Speech Emotion Recognition Using Deep Learning on audio recordings

S.Suganya and **E.Y.A.Charles**
suganyasuven@gmail.com
charles.ey@univ.jfn.ac.lk

Abstract

Speech emotion recognition plays a prominent role in human-centred computing. However, It is still unclear that, which features of a human speech are robust enough to distinguish emotions. This work proposes an end-to-end deep learning approach which applies deep neural network on a raw audio recording directly. Proposed model was assessed on USC-IEMOCAP and EmoDB and obtained accuracy of 68.6% for IEMOCAP and 85.62% for EmoDB.

Objective

To propose a method that applies deep neural network to raw waves directly to perform emotion recognition.

Proposed Approach

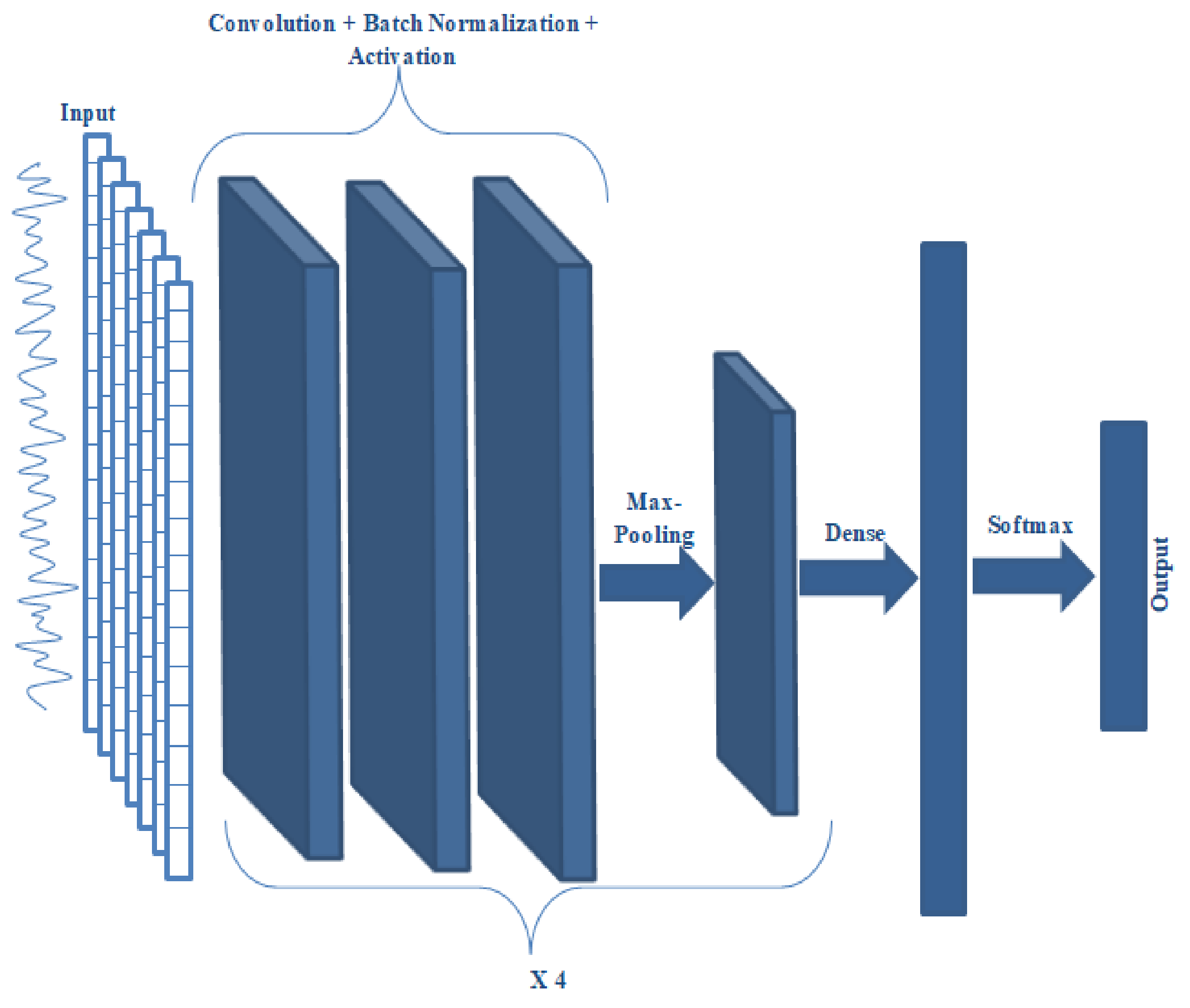

Figure 2 Flow diagram of Proposed method

- In most of the studies, MFCC features and spectrograms are used for the experiments. However it is still challenging to choose an optimal feature set for this task and time consuming procedure.
- Aims to develop a deep neural network which takes raw waveforms that represented as a long vector of values as input, instead of handcrafted features or spectrograms.
- Speech recordings in both IEMOCAP and EmoDB datasets were sampled at 16 kHz for this study.
- Proposed model contains seven convolutional layers, one fully connected layer and a softmax layer.

