Identification of emotions from speech using Deep Learning

by
Hasnain Hussain

hasnainhussain07@gmail.com

Abstract: The aim of the paper is about the detection of the emotions elicited by the speaker while talking. As an example, speech produced in a state of fear, anger, or joy becomes loud and fast, with a higher and wider range in pitch, whereas emotions such as sadness or tiredness generate slow and low-pitched speech. Detection of human emotions through voice-pattern and speech-pattern analysis has many applications such as better assisting human-machine interactions. In particular, we are presenting a classification models of emotions elicited by speeches based on deep neural networks (CNNs), Support Vector Machine (SVM), Multilayer Perceptron (MLP) Classification based on acoustic features such as Mel Frequency Cepstral Coefficient (MFCC).The models have been trained to classify seven different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprise). Our evaluation shows that the proposed approach yields accuracies of 86%, 84% and 82% using CNN, MLP and SVM respectively, for 7 emotions using Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset and Toronto Emotional Speech Set (TESS) Dataset.

Keywords: Emotion detection, deep learning, machine learning, classification, mel-frequency cepstral coefficients, CNN, RAVDESS, TESS, SVM, MLP.

1. INTRODUCTION

Human communication through the spoken language is the base for information exchange. It also finds use in numerous practical applications in various fields like Business Process Outsourcing (BPO) Centre and Call Centre to detect the emotion useful for identifying the happiness of the customer about the product, to enhance the speech interaction, to solve various language ambiguities and adaption of computer systems according to the mood and emotion of an individual.

In the proposed models the aim is to identify only the emotion that has a greater value in the audio track. Different approaches have been tried to have a machine classify feelings like computer vision or text analytics. In this work, our goal is to use pure audio data considering MFCC [4].

Rest of the paper is organized as follows: Section 2 states the related work in the area of speech emotion recognition. The implementation details including proposed methodology, algorithms, dataset are discussed in Section 3. System architecture is discussed in Section 4. The results are summarized in Section 5 followed by conclusions in Section 6.

2. RELATED WORK

In this field of research, many classification strategies have been introduced over the years. One such program, proposed by Iqbal et al. [1] used the Gradient Boosting, KNN and SVM to work on granular partitioning in the RAVDESS data to identify differences based on gender getting an overall accuracy between 40% and 80% depending on the specific task. Three types of datasets were created including male recordings only, female recordings and mixed. In RAVDESS (male) dataset, SVM and KNN have 100% recognition for all the anger and neutrality, but in the excitement and sadness Gradient Boosting did better than SVM and KNN. In RAVDESS (female) dataset, SVM achieves 100% accuracy with the same anger as half the males. KNN performance was also good for anger and neutrality with accuracy of 87% and 100% respectively. KNN performance was found to be poorer in happiness and grief as compared to other category of travelers. Of the combined male and female data rates, the performance of SVM and KNN was significantly better with anger and neutrality than Gradient Boosting. KNN's performance was really bad for happiness and sadness. The average performance of the classifiers in the male dataset is better than the female dataset without SVM. For aggregated data, SVM gains higher accuracy than gender data sets. Another approach introduced by Jannat et al. [2] found 66.41% accuracy in audio data and 90% accuracy for mixing audio and video data. In particular, given the already processed image data including faces and audio waveforms, the authors trained 3 different depth networks: one net only applies to image data, only one to fixed audio waveforms, and one third to both data and waveform data. One of the earliest methods using the dataset RAVDESS, but it only distinguished it from other available emotions [8]. Three types of song sharing algorithms were suggested: a simple model had been developed, a single work area model and a multi-task height model were developed. A simple model used a single, independent domain classifier. Two hierarchical types were used domain during training. The single function model trained different classifications for each domain. The multidisciplinary model trained a multidisciplinary partner to jointly share emotions across both domains. In the test phase, the test data was segmented based on the predicted domain. Data was analyzed using a classifier that matched the estimated domain. The work had been carried out with the adoption of a directed acyclic graph SVM (DAGSVM) [7].

