Speech emotion recognition using Convolutional neural networks

Elnaz Khaveh

1. Introduction

Speaking has been considered the strongest and fastest communication tool between humans. Replacing humans with machines in various areas in this era has led researchers to think of speech as an efficient way of interacting with them. The importance of recognizing and analyzing human voices by machines is undeniable in the human-computer interaction (HCI)[1] because great information such as age, gender, nationality, and the speakers' emotions can be identified from the voice. This paper focuses on extracting the speakers' emotions from their acoustic features.

Figure 1. Speech emotion recognition

Speech Emotion Recognition (SER) is a highly complex task; because it is unclear whether the emotions are natural or just the person is acting that way. So, distinguishing between these two can be very challenging, but in what follows, we can see some essential real-world applications of SER:

1) Call center: Providing the operators with SER to record the satisfaction or dissatisfaction of the customers[2].
2) Robots: SER is used for the human-robot interaction[3].
3) Lie detection: Detect the emotions of the criminals from their speech in courts[1].
4) Psychotherapy: In couple therapy or individuals, SER is used to diagnose the issue[4].
5) Baby caring centers: At very advanced levels, detect the emotion of kids when they cry to provide them with their needs in baby care centers[5].
6) Voice assistant: new applications like “Siri” and “Alexa” which talk to the people, should identify their emotions to interact with them naturally. [6].

Researchers have put great effort into this area, due to the mentioned facts, in recent years. The task was done by extracting various verbal and acoustic features and using various models and datasets. For example, Lin et al.[7] used the classical machine learning model like the Hidden Markov Model (HMM) and Support Vector Machines (SVM) and the short-term autocorrelation method for the feature extraction with the accuracy of 88.9% on the DES database. In another paper [8] we have a Gaussian mixture model (GMM) for modeling and fundamental frequency, time, amplitude, cepstrum for the feature extraction on the CASIA dataset with seven classes, a recognition rate of 73%, and a significant number of other papers in this research field.

1.1. Problem statement

This paper aims to recognize speech emotion from a deep learning approach. A Convolutional Neural Network is used to do so. The first step in every classification task is feature extraction, with numerous features for voice [figure 2]. For this task MFCC is used as feature. The RAVDESS dataset is used for this classification, which is a multimodal database of emotional speech and song. The included emotions are calm, happy, sad, angry, fearful, surprised, and disgusted.

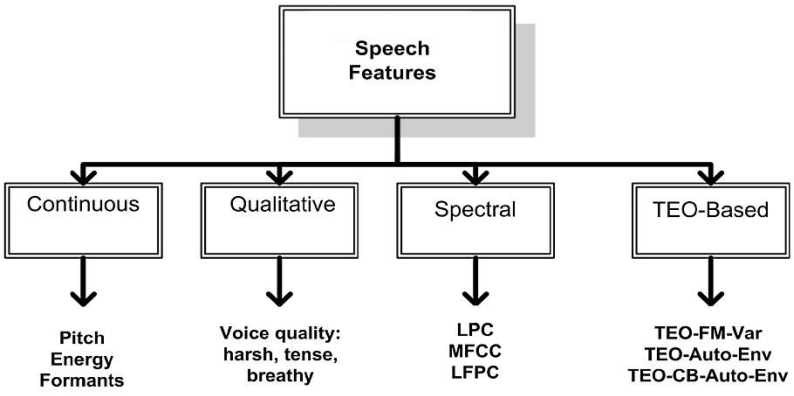

Figure 2.Diagram of Speech features

1.2. Research hypothesis

Since CNN plays an essential role in recognition tasks such as face recognition, handwriting recognition, image recognition, etc., we can expect that this voice recognition also performs well because the features extracted from the speech can be considered images.

1.3. Research contributions

In case of finding a highly accurate model, we may use it in the areas which need high accuracy and are more sensitive like training machines to see the situation of couples in

a relationship in serious issues or diagnose them based on their emotions and using it in courts to detect the lies of the criminals and make decisions instead of the judges.

In the following sections, the details of this research have been explained. First related works are discussed in the next part, then the dataset, the model's architecture, and voice features are described in materials and methods. After that, the results of this deep learning algorithm are shown with discussions, and finally, we can conclude from these results in the conclusion section.

