Original Article

Analyzing Vocal Patterns To Determine Emotions Using Machine Learning

U.Udayakumar[1*], Surya Susan Thomas[2], J.Jebamalar Tamilselvi[3]

[1*]Assistant Professor, Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology, Ramapuram Campus, Chennai – 600089, Tamilnadu, India,
udaya.vijay13@gmail.com
[2]Assistant Professor, Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology, Ramapuram Campus, Chennai – 600089, Tamilnadu, India susann.research@gmail.com
[3]Associate Professor, Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology, Ramapuram Campus, Chennai – 600089, Tamilnadu, India jebamalj@srmist.edu.in

***Corrresponding Author:** Dr Rachit Garg

*Assistant Professor, Department of Computer Science, College of Science and Humanities, SRM Institute of Science and Technology, Ramapuram Campus, Chennai – 600089, Tamilnadu, India,
udaya.vijay13@gmail.com
Doi:10.47750/Pnr.2022.13.S08.528

Abstract

The possibility to fluctuate vocal sounds to deliver discourse is one of the significant highlights which separate people from other living creatures. It is possible to distinguish human feeling by a few credits like pitch, tone, tumult, and vocal tone. It has frequently been seen that people express their feelings by changing different vocal credits during discourse age. Subsequently, recognizable proof of human feelings utilizing voice and discourse examination has a reasonable chance and might be gainful in working on human conversational abilities. One can follow an algorithmic methodology for discovery and investigation of human feelings with the assistance of voice and discourse handling. The proposed approach has been created with the goal of consolidation with advanced AI frameworks for further developing human-PC collaborations with the assistance of AI models SVM (Support Vector Machine) and CNN (Convolution Neural Networks) by the extraction of MFCC features.

Keywords: Conventional Neural Networks (CNN), Support Vector Machine (SVM), MFCC(Mel Frequency Cepstral Coefficient)

INTRODUCTION

Detecting emotions is one of the most important marketing strategy in today's world. One could personalize different things for an individual specifically to suit their interest. For this reason, it was decided to do a research where one could detect a person's emotions just by their voice which will let us manage many AI related applications. Some examples could be including call centers to play music when one is angry on the call. Another could be a smart car slowing down when one is angry or fearful. As a result this type of application has much potential in the world that would benefit companies and also even safety to consumers. speech data which is available in three different format : Audio Visual – Video with speech, Speech – Audio,Visual – Video only. The Audio only zip file dealing with finding emotions from speech. The zip file consisted of around 1500 audio files which were in wav format. Each audio file has a unique identifier at the 6th position of the file name which can be used to determine the emotion the audio file consists.Five different emotions are included in the dataset comprising of Calm, Happy, Sad, Angry and Fearful. Librosa library in Python to process and extract features from the audio files. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. Using the librosa library we were able to extract features i.e MFCC(Mel Frequency Cepstral Coefficient). MFCCs are a feature widely used in automatic speech and speaker recognition. We also separated out the females and males voice by the using the identifiers provided in the website. This was because as experiment we found out that separating male and female voices increased by 15%. It could be because of the pitch of the voice was affecting the results.

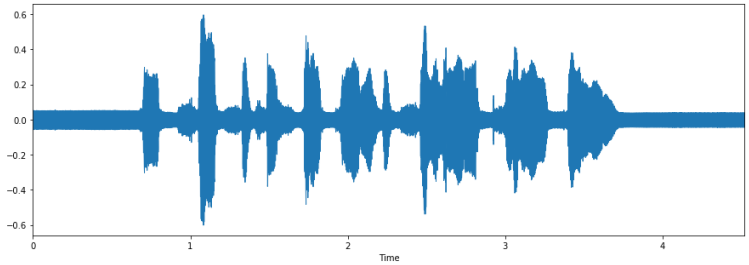


Fig 1: *Automatic speech and speaker recognition*

Each audio file gave us many features which were basically array of many values. These features were then appended by the labels which we created in the previous step. The next step involved dealing with the missing features for some audio files which were shorter in length. We increased the sampling rate by twice to get the unique features of each emotional speech. We have not increase the sampling frequency even more since it might collect noise thus affecting the results.