CNN model

M9
Input : 96000 x 1 time-domain waveform
[80/4, 64]
Max_pooling : 4 x 1 (output : 6000 x 64x n)
[3/1, 128] x 2
Max_pooling : 4 x 1 (output : 1500x 64x n)
[3/1, 256] x2
Max_pooling : 4 x 1 (output : 375x 128 x n)
[3/1, 512] x 2
Global average pooling (output : (1 x 512 x n)
FC(1024)
Softmax

([80/4, 64] denotes a convolutional layer with 64 filters and kernel size 80 with stride 4

Figure 2 Proposed model

Dataset

To evaluate our methodology, the Berlin Database of Emotional Speech (EmoDB) database [1] and the Interactive Emotional Motion Capture dataset (IEMOCAP) [2] published by the University of Southern California, are used to train and evaluate the proposed CNN model. The following figures illustrate the distribution of classes in the the datasets.

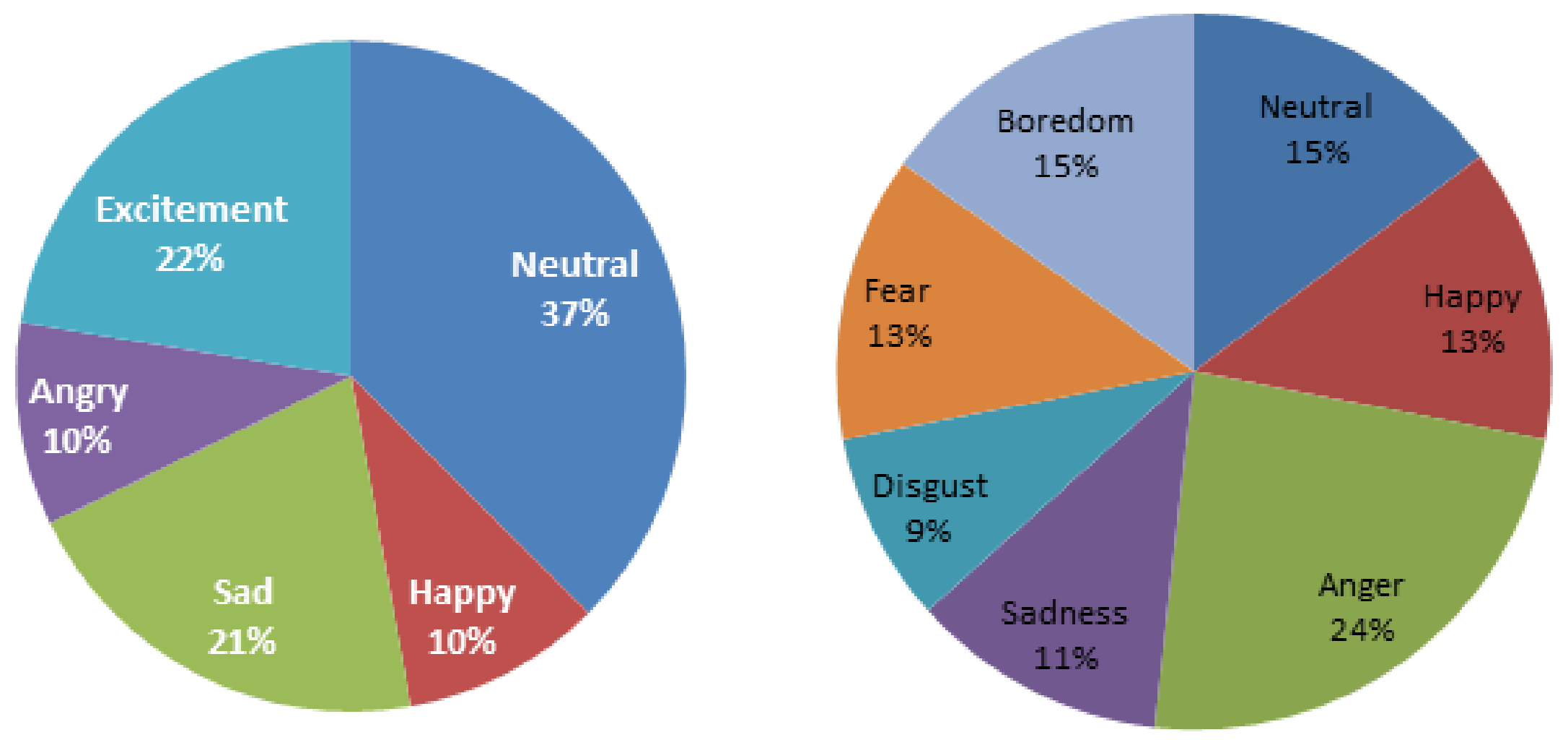

Figure 3 Distributions of datasets

Experiments

- For IEMOCAP dataset, experiments conducted for different sets of emotion classes,
  - [Anger, Happiness, Sadness, Neutral]
  - [Anger, Excitement, Sadness, Neutral]
  - [Anger, Sadness, Neutral]
- For EmoDB database, experiments conducted over 7 emotions.
  - [Anger, Happiness, Sadness, Neutral, Fear, Disgust and Boredom]

Experimental Results

- IEMOCAP
  - Emotion Classes
    - [Anger, Happiness, Sadness, Neutral] – 68.6%
    - [Anger, Excitement, Sadness, Neutral] – 64.3%
    - [Anger, Sadness, Neutral] – 79.3%
- EmoDB
  - [Anger, Happiness, Sadness, Neutral , Fear, boredom, Disgust] - 85.62%

Confusion Matrices

- *For EmoDB*

Class Labels	Neutral	Happiness	Sadness	Anger	Disgust	Fear	Boredom
Neutral	91.7	0	4.2	4.2	0	0	0
Happiness	4.8	47.6	0	42.9	0	4.8	0
Sadness	0	0	100	0	0	0	0
Anger	2.6	5.3	0	92.1	0	0	0
Disgust	0	7.1	0	7.1	78.6	0	7.1
Fear	5.3	5.3	0	0	0	89.5	0
Boredom	4.0	0	0	0	0	4.0	92.0

- *For IEMOCAP*
  - [Anger, Happiness, Sadness, Neutral]

Class Labels	Anger	Happiness	Neutral	Sadness
Anger	59.2	3.1	36.0	1.7
Happiness	11.2	14.3	69.2	5.2
Neutral	4.7	4.1	79.9	11.3
Sadness	1.8	1.6	20.2	76.4

  - [Anger, Excitement, Sadness, Neutral]

Class Labels	Anger	Excitement	Neutral	Sadness
