SPEECH RECOGNITION USING MFCC FEATURE

Presentation by :

SUMAN DHAKAL

KATHMANDU UNIVERSITY

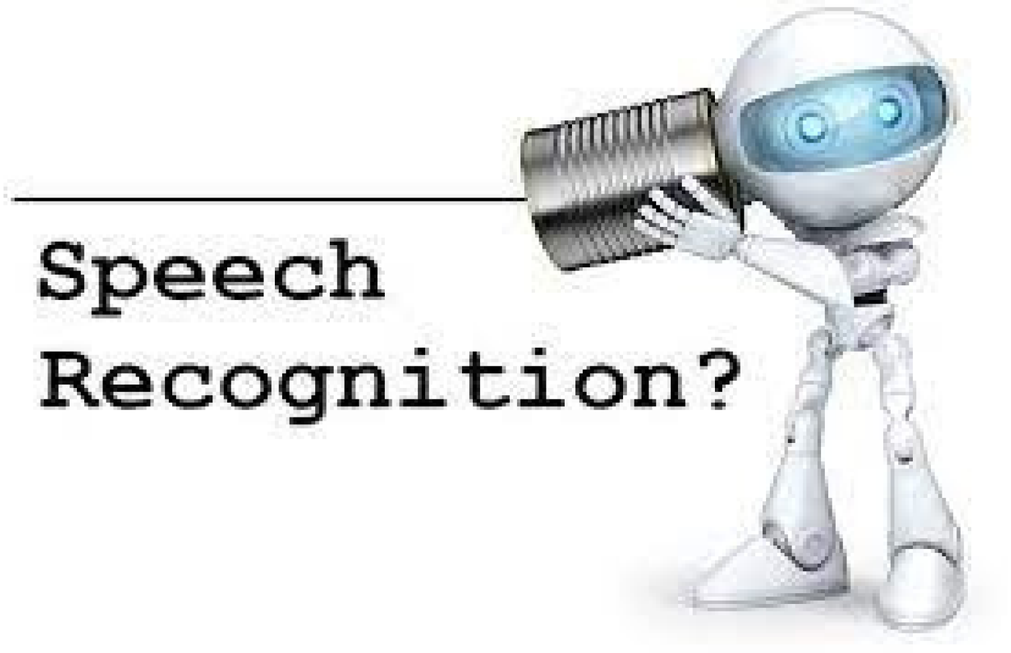
OVERVIEW

- INTRODUCTION
- OBJECTIVE
- MOTIVATION
- LITERATURE REVIEW
- METHODOLOGY
- CONCLUSION
- REFERENCE

INTRODUCTION

- Human-machine interaction is widely used nowadays in many applications. One of the mediums of interaction is speech. The main challenge in human-machine interaction is the detection of emotion from speech. Emotion can play an important role in decision-making.
- Emotion can be detected from different physiological signals also. If emotion can be recognized properly from speech then a system can act accordingly. Emotion identification can be done by extracting the features or different characteristics from the speech and training needed for many speech databases to make the system accurate.
- An emotional speech TESS dataset is selected then emotion-specific features are extracted from those speeches and finally an MLP classification model is used to recognize the emotions.

OBJECTIVES



OBJECTIVE 01

To build a model to recognize emotion from speech using the librosa and sklearn libraries and the TESS dataset.

OBJECTIVE 02

To present a classification model of emotion elicited by speeches based on deep neural networks MLP Classification based on acoustic features such as Mel Frequency Cepstral Coefficient (MFCC). The model has been trained to classify seven different emotions(happy, fearful, disgust, angry, neutral, surprised,sad).

MOTIVATION

- As human beings speech is amongst the most natural way to express ourselves. We depend so much on it that we recognize its importance when resorting to other communication forms like emails and text messages where we often use emojis to express the emotions associated with the messages. As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today's digital world of remote communication.
- Emotion detection is a challenging task, because emotions are subjective. There is no common consensus on how to measure them. We define a Speech Emotions Recognition system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them.