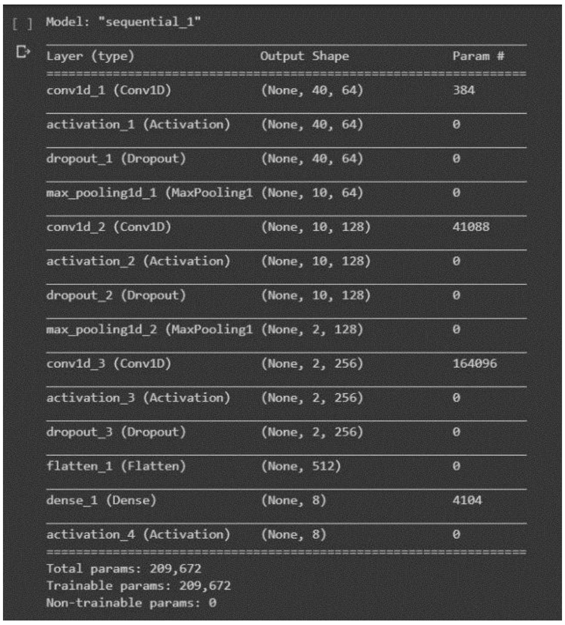
3. IMPLEMENTATION DETAILS

3.1 METHODOLOGY

The classification models of emotion recognition proposed here are based on deep learning strategy based on CNN, SVM classifier and MLP classifier. The key idea is considering the MFCC [4] commonly referred to as the ”spectrum of a spectrum”, as the only feature to train the model. MFCC is a different interpretation of the Mel-frequency cepstrum (MFC), and it has been demonstrated to be the state of the art of sound formalization in automatic speech recognition task [5]. The MFC coefficients have mainly been used as the consequence of their capability to represent the amplitude spectrum of the sound wave in a compact vectorial form. As described in [4], the audio file is divided into frames, usually using a fixed window size, in order to obtain statistically stationary waves. The amplitude spectrum is normalized with a reduction of the ”Mel” frequency scale. This operation is performed for empathizing the frequency more meaningful for a significant reconstruction of the wave as the human auditory system can perceive. For each audio file, 40 features have been extracted. The feature has been generated converting each audio file to a floating-point time series. Then, a MFCC sequence has been created from the time series. The MFCC array has been transposed and the arithmetic mean has been calculated on its horizontal axis.

3.2 ALGORITHMS

The deep neural network (CNN) designed for the classification task is reported operationally in Fig. 1. The network is able to work on vectors of 40 features for each audio file provided as input. The 40 values represent the compact numerical form of the audio frame of 2s length. Consequently, we provide as input a number of training files of size (40 x 1) on which we performed one round of a 1D CNN with a ReLU activation function [6], dropout of 20% and a max-pooling function 2 x 2. The rectified linear unit (ReLU) can be formalized as g(z) = max{0, z}, and it allows us to obtain a large value in case of activation by applying this function as a good choice to represent hidden units. Pooling can, in this case, help the model to focus only on principal characteristics of every portion of data, making them invariant by their position. We have run the process described once more by changing the kernel size. Following, we have applied another dropout and then flatten the output to make it compatible with the next layers. Finally, we applied one Dense layer (fully connected layer) with a softmax activation function, varying the output size from 640 elements to 8 and estimating the probability distribution of each of the classes properly encoded (Neutral=0; Calm=1; Happy=2; Sad=3; Angry=4; Fearful= 5; Disgust=6; Surprised=7).



```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_1 (Conv1D)            (None, 40, 64)            384
_________________________________________________________________
activation_1 (Activation)    (None, 40, 64)            0
_________________________________________________________________
dropout_1 (Dropout)          (None, 40, 64)            0
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 10, 64)            0
_________________________________________________________________
conv1d_2 (Conv1D)            (None, 10, 128)           41088
_________________________________________________________________
activation_2 (Activation)    (None, 10, 128)           0
_________________________________________________________________
dropout_2 (Dropout)          (None, 10, 128)           0
_________________________________________________________________
max_pooling1d_2 (MaxPooling1 (None, 2, 128)            0
_________________________________________________________________
conv1d_3 (Conv1D)            (None, 2, 256)            164096
_________________________________________________________________
activation_3 (Activation)    (None, 2, 256)            0
_________________________________________________________________
dropout_3 (Dropout)          (None, 2, 256)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 8)                 4104
_________________________________________________________________
activation_4 (Activation)    (None, 8)                 0
=================================================================
Total params: 209,672
Trainable params: 209,672
Non-trainable params: 0
```

Fig. 1. Detailed description of the architecture of the proposed classifier

2. **Multilayer Perceptron (MLP)** is a class of feedforward artificial neural network (ANN). MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable.

3. **Support Vector Machine (SVM)** is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Data can be scaled before applying to an SVM classifier to avoid attributes in greater numeric ranges while processing it. Scaling also serves the purpose of avoiding some numerical difficulties during calculation.

3.3 DATASET

For this task, the dataset is built using 5252 samples from:

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset[3] and Toronto emotional speech set (TESS) dataset[8]

The samples include:

1440 speech files and 1012 Song files from RAVDESS. This dataset includes recordings of 24 professional actors (12 females, 12 males), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions.

2800 files from TESS. A set of 200 target words were spoken in the carrier phrase "Say the word ____' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 stimuli in total. Two actresses were recruited from the Toronto area. Both actresses speak English as their first language, are university educated, and have musical training. Audiometric testing indicated that both actresses have thresholds within the normal range.

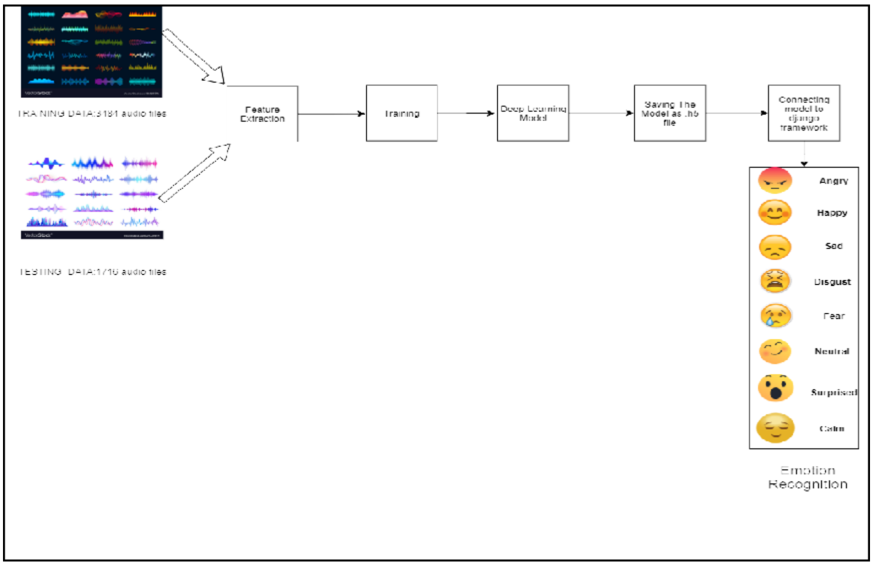
The classes the model wants to predict are the following: (0 = Neutral, 1 = Calm, 2 = Happy, 3 = Sad, 4 = Angry, 5 = Fearful, 6 = Disgust, 7 = Surprised). This dataset is skewed as there is **not a calm class** in TESS. Hence there is less data for that particular class and this is evident when observing the classification report.

Filename identifiers

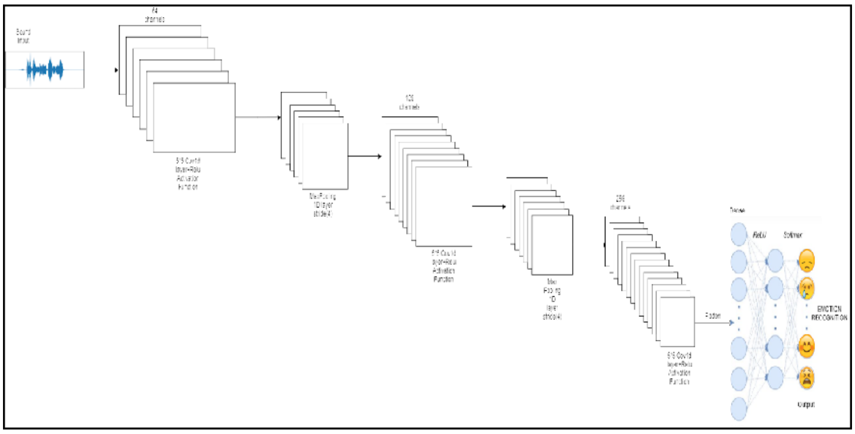
Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
Vocal channel (01 = speech, 02 = song).
Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
Repetition (01 = 1st repetition, 02 = 2nd repetition).
Actor (01 to 24. Odd-numbered actors are male, even-numbered actors are female).

