Emotion Recognition using Supervised Learning

Blanca Bastardés Climent blancabc@kth.se
Osheen Sharma osheens@kth.se

Abstract

The use of Automatic Speech Emotion Recognition(ASER) is a current research topic with a wide range of applications which provides information from the personality and psychological state of the speaker. The aim of this project is to explore different supervised automatic models to classify 8 emotional states and provide a comparison between the performance of Machine Learning (ML) and Deep Learning (DL) algorithms. Mel-frequency cepstral coefficients (MFCC) were extracted from 1440 speech samples from RAVDESS database and used as speech features to train the models. As it was expected, a Convolutional Neural Network (CNN) as a DL approach achieved better classification, resulting in test accuracy as 65.28% as compared to Random Forest classifier (60.41%) & Support Vector Machine (52.08%). Moreover, two more tests were conducted. First test was to explore the effect of gender on the learning & second test was to make sure that the learning of the model is not influenced by the speakers while predicting the emotions. Further improvements could be based on the addition of noise to the signals to increase the variability in the training and thus be able to obtain a more generalized model. In addition, evaluating unsupervised models would be of great interest to study the number of groups the algorithm can find without using labels.

Keyword: ASER, supervised, MFCC, CNN, classification, unsupervised

1 Introduction

Emotions play an important role in human interactions but they are a difficult concept to define. Several different approaches are proposed to recognize emotional states such as facial expressions, physiological signals and speech, between others.

Speech and Speaker Recognition (SSR) is one such approach that has the ability to recognize words and phrases from human speech along with facial expressions of the speaker. The first speech recognition experiment was recorded in 1000 A.D. which laid the foundation of the technology using natural language processing as an input [1]. Later, Bell Laboratories worked on recognizing the numbers in the speech. The success in the technology made the researchers to contribute in the field for the improvement and defining the application in various domains. Some interesting applications are: health care, computer games, E-learning, banking etc. [2]

To perform the speech analysis various features need to be extracted from the signal which will enable the machine to differentiate between different speeches and emotions. These features contain important information like intensity, frequency, Linear Prediction Cepstrum Coefficients (LPCC) and Mel-frequency cepstrum coefficients (MFCC) [2].

In this project work the mean of the 13 and 40 MFCC was used as features, discarding time component, and implement different supervised learning models that can recognize 8 different emotions from 24 different speakers. To achieve this two machine learning algorithms: Support Vector Machine and Random Forest were implemented. Additionally, to study the impact of neural networks on speech database a Convolutional Neural Network(CNN) was implemented.

2 Literature Review

Emotion recognition has become a topic of interest in recent years. In the effective contribution towards this approach F.A. Russo et al. implemented linear and neural network models to infer the emotion (valence and arousal) that can be induced in the listener after listening to music and reported the accuracy of 89.75% for valence and 88.92% for arousal [3]. Parente et al. studied the reasons for successful implementation of speech recognition technology in the healthcare domain by interviewing the healthcare professionals [4]. Other healthcare application that was reported by Henricks et al. in pathology lab for the documentation purpose [5]. In the research work by A.Iqbal et al. the authors have classified 7 different emotions using 25 MFCC as features and implemented SVM, Random Decision Forest and Gradient Boosting to predict the emotions. In the paper they reported the highest accuracy of 81.05% which was achieved by Random Forest [6]. Shaqra et al. studied how the age and gender affect the emotion classification task. Hierarchical classification models were used concluding that training separated classifiers for each gender and age group leads to better results than having one model for both genders and ages [7]. Recently, research group from Bangladesh Adib A. Zamil et al merged two database, Emo-DB and RAVDESS to classify emotions from speech signals. They used MFCC as features and trained Logistic Model Tree to classify the emotions. From the study, it was concluded that for few emotions the model performance was promising but for some cases the model was confusing the emotions with other. The highest accuracy was reported as 70% [8].

Many are the researchers who opt for ML algorithms to perform emotion classification, however, DL approaches are also explored such as in the study from Huang et al. They used SVM as a ML approach and a CNN as a DL approach. The best accuracy achieved with SVM was 48.11% whereas with CNN was 85% which proves that DL approaches seem to be promising [9]. An interesting study was conducted by Tripathi et al. in which they made use of transcriptions along with MFCC to help the emotion recognition using DL. The models were trained separately on each features as well as in combination. The highest accuracy reported was 76.1% when the model was trained on both MFCC as text as features. This study shows that the higher the number of features, the better is the performance of the CNN model [10].