LITERATURE REVIEW

[1] Speech Emotion Recognition using Neural Network and MLP Classifier (Jerry Joy, Aparna Kannan, Shreya Ram, S. Rama)

- **MLP Classifier**
- **5 features extracted- MFCC, Contrast, Mel Spectrograph Frequency, Chroma and Tonnetz**
- **Accuracy 70.28%**

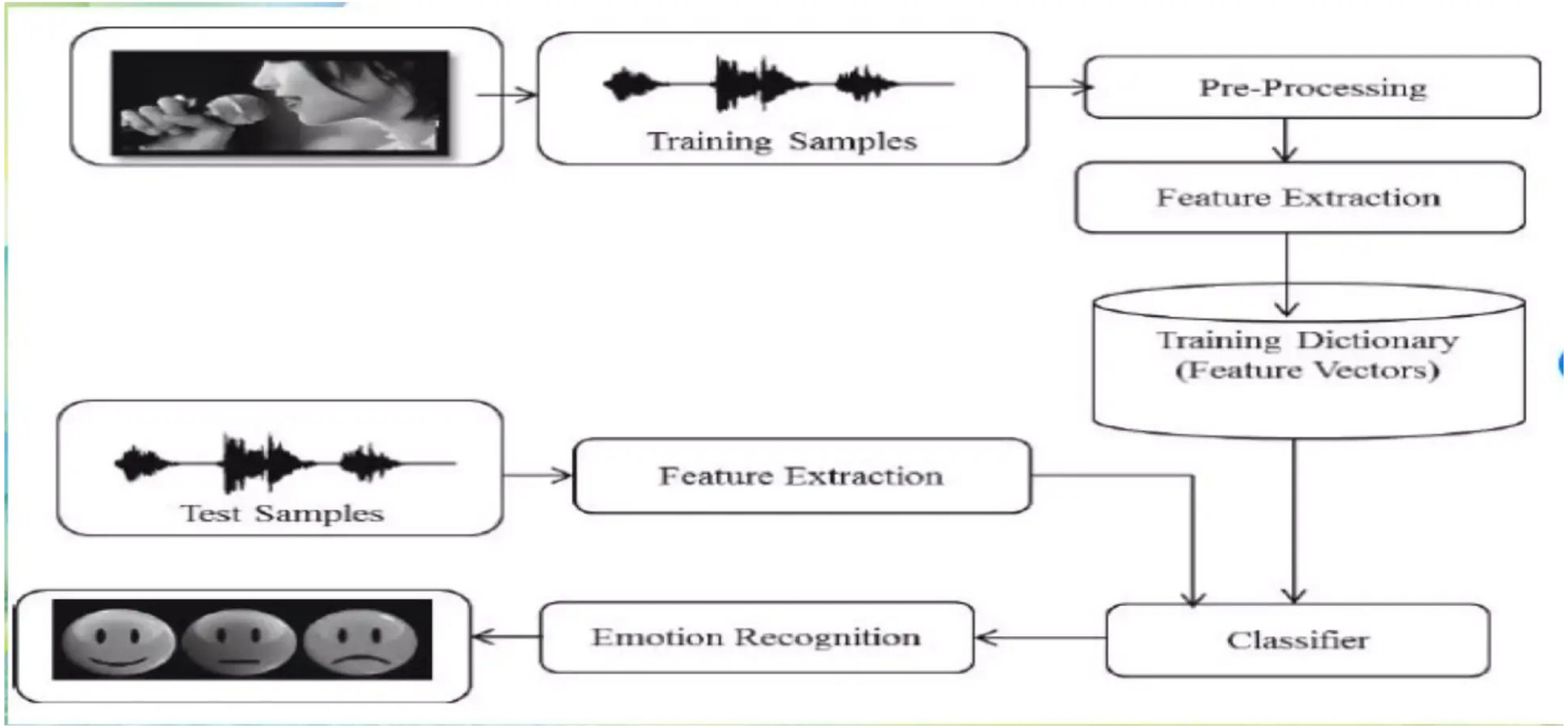
[2] Voice Emotion Recognition using CNN and Decision Tree (Navya Damodar, Vani H Y, Anusuya M A.)

- **Decision Tree, CNN**
- **MFCCS extracted**
- **Accuracy 72% CNN, 63% Decision Tree**