2. Related works

SER has been one of the fields on which the researchers focused in recent years. The available data and modeling approaches for this classification have become updated, so more and more papers are being published in this area, some of which are mentioned in the following.

In 2015, Qin et al. [9] did an SER task by extracting acoustic and lexical features using an SVM model because, among the classic machine learning models, SVM has a better generalization performance. Also, for classifying the emotions, they used both acoustic and verbal features. What is done in this project is using CNN and RNN as the deep learning model and eight classes of emotions instead of four. So, a more comprehensive range of emotions can be recognized. Also, only spectral features are extracted rather than both lexical and acoustic because the lexical representation of the words cannot express the emotions accurately, and actors might use happy words with a sad emotion, for instance. There are many other papers in which they used SVM for SER [2, 10], but using a deep learning algorithm like CNN might capture the features of speech much better, enhancing the results. Also, some research has been done with CNN in this area [11], but mostly used MFCC or spectrograms as features. In this paper, a combination of both is proposed beside each to see whether they modify the classification task or not. A 3D convolutional neural network was proposed for SER in a more recent task [12].

The database used for speech emotion recognition is also of high importance. In [13], the Berlin dataset was used with a CNN model, which achieved an accuracy of 84.3%. However, the problem is that the Berlin database is a minimal database of only four actors. When the training data is from the same person and same voice, the model might perform very well on that, but it cannot generalize to other new voices. In another study [12], a CNN-based model is used for speech emotion classification on the SAVEE database, which is from 4 male actors aged between 27 and 31. This is a tiny sample for the recognition task because it is restricted to some actors, gender, and age group. Therefore, this model might show less accurate scores for older adults or females.

In [14], two issues have been addressed; first complexity cost of SER is reduced, and second, the performance is improved using a lightweight CNN model and a plain kernel with a modified pooling strategy. They used spectrograms to recognize hidden emotional features for feature

extraction, which make this study outperform the other studies. On the other hand, they reduced the number of layers for the cost complexity, which might harm the performance of larger datasets.

Another paper addressed the issue of noise in speech signals[15]. There is a CNN-based framework for SER and using spectrograms as feature extractions in this paper. The CNN architecture has input layers, convolutional layers, and fully connected layers followed by a SoftMax classifier. The positive point of this paper is that they used a dynamic adaptive threshold technique to remove noise and silent signals from speech signals. Then they are converted to spectrograms to increase the accuracy and decrease the computational complexity of the proposed model, so our model is inspired somehow by this paper in pre-processing of the data.

3. Materials and methods

3.1. Dataset

RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) is a multi-modal database for speech and song. Multi-modal means consisting of audio-only, face-only, and audio-face files in which 24 professional actors (12 female, 12 male) vocalize two sentences in a neutral North American accent. The sentences are “Kids are talking by the door” and “Dogs are sitting by the door”.

The emotions calm, happy, sad, neutral, angry, fearful, surprised, and disgusted are included in the speech part of this database, and the songs are happy, sad, fearful, calm, and angry. For this paper, audio-only speech files are used. Each actor says each sentence in 2 different levels of intensity except the neutral emotion, which makes the database imbalanced [figure 3] because the number of files for this emotion is exactly half of the others, and it is removed from the dataset before feeding it to the model.

RAVDESS was preferable for this task compared to other speech datasets due to the number of actors and the fact that it is gender balanced. Also, it covers a wider variety of emotions.

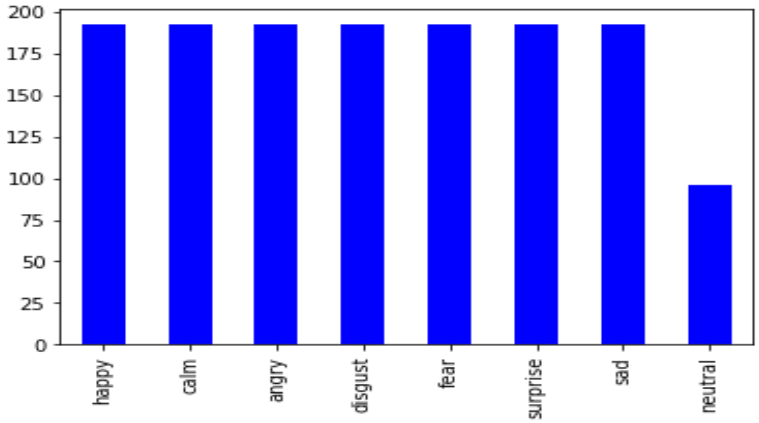

Figure 3.Distribution of emotions

3.1.1. Loading the dataset

To load the dataset, the Librosa package is used, which is a package to load and analyze the audio files and music [16]. It loads the audio files as floating-point time series.