Speech Emotion Recognition

Speech Emotion Recognition (SER) is a system that can identify the emotion of different audio samples. From the description, this task is similar to text sentiment analysis, and both also share some applications since they differ only in the modality of the data – text versus audio. Like sentiment analysis, you can use speech emotion recognition to find the emotional range or sentimental value in various audio recordings such as job interviews, caller-agent calls, streaming videos, and songs. Moreover, even music recommendation or classification systems can cluster songs based on their mood and recommend curated playlists to the user. It is safe to assume that the complex algorithms of Spotify and YouTube also have an SER component that helps in music recommendations. From a machine learning perspective, speech emotion recognition is a classification problem where an input sample (audio) needs to be classified into a few predefined emotions. Of course, the challenge in this problem goes beyond technical – how does one even define emotion and consistently decide the class given an audio sample that can be ambiguous to even humans. The issue is more pressing for dataset creators, but it also becomes essential while evaluating a trained model. Further below, we will see that our dataset contains two similar-sounding emotions, “calm” and “neutral,” which can be tricky for even humans to ascertain in ambiguous cases. Meanwhile, “angry” and “happy” have prominent differences that the model can quickly learn. Human speech contains several features that the listener interprets to unpack the rich information transmitted by the speaker. The speaker also inadvertently shares tone, energy, speed, and other acoustic properties, which helps capture the subtext or intention and literal words.

Proposed Model

RNN/LSTMs: The models perform computations on a timestep sequence, meaning they can remember past data from the same sample while processing the next timestamp. Numeric features are fed to a neural network that generates an output logit vector. The output features can be mapped to text data using a decoding technique such as HMMs or Connectionist Temporal Classification (CTC).

Attention-based models: These are now the most used models for any task that involves mapping two data formats. An attention-based model can use previously predicted sequences and learn the mapping of new ones using an encoder-decoder approach.

Listen-Attend-Spell (LAS): This was one of the first approaches to combine the above two methods by creating an encoder that learns features using bidirectional LSTMs. Next, the decoder is designed to be an attention-based unit that learns from the learned representation of the encoder to produce an output probability for the next character sequence. For the classification problem of Speech Emotion Recognition, LSTMs or their more complicated versions are used when dealing with MFCCs as time-series data. They capture the changes in features over time for a given speech sample and model the behavior to predict the emotion class.

MLP Model: The MLP model we created had a very low validation accuracy of around 25% with 8 layers, softmax function at the output, batch size of 32 and 550 epochs.

Experimental Results

The data contains 3-second audio clips spoken of the same two sentences by 24 different actors over an emotional range of 7 emotions. Moreover, 12 male and 12 female actors give the data a more diverse and challenging range. Thus there are a total of 1440 samples.

- 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)

Analysis:

MLP Model: The MLP model we created had a very low validation accuracy of around 25% with 8 layers, softmax function at the output, batch size of 32 and 550 epochs.

```
In [60]: train[255:265]
```

Out[60]:

	4	5	6	7	8	9	...	121	122	123	124	125	126	127	128	129	0
255	0.582	0.243815	0.234133	0.220812	0.222221	0.232087	...	0.248799	0.253912	0.260256	0.257698	0.258209	0.256242	0.255648	0.255648	0.255701	angry
256	0.521	0.285065	0.291352	0.303514	0.308232	0.328804	...	0.234485	0.228035	0.216631	0.214859	0.212437	0.213037	0.218348	0.223208	0.224450	fearful
257	0.765	0.108862	0.103840	0.101478	0.107730	0.103912	...	0.066940	0.036635	0.027208	0.036532	0.053178	0.065569	0.057186	0.039764	0.021314	angry
258	0.141	0.074467	0.089486	0.088280	0.092139	0.093846	...	0.054423	0.053604	0.055540	0.058426	0.060729	0.068808	0.088886	0.098216	0.090357	sad
259	0.724	0.281591	0.296421	0.285957	0.260214	0.257237	...	0.299710	0.291853	0.291916	0.299710	0.299710	0.299710	0.287766	0.252755	0.243608	happy
260	0.779	0.330779	0.330779	0.330779	0.330779	0.330779	...	0.288739	0.287423	0.283312	0.291878	0.305482	0.321055	0.327999	0.301280	0.300456	calm
261	0.433	0.169379	0.171645	0.179289	0.190308	0.182795	...	0.149075	0.147707	0.159900	0.184663	0.187635	0.168762	0.149145	0.130382	0.120786	neutral
262	0.036	0.238554	0.242728	0.229463	0.228398	0.243454	...	0.223064	0.207814	0.210600	0.210909	0.202713	0.192792	0.192630	0.195298	0.187149	happy
263	0.079	0.326079	0.305091	0.284397	0.274060	0.266039	...	0.156601	0.185422	0.202734	0.204833	0.213753	0.221158	0.222267	0.185138	0.151496	sad
264	0.975	0.172604	0.173216	0.167372	0.168891	0.178888	...	0.205757	0.200951	0.197044	0.193599	0.208915	0.228052	0.219472	0.205900	0.201549	surprised

Fig 2: *MLP model*

We built a Multi Perceptron model, LSTM model and CNN models. The MLP and LSTM were not suitable as it gave us low accuracy. As our project is a classification problem where were categorize the different emotions.