METHODOLOGY

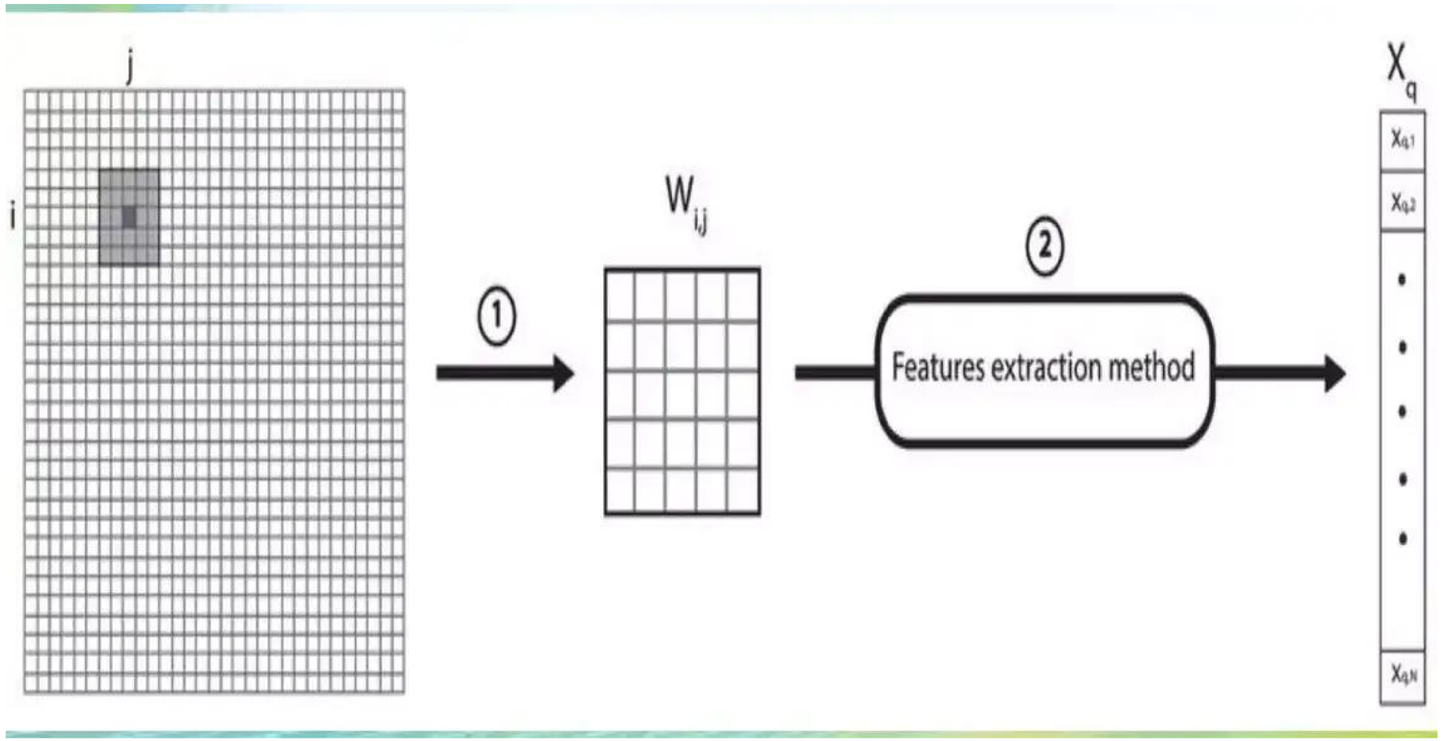
PREPROCESSING → FEATURE EXTRACTION → CLASSIFICATION

SYSTEM FLOW DIAGRAM



FREATURE EXTRACTION

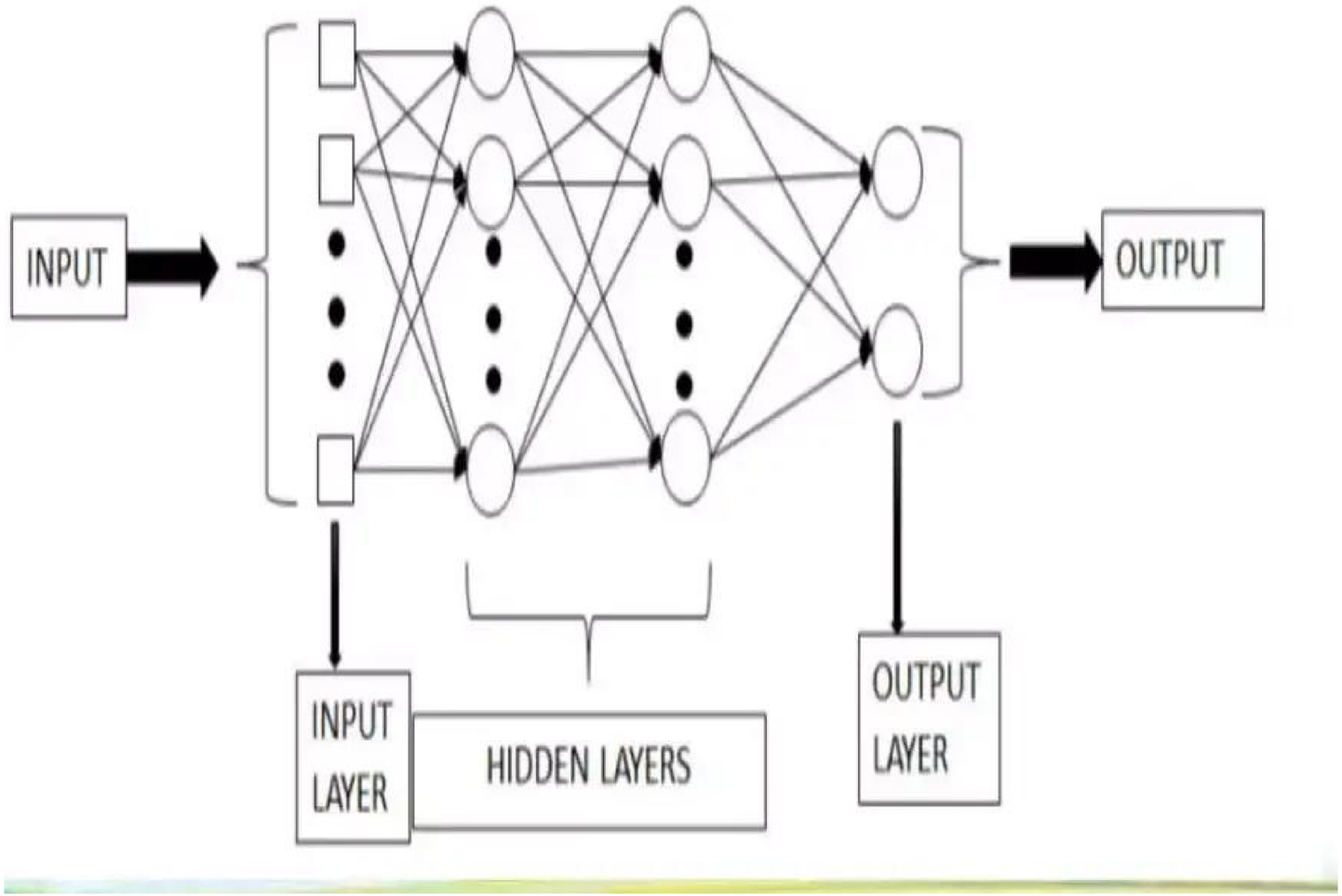
- Extract the feature from audio file
- Used to identify How we speak



- PITCH
- LOUDNESS
- RHYTHM,ETC

CLASSIFICATION

Match the feature with corresponding emotions



Multi-Layer Perceptron Classifier

- A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN).
- MLP consists of at least three layers of nodes- input layer, hidden layer and output layer.
- MLPs are suitable for classification prediction problems where inputs are assigned a class or label.