3.1.2. Pre-processing

The audio files in RAVDESS do not have the same length, and in some parts, they have unnecessary silence, so we need some steps, which will be explained in what follows, before feeding them to the model. The wave plots below show the steps of processing (Figures 4,5,6 and 7):

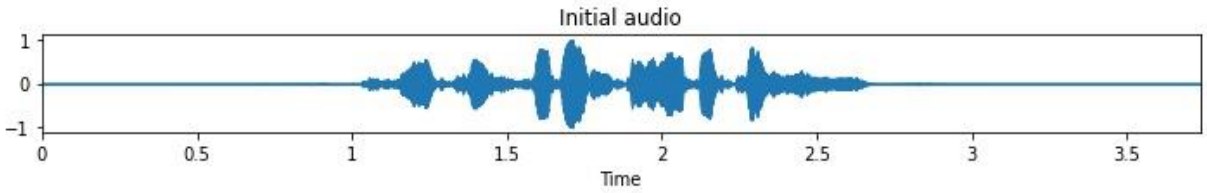

*Figure 6.*Initial audio waveplot.

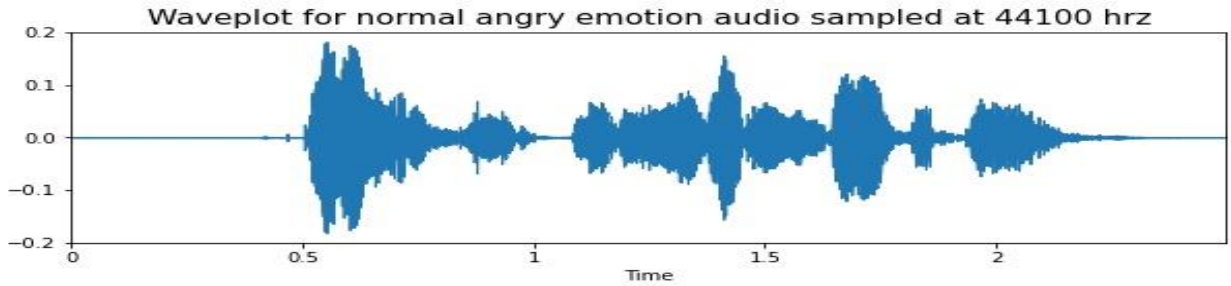

Figure 5. Normalized waveplot

The first step is to normalize the data but normalizing the audio sounds is different from other databases. Normalizing the audio means increasing the volume of all the files to a maximum target in dBFS (A unit to measure the amplitude levels). This helps to boost the volumes and have clearer audio.

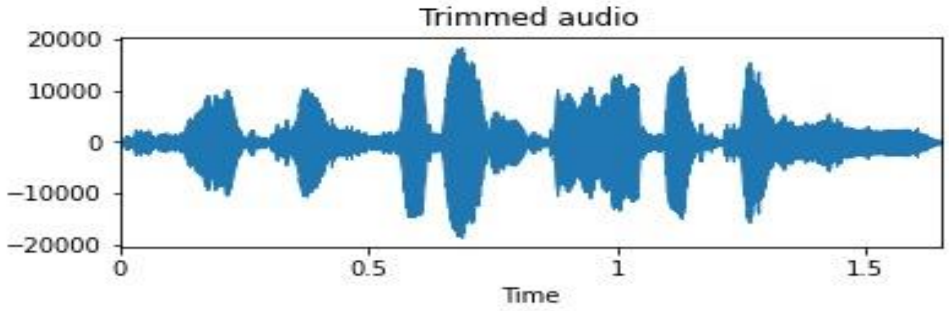

Figure 7. Trimmed waveplot.

