**SIT 789**

**Task 8.2: Speech emotion recognition using spectral features**

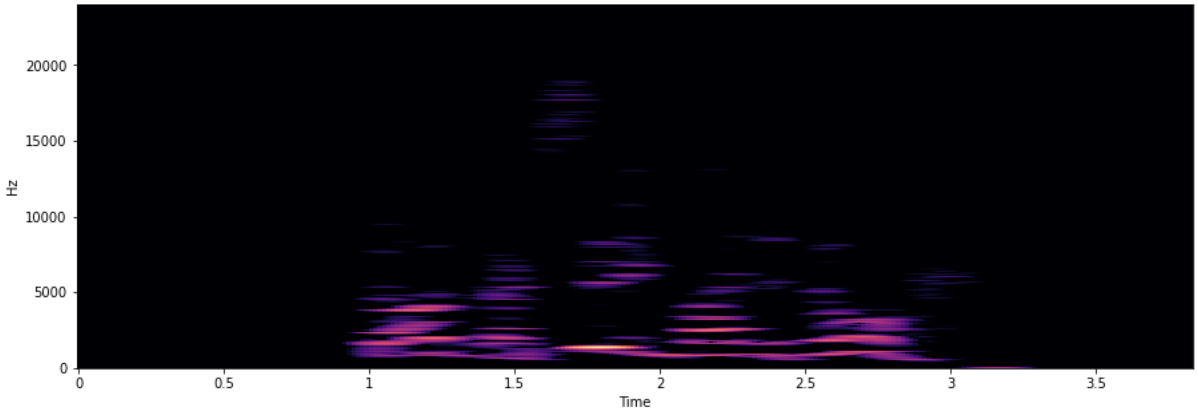
## 2. Speech emotion recognition using spectral features

```
############################################################
Testing spectral feature: sc_mel
SVM accuracy for spectral feature: sc_mel  is:  0.4140625
SVM confusion matrix
 [[18  5  3  6]
 [ 6 16  5  5]
 [ 4  6 13  9]
 [ 6  7 13  6]]
############################################################
Testing spectral feature: sbw_mel
SVM accuracy for spectral feature: sbw_mel  is:  0.4921875
SVM confusion matrix
 [[15  3  7  7]
 [ 3 17  3  9]
 [ 7 11 12  2]
 [ 4  6  3 19]]
############################################################
Testing spectral feature: sbe_mel
SVM accuracy for spectral feature: sbe_mel  is:  0.5859375
SVM confusion matrix
 [[25  5  2  0]
 [ 6 24  1  1]
 [ 8  9 15  0]
 [ 7  5  9 11]]
############################################################
Testing spectral feature: sfm_mel
SVM accuracy for spectral feature: sfm_mel  is:  0.25
SVM confusion matrix
 [[ 0  0  0 32]
 [ 0  0  0 32]
 [ 0  0  0 32]
 [ 0  0  0 32]]
############################################################
Testing spectral feature: re_mel
SVM accuracy for spectral feature: re_mel  is:  0.390625
SVM confusion matrix
 [[10 10 10  2]
 [ 3 20  5  4]
 [ 8  6 13  5]
 [ 9  6 10  7]]
############################################################
Testing spectral feature: se_mel
SVM accuracy for spectral feature: se_mel  is:  0.3671875
SVM confusion matrix
 [[13 10  3  6]
 [ 5 19  3  5]
 [11  7 10  4]
 [ 9 12  6  5]]
############################################################
```

