International Journal of Research in Engineering, Science and Management
Volume 4, Issue 8, August 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Face Emotion and Audio Analysis using Machine Learning

A. Ashwin Siva[1*], C. Infan Chelsea[2], K. Kishore[3], A. Thiyagarajan[4]
[1,2,3]*Student, Department of Information Technology, Sri Venkateswara College of Engineering , Sriperumbudur, India*
[4]*Assistant Professor, Department of Information Technology, Sri Venkateswara College of Engineering , Sriperumbudur, India*

***Abstract*: Human emotions are mental states of feelings that arise spontaneously rather than through conscious effort and are accompanied by physiological changes in them which implies changes in their audio and way of speech. Some of critical emotions are Normal, happy, anger, sadness, fear and enthusiast. Emotion recognition from audio signal requires Audio feature extraction and visualizations, Training the model for accuracy calculation, Implementation process of CNN model, and classification of speech emotions. The feature vector consists of elements of the audio signal which characterize speaker specific features such as tone, pitch, energy, which is crucial to train the classifier model to recognize a particular emotion accurately. In order to computer modeling of human's emotion, a plenty of research has been accomplished. But still it is far behind from human vision system. In this paper, we are providing better approach to predict human emotions (Frames by Frames) using deep Convolution Neural Network (CNN) and how emotion intensity changes on the audio from low level to high level of emotion. In this algorithm, FERC-2013 database has been applied for training. The assessment through the proposed experiment confers quite good result and obtained accuracy may give encouragement to the researchers for future model of computer based emotion recognition system.**

***Keywords*: Artificial Intelligence, Machine Learning, Recurrent Neural Network, Transfer learning.**

1. Introduction

This section gives a detailed introduction about the domain and also discusses the various existing solutions available along with their disadvantages. It also emphasizes and highlights the need for the proposed system. Speech Emotion Recognition (SER) is an active area of research and a better way to communicate using among Human–Computer Interaction (HCI). Speech signals play an important role in various real-time HCI applications, such as clinical studies, audio

surveillance, lies detection, games, call centers, entertainment, and many more. However, the existing SER techniques still have some limitations, which include robust feature selection and advanced machine learning methods, for an efficient system. Thus, researchers are still working to find a significant solution in order to choose the right features and advance the Artificial Intelligence (AI) based classification techniques. Similarly, the background noise in a real-world voice could also be dramatically effective on the machine learning system. Nevertheless, the development of a decent speech-based emotion recognition system can easily increase the user experience in different areas with the HCI, such as AI cyber security and mobile health (mHealth). The AI model has better potential than classical model to recognize the emotional state of the speaker from their signals during speech and shows a considerable impact on the SER in order to imitate these emotions. Deep learning and the AI performed a significant improvement in the mobile health assistance field as well as increased their performance. Nowadays, the researchers utilize the deep learning approaches in order to solve the recognition problems, such as voice recognition, emotion recognition, gesture recognition, face recognition, and image recognition.

A. Overview

Speech Emotion Recognition (SER) is an active area of research and a better way to communicate using among Human–Computer Interaction (HCI). Speech signals play an important role in various real-time HCI applications, such as clinical studies, audio surveillance, lies detection, games, call centers, entertainment, and many more. However, the existing SER techniques still have some limitations, which include robust features election and advanced machine learning methods, for an efficient system. Thus, researchers are still working to find a significant solution in order to choose the right features and advance the Artificial Intelligence (AI) based classification techniques. Similarly, the background noise in a real-world voice could also be dramatically effective on the machine learning system. Nevertheless, the development of

a decent speech-based emotion recognition system can easily increase the user experience in different areas with the HCI, such as AI cyber security and mobile health (mHealth).The AI model has better potential than classical model to recognize the emotional state of the speaker from their signals during speech and shows a considerable impact on the SER in order to imitate these emotions. Deep learning and the AI performed significant improvement in the mobile health assistance field as well as increased their performance. Nowadays, the researchers utilize the deep learning approaches in order to solve the recognition problems, such as voice recognition, emotion recognition,

*Corresponding author: sivaashwin12345@gmail.com

gesture recognition, face recognition, and image recognition. The development of machine learning methods provides a new direction for addressing the problem demonstrated above. Typically, since the powerful ability of Recurrent Neural Network (RNN) to learn high abstract feature representations, it has been widely applied to the area of SER compared the accuracy of RNN and traditional models like SVM in speech emotion recognition, and found that RNN could effectively extract features and was more prominent in two-dimensional signal modeling.

B. Python

If you Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. It was created by Guido van Rossum during 1985- 1990.Like Perl, Python source code is also available under the GNU General Public License(GPL). This tutorial gives enough understanding on Python programming language. Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

Python is Interpreted − Python is processed at runtime by the interpreter. You do not need to compile your program before executing it. This is similar to PERL and PHP.

Python is Interactive − you can actually sit at a Python prompt and interact with the interpreter directly to write your programs.

Python is Object-Oriented: Python supports Object-Oriented style or technique of programming that encapsulates code within objects.

Python is a Beginner's Language: Python is a great language for the beginner-level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

1) Python features

- Easy-to-learn − Python has few keywords, simple structure, and a clearly defined syntax. This allows the student to pick up the language quickly.
- Easy-to-read − Python code is more clearly defined and visible to the eyes.
- Easy-to-maintain − Python's source code is fairly easy-to-maintain.
- A broad standard library − Python's bulk of the library is very portable and cross-platform compatible on UNIX, Windows, and Macintosh.

C. Syntax and Semantics

It has been suggested that this article be merged with Python syntax and semantics. Python is meant to be an easily readable language. Its formatting is visually uncluttered, and it often uses English keywords where other languages use punctuation. Unlike many other languages, it does not use curly brackets to delimit blocks, and semicolons after statements are allowed but are rarely, if ever, used. It has fewer syntactic exceptions and special cases than C or Pascal.

D. Machine Learning

Machine learning is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence. Machine learning algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so. Machine earning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or unfeasible to develop conventional algorithms to perform the needed tasks. A subset of machine learning is closely related to computational statistics, which focuses on making predictions using computers; but not all machine leaning is statistical learning. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning. Data mining is a related field of study, focusing on exploratory data analysis through unsupervised learning. In its application across business problems, machine learning is also referred to as predictive analytics.

The figure 1 involves the process of machine learning where the data is collected first then the data is pre-processed to clean the data next the model is trained with the given dataset and finally the model is evaluated using the accuracy of the model. The model if has lower accuracy the model can be improved to produce higher accuracy.

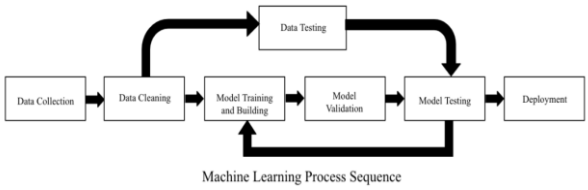

Fig. 1. Machine Learning Process

E. Indentation

Python uses whitespace indentation, rather than curly brackets or keywords, to delimit blocks. An increase in indentation comes after certain statements; a decrease in indentation signifies the end of the current block. Thus, the program's visual structure accurately represents the program's semantic structure. This feature is sometimes termed the off-side rule, which some other languages share, but in most languages indentation doesn't have any semantic meaning.

F. RNN

A Recurrent Neural Network (RNN) is a class of artificial neural networks where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. Derived from feedforward neural networks, RNNs can use their internal state (memory) to process variable length sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition.

The term "recurrent neural network" is used indiscriminately to refer to two broad classes of networks with a similar general structure, where one is finite impulse and the other is infinