SPEECH EMOTION RECOGNITION USING DEEP LEARNING

U D Prasan[1*], R Narayana Rao[2]

[1]Department of Computer Science and Engineering, Aditya Institute of Technology and Management, Tekkali - 532201, India.
[2]Dept of CSE, Potti Sriramulu College of Engineering and Technology
* Corresponding author: udprasanna@gmail.com

Abstract

Speech Emotion Recognition (SER) is the act of recognizing human emotion from speech. Automatic speech recognition is an active field of study in artificial intelligence and machine learning whose aim is to generate machines that communicate with people via speech.Speech is an information-rich signal that contains paralinguistic information that is conveyed by speech. A model is designed that could recognize the emotion in a speech sample. Emotion recognition is done using Support Vector Machine (SVM) as well as Multi-Layer Perceptron (MLP) Neural Network. The training for SVM was much faster when compared to MLP.
In this project, we implement model using MLP classifier using voice quality features extracted from the RAVDESS -Ryerson Audio-Visual Database of Emotional Speech and Song Database. We will load the data, extract features from it, then split the dataset into training and testing sets. Then we will initialize an MLP Classifier and train the model. Finally, we'll calculate the accuracy of our model.

Keywords: Speech Emotion Recognition, Automatic speech recognition, artificial intelligence and machine learning, Support Vector Machine, Multi-Layer Perception, RAVDESS

Introduction

Speech is the fast and best normal way of communicating amongst human. Speech is series sequence of words of pre-established language and it is an essential medium for communication. Speech technology is a computing technology that empowers an electronic device to recognize, analyse and understand spoken words or audio. Although there is a significant improvement in speech recognition but still researchers are away from natural interplay between computer and human since computer is not capable of understanding the human emotional state.

At present, speech emotion recognition was an emerging crossing field of artificial intelligence. Developing machines that understand paralinguistic information, such as emotion, facilitates the human- machine communication as it makes the communication more clear and natural. Speech emotion recognition was a technology that extract emotional feature from speech signals by computer and contrasts and analyses the characteristic parameters and the emotional change acquired. Finally, the law of speech and emotion was concluded and speech emotional states were judged according to the law.

The research was widely applied in human-computer interaction, interactive teaching, entertainment, security fields, and so on. Speech emotion processing and recognition system was generally composed of three parts, which were speech signal acquisition, feature extraction, and emotion recognition. Different features of speech such as speaking rate, intonation, energy, formant frequencies, frequency (pitch), loudness, duration, and spectral Characteristic is used to encode emotion in speech.

Literature Survey

[1] A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM Authors: Chenchen Huang, Wei Gong, Wenlong Fu, and Dongyu Feng Published on 12 August 2014.

This paper selected 1200 sentences which contain sadness, anger, surprise, and happiness, four basic emotions for training and recognition. This paper used 40% of the voice data for training and 60% of the voice data for testing. The experimental group is speech emotion recognition model established by deep belief network via extracting phonetic characteristics. The control group is established by extracting traditional speech feature parameters to the speech emotion recognition model under the condition of the same emotional speech input. At last, contrasted and analyzed the experimental data for the conclusion.

[2] Speech Emotion Recognition Using Support Vector Machine
Authors: Yashpalsing Chavhan Student VIT, Pune India, M. L. Dhore Professor VIT, Pune India Pallavi Yesaware Student VIT, Pune India In this paper Berlin emotion database of German language is used for feature extraction. MFCC and MEDC features are extracted from a speech files in .wav format. Berlin Emotion database contains 406 speech files for five emotion classes. Emotion classes Anger, sad, happy, neutral, fear are having 127, 62, 71, 79 and 67 speech utterance respectively. The LIBSVM is trained on MFCC and MEDC feature vectors using RBF and Polynomial kernel functions It is also observed that results from LIBSVM by using RBF and Polynomial kernel function are 93.75% and 96.25% respectively Regarding LIBSVM using RBF and Polynomial kernels it is observed that by changing the parameters of a kernel functions better results can be obtain.

[3] Classification on Speech Emotion Recognition -A Comparative Study
Authors: Theodoros Iliou, Christos-Nikolaos Anagnostopoulos .This paper it describes a comparative analysis of four classifiers for speech signal emotion recognition. Recognition was performed on emotional Berlin Database. This work focuses on speaker and utterance (phrase) dependent and independent framework. One hundred thirty three (133) sound/speech features have been extracted from Pitch, Mel Frequency Cepstral Coefficients, Energy and Formants Concluding this paper, the 35-input vector, seems to be quite promising for speaker independent recognition in terms of high and low arousal emotions when tested in Berlin database. Nevertheless, this vector is not sufficient enough to describe the intra class variations of the two hyper classes. The major finding in this work is that PNN classifier achieved almost perfect classification (94%) in speaker dependent emotion recognition.

[4] A Comparative Study On Speech Emotion
Authors: Anushka Sandesara, Shilpi Parikh, Pratyay Sapovadiya, Mrugendrasinh Rahevar In this paper, aim is to gather and provide comprehensive information about various significant methods which have been developed and determined the best suited methods for speech emotion recognition. Information regarding particular technology, its techniques, their limitations, type of result generated is required before implementing any method which is provided in this paper.

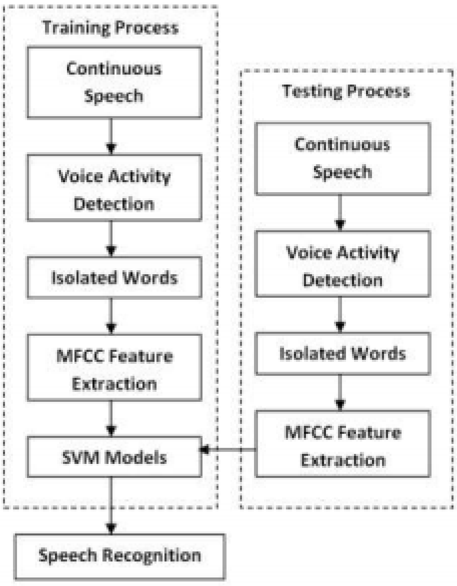
[5] Human Speech Emotion Recognition

Authors: Maheshwari Selvaraj, Dr.R.Bhuvana, S.Padmaja,Assistant Professor, Department of Computer Application, Department of Software Application, In this paper, the concept implemented was emotion recognition using MFCC approach using Radial basis function network. Support vector machine is used to classifying the gender in this work. Gender speech classifier is based on pitch analysis. MFCC approach for emotion recognition from speech is a stand-alone approach which does not require calculation of any other acoustic features and produce more accurate results. Hence proved that the Radial basis function network recognize emotions more accurately than the Back Propagation Network.

Existing System

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. SVM is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. Simply put, it does some extremely complex data transformations, then figures out how to seperate your data based on the labels or outputs you've defined



Disadvantages:

□ It doesn’t perform well when we have large data set because the required training time is higher

□ It also doesn’t perform very well, when the data set has more noise i.e. target classes are overlapping

□ SVM doesn’t directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.

Proposed System

Multi Layer Perceptron (MLP)

Multilayer perceptron MLP is a deep learning method. A multilayer perceptron (MLP) is a feed forward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers. MLP uses backpropogation for training the network. A multilayer perceptron is a neural network connecting multiple layers in a directed graph, which means that the signal path through the nodes only goes one way. Each node, apart from the input nodes, has a nonlinear activation function.

Construction of MLP Classifier[4]:
Step 1: Initialization of MLP Classifier by defining and initialization of required parameters.
Step 2: Train the Neural Network with the provided data.
Step 3: Prediction of output
Step 4: Calculating the accuracy

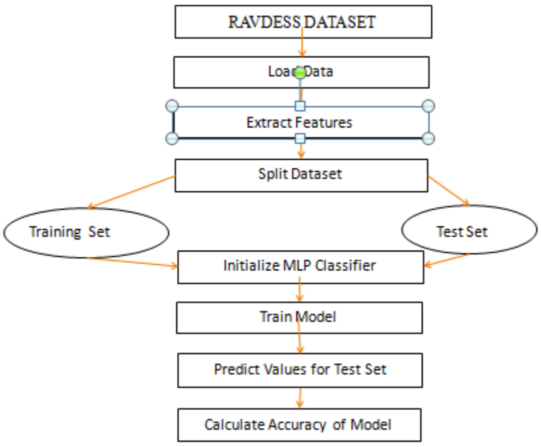
Salient points of Multilayer Perceptron (MLP) in Scikit-learn .There is no activation function in the output layer. For regression scenarios, the square error is the loss function, and cross-entropy is the loss function for the classification It can work with single as well as multiple target values regression. Unlike other popular packages, likes Keras the implementation of MLP in Scikit doesn't support GPU. We cannot fine-tune the parameters like different activation functions, weight initializers etc. for each layer.

