
Now, please notice this is actually a very common technique which is not just limited to speech, but it is applied to vision and text features as well. Other approaches are the min-max approach. So, you set the bound the minimum and the maximum value across the whole data set.

And then you will map the whole of the features from all the speech data points based on this minimum and maximum. So, you are just you either stretching or you are actually squeezing the feature values for a particular data point. Then friend's others are based on normalization of the distribution. So, you will like to have a normal distribution for the feature which you are extracting from speech.

Now, here is a problem with these standard you yet very commonly used features. The problem is these feature normalization techniques can adversely affect the emotional discrimination of the features. Since what we are doing is our observation is due to the differences in the different speakers, we observe that the observable difference between the feature for different emotion that varies.

So, we are applying some normalization techniques. However, since this normalization technique is generally applied across the whole data set then what it can do is, it can sometimes reduce the discriminative aspect of the feature for certain subjects as we are comparing all the subjects together.

(Refer Slide Time: 14:40)



Iterative Feature Normalization

1. Applying a single normalization across the entire corpus can adversely affect the emotional discrimination of the features.
2. Estimating the features using only neutral (non-emotional) samples: IFN (Busso et. al. 2011)

Features Normalization

Automatic emotion speech detection

Emotional speech

Neutral speech

Have labels changed?

Yes

No

Estimating parameters for normalization

So, to mitigate this now I will mention about one simple yet very effective technique for feature normalization which is called the Iterative Feature Normalization. Now, the motivation for IFN iterative feature normalization is as follows. As we have seen that the applying a single normalization across the entire corpus can adversely affect the emotional discrimination of the features.

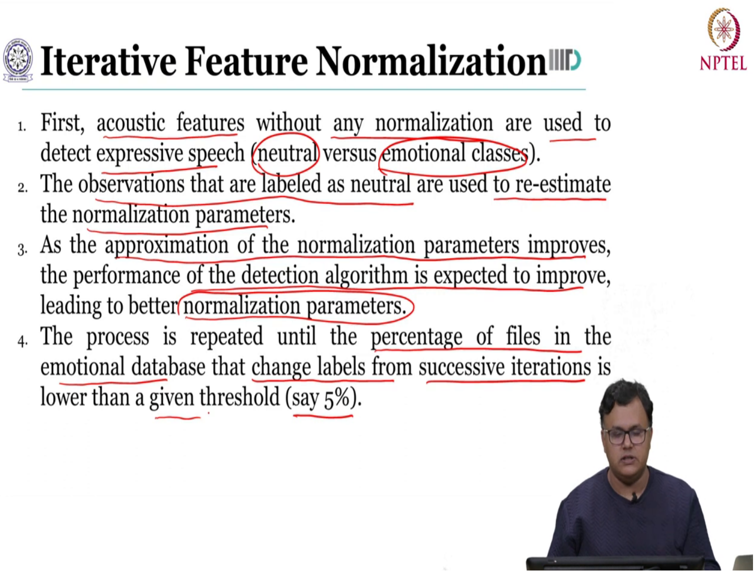
Therefore, let us try to work on this aspect ok, which is applying the normalization on the entire data set. So, what Carlos Busso and his team proposed in 2011 was that let us estimate the feature using only the neutral non-emotional samples. So, let me have a baseline let me have a reference and what can be a good reference? Well, identifying the neutral emotion utterance in audio sample.

Now, once I know what is the neutral utterance or neutral sample in my data for a subject, I can treat that as a baseline to normalize the other samples, ok. Now, friend's similar methods are also used in facial expression analysis with videos as well. Wherein one could modify the geometric features of a face based on the neutral expression. So, typically a neutral expression you will observe the lips are closed.

So, the difference between the points the distance is less very less negligible. So, that is used as a baseline for comparing with let us say when the mouth is wide open, ok. So, now let us come back to speech, ok. So, we are going to talk about the iterative feature normalization. You get the input audio signal we are interested in feature normalization. So, what do we do?