4. SYSTEM ARCHITECTURE

4.1 SYSTEM DESIGN FLOWCHART


TRAINING DATA:3181 audio files
TESTING DATA:1716 audio files
Feature Extraction
Training
Deep Learning Model
Saving The Model as .h5 file
Connecting model to django framework
Angry
Happy
Sad
Disgust
Fear
Neutral
Surprised
Calm
Emotion Recognition

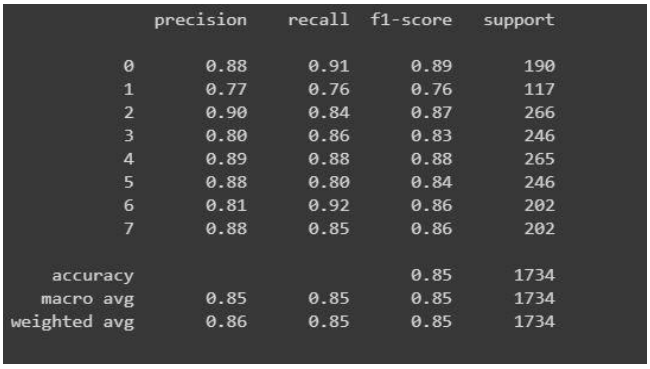
4.2 CNN LAYER DESCRIPTION


Sound Input
64 channels
5*5 Conv1d layer+Relu Activation Function
MaxPooling 1D layer stride(4)
128 channels
5*5 Conv1d layer+Relu Activation Function
Max Pooling 1D layer stride(4)
256 channels
5*5 Conv1d layer+Relu Activation Function
Flatten
Dense
ReLU
Softmax
EMOTION RECOGNITION
Output

5. DISCUSSION OF RESULTS

The results obtained from the evaluation phase show the effectiveness of the model compared to the baselines and the state of the art on the **RAVDESS** and **TESS** dataset. In particular, Table I. shows the values of precision, recall and F1 obtained for each of the emotional classes. These results show us that precision and recall are very balanced, allowing us to obtain F1 values distributed around the value **0.85** for almost all classes. The small variability of F1 results point out the robustness of the model that effectively manages to correctly classify emotions in eight different classes. The classes ”Calm” and ”Disgust” are the ones in which the model is less accurate, but this result does not surprise us because it is known in the Introduction that they are the most difficult classes to identify not only by speech but also while observing facial expression or analysing written text [15]. In order to evaluate the effectiveness of the classification of emotions proposed in this work, we decided to compare it with the results obtained from two other algorithms namely SVM and MLP classifier. The results shown in Tab II allow us to observe how the F1 values of our model are **better** than baselines and competitors on all classes. However, it is necessary to point out that the drop in performance is minimal and also done to prevent **overfitting**, and it is well known that as the number of classes increases, the classification task becomes more complex and loses its accuracy. Nevertheless, the **CNN-MFCC** model proposed here manages to obtain a score of F1 that, on average, is equivalent to that of the two jobs we have been confronted with. A further index of model reliability can be found in Fig. 2 and Fig. 3. In the first one, it is possible to observe how the value of loss (error in the accuracy of the model) tends to decrease both on the test set and on the training set up to the 200th epoch. The decrease is less evident from the 100th epoch but still perceptible. In Fig. 3, it is reported the average value of accuracy on all the classes that, to the contrary of the loss, increases with the increases of the ages. Such values of loss and accuracy do not differ much among the training and test dataset, allowing us to affirm that the model does not turn out to be overfitted while training. The consequence of this is, in fact, in line with the F1 scores previously observed.

Table I. Results of the CNN model on the test set per each class.



```
              precision    recall  f1-score   support

           0       0.88      0.91      0.89       190
           1       0.77      0.76      0.76       117
           2       0.90      0.84      0.87       266
           3       0.80      0.86      0.83       246
           4       0.89      0.88      0.88       265
           5       0.88      0.80      0.84       246
           6       0.81      0.92      0.86       202
           7       0.88      0.85      0.86       202

    accuracy                           0.85      1734
   macro avg       0.85      0.85      0.85      1734
weighted avg       0.86      0.85      0.85      1734
```

Table II..F1-Score for each class compared to the baselines (SVM, MLP) and the state of art

CLASS	**MLP**	**SVM**	**CNN**
SAD	0.81	0.82	0.80
ANGRY	0.89	0.91	0.89
HAPPY	0.82	0.84	0.90
DISGUST	0.80	0.81	0.81
SURPRISE	0.80	0.87	0.88
NEUTRAL	0.89	0.93	0.88
CALM	0.75	0.64	0.77
FEAR	0.84	0.81	0.88

