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

Kutay Uğurlu 2232841

February 5, 2022

Contents

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, evaluation of the models with other datasets and the discussion on the results.

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.

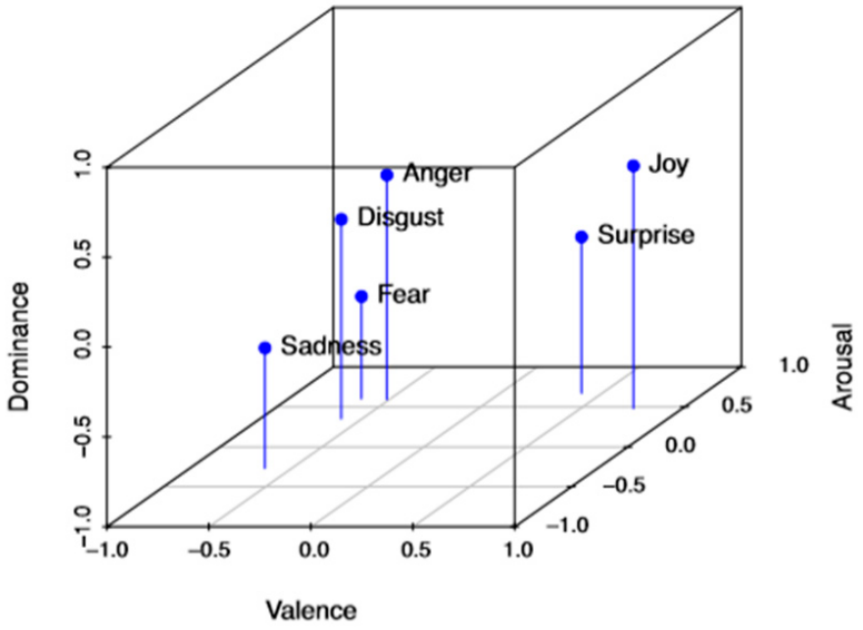


Figure 1: Categorical Emotions in VAD space [4]

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

This section composes of two sections describing the implementation of the original authors and slight changes made.

3.1 Original Paper

3.1.1 Data

3.1.1.1 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].

3.1.1.2 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
- f_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 frames are labelled as silent by being compared to a threshold.

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

and

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

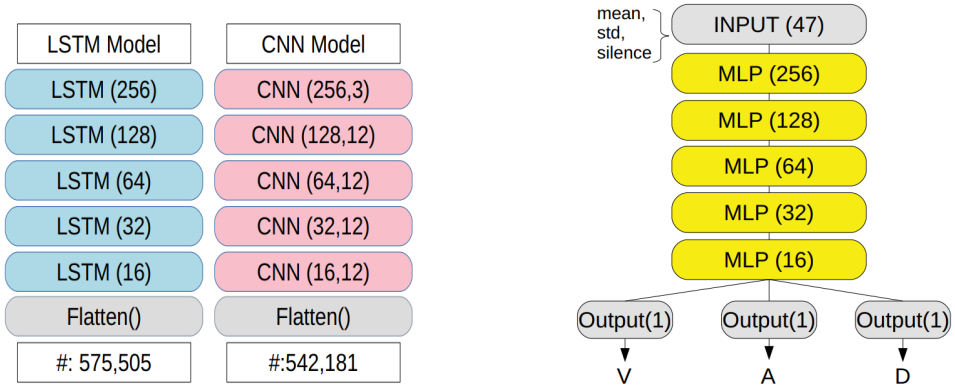
where the factor 0.3 is determined empirically.

3.1.1.3 Scaling

Both the data and the labels that were originally labelled in the range $[-5, 5]$ are scaled to the range [-1,1] in ReLU activated CNNs and LSTM networks.

3.1.2 Models

The benchmark models which authors use to compare the model performance are composed of CNN and LSTM units. The authors use the same number of units and layers in the model. In Figure 2, it can be seen that the models have approximately the same number of trainable parameters.
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].