We use automatic emotion speech detection and that gives us two labels indicating if it is neutral speech and if it is emotional speech. Once we know what is emotional speech, we then use that to normalize the parameters. We do this iteratively and then we get the ideal normalization.

(Refer Slide Time: 17:24)



Iterative Feature Normalization

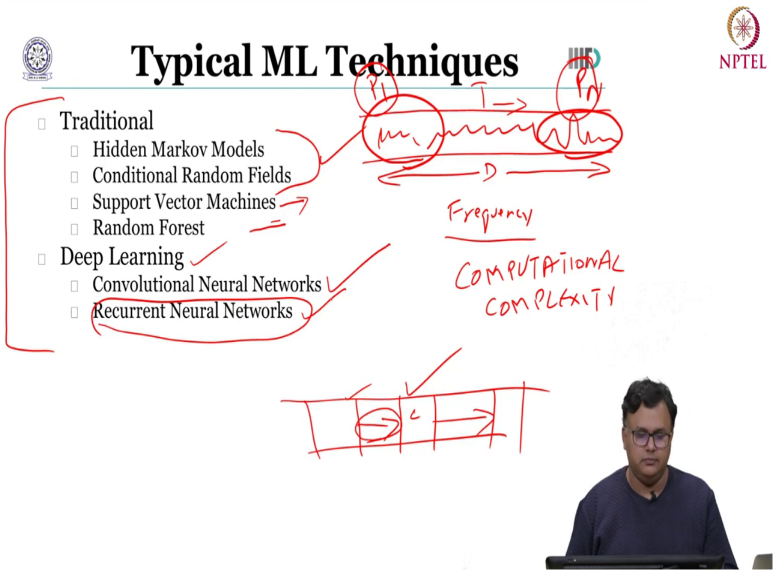
1. First, acoustic features without any normalization are used to detect expressive speech (neutral versus emotional classes).
2. The observations that are labeled as neutral are used to re-estimate the normalization parameters.
3. As the approximation of the normalization parameters improves, the performance of the detection algorithm is expected to improve, leading to better normalization parameters.
4. The process is repeated until the percentage of files in the emotional database that change labels from successive iterations is lower than a given threshold (say 5%).

So, here are the steps one by one. First, we take the acoustic features could be any of a feature friend's your MFCCs your F naughts and so forth without any normalization. So, we do not apply any Z-cross normalization or any min-max approach. Now, we use these features and we detect the expressive speech. Essentially, which part of the speech is neutral which is showing some emotion so, kind of a binary problem binary class problem.

Now, the these observations which are labeled as neutral. So, those part of the speech or the speech sample themselves which are labeled as neutral, they are used to re-estimate the normalization parameters. As now the approximation of normalization parameter improves, the performance of the detection algorithm is expected to improve. Now, this in turn is going to give you better normalization parameters.

So, this is an iterative process and this process is repeated until certain percentage of files in the emotional database they have now got change label from successive iteration. Now, this is give from the original work this is given a threshold of 5 percent. But you know you can empirically vary that as well. So, what you are saying you get neutral change parameters, reiterate now again detect neutral and expressive speech and get the newer parameters and run this iterative process.

(Refer Slide Time: 19:04)



Typical ML Techniques

- Traditional
  - Hidden Markov Models
  - Conditional Random Fields
  - Support Vector Machines
  - Random Forest
- Deep Learning
  - Convolutional Neural Networks
  - Recurrent Neural Networks

Now, friends what we have done? We said we have a bunch of features which could be extracted for analysis of the acoustic signal, we then discussed about the feature normalization part. After that we can learn different machine learning techniques. Now, here I am mentioning the commonly used machine learning techniques. Now, the reason to discuss this is as follows.

Based on how we are extracting the audio feature, let us say you know this is your timeline and we have some audio feature which we are extracting, ok. The duration of the window the frequency of occurrence of the input so, at what frequency am I getting the data. And let us say in this signal if I was to consider this part and call it part P 1 and call this part as P of N, how important it is for me to correctly be able to predict the emotion for P of N with or without prior information let us say P of 1?