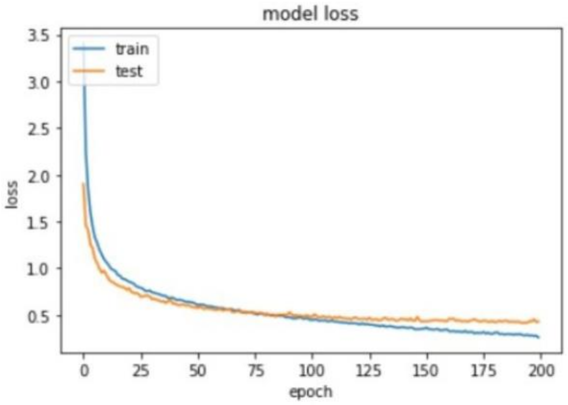


Fig. 2. Trend of the cost function of our deep learning model over 200 epochs

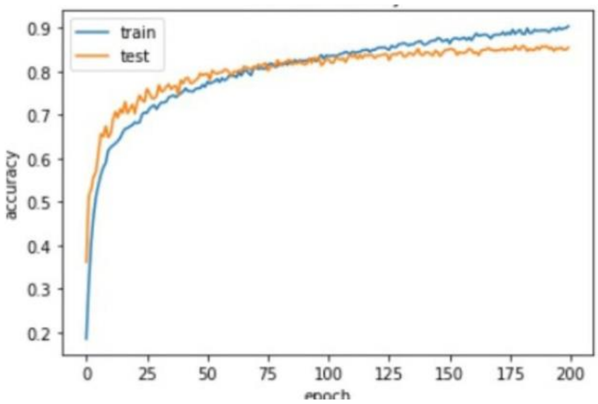


Fig. 3. Trend of the accuracy of our deep learning model over 200 epochs.

6. CONCLUSION

In this work, we presented an architecture based on deep neural networks for the classification of emotions using audio recordings from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Toronto Emotional Speech Set (TESS) .The model has been trained to classify seven different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised) and obtained an overall F1 score of 0.85 with the best performances on the Happy class (0.90) and worst on the calm class (0.77). To obtain such a result, we extracted the MFCC features (spectrum of-a-spectrum) from the audio files used for the training. On the above representations of input data, we trained a deep neural network that uses 1D CNNs, max-pooling operations and Dense Layers to estimate the probability of distribution of annotation classes correctly. The approach was tested on the data provided by the RAVDESS dataset. As a baseline for our task, we considered a MLP Classifier trained on the same dataset achieving an average F1 score of 0.84 over the 8 classes. After the MLP classifier, we trained a SVM classifier that achieved an F1 score of 0.82. Our final choice was a deep learning model that obtained a F1 score of 0.86 on the test set. The good results obtained suggest that such approaches based on deep neural networks are an excellent basis for solving the task. In particular, they are general enough to work in a real application context correctly and previous versions of this work were all developed using only the RAVDESS dataset and TESS has been added recently. Also, the previous versions of this work used audio features extracted from the videos of the RAVDESS dataset. This particular part of the pipeline has been removed because it was shuffling very similar files in the training and test sets, boosting the accuracy of the model as a consequence (overfitting).

REFERENCES:

[1] Iqbal, A. and Barua, K. A real-time emotion recognition from speech using gradient boosting. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (2019), IEEE, pp. 1–5

[2] Jannat, R., Tynes, I., Lime, L. L., Adorno, J., and Canavan, S. Ubiquitous emotion recognition using audio and video data. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (2018), ACM, pp. 956–959.

[3] LIVINGSTONE, S. R., AND RUSSO, F. A. : The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one 13, 5 (2018), e0196391.

[4] Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In ISMIR (2000), vol. 270, pp. 1–11.

[5] Muda, L., Begam, M., and Elamvazuthi, I. : Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques. arXiv preprint arXiv:1003.4083 (2010).

[6] Nair, V., and Hinton, G. E. : Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (2010), pp. 807–814

[7] Platt, J. C., Cristianini, N. and Shawe-Taylor, J. : Large margin dags for multiclass classification. In Advances in Neural Information Processing Systems 12, S. A. Solla, T. K. Leen, and K. Muller, Eds. MIT Press, 2000, pp. 547–553

[8] Toronto emotional speech set (TESS) (https://tspace.library.utoronto.ca/handle/1807/24487)