Speech Emotion Recognition

Why Speech to Emotion directly?

- The main problem with traditional systems is that the **errors** occurred while **transcribing audio** is **propagated** and that affects the emotion recognition task.

- We **lose** important **acoustic features** like pitch, loudness etc.

- **Confusing emotional contexts** in some cases. For example, the phrase, ‘What a man!’ can indicate surprise or even disgust.

Motivation

- A human lacking essential brain function or having a malfunction on the neural subsystem responsible for handling emotions cannot efficiently perform decision-making.

- Computers are no longer just logical computing machines. With the advent of human machine interaction, apart of ‘**what is said**’ and ‘**who said it**’, ‘**how it is said**’ plays a key role for effective human-machine communication and for the machine to react properly.

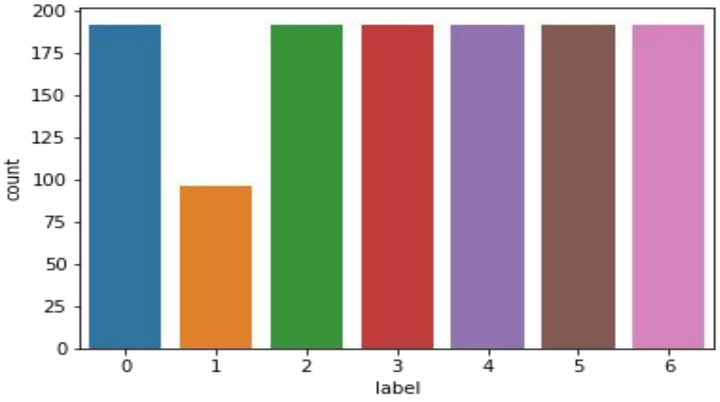
- Other applications include the **psychiatric diagnosis**, **intelligent toys** and **lie detection** in call centers as well as **evaluation of mental state** of the driver for the start of his/her day.

Datasets Used

- Name: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)
- For our project, only the **audio tracks** have been used.

- Description about the audio dataset:
  - Gender balanced consisting of **24** professional actors.
  - 7 emotions in total: **calm, happy, sad, angry,fearful, surprise, and disgust expressions.**
  - Each expression is produced at two levels of emotional intensity, with an additional neutral expression.
  - **1250** samples in total.

Count Plot(emotion vs count)

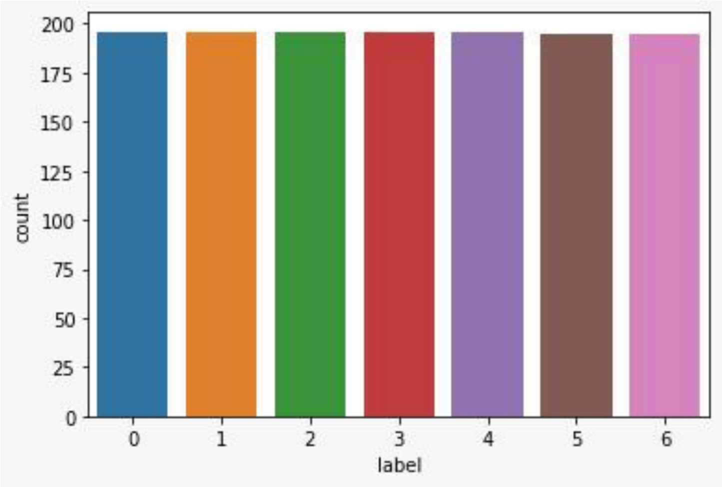


Datasets Used

- Name: The Toronto Emotional Speech Set (TESS)

- Description about the audio dataset:
  - Gender balanced consisting of **2** professional actresses (aged 24 and 64 years).
  - 7 emotions in total: **anger, disgust, fear, happiness, surprise, sadness, neutral**.
  - Audiometric testing indicated that both actresses have thresholds within the normal range.
  - **1370** samples in total.

Count Plot(emotion vs count)



MFCC Feature Extraction

- The signal has a sampling frequency of **16KHz.**
- Used Scipy function to load the audio file and extract the features.

Steps:

- Apply **pre-emphasis** on the signal to amplify the high frequencies.
  - Help in **balancing the magnitude**
  - **Improve signal-to-noise(SNR) ratio**.
- After pre-emphasis, split the signal into short time frames and apply a **Fourier transform over the short time frame**.
  - Frame size: **25ms**, stride: **10ms**
- After slicing the signal into frames, we apply a window function(Window function) to each frame.

MFCC Feature Extraction (Contd.)

Steps:

- We apply **N-point FFT**(Fast Fourier Transform) on each frame to calculate the **frequency spectrum** and then compute the **power spectrum**.
- To computer Filter banks, **Triangular filters** are applied, typically **40** filters, to extract the frequency bands.
- Apply **Discrete Cosine Transform**(DCT)
  - To decorrelate the filter bank coefficients
  - Yield a compressed representation of the filter banks .
- The resulting coefficients are the 40 dimensional MFCC features.

Data Augmentations

- **Noise Injection:** It simply adds some random value or random background music into data.

- **Changing pitch:** Pitch is the fundamental period of the speech signal. We used **librosa** library which changes the pitch randomly.

- **Adding Reverb:** It is the reflection of sound waves created by the superposition of echoes. Used **pysndfx** library for this.

- **Changing Speed:** Used **cv2 resize** function which slightly increases or decreases the audio speed.

Model used

- We split the dataset in the ratio **80:20** for train and test sets respectively.
- We trained two **Random Forest Classifiers** with **100** estimators each.
  - One was trained on the **raw dataset** examples. The other was trained on **augmented** samples.
  - Each of the augmentations was applied on the original signal itself, one at a time.
- After augmenting the data, the size of the dataset increased from **1250** samples to **6250** samples.

Results on RAVDESS

- The confusion matrices, the classification reports and accuracies for the **test sets**(using Random Forest with and without augmentation)

Raw Data

```
Raw test data
Classification Report
              precision    recall  f1-score   support

           0       0.47      0.66      0.55        32
           1       0.67      0.43      0.52        14
           2       0.84      0.75      0.79        48
           3       0.57      0.51      0.54        41
           4       0.48      0.67      0.56        33
           5       0.76      0.66      0.70        38
           6       0.59      0.50      0.54        44

    accuracy                           0.61       250
   macro avg       0.62      0.60      0.60       250
weighted avg       0.63      0.61      0.62       250

Confusion Report
[[21  1  0  3  1  4  2]
 [ 2  6  2  0  4  0  0]
 [ 2  0 36  3  4  1  2]
 [ 7  1  1 21  6  1  4]
 [ 3  1  1  4 22  0  2]
 [ 3  0  1  4  0 25  5]
 [ 7  0  2  2  9  2 22]]


Test accuracy: 0.612
```

Augmented Data

```
Augmented test data
Classification Report
              precision    recall  f1-score   support

           0       0.64      0.73      0.69       187
           1       0.91      0.53      0.67       109
           2       0.77      0.81      0.79       193
           3       0.71      0.71      0.71       198
           4       0.67      0.75      0.71       186
           5       0.77      0.82      0.79       180
           6       0.82      0.71      0.76       195

    accuracy                           0.74      1248
   macro avg       0.76      0.72      0.73      1248
weighted avg       0.75      0.74      0.73      1248

Confusion Report
[[137   0   6   8  20   7   9]
 [ 14  58  11   9  14   2   1]
 [  7   1 157  10   4  12   2]
 [ 13   1   8 141  13  18   4]
 [ 18   3   5   6 140   2  12]
 [  9   0   8  10   3 147   3]
 [ 15   1  10  14  14   3 138]]


Test accuracy: 0.7355769230769231
```

Results on TESS

- The confusion matrices, the classification reports and accuracies for the **test sets**(using Random Forest with and without augmentation)