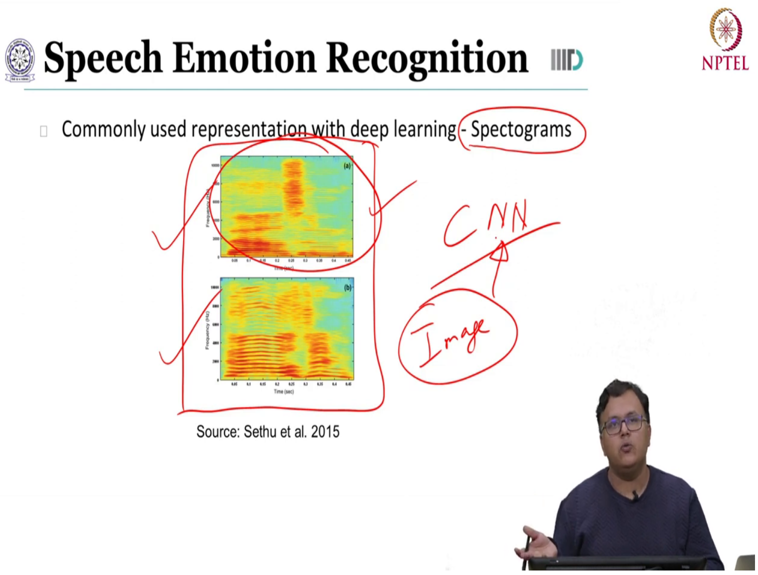
So, how much of the prior information is required? Ok. In other words, how much temporal variation is required across the windows, how much history do you need? Ok. So, that is based on the use case. And also, would be one of the primary factors along with the frequency time duration and also let me add here the computational complexity, ok. To decide which machine learning technique are you going to use.

So, commonly in speech analysis the state-based machines they have been used the hidden Markov models, the condition random fields also researchers have used, support vector machines at random forest wherein they will first compute the feature from the whole of the sample mostly you know you will take the whole of the sample speed sample extract the features and then either run a support vector machine or random forest.

Recently we have also seen researchers are using deep learning based techniques as well. So, either you can use your convolutional neural network or your recurrent neural networks. Now, obvious question comes well, if you want to use the convolutional neural network you would need let us say an image like feature, right. So, we will come to this in the coming slides.

Your RNN based learning the motivation is very similar to this. You have your P signal you divide it into chunks and let us say you want to learn the feature from the chunks and also want to understand the information from the prior right from the background. So, let us say I am here in the cell and I would not only analyse the feature for this a chunk, but I also want some learning from the background, right. So, that is how you would commonly use RNNs as well.

(Refer Slide Time: 22:49)



Speech Emotion Recognition

- Commonly used representation with deep learning - Spectograms

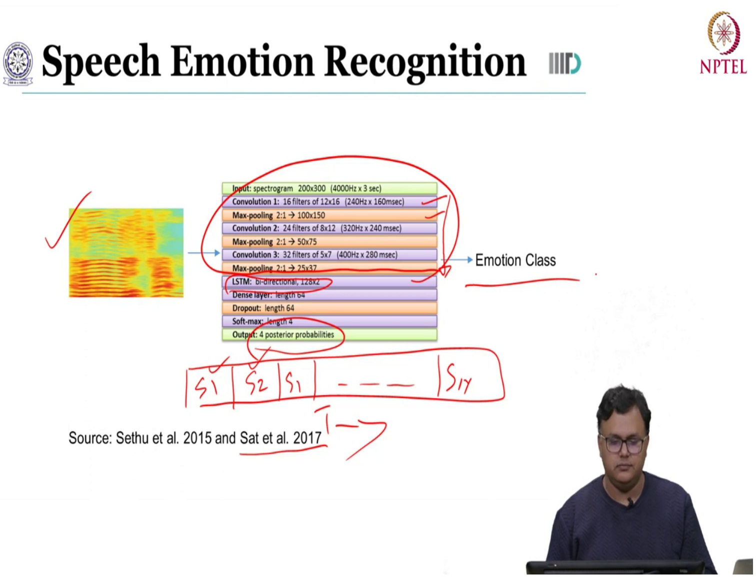
Source: Sethu et al. 2015

Now, coming to the feature which I have been talking about for which is commonly used in your convolutional neural networks and has been shown to have high quality highly accurate speech-based emotion recognition is the representation of the audio signal in the form of spectrograms.

