EE583 Pattern Recognition Project
Inspection of *Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition*

Kutay Uğurlu 2232841

February 4, 2022

1 Introduction

Speech emotion recognition is the task of extracting the emotion category from either text or audio data. It has already some applications in security, medicine, entertainment and education [1].
This project report investigates the idea that Atmaja *et al.*stated in *Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition* [2]. The authors of the study discusses the need of utilizing modern computation units, such as Long Short Term Memory(LSTMs) and Convolutional Neural Networks(CNNs), in the neural networks for the task of dimensional speech emotion recognition.
The organization of the report is as follows:

- The problem and the proposed solutions to it are briefly introduced in this section.
- The theory of dimensional and categorical emotions are mentioned in Section 2.
- The methods, datasets and neural network architectures used in the implementation of the speech emotion recognition can be seen in Section 3.
- Section 4 is left for the definition of the evaluation metrics utilized in the study, the reproduced results, and evaluation of the models with other datasets.

2 Theory

The idea that emotion categories can be separated based on some emotional dimensions dates back to 1979. Russel argued in [3] that categorical emotions, namely sadness, anger, joy, etc., can be classified by the values they represent in three dimensions: Valence, Arousal and Dominance.

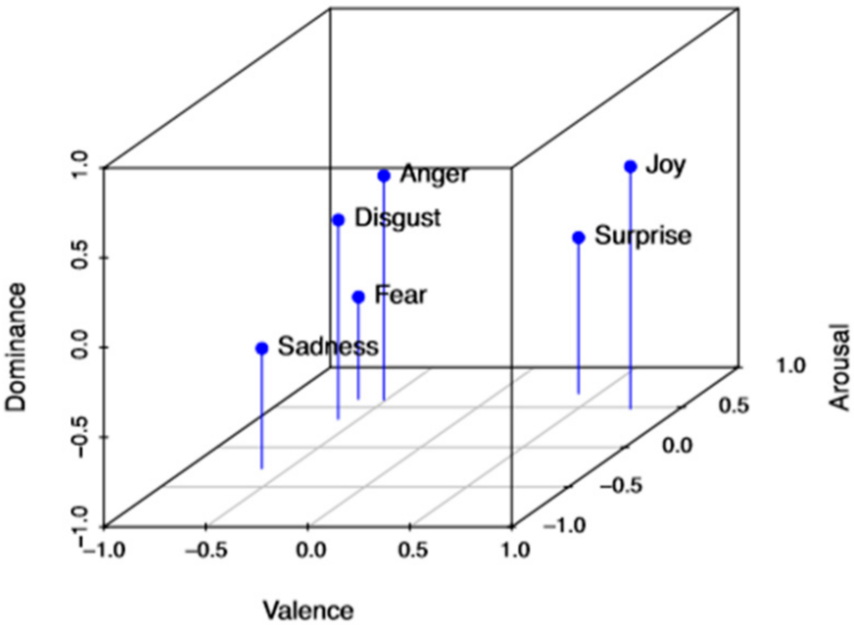

Figure 1: Categorical Emotions in VAD space [4]

Table 1: VAD Dimensions of 6 basic emotions [4].

	Valence	Arousal	Dominance
Anger	-0.43	0.67	0.34
Joy	0.76	0.48	0.35
Surprise	0.4	0.67	-0.13
Disgust	-0.6	0.35	0.11
Fear	-0.64	0.6	-0.43
Sadness	-0.63	0.27	-0.33

3 Implementation

3.1 Data

3.1.1 Datasets

There are mainly two datasets utilized in the paper.

1. **IEMOCAP(The Interactive emotional Dyadic Motion Capture database):** 12 hours of speech data consisting of 10039 utterances is used [5].

2. **MSP-IMPROV:** 18 hours of speech data consisting of 8438 utterances is used [6].

3.1.2 Data preprocessing

The shared data in [2] are already preprocessed and the scripts and tools the authors utilized was not explicitly shared. However, the types of features Atmaja *et al.* used is shared in the paper. The audio in the dataset are used to extract 47 features per utterance. These features are obtained in a 2-level process. First, the Low Level Descriptors defined in [7] are calculated by the `opensmile` software. These Low Level Descriptors are as follows:

- Intensity
- Alpha ratio
- Hammarberg index
- Spectral slope 0-500 Hz
- Spectral slope 500-1500 Hz
- Spectral flux
- 4 MFCCs
- { = 0
- jitter
- shimmer
- Harmonics-to-Noise Ratio (HNR)
- Harmonic difference H1-H2
- Harmonic difference H1-A3
- F1
- F1 bandwidth
- F1 amplitude
- F2
- F2 amplitude
- F3
- F3 amplitude

Then, 47 features, High Statistical Functions of these 23 features are calculated in two sets by utilizing the mean and standard deviation. In addition, authors defined an extra feature: Silence. The silence is defined as the ratio of the silent frames per utterance.

$$S = \frac{N_s}{N_t} \tag{1}$$

where N_s is the number of silent frames and N_t is the number of total frames. The framed are labelled as silent by being compared to a threshold.

$$Threshold = 0.3 \times X_{RMS} \tag{2}$$

and

$$X_{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x[i]^2} \tag{3}$$

where the factor 0.3 is determined empirically.

3.2 Models

The multiplayer perceptron models are implemented with Scikit-learn's MLP Regressor [8] and the remaining CNN and LSTM models are implemented with Tensorflow [9].

4 Results and Discussion

References

[1] L. Cen, F. Wu, Z. L. Yu, and F. Hu, "A real-time speech emotion recognition system and its application in online learning", p. 27–46, 2016.

[2] B. T. Atmaja and M. Akagi, "Deep multilayer perceptrons for dimensional speech emotion recognition", *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, p. 325–331, 2020.

[3] J. A. Russell, "Affective space is bipolar.", *Journal of personality and social psychology*, Vol. 37, No. 3, p. 345, 1979.

[4] O. Blan, G. Moise, L. Petrescu, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Emotion classification based on biophysical signals and machine learning techniques", *Symmetry*, Vol. 12, No. 1, p. 21, 2020.

[5] C. Busso, M. Bulut, C.-C. Lee, "Iemocap: Interactive emotional dyadic motion capture database", *Language resources and evaluation*, Vol. 42, No. 4, p. 335–359, 2008.

[6] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception", *IEEE Transactions on Affective Computing*, Vol. 8, No. 1, p. 67–80, 2016.

[7] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor", in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, p. 1459–1462.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, Vol. 12, p. 2825–2830, 2011.

[9] Martín Abadi, Ashish Agarwal, Paul Barham, *TensorFlow: Large-scale machine learning on heterogeneous systems*, Software available from tensorflow.org, 2015. [Online]. Available: `https://www.tensorflow.org/`.