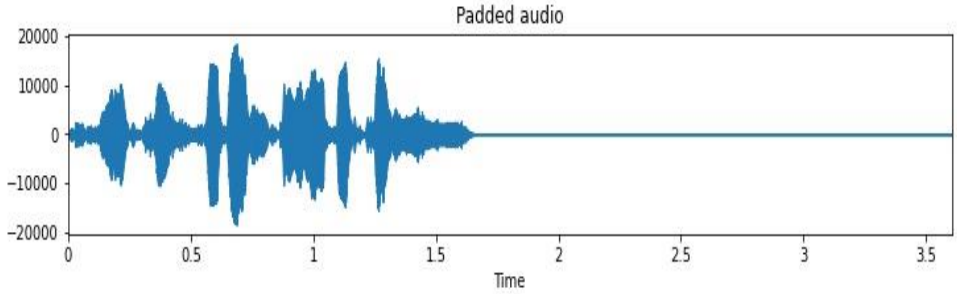

Figure 4. Padded waveplot.

By trimming, we remove the silence from the beginning and the end of the files, and by padding, the audio files are filled with zeros (in this case, silence) to have the same length audios, and now they are ready to give it to the model.

3.2. Feature extraction

When we talk, we change the air pressure in front of us, which are considered waves, and form the wave plot. The waves per second are the frequency.

3.2.1. Mel scale

Before going to the concept of MFCC, knowing what Mel-scale is, is required.
It is the logarithmic transformation of the frequency of a signal, and it is calculated from the following formula:

$$m = 1127 \cdot \log\left(1 + \frac{f}{700}\right)$$

Mel Frequency Cepstral Coefficients is a very complex concept, but to say briefly, they are features in speech recognition that can be calculated in the following steps:

1) Frame the signal into short frames
2) Map the powers to the Mel scale
3) Take the logs of the powers at each of the Mel frequencies
4) Take the discrete cosine transform of the logs of powers
5) The amplitude of the resulting spectrum is the MFCC

3.3. Model selection

3.3.1. Convolutional neural networks

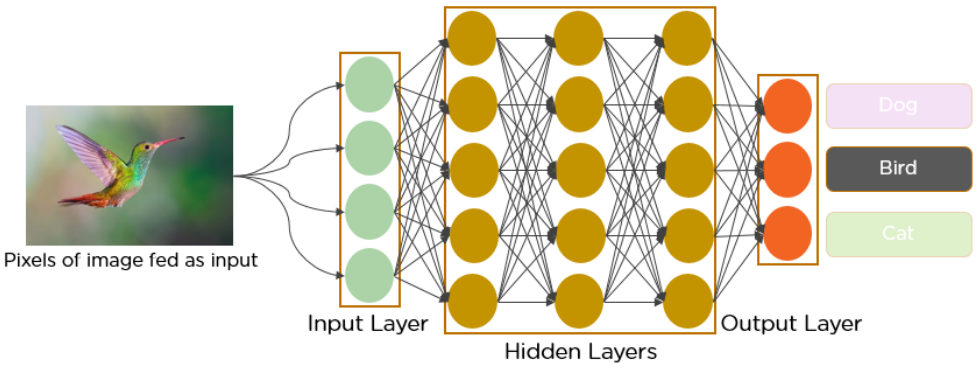

Figure 8.CNN[17]

CNN is a deep learning algorithm to classify images which can also be considered as matrices of pixels. The layers of CNN are arranged in a way that first they extract the simple features of the image (lines, curves, etc.), and then in the following layers, they extract the more complex features. This algorithm converts the picture to a picture with less dimension to simplify the classification process.

3.3.2. Architecture of the model

The pre-processed data was split to train test and validation. First, it was split into 30 percent for test and validation and 70 percent for the train set. Then from the 30 percent again, we split to 30 percent of test and 70 percent of validation, and we feed these data to the model explained in the following:

The first layer is a Conv1D, with the 512 and 20 as the filter and kernel sizes, respectively, and the activation function “Relu.”
The second layer is a Maxpooling1D with size 8.
Then the input is flattened in the flattened layer.
The last two layers are a dense, size seven, and a SoftMax layer, respectively.
The architecture is shown in figure 9.