Algorithm
□ Class MLP Classifier implements a multi-layer perceptron (MLP) algorithm that trains using Backpropagation
□ MLP trains on two arrays: array X of size (n_samples, n_features), which holds the training samples represented as floating point feature vectors; and array y of size (n_samples,), which holds the target values (class labels) for the training samples:
Step1: import MLPClassifier from sklearn.neural_network
Step 2: Assign [[0., 0.], [1., 1.]] to X (X- [[0., 0.], [1., 1.]])
Step 3: Assign [0, 1] to y (y -[0, 1])
Step 4: Assign MLP Classifier (solver='lbfgs', alpha=1e-5, ... hidden_layer_sizes=(5, 2), random_state=1) to clf
Step 5: clf.fit(X, y)
Step 6: After fitting (training), the model can predict labels for new samples:
clf.predict([[2., 2.], [-1., -2.]]) array([1, 0])

Advantages and Disadvantages
The advantages of Multi-layer Perceptron are:
□ Capability to learn non-linear models.
□ Capability to learn models in real-time (on-line learning) using partial_fit.

The disadvantages of Multi-layer Perceptron (MLP) include:
□ MLP with hidden layers have a non-convex loss function where there exists more than one local minimum. Therefore different random weight initializations can lead to different validation accuracy.
□ MLP requires tuning a number of hyperparameters such as the number of hidden neurons, layers, and iterations.
□ MLP is sensitive to feature scaling.

Architecture



Overall Architecture

The above architecture describes the process of building the model. Initially, Make Necessary Imports-such as Librosa, Soundfile, numpy, os, glob, pickle, Sklearn. Take Dataset named RAVDESS contains emotions - calm, anger, neutral, happy, sad, fearful, disgust, surprised. Load data from RAVDESS dataset and Extract features from it. Feature Extraction is done using feature extraction techniques : MFCC, Chroma, Mel. Split extracted dataset into Training set and Testing set and find number of features extracted. In a dataset, a training set is implemented to build up a model, Data points in the training set are excluded from the test set. And test set is to validate the model built.

Initialize Model: MLP Classifier and fit/Train it. Predict values for test set. Calculate Accuracy of the model.

RAVDESS Dataset

Datasets: A collection of instances is a dataset and when working with machine learning methods we typically need a few datasets for different purposes.The RAVDESS is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing 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. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats. The set of 7356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America.[9] A further set of 72 participants provided test-retest data. High levels of emotional validity and test-retest intrarater reliability were reported. Corrected accuracy and composite "goodness" measures are presented to assist researchers in the selection of stimuli. Training Dataset A dataset that we feed into our machine learning algorithm to train our model.

Testing Dataset A dataset that we use to validate the accuracy of our model but is not used to train the model. It may be called the validation dataset.

System Requirements and Technical Discussions

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Its high-level built in data structures, combined with dynamic typing and dynamic binding, make it very attractive for Rapid Application Development. Python combines remarkable power with very clear syntax.

Sklearn:
Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, and clustering and dimensionality reduction via a consistence interface in Python

Librosa:
Librosa is a Python package for music and audio analysis. Librosa is basically used when we work with audio data like in music generation (using LSTM's), Automatic Speech Recognition.

Matplotlib:

It is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

PIP:
pip is a de facto standard package-management system used to install and manage software packages written in Python. Many packages can be found in the default source for packages and their dependencies — Python Package Index. Most distributions of Python come with pip preinstalled.

PIP:

pip is a de facto standard package-management system used to install and manage software packages written in Python. Many packages can be found in the default source for packages and their dependencies — Python Package Index. Most distributions of Python come with pip preinstalled

Sound File:

SoundFile is an audio library based on libsndfile, CFFI and NumPy.Soundfile can read and write sound files. File reading/writing is supported through libsndfile, which is a free, cross-platform, open-source (LGPL) library for reading and writing many different sampled sound file formats that runs on many platforms including Windows, OS X, and Unix.

Chroma:
The term chroma feature or chromagram closely relates to the twelve different pitch classes. Chroma-based features, which are also referred to as "pitch class profiles", are a powerful tool for analyzing music whose pitches can be meaningfully categorized (often into twelve categories) and whose tuning approximates to the equal-tempered scale. One main property of chroma features is that they capture harmonic and melodic characteristics of music, while being robust to changes in timbre and instrumentation.

Mel spectrum:

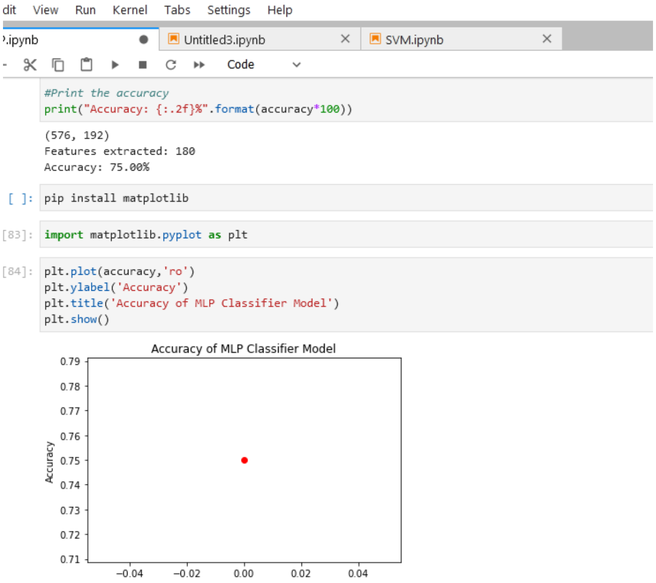
Mel spectrum is computed by passing the Fourier transformed signal through a set of band-pass filters known as Mel-filter bank. A Mel is a unit of measure based on the human ears perceived frequency. It does not correspond linearly to the physical frequency of the tone, as the human auditory system apparently does not perceive pitch linearly.

MFCC:

Mel Frequency Cepstral Coefficents (MFCCs) is a way of extracting features from an audio. The MFCC uses the MEL scale to divide the frequency band to sub-bands and then extracts the Cepstral Coefficents using Discrete Cosine Transform (DCT). MEL scale is based on the way humans distinguish between frequencies which makes it very convenient to process sounds.

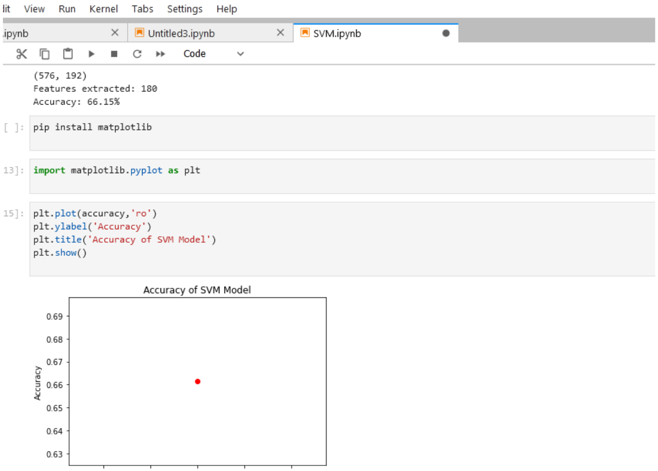
Results and Discussions

The graph showing the accuracy of MLP Classifier. The accuracy generated in the model developed using MLP classifier is 75%.



```
#Print the accuracy
print("Accuracy: {:.2f}%".format(accuracy*100))

(576, 192)
Features extracted: 180
Accuracy: 75.00%

[ ]: pip install matplotlib

[83]: import matplotlib.pyplot as plt

[84]: plt.plot(accuracy,'ro')
plt.ylabel('Accuracy')
plt.title('Accuracy of MLP Classifier Model')
plt.show()
```

Output of MLP Classifier



```
(576, 192)
Features extracted: 180
Accuracy: 66.15%
```

```
[ ]: pip install matplotlib
```

```
13]: import matplotlib.pyplot as plt
```

```
15]: plt.plot(accuracy,'ro')
     plt.ylabel('Accuracy')
     plt.title('Accuracy of SVM Model')
     plt.show()
```

Output of SVM

The graph showing the accuracy of SVM. The accuracy generated in the model developed using SVM is 66.15%. By observing the graphs of both proposed system and existing system, i.e. MLP Classifier and SVM, we can conclude that MLP Classifier (proposed system) is better one.

Conclusion And Future Scope

The emerging growth and development in the field of AI and machine learning have led to the new era of automation. Most of these automated devices work based on voice commands from the user. Many advantages can be built over the existing systems if besides recognizing the words, the machines could comprehend the emotion of the speaker (user). Some applications of a speech emotion detection system are computer-based tutorial applications, automated call center conversations, a diagnostic tool used for therapy and automatic translation system.