Anger	40.1	24.9	34.3	0.7
Excitement	11.2	45.8	39.8	3.2
Neutral	2.4	11.0	75.8	10.8
Sadness	0.9	1.6	22.8	74.7

  - [Anger, Sadness, Neutral]

Class Labels	Anger	Neutral	Sadness
Anger	62.6	35.3	2.1
Neutral	3.7	82.2	14.1
Sadness	1.8	17.1	81.1

Analysis

- In IEMOCAP,
  - Neutral and sadness classes shows high true positive.
  - Happiness and anger are more classified as Neutral emotions.
  - According to the results, it can be observed that the correlation between anger, happiness and neutral are less compared to Happiness and Excitement emotion classes.
- In EmoDB
  - All emotions except happiness showed high class accuracy.
  - For emotion class sadness, the model achieved 100% accuracy and happiness was heavily confused with anger emotion.

Discussion & Conclusion

- Usage of Mel-scale spectrograms on a deep CNN and combinations of CNN and LSTM achieved a recognition rate in between 62 – 70% .
- In a recent study, phoneme features is combined with the spectrogram features achieved a accuracy of 73.9%.
- Thus, it can be noticed that achieved results are close to the accuracy of CNNs on spectrogram.
- Since the computation of spectrogram is costly and time consuming task, it can be concluded that the proposed approach is highly feasible for the emotion recognition task.

Reference

[1]. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of German emotional speech,” *INTERSPEECH*, pp.1517–1520, 2005.

[2]. C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[3]. G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, S. Zafeiriou, and B. Schuller, “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,” *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204, 2015.

[4]. W. Dai, C. Dai, S. Qu, J. Li, and S. Das, “Very deep convolutional neural networks for raw waveforms,” *2017 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, pp. 421–425, 2017.

[5]. M. B. Mustafa, A. M. Yusoof, Z. M. Don, and M. Malekzadeh, “Speech emotion recognition research: An analysis of research focus,” *International Journal of Speech Technology*, vol. 21, pp. 137–156, 2018.