3 Methods

This section discusses about the materials and tools that were used to accomplish the aim of the project and also the implementation of the neural network and machine learning algorithms for training the models. The inputs for all the classification methods are the mean of the MFCC of each sample's frames.

3.1 Dataset

To conduct the project for this course we used RAVDESS database which was created by Steven R. Livingstone et al. in North American English language [11]. The database consists of 24.8GB file in which the audio clip was in *.wav* format, audio+video clip in .mp4 format and only video clip with no sound. Professional 24 different actors were chosen where the even number actors were female and odd number actors were male. Each actor performs two level of emotion intensity: normal & strong and two statements were recorded with seven syllables in length and were matched in word frequency. The recording was done on 8 different emotions: neutral, calm, happy, sad, angry, fearful, surprise, and disgust.

For this project, only samples from the "audio-only" modality were used. To extract the MFCC of each sample `librosa` python package was used. The dataset was divided into training, validation and testing using the library from python `sklearn.model_selection`. In the beginning 20% of the dataset was used as test set but later only 10% of the dataset is used for testing and it was observed that the accuracies improved as more data was given for training. The MFCC were computed from the given audio speech and was used as input to the training models whereas the emotions were used as the labels. For this work all experiments have been conducted using `keras`, `sklearn` and `pandas` as python libraries.

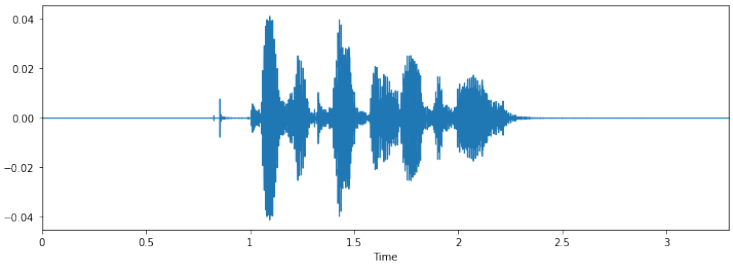


Figure 1: Example of a speech sample.

3.1.1 Aim

The aim of this project was to predict the emotions from the speech. In order to accomplish the main goal, two different approaches were followed which are discussed in detail in Section 3.2 and 3.3.

3.2 Machine Learning in ASR

The task of recognising speech for a machine is not as easy as people might think. It is much more than simply listening to the words. The use of Machine Learning (ML) allows computers to reproduce more and more things that humans can do. In Automatic Speech Recognition (ASR), by providing many speech samples to the machine it will be able to learn and identify patterns from it to perform a classification. For this process, there are the ML techniques employed for ASR.

Support Vector Machines (SVMs) is a method which has been proved to be able to cope with hard classification problems. SVMs are discriminative classifiers based on maximizing the separation between categories. They can efficiently perform both linear and non-linear classification, depending on the kernel specified.

Random Forest (RF) is a classification method based on decision tree theory. The difference with the conventional decision trees is the addition of randomness when selecting the features, whereas decision trees give higher importance to some features.

3.3 Deep Learning in ASR

Deep Learning (DL) is branch of ML which has shown top performance in many research fields such as natural language processing, image analysis and computer vision etc. Speech and Speaker Recognition is no exception. The learning can be divided into supervised, semi-supervised and unsupervised. CNNs are a type of neural networks which have prior knowledge of objects and features to compensate for less amount of data. Basic CNN architecture consists of Convolutional layer, Activation function, Pooling layer and the the fully connected layer also known as dense layer.

For this project we use basic CNN architecture with 2 Convolutional layer with filter size of 8 followed by Max Pooling layer with filter size 5. The Activation function used for each layer was `relu`. In the end 3 Dense layers were added followed by Dropout of 0.3 in order to avoid over-fitting during the training of the model. The last Activation layer was chosen as `softmax`. The summary of the model architecture can be seen in Table 1.

4 Experiments

The main focus of this project was on emotion recognition. Both ML and DL approaches were implemented and a comparison between their performance was carried out. SVMs and Random Forest classifiers were used as ML methods; and a CNN as a DL approach. In addition, two more experiments was carried out to explore the data, as the effect of the gender in the learning process and splitting of the data.

Layer	Output Shape	Parameters
Conv1D	(None, 40, 128)	1152
Activation	(None, 40, 128)	0
MaxPooling	(None, 8, 128)	0
Conv1D	(None, 8, 256)	262400
Activation	(None, 8, 256)	0
MaxPooling	(None, 1, 256)	0
Flatten	(None, 256)	0
Dense	(None, 256)	65792
Activation	(None, 256)	0
Dense	(None, 256)	65792
Activation	(None, 256)	0
Dropout	(None, 256)	0
Dense	(None, 8)	2056
Activation	(None, 8)	0
Total		397,192

Table 1: Model Architecture

The first step before implementing and feeding the classifiers and the neural network, was the feature extraction from the data samples. The MFCC were chosen as features for this project. Different number of features were tested. In the first trial 13 MFCC coefficients were tested and to compare the model performance the second trial was made with 40 MFCC coefficients.

Table 2: Classification results from the different labels and models.

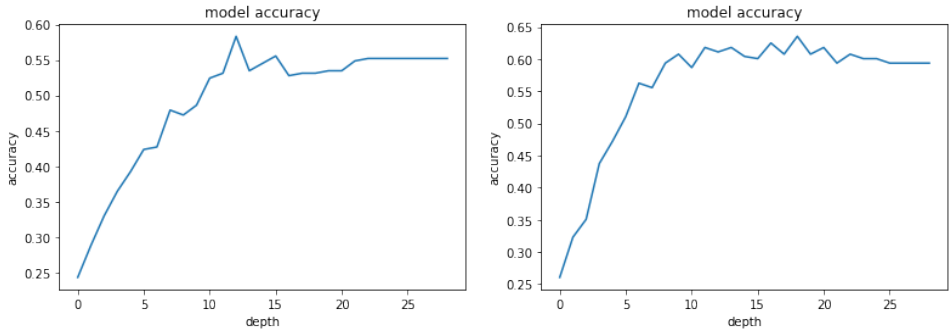
		13 MFCC			40 MFCC		
Label	Model	Train	Validation	Test	Train	Validation	Test
Emotions	SVM			0.4131			0.5208
Emotions	RF			0.5520			0.6041
Emotions	CNN	0.7230	0.5385	0.5208	0.9710	0.7154	0.6528

4.1 Emotion Recognition

First, the 13 MFCC extracted from the speech samples were fed into the SVM classifier. From all the kernel functions, `linear` was the one chosen giving an accuracy of 41.31%. Random Forest was the second classifier tested. In this case, the criterion selected to measure the quality of a split was `gini impurity` and the limit of depth was set to 30. The accuracy achieved with this classifier was 55.20%. Same process was repeated for 40 MFCC coefficients and the performance were compared. It was observed that higher MFCC coefficients resulted in higher accuracy as illustrated in Table 2.

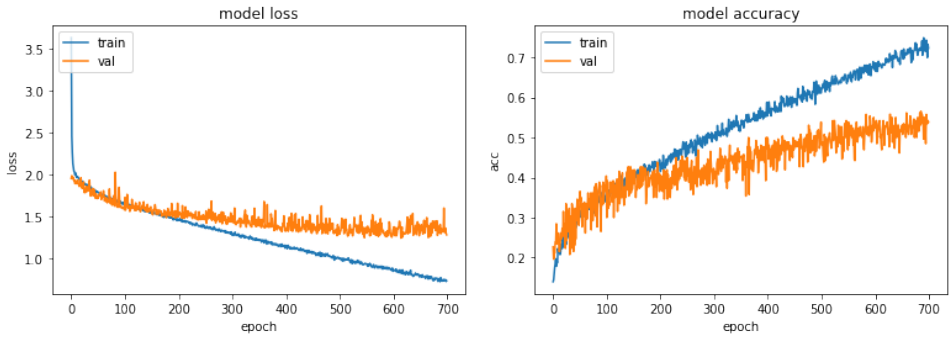
Further, the CNN was trained for 700 epochs where the MFCC were used as inputs and the emotions were used as labels to predict and evaluate the performance of the network. The optimizer used in this experiment was `RMSprop` with learning rate of 3e-5. As in our project we have multi label classification problem hence, the loss was chosen as `sparse_categorical_crossentropy`. The model was trained for both 13 MFCC and 40 MFCC coefficients. A comparison of the performance of the different experiments to classify the speech samples based on the emotions can be seen in Table 2. The best performance was achieved by the CNN (65.28%) trained on 40 MFCC. Hence, the further experiments were conducted on CNN model architecture.

Figures 3 and 4 show the model loss and accuracy for both 13 and 40 MFCC. It can be seen that the model trained on 40 MFCC gives better curves, which means that a higher number of MFCC helps the learning process. Figure 5 shows the confusion matrix of the predictions of the emotions with a CNN as classifier and 40 MFCC as features. It can be seen that there were not many mismatches in general, although some emotions were better recognised.