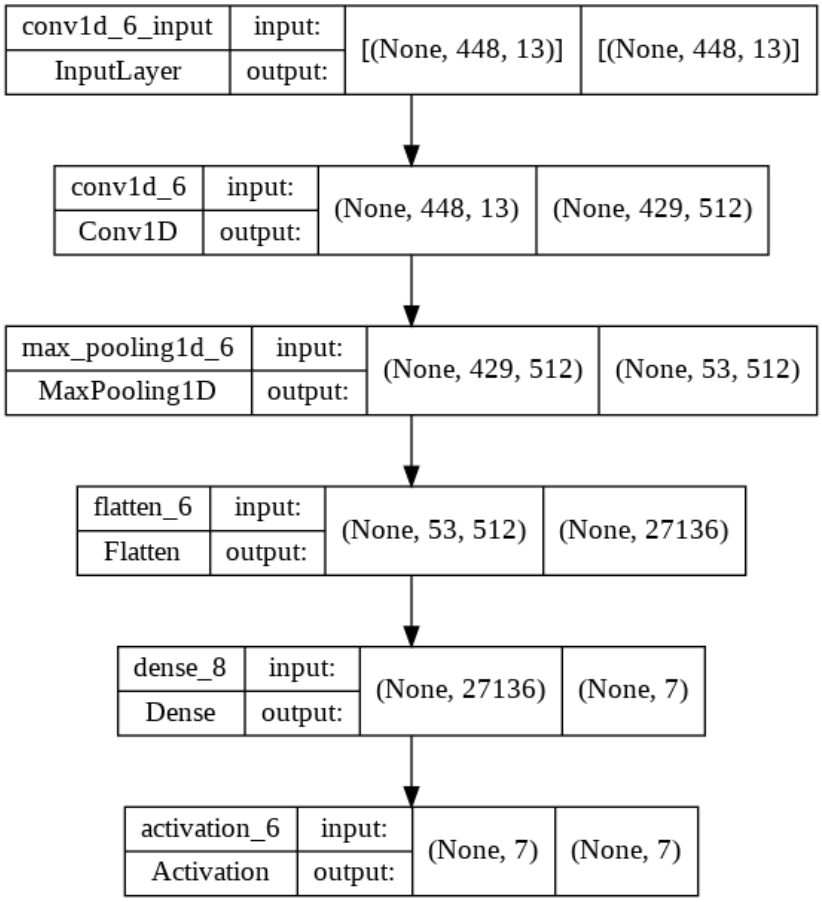

Figure 9.Model architecture

The optimizer for this algorithm is RMSprop, with 0.0005 as the learning rate. Also, to avoid overfitting, there are several methods. In this study, early stopping is used for this purpose, which stops the training process when the monitored metric has stopped improving. The parameters for early stopping are as follows:

Monitor = “val_loss”

Patience = 30 (stop training after 30 epochs with no improvement)

Min_delta = 0.0001 (minimum amount of change, if the changes after 30 epochs are less than 0.0001, training will be stopped.)

To compile the model, the loss is the “categorical_crossentropy”; for metrics, the “categorical_accuracy” is used. Finally, we fit the model with batch_size = 32 and epochs = 100. The hyperparameters for this model were set manually by trying many of them and choosing the one with the best results.

4. Results and discussions

4.1. Result

The measurements used for this model are accuracy, precision, recall, f1_score, and ROC_AUC, which can be seen in the tables below:

Table 1.Accuracy and loss for test and validation sets

	Loss	Accuracy
Validation	4.8262	0.6157
Test	4.3436	0.5935

Table 2.Precision, recall, f1_score, ROC_AUC

	precision	Recall	F1-score	ROC-AUC
macro	0.59930	0.59422	0.58515	0.8922
weighted	0.63454	0.59349	0.60440	0.8906

In the next 2 plots we can see the loss and accuracy of the model for the training and validation sets.