In this thesis, the steps of building a speech emotion detection system were discussed in detail and some experiments were carried out to understand the impact of each step. Initially, the limited number of publically available speech database made it challenging to implement a well-trained model. Next, several novel approaches to feature extraction had been proposed in the earlier works, and selecting the best approach included performing many experiments. Finally, the classifier selection involved learning about the strength and weakness of each classifying algorithm with respect to emotion recognition.

At the end of the experimentation, it can be concluded that an integrated feature space will produce a better recognition rate when compared to a single feature. The literature in speech emotion detection is not very rich and researchers are still debating what features influence the recognition of emotion in speech. There is also considerable uncertainty as to the best algorithm for classifying emotion, and which emotions to class together. We reviewed and discussed various speeches emotional recognition systems based approaches. We also compare its performance in terms of classifier, features, recognition rate, and datasets. Well-design classifiers have obtained high classification accuracies between different types of emotions. We proposed an emotion recognition model using MLP classifier and generated accuracy of the model. We proved that MLP classifier is better method than SVM by calculating accuracy of both the models. MLP Classifier generated 75% accuracy whereas SVM generated 66.15% accuracy. From obtained graphs of SVM and MLP classifier, we can conclude MLP classifier is better than SVM. In Future Advancements, some more deep learning techniques like CNN can be implemented.Generate more efficient models with high accuracy.

References

[1]Yashpalsing Chavhan Student VIT, Pune India,M. L. Dhore Professor VIT, Pune India ,Pallavi Yesaware Student VIT, Pune India,Speech Emotion Recognition Using Support Vector Machine, Article in International Journal of Computer Applications ,Published on February 2010.

[2] Chenchen Huang, Wei Gong, Wenlong Fu, and Dongyu Feng'A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM', Published on 12 August 2014, Hindawi Publishing Corporation Mathematical Problems in EngineeringVolume 2014, Article ID 749604.

[3]Theodoros Iliou, Christos-Nikolaos Anagnostopoulo Cultural and Communication Department,University of the Aegean, 'Classification on Speech Emotion Recognition -A
Comparative Study' International Journal on Advances in Life Sciences, vol 2 no 1 & 2, year 2010.

[4] Anushka Sandesara, Shilpi Parikh, Pratyay Sapovadiya, Mrugendrasinh Rahevar UG Student, U & P U. Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Science and Technology, CHARUSAT, Changa, India 4Assistant Professor, U & P U. Patel Department of Computer Engineering, Chandubhai S. Patel Institute of Science and Technology, CHARUSAT, Changa, India A Comparative Study On Speech Emotion International Journal of Research in Engineering, Science and Management Volume-3, Issue-11, November-2020.

[5] Maheshwari Selvaraj, Dr.R.Bhuvana, S.Padmaja,Assistant Professor, Department of Computer Application, Department of Software Application, A.M.Jain College,Chennai, India,Assistant Professor, School of Computing Sciences,Vels University, Chennai,India Human Speech Emotion Recognition, International Journal of Engineering and Technology (IJET).

[6] Nithya Roopa S., Prabhakaran M, Betty.P, Speech Emotion Recognition using Deep
Learning, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-4S, November 2018.

[7] Darshan K.A, Dr. B.N. Veerappa**,**U.B.D.T. College of Engineering, Davanagere, Karnataka, India ,Dr. B.N. Veerappa, Department of Studies in Computer Science and Engineering, U.B.D.T. College of Engineering, Davanagere, Speech Emotion Recognition**,** International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 09 | Sep 2020.

[8] Theodoros Iliou, Christos-Nikolaos Anagnostopoulos, Cultural Technology and Communication Department University of the Aegean, Classification on Speech Emotion Recognition - A Comparative Study, International Journal on Advances in Life Sciences, vol 2 no 1 & 2, year 2010, http://www.iariajournals.org/life_sciences/

[9] Babak Joze Abbaschian , Daniel Sierra-Sosa and Adel Elmaghraby, Computer Science and Engineering Department, University of Louisville, Louisville, KY 40292, USA, Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models, Published on 10 February 2021, Sensors 2021, 21, 1249,https://www.mdpi.com/journal/sensors

[10] Sundarprasad, Neethu, Speech Emotion Detection Using Machine Learning Techniques (2018), San Jose State University ,SJSU Scholar Works ,Master's Projects.