So, this is essentially the visualization of the frequency. So, you can see frequency against time also it gets you the amplitude. Now, what do you see here friends is there are two spectrograms from the same subject spectrogram one is when the speech was neutral and spectrogram two is when the speech was angry. You can very well see the differences in these two visualizations.

And since we are able to visualize this, I can treat a spectrogram as an image. You know I can assume that this is actually an image. Now, if this is an image then I can use a convolutional neural network to train with a spectrogram as an input and the output would be the emotion classes.

(Refer Slide Time: 24:18)



Speech Emotion Recognition

Input: spectrogram 200x300 (4000Hz x 3 sec)
Convolution 1: 16 filters of 12x16 (240Hz x 160msec)
Max-pooling 2:1 → 100x150
Convolution 2: 24 filters of 8x12 (320Hz x 240 msec)
Max-pooling 2:1 → 50x75
Convolution 3: 32 filters of 5x7 (400Hz x 280 msec)
Max-pooling 2:1 → 25x37
LSTM: bi-directional, 128x2
Dense layer: length 64
Dropout: length 64
Soft-max: length 4
Output: 4 posterior probabilities

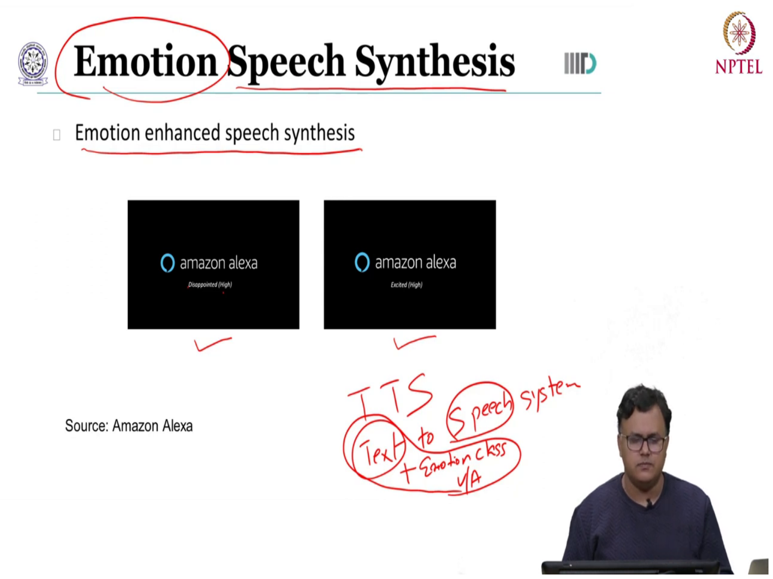
→ Emotion Class

Source: Sethu et al. 2015 and Sat et al. 2017

So, here is an example of a work by Sat and others where they proposed this emotion recognition system wherein you have your spectrogram as input. And then similar to your traditional deep convolutional neural network you have your series of convolutions, max pooling and so forth and to induce the time window as well. So, essentially you have a bi-directional LSTM. So, what you are doing?

You are saying well this was my audio signal, this is time I have my S1 spectrogram one S2 spectrogram two. Of course, you could have overlapping windows as well, but this is just for visualization S3 and so forth till S of N, right. So, these are all spectrogram you input this here you get the feature representation for each spectrogram and then you are using a recurrent neural network to finally, predict the emotion categories ok, these emotion classes.