(a) Benchmark model architectures (b) Proposed Multilayer MLP architecture

Figure 2: Model architectures utilized in [2]

3.1.2.1 Loss Function: CCC Loss

To optimize the network, Concordance Correlation Coefficient in Eqn. 4 is used.

$$CCC = \frac{2\rho\sigma_x\sigma_y}{{\sigma_x}^2 + {\sigma_y}^2 + (\mu_x - \mu_y)^2} \tag{4}$$

Then, the loss can be found as $CCCL = 1 - CCC$.
The total loss with respect to this loss is defined to be the sum of loss of individual classes follows:

$$CCCL_T = CCCL_V + CCCL_A + CCCL_D \tag{5}$$

3.2 Changes made

3.2.1 Data

3.2.1.1 Datasets

The authors of the study made use of the most of the databases whose utterances are labelled by multiple annotators in the Valence-Arousal-Dominance dimension. Although there are many available datasets on Speech Emotion Recognition, only a few of them are labelled in the same manner. However, these datasets were not available for open access. The remaining available datasets are either labelled categorically, *i.e.* sad, joy, anger etc. or labelled only in 2 dimension: Valence and Arousal. Since the latter option requires one to re-train the model and does not provide one with the chance of evaluating the pretrained model with the new test data, categorical dataset, MELD [2], was used for the evaluation. The dataset consists of utterances in mp4 format from the TV series "Friends" and the categorical emotions regarding these utterances. The number of labeled utterance is 2610. However, the number of utterances labelled as "neutral" in the emotions, which corresponds to the origin of the Valence-Arousal-Dominance space, was exceedingly higher than the others, and it was causing problems in both class-imbalance sense and in the learning part of the regression networks, hence those samples were dropped from the dataset.

3.2.1.2 Data Preprocessing

To augment the dimensional, *i.e.* numerical, labels from the categorical labels. This is conducted by mapping the categorical labels to points in the Valence-Arousal-Dominance space via the mapping given in Table 1 first, and perturbing the points with normally distributed noise with standard deviation 0.01, to reflect the effect of multiple annotators. To get the time series data in WAV format in the same datarate of 44.1 kHz from MP4 format, `FFMPEG` is utilized.
The whole data formation process can be seen in Figure 3.

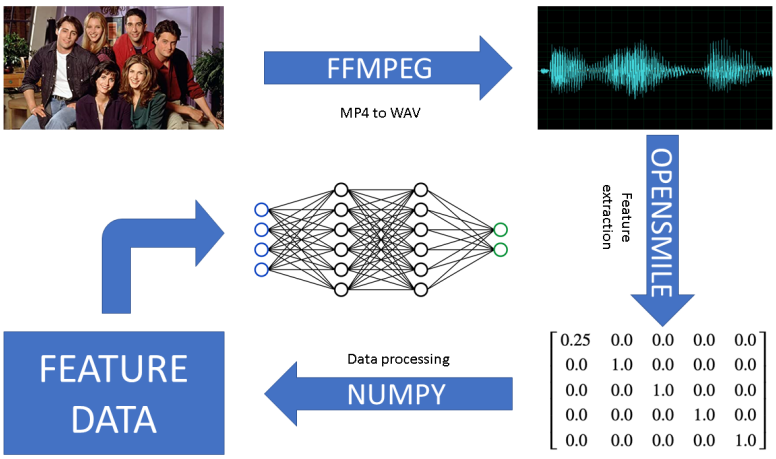


Figure 3: The data formation steps used to augment data from MELD dataset

3.2.2 Models

The pre-trained models are evaluated using the new test data. Furthermore, to make better conclusions on the models' capability to fit the data, the data is split to train, test and validation sets and the model is evaluated after being retrained. The architectures of the model were not changed. The only change made was the number of training and patience epoch for early stopping.

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/`.