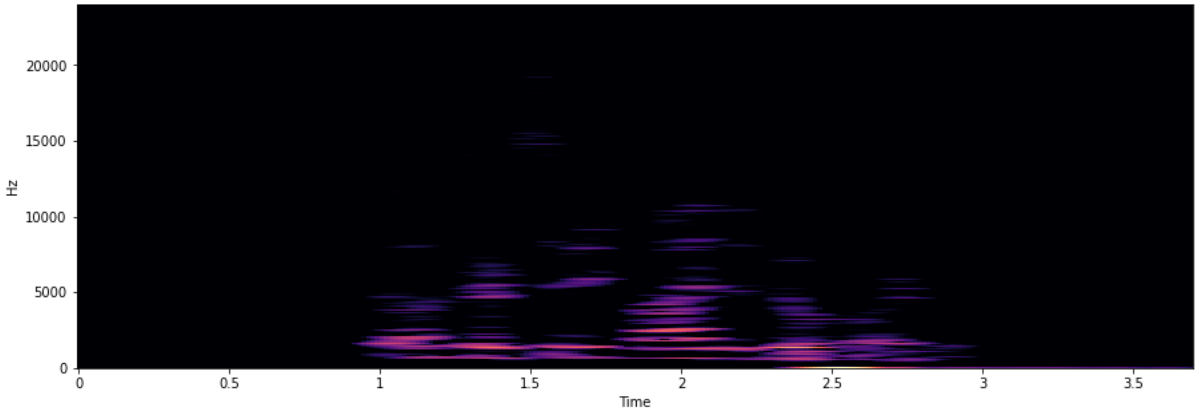
We can see that using Spectral Band Energy (SBE), we get maximum accuracy of 58.5 using SVM with C = 0.1. While the SBE performs best on the given dataset, we can also see that Spectral Flatness Measure (SFM) is the worst performing spectrogram method on this dataset.

**Some examples of mel-scale spectrogram visualisations. (These clips are selected randomly from the training set).**

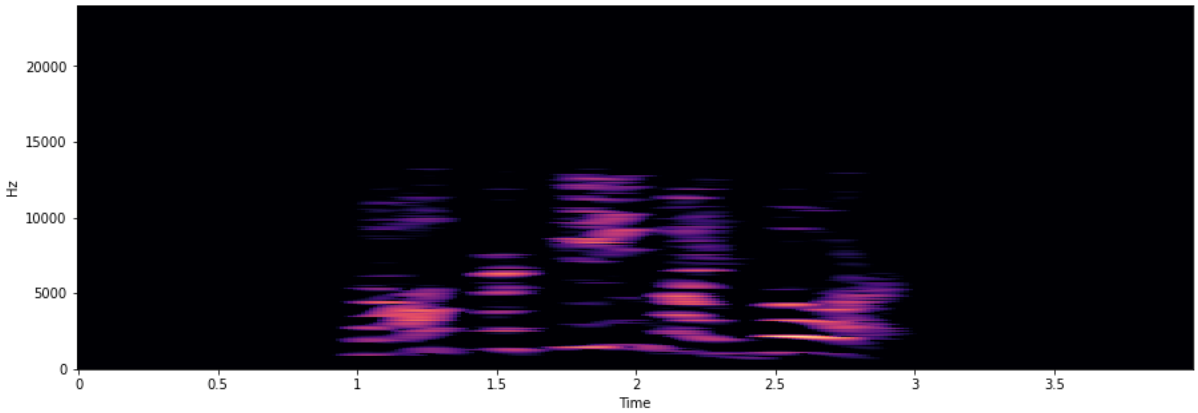
EmotionSpeech/Train/Happy/03-01-03-02-02-02-03.wav