(Refer Slide Time: 25:24)



Emotion Speech Synthesis

- Emotion enhanced speech synthesis

Source: Amazon Alexa

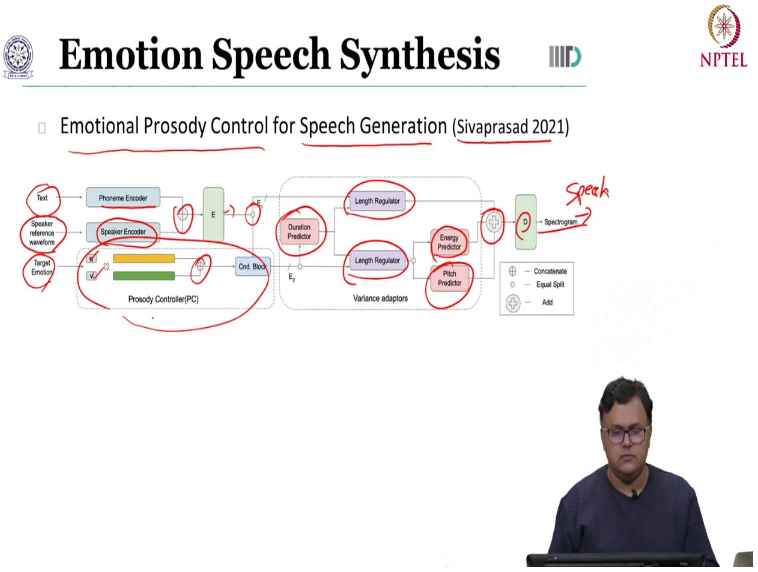
Now, with this we have seen how emotion is predicted using speech signal and friends with the tutorial which is after this lecture. So, for this very week you will also see a in detail example of how to create a simple speech-based emotion recognition system starting from getting a dataset and then extracting different features and trying out different classifiers.

So, till this is the recognition part which falls under the affect sensing step of affective computing. Now, let me give you an example of speech synthesis where emotion is also added. So, here we have two samples which are generated using Amazon's Alexa. The first one is your speech synthesis where the subject is disappointed let us play that. I am playing a single hand and what looks like a losing game.

The second one is when the speaker sounds excited let us hear this one. I am playing a single hand and what looks like a losing game. By listening to these two samples one can easily tell what is the emotion reflected, right. So, what this means is when we want to generate speech. So, you have a text to speech system which takes into input the text generates the required speech for emotional speech synthesis along with the text input.

We also need to give as an input the emotion class or could be let us say the valence arousal intensities. Now, with these two inputs to your TTS one would then get emotion enhanced speech synthesis. So, this is a very nascent area and we see that there is a lot of work going on to generate emotion enhanced speech.

(Refer Slide Time: 28:09)



Emotion Speech Synthesis

- Emotional Prosody Control for Speech Generation (Sivaprasad 2021)

Now, here is an example by Sivaprasad and others who generate speech with emotional prosody control. So, let us look at the framework the system which is proposed. So, what we have here is first the text input, this is the text against which they are going to generate the speech.

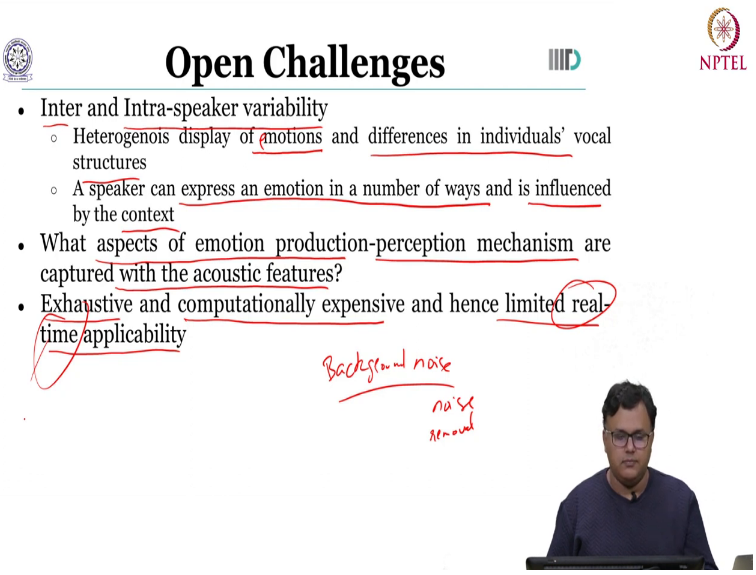
Then we have the second input which is the speaker style. So, this is the speaker reference waveform and third friends is your target emotion the value for arousal and valence. Now, text goes into a phoneme encoder. So, you extract a representation for the required phonemes for the input text statement.