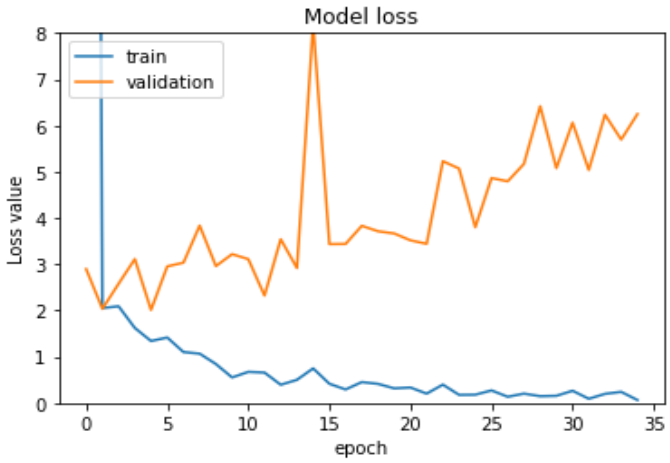

Figure 11.Loss of the model

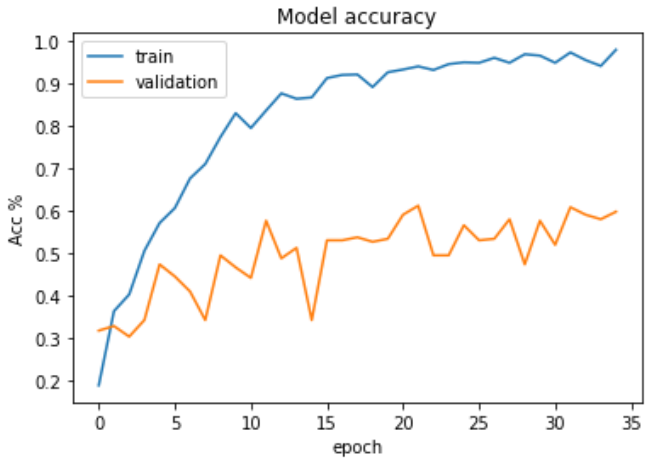

Figure 10.Accuracy of the model

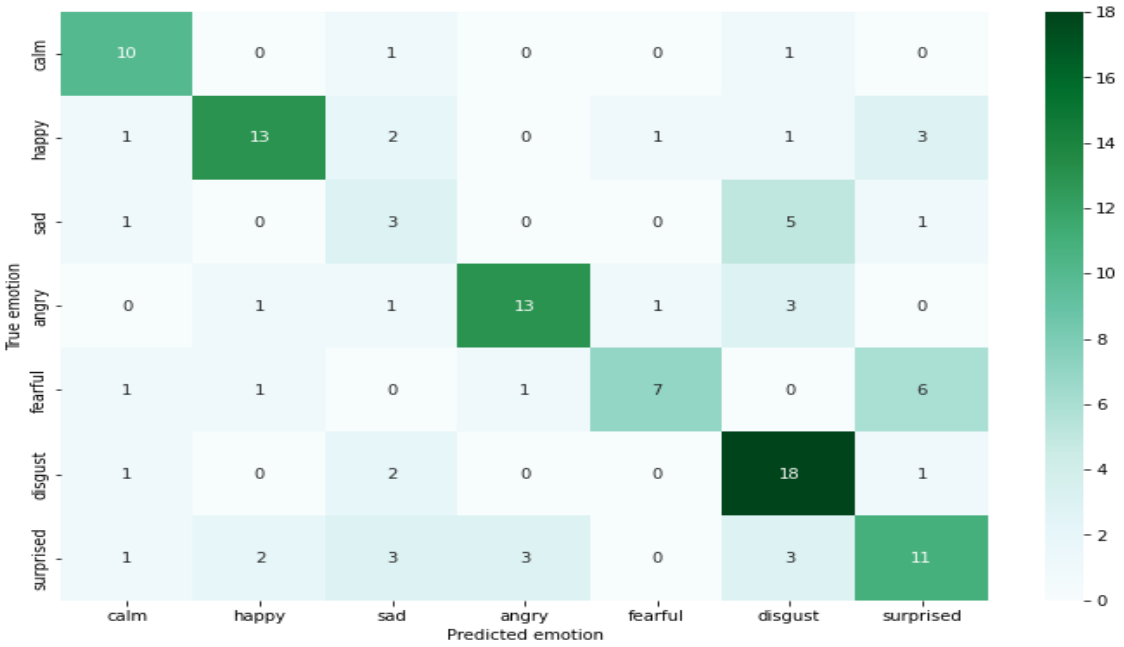

Figure 12.Confusion matrix

4.2. Discussion

For this task, we tried several architectures, and we have some observations. Firstly, although a regularization technique is used, we can see that the model is over-fitting, even if the parameters of the early stopping change. The dropout layer was also used in one of the architectures but did not enhance the performance.