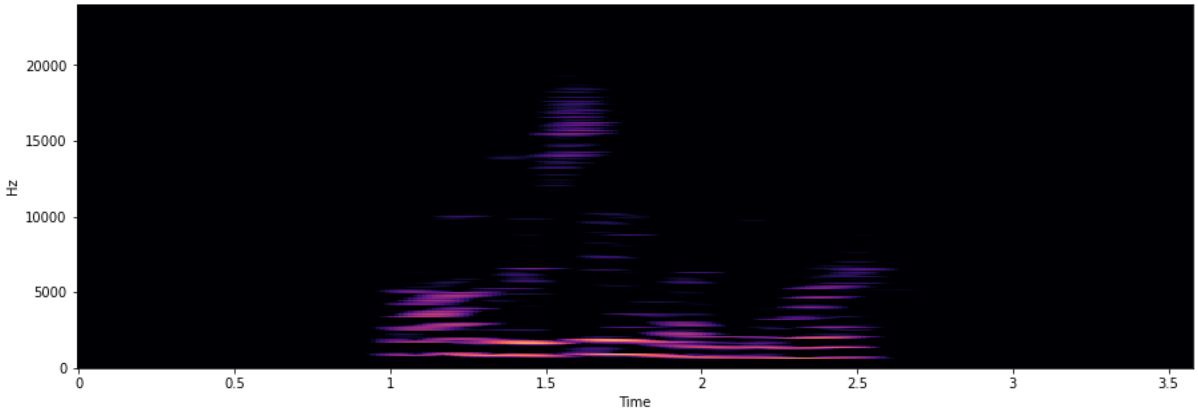

1 : 128 : EmotionSpeech/Train/Sad/03-01-04-02-02-02-03.wav




3 : 128 : EmotionSpeech/Train/Angry/03-01-05-02-02-02-02.wav

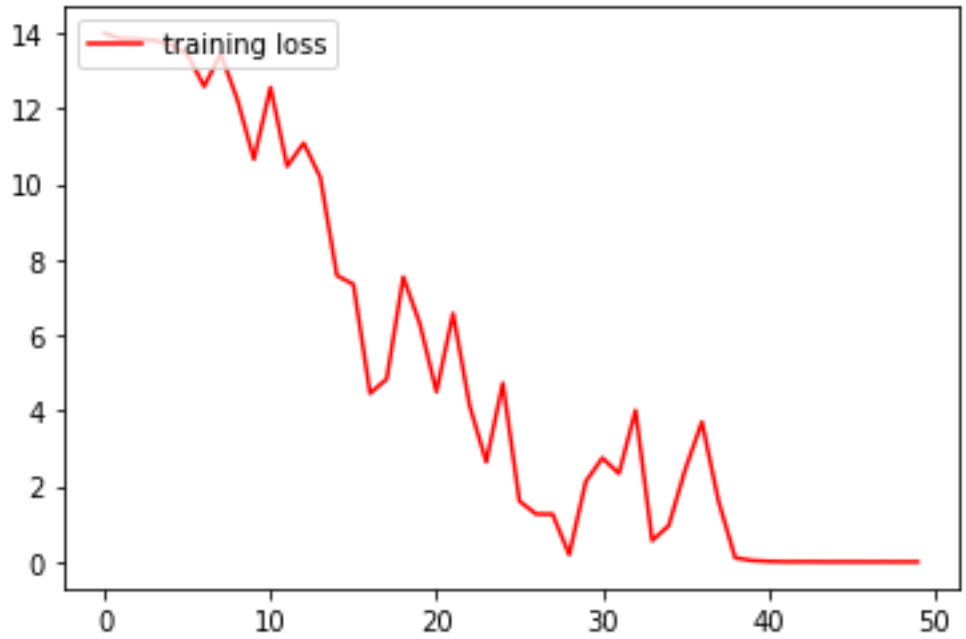


5 : 128 : EmotionSpeech/Train/Calm/03-01-02-01-02-02-04.wav




## 3. Speech emotion recognition using deep learning

## Training loss for the given dataset




## Evaluation metrics using trained model.

Accuracy of Angry : 68 %
Accuracy of  Calm : 56 %
Accuracy of Happy : 37 %
Accuracy of  Sad : 40 %

**Confusion matrix:**

[[22  1  6  3]
[ 4 18  1  9]
[ 4  7 12  9]
[ 7  1 11 13]]

Classification report

```
print(classification_report(groundtruth_labels_entire, predicted_labels_entire))
```

```
              precision    recall  f1-score   support

           0       0.59      0.69      0.64        32
           1       0.67      0.56      0.61        32
           2       0.40      0.38      0.39        32
           3       0.38      0.41      0.39        32

    accuracy                           0.51       128
   macro avg       0.51      0.51      0.51       128
weighted avg       0.51      0.51      0.51       128
```

We can see that the deep learning architecture performs well and if we train deeper networks then we can achieve higher accuracy.

**Do you think that we should apply this technique in this task to improve the performance of emotion recognition?**

While some data augmentation techniques might make the network more robust, it is essential that data augmentation is done with care as some data augmentation techniques might change the spectrogram which might lead to poor performance. Thus, only the augmentation techniques which do not alter the shape of spectrogram should be used.