Then the speaker waveform is input into a speaker encoder it extracts the characteristics of the speaker essentially the style with which one speaks. These are concatenated input into an encoder. In parallel we have the AV the arousal and valence vectors which are concatenated and later we are concatenating that with the input from the phoneme and speaker encoder.

There are then duration predictor to predict the duration of the samples of the word which are going to generate. And further this is input into a length regulator where the energy of a required word is predicted, the pitch is predicted and in parallel the length regulator and the output of the energy predictor and predictor are concatenated are added into a decoder.

So, this is your decoder. Now friends this gives you a spectrogram the visualization of your frequencies. And from this we can generate speech which is controlled by the emotional state so, the target emotion. So, what this means is from a bird's eye view the system would require the emotion and a representation for emotion and a series of encoders and decoders.

(Refer Slide Time: 30:37)



Open Challenges

- Inter and Intra-speaker variability
  - Heterogenois display of emotions and differences in individuals' vocal structures
  - A speaker can express an emotion in a number of ways and is influenced by the context
- What aspects of emotion production-perception mechanism are captured with the acoustic features?
- Exhaustive and computationally expensive and hence limited real-time applicability

Background noise

noise
removed

Now, let us look at some of the open challenges in speech-based affect analysis. The first is inter and intra speaker variability. So, we have different speaking styles you add to it. Let us say people coming in from different cultures speaking the same language, ok. Let us say you have an Indians descent person and Asian descent person or Caucasian and all three are speaking English.

So, these are let us say the ethnicities of the subjects in your data set. Now, this is going to lead to a lot of variation even though everyone is speaking the same language because of different difference pronunciation different style in which we speak, right. So, for generalization of affect analysis system which is based on speech we would require to have generic representations.

Now, we will observe that when we are want to have these generic representations they will also have to be agnostic or at least would have analysed and then extracted generic representations for the different display of emotions and the differences in individuals vocal structures, ok. Even though let us say you have an Indian speakers they will have different vocal structures which will lead to the variability in the data which is captured.

Further a speaker can express an emotion in n number of ways and is influenced by the context, ok. Now, the context can again here be where the person is with whom the person is interacting. So, the emotion would be reflected in different ways. Let us say when you say the same statement to a friend and compare that while you are speaking the same statement to let us say an elder, right.

So, the same statement would have a difference in either the intensity of emotion or could have different emotions altogether. This adds to the variability right the content the linguistic remains the same, but the emotion changes. Further the second open challenges what are the aspects of emotion production-perception mechanism which are captured with the acoustic feature?

We have seen different features are proposed for analysis. Some features would extract a particular attribute of your signal and when you are trying to understand trying to perceive the emotion one feature could be better let us say for arousal the other could be better for valence, right. And how do you actually choose the right balance? Maybe fusion that could be one.