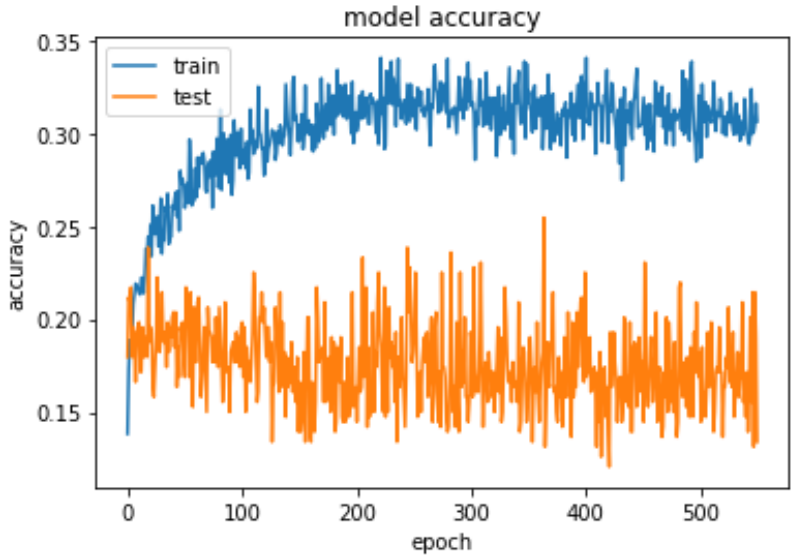


Fig 3: *Training and Testing the CNN-LSTM Network*

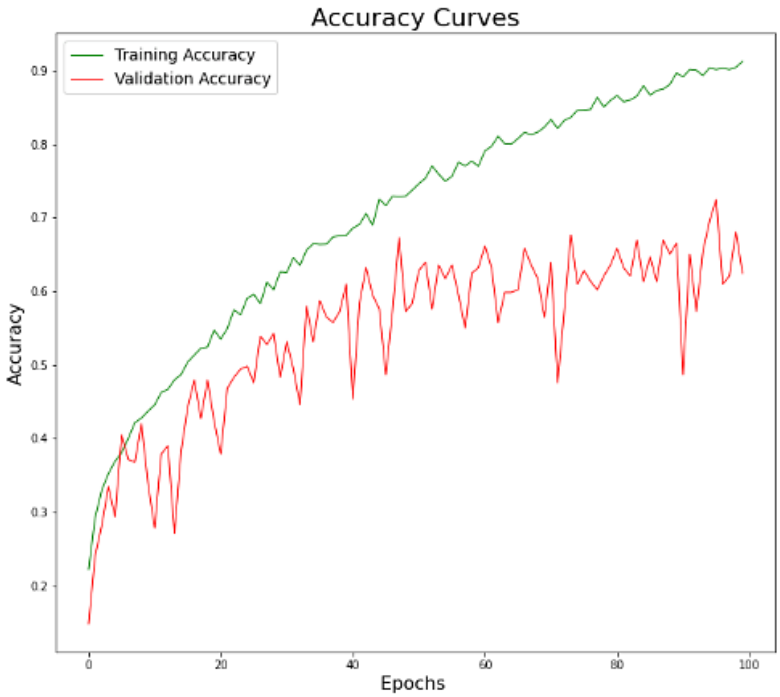
```
model.compile(optimizer=SGD(lr=0.01, decay=1e-6, momentum=0.8),
              loss='categorical_crossentropy', metrics=['accuracy'])

history = model.fit(X_train, y_train,
                    batch_size=64,
                    epochs=100,
                    validation_data=(X_test, y_test))
```

The final test and train results were as follows:

```
loss: 0.2290 - accuracy: 0.9212 - val_loss: 1.3690 - val_accuracy: 0.6691
```

So we see a slight improvement in training than before. While the validation accuracy has dropped, we have scope for improvement given the ample size of the dataset and superior model complexity. We did not have this scope in the earlier model as it risked overfitting by increasing the number of layers. For more insight into the training and testing, plot the loss and accuracy curves from model history as we did before. The loss curves show how validation loss fluctuates highly. An early stopping strategy can help stop the training at the best loss or accuracy performance if the performance does not improve after a fixed number of steps.



Conclusion

After building numerous different models, it has been found that this model gives the best output for the emotion classification problem. A validation accuracy of 70% is achieved with this model. The model could perform better if provided with more test data and the proposed model also skillfully distinguished between a male and a female voice. Predictive analysis of the proposed model is noteworthy. In the future , a sequence-to-sequence model to generate voice based on different emotions could be built to enhance this work.

References

1. S. Rama, Survey on Speech Emotion Recognition using Neural Network and MLP Classifier, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India.
2. S. Lalitha, Abhishek Madhavan, Bharath Bhushan, Srinivas Saketh, Survey on Speech Emotion recognition, Published in: 2014 International Conference on Advances in Electronics Computers andCommunications.
3. Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof and Mohamed Ali Mahjoub, Survey on Speech Emotion Recognition: Methods and Cases Study, University of Maine, Le Mans University, France, LATIS Laboratory of Advanced Technologies and Intelligent Systems, University of Sousse, Tunisia, Higher Institute of Applied Sciences and Technology of Sousse, University of Sousse, Tunisia.
4. Building a Vocal Emotion Sensor with Deep Learning https://towardsdatascience.com/building-a-vocal-emotion-sensor-with-deep-learning-bedd3de8a4a9
5. A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling, Sadil Chamishka, Ishara Madhavi, Rashmika Nawaratne, Damminda Alahakoon, Daswin De Silva, Naveen Chilamkurti & Vishaka Nanayakkara.
6. ANALYSING VOCAL PATTERNS TO DETERMINE EMOTIONS USING LSTM Dr. Suprava Patnaik,Ritika Kaushal,Shivani Kadam,Dipti Kathayat,Vaibhav Nardekar
7. Automatic Speech Emotion Recognition Using Machine Learning Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Cleder
8. Reza Chu, Speech Emotion Recognition with Convolutional Neural Network