Building the MLP Classifier involves the following steps-

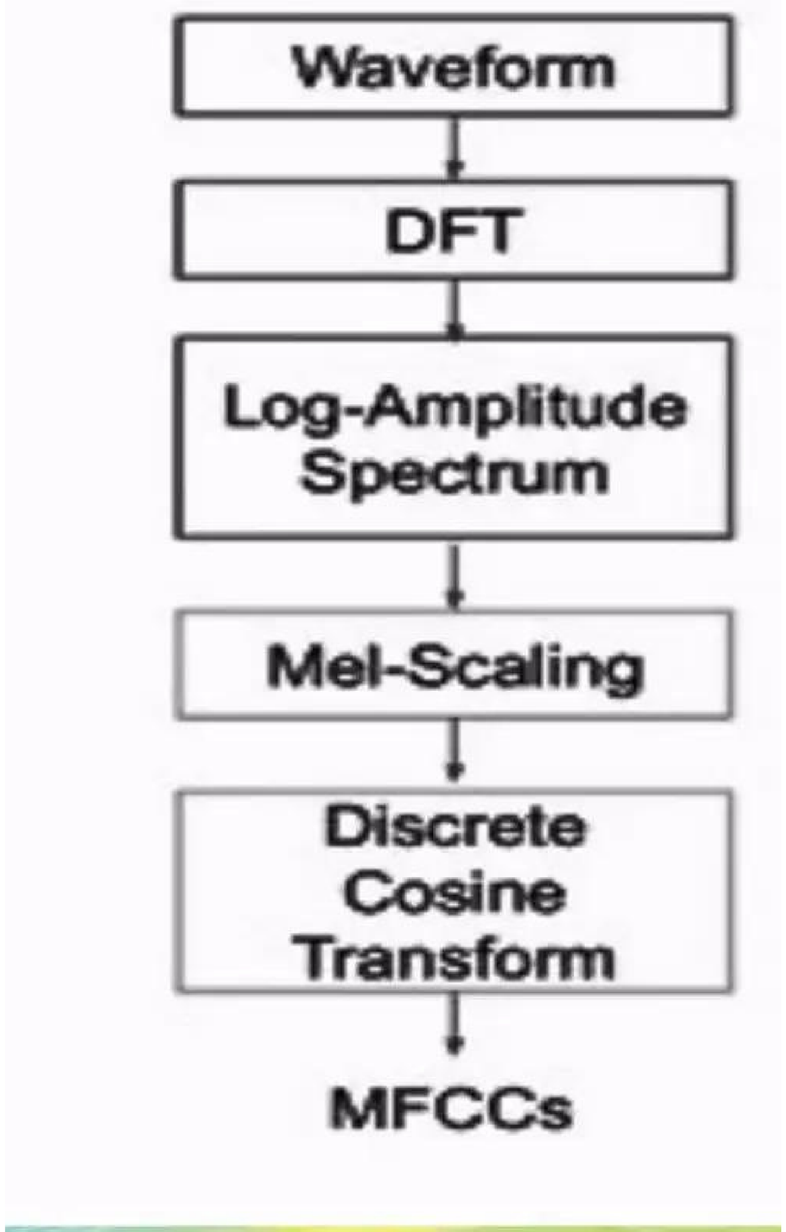
1. Initialisation MLP Classifier.
2. Neural Network.
3. Prediction.
4. Accuracy Calculation.

Feature Extraction

From the Audio data we have extracted a key features which have been used in this, namely:

- MFCC (Mel Frequency Cepstral Coefficients)

MFCC (Mel Frequency Cepstral Coefficients)



ACCURACY

```
3]: print("Accuracy of our model on test data : " , model.evaluate(x_test,y_test)[1]*100 , "%")

epochs = [i for i in range(40)]
fig , ax = plt.subplots(1,2)
train_acc = history.history['accuracy']
train_loss = history.history['loss']
test_acc = history.history['val_accuracy']
test_loss = history.history['val_loss']

fig.set_size_inches(20,6)
ax[0].plot(epochs , train_loss , label = 'Training Loss')
ax[0].plot(epochs , test_loss , label = 'Testing Loss')
ax[0].set_title('Training & Testing Loss')
ax[0].legend()
ax[0].set_xlabel("Epochs")

ax[1].plot(epochs , train_acc , label = 'Training Accuracy')
ax[1].plot(epochs , test_acc , label = 'Testing Accuracy')
ax[1].set_title('Training & Testing Accuracy')
ax[1].legend()
ax[1].set_xlabel("Epochs")
plt.show()
```

```
66/66 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.9585 - loss: 0.2226
Accuracy of our model on test data :  95.4285740852356 %
```

CLASSIFICATION MATRIX

	precision	recall	f1-score	support
angry	0.97	0.97	0.97	308
disgust	0.92	0.95	0.94	291
fear	0.97	0.97	0.97	303
happy	0.96	0.93	0.94	310
neutral	0.96	0.95	0.96	322
sad	0.98	0.99	0.98	279
surprise	0.93	0.92	0.92	287
accuracy			0.95	2100
macro avg	0.95	0.95	0.95	2100
weighted avg	0.95	0.95	0.95	2100

CONCLUSION

- The proposed model achieved an accuracy of 95.42%.
- Angry was the best identified emotion.

REFERENCE

[1] Jerry Joy, Aparna Kannan, Shreya Ram, S. Rama Speech Emotion Recognition using Neural Network and MLP Classifier, IJESC, April 2020.

[2]Navya Damodar, Vani H Y, Anusuya M A. Voice Emotion Recognition using CNN and Decision Tree. International Journal of Innovative Technology and Exploring Engineering (IJITEE), October 2019.

[3]TESS Dataset: https://zenodo.org/record/1188976#.X5r20ogzZPZ

[4]MLP/CNN/RNN Classification: https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/

[5]MFCC:https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

Thank You So Much!