Raw Data

```
Raw test data
Classification Report
              precision    recall  f1-score   support

           0       0.97      0.95      0.96        37
           1       0.97      1.00      0.99        39
           2       0.96      0.90      0.93        52
           3       0.87      0.97      0.92        35
           4       1.00      0.97      0.99        38
           5       1.00      1.00      1.00        28
           6       1.00      1.00      1.00        45

    accuracy                           0.97       274
   macro avg       0.97      0.97      0.97       274
weighted avg       0.97      0.97      0.97       274

Confusion Report
[[35  1  1  0  0  0  0]
 [ 0 39  0  0  0  0  0]
 [ 0  0 47  5  0  0  0]
 [ 0  0  1 34  0  0  0]
 [ 1  0  0  0 37  0  0]
 [ 0  0  0  0  0 28  0]
 [ 0  0  0  0  0  0 45]]


Test accuracy: 0.9671532846715328
```

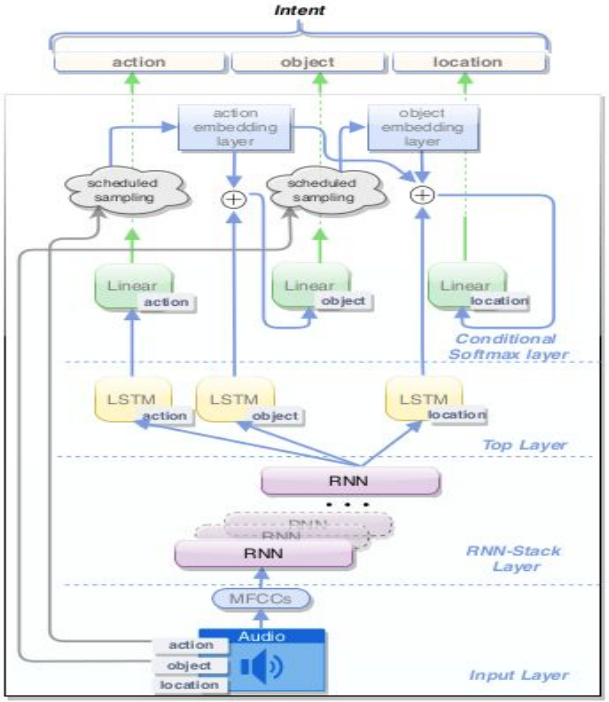
Augmented Data

```
Augmented test data
Classification Report
              precision    recall  f1-score   support

           0       0.98      0.99      0.98       194
           1       1.00      0.98      0.99       189
           2       0.97      0.97      0.97       208
           3       0.97      0.97      0.97       190
           4       0.99      1.00      0.99       191
           5       0.99      0.99      0.99       196
           6       1.00      0.99      0.99       202

    accuracy                           0.99      1370
   macro avg       0.99      0.99      0.99      1370
weighted avg       0.99      0.99      0.99      1370

Confusion Report
[[192   0   2   0   0   0   0]
 [  1 186   0   0   2   0   0]
 [  2   0 202   4   0   0   0]
 [  0   0   4 185   0   1   0]
 [  0   0   0   0 191   0   0]
 [  0   0   0   0   0 195   1]
 [  1   0   0   1   0   0 200]]


Test accuracy: 0.9861313868613139
```

Emotion Recognition using Bidirectional LSTM[3]

(**Ref**. E. Palogiannidi, I. Gkinis, G. Mastrapas, P. Mizera and T. Stafylakis, "End-to-End Architectures for ASR-Free Spoken Language Understanding," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7974-7978, doi: 10.1109/ICASSP40776.2020.9054314.)

- Uses 40 dimensional **MFCC features** extracted every **10ms**.
- The model performs very well on **Spoken Language Understanding** task, achieving an accuracy of **98.85%**.
- The model comprises of three parts:
  - RNN stack
  - Representation layer
  - Classifier layer



References

1. Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

2. M. A. Hossan, S. Memon and M. A. Gregory, "A novel approach for MFCC feature extraction," 2010 4th International Conference on Signal Processing and Communication Systems, Gold Coast, QLD, 2010, pp. 1-5, doi: 10.1109/ICSPCS.2010.5709752.

3. E. Palogiannidi, I. Gkinis, G. Mastrapas, P. Mizera and T. Stafylakis, "End-to-End Architectures for ASR-Free Spoken Language Understanding," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7974-7978, doi: 10.1109/ICASSP40776.2020.9054314.

Thank You