The more complex the architecture, the worse the model performs, and a smaller number of layers was better. Because of the dataset's size, the smaller datasets need more simple architectures.

Regarding the results, we can see that the most misclassification happens when we have fearful emotions, but the model predicts it as surprising. The second one is when we have sad, but the model classifies it as disgust.

Overall, we can accept the initial hypothesis that CNN performs well for speech recognition tasks, although the accuracy is not that high here, and we had over-fitting. In further research, we will modify the architecture to have better accuracy and prevent over-fitting.

4.2.1. Future work

The first problem was the dataset size; maybe with some data augmentation or combining two or more datasets, we can address this issue. Also, it was so biased because all were with North American accents, so the model might not predict that well for other accents.

Secondly, it would be great if we had some real-time voice recognition to predict the emotions of any speech given at the time. Further work should be done to tackle the over-fitting issue and increase the accuracy and other evaluation metrics. Some grid searches can also find the best hyper parameters instead of doing them manually.

Finally, we are going to ensemble the model for this paper with an RNN model to see the results and check if they perform better individually or not.

5. Conclusion

Speech has been the strongest communication tool between humans, and nowadays, we consider it a tool to communicate with machines. Recognizing the emotions behind the words while speaking helps us in numerous areas, as mentioned in the introduction. Convolutional neural networks perform this classification task very well because they extract the features and recognize the patterns.

References

[1] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems,* vol. 52, no. 3, pp. 1467-1478, 2013.

[2] A. B. Ingale and D. Chaudhari, "Speech emotion recognition," *International Journal of Soft Computing and Engineering (IJSCE),* vol. 2, no. 1, pp. 235-238, 2012.

[3] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "On the robustness of speech emotion recognition for human-robot interaction with deep neural networks," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018: IEEE, pp. 854-860.

[4] M. Nasir, B. R. Baucom, P. Georgiou, and S. Narayanan, "Predicting couple therapy outcomes based on speech acoustic features," *PloS one,* vol. 12, no. 9, p. e0185123, 2017.

[5] S. Yamamoto, Y. Yoshitomi, M. Tabuse, K. Kushida, and T. Asada, "Detection of baby voice and its application using speech recognition system and fundamental frequency analysis," in *Proceedings of the 10th WSEAS international conference on Applied computer science*, 2010, pp. 341-345.

[6] S. Kavitha, N. Sanjana, K. Yogajeeva, and S. Sathyavathi, "Speech Emotion Recognition Using Different Activation Function," in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2021: IEEE, pp. 1-5.

[7] Y.-L. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," in *2005 international conference on machine learning and cybernetics*, 2005, vol. 8: IEEE, pp. 4898-4901.

[8] S. Xu, Y. Liu, and X. Liu, "Speaker recognition and speech emotion recognition based on GMM," in *3rd international conference on electric and electronics*, 2013: Atlantis Press, pp. 434-436.

[9] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2015: IEEE, pp. 4749-4753.

[10] https://zenodo.org/badge/DOI/10.5281/zenodo.6759664.svg/DOI: https://doi.org/10.5281/zenodo.6759664

[11] N. R. Kanth and S. Saraswathi, "Efficient speech emotion recognition using binary support vector machines & multiclass SVM," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, 2015: IEEE, pp. 1-6.

[12] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*, 2017: IEEE, pp. 1-5.

[13] N. Hajarolasvadi and H. Demirel, "3D CNN-based speech emotion recognition using k-means clustering and spectrograms," *Entropy,* vol. 21, no. 5, p. 479, 2019.

[14] A. B. A. Qayyum, A. Arefeen, and C. Shahnaz, "Convolutional neural network (CNN) based speech-emotion recognition," in *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, 2019: IEEE, pp. 122-125.

[15] T. Anvarjon and S. Kwon, "Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features," *Sensors,* vol. 20, no. 18, p. 5212, 2020.

[16] S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors,* vol. 20, no. 1, p. 183, 2019.

[17] https://www.analyticsvidhya.com/blog/2021/05/convolutional-neural-networks-cnn/