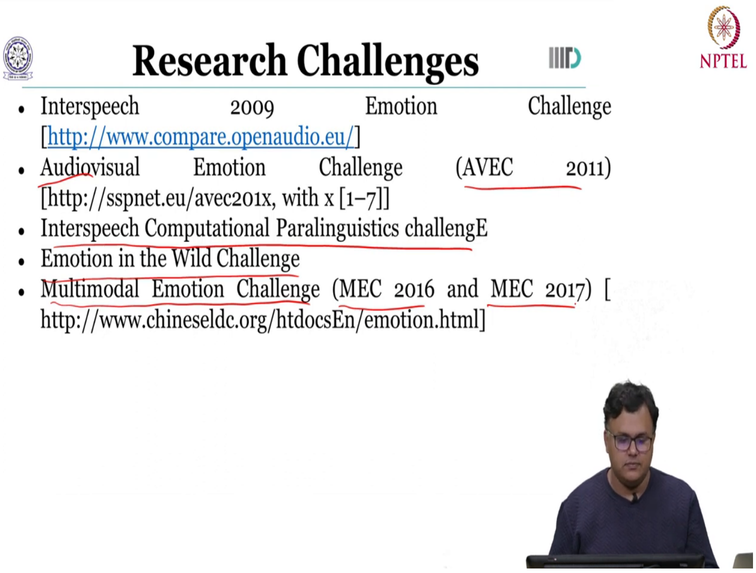
The third open challenge friends is that the speech based affect recognition can be exhaustive and can be computationally expensive. And hence it can have limited real time applicability. We have seen in lecture 1 for speech the trade off right, between the window of sample duration and how much real time it needs to be.

So, if you have a longer duration sample it can be more computational, but that may have far more detailed information which could be required for an accurate prediction. However, if you have a smaller duration sample that could have lesser information, but could be computed closer to real time, right. So, this is an open challenge that if you want the rich information how can we do that in closer to real time.

Now, along with this one more aspect which will come into the picture is, let us say even if you are having a window of audio which is giving you features which are good enough for predicting the emotion of the person in that very time stamp. There could be things such as background noise in that sample let us say background music is there where the subject is. So, we would require a noise removal step as well before feature extraction or it could be that you could have a noise removal step after feature extraction, right.

So, this could also affect the computational aspect and I have it real time.

(Refer Slide Time: 35:33)



Research Challenges

- Interspeech 2009 Emotion Challenge [http://www.compare.openaudio.eu/]
- Audiovisual Emotion Challenge (AVEC 2011) [http://sspnet.eu/avec201x, with x [1–7]]
- Interspeech Computational Paralinguistics challengE
- Emotion in the Wild Challenge
- Multimodal Emotion Challenge (MEC 2016 and MEC 2017) [http://www.chineseldc.org/htdocsEn/emotion.html]

Now, let us look at some of the research challenges friends. The very commonly used research challenge benchmarking platform is the Interspeech 2009 Emotion Challenge. The other one which is very commonly used in the community is your Audiovisual Emotion Challenge which is the AVEC challenge which will have several tasks for emotion, but it is also used for multi-model emotion recognition as well audiovisual.

But the audio here is also very rich. Then we also have the Interspeech Computational Paralinguistic Challenge. Now, this one is actually a very commonly used benchmark in the speech-based affect recognition community and this has been running for years with different tasks related to affect. Then we also have the Emotion in the Wild Challenge. So, motion recognition in the wild.

Here you have audio and video, but audio itself is a combination of the background music, background noise and the speaker's voice. So, that is also used for understanding of affect from audio. So, these are openly available resources which anyone could access based on the different licensing agreements with these resources. And then you can use them for creating evaluating speech-based emotion recognition works.

Now, along with this I will also like to mention other Multimodal Emotion Challenge which is the MEC and MEC 16 and 2017 challenge which is in the Chinese language. So, friends with this we come to the end of the second lecture of speech-based emotion recognition. In this we briefly touched upon the features which are hand-engineered features and have been commonly used in speech analysis the prosody-based features and then your MFCC's.

Later on, we talked about an important step in speech analysis for emotion prediction which is your normalization of the feature. To this end we looked at the iterative feature normalization technique. Then we looked at the different machine learning techniques which are used for the affect prediction and later on we touched upon the concept of speech synthesis where emotion is also induced into the generated speech.

Thank you.