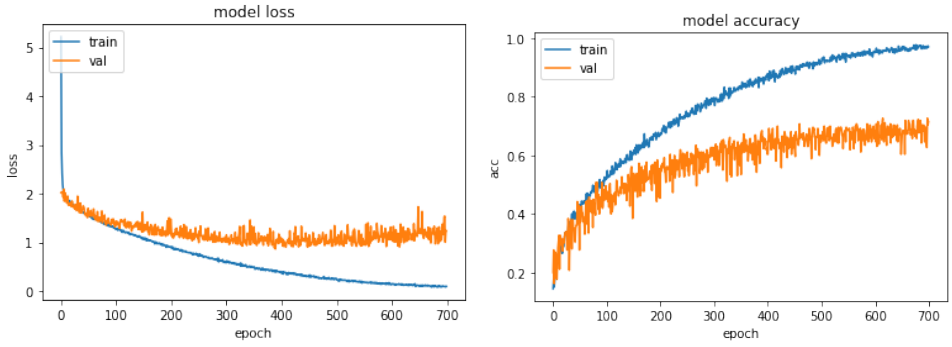
(a) 13 MFCC (b) 40 MFCC

Figure 2: Evolution of the accuracy with the depth using Random Forest for 13 and 40 MFCC respectively.



(a) (b)

Figure 3: Results of the CNN trained on 13 MFCC as features to classify emotions.



(a) (b)

Figure 4: Results of the CNN trained on 40 MFCC as features to classify emotions.

4.2 Gender Effect

To understand the effect of gender when training the model, CNN was trained separately on both male and female gender. The prediction was better achieved on male voices as compared to female voices. The results of the accuracy assessment can be found in Table 3 under 'Without Isolating' column.

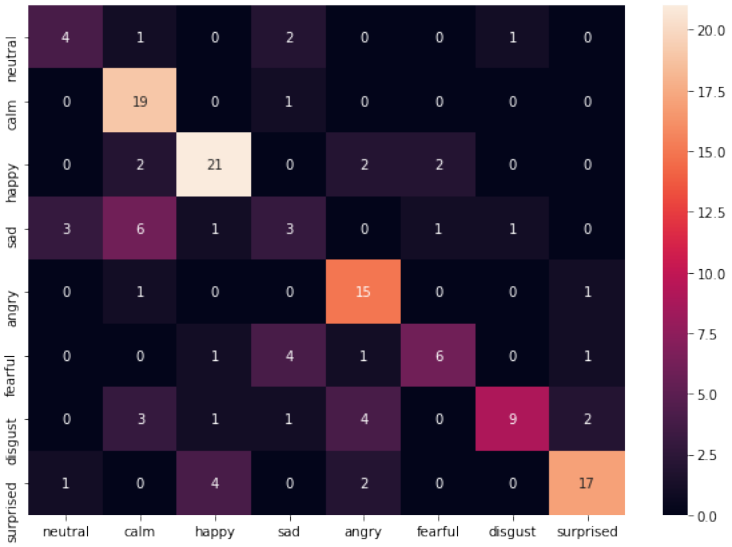


Figure 5: Confusion matrix of the predictions of the emotions after using the CNN as classifier and 40 MFCC as features.

4.3 Isolating Actors for Test Dataset

A different approach was implemented to split the dataset in a way that some actors were isolated from the train validation set and were kept only as test set. This was done in order to study the impact of splitting the dataset and to make sure that the model is not influenced by the speakers while predicting the emotions. The splitting was done for both the genders separately. As mentioned before there are 24 actors in the dataset where even number of actors are female and odd number of actors are male resulting in 12 actors for each category respectively. The splitting of data was done in such a way that last two actors from the list of 12 actors for each gender were kept unseen from the training of the model and were only used for testing the model performance. To this end, CNN model was trained with 40 MFCC for 500 epochs with the same hyper-parameter settings as mentioned in previous Section 4.1. The following Table 3 illustrates the accuracy achieved after the training on isolated dataset under the column 'Isolating'.

Table 3: Classification results from isolated and non-isolated dataset.

		Without Isolating			Isolating		
Label	Model	Train	Validation	Test	Train	Validation	Test
Female Actors	CNN	0.865	0.677	0.583	0.873	0.750	0.283
Male Actors	CNN	0.927	0.715	0.736	0.900	0.675	0.500

5 Conclusion

In this project we have tested three different types of supervised learning methods for predicting human emotions from the speech signals. In total one neural network and two machine learning algorithms were implemented and the results were compared as shown in Section 4. The neural network performed better as compared to ML algorithms. The highest testing accuracy was reported as 65.28% for 40 MFCC. However from the two ML algorithms implemented, the Random Forest performed best. As Random Forest is easy to implement and it has less parameters to tune, it has a higher probability to find the optimal classification of the data, giving higher accuracy (60.41%) as compared to SVM (52.08%) in this application. Also, it was observed that higher number of MFCC lead to higher model accuracies as the model has more features which enable the network to learn better. Additionally, to explore more the understanding of deep neural networks two more tests were performed to predict emotions based on gender from the speech signal. The difference

between the tests was the way of splitting the data, where in one of them few actors were isolated as the test set so that the network does not see them while training. It was interesting to see how the network differentiates between male and female voices. However, we can also conclude from the Table 3 that when the test set has samples from speakers that were not trained initially by the network, the accuracies decreased significantly. This assessment provides with the knowledge that the previous models, which were trained without isolating few actors as the test set, might be biased towards speakers.

6 Future Work

As the data doesn't have much variations, it will be interesting to add noise to the signals in order to generalize the learning process and make the model compatible to learn from the variations in the signals. Another solution would be to merge other datasets to increase the number of samples and variability. Also exploring of other neural networks like Long Short Term Memory Loss (LSTM), Recurrent Neural Network (RNN), among others, would give better understanding and comparison between different architectures. Additionally, implementation of unsupervised learning methods can also be added as one of the experiments since they have the advantage of feeding the raw data without the label and have less complexity. This will provide with the information of which patterns are learnt from the model when the labels are not given to the network unlike supervised learning. Having no labels the model is not forced to have specific number of classes to perform the classification.

References

[1] *Speech Recognition Software: Past, Present Future*. 2018. URL: https://www.globalme.net/blog/speech-recognition-software-history-future/.

[2] Leila Kerkeni et al. "Automatic Speech Emotion Recognition Using Machine Learning". In: *Social Media and Machine Learning*. Ed. by Alberto Cano. Rijeka: IntechOpen, 2020. Chap. 2. DOI: 10.5772/intechopen.84856. URL: https://doi.org/10.5772/intechopen.84856.

[3] Frank A. Russo, Naresh N. Vempala, and Gillian M. Sandstrom. "Predicting musically induced emotions from physiological inputs: linear and neural network models". eng. In: *Frontiers in psychology* 4 (Aug. 2013). PMC3737459[pmcid], pp. 468–468. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00468. URL: https://doi.org/10.3389/fpsyg.2013.00468.

[4] Ronaldo Parente, Ned Kock, and John Sonsini. "An analysis of the implementation and impact of speech-recognition technology in the healthcare sector". eng. In: *Perspectives in health information management* 1 (June 2004). PMC2047322[pmcid], pp. 5–5. ISSN: 1559-4122. URL: https://pubmed.ncbi.nlm.nih.gov/18066385.

[5] Walter H. Henricks et al. "The Utility and Cost Effectiveness of Voice Recognition Technology in Surgical Pathology". In: *Modern Pathology* 15.5 (May 2002), pp. 565–571. ISSN: 1530-0285. DOI: 10.1038/modpathol.3880564. URL: https://doi.org/10.1038/modpathol.3880564.

[6] A. Iqbal and K. Barua. "A Real-time Emotion Recognition from Speech using Gradient Boosting". In: *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. 2019, pp. 1–5.

[7] Ftoon Abu Shaqra, Rehab Duwairi, and Mahmoud Al-Ayyoub. "Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models". In: *Procedia Computer Science* 151 (2019). The 10th International Conference on Ambient Systems, Networks and Technologies (ANT 2019) / The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40 2019) / Affiliated Workshops, pp. 37–44. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2019.04.009. URL: http://www.sciencedirect.com/science/article/pii/S1877050919304703.

[8] A. A. A. Zamil et al. "Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames". In: *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)*. 2019, pp. 281–285.

[9] Andrew Huang and Puwei Bao. *Human Vocal Sentiment Analysis*. 2019. arXiv: 1905.08632 [eess.AS].

[10] Suraj Tripathi et al. *Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions*. 2019. arXiv: 1906.05681 [eess.AS].

[11] Steven R. Livingstone and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". In: *PLOS ONE* 13.5 (May 2018), pp. 1–35. DOI: 10.1371/journal.pone.0196391. URL: https://doi.org/10.1371/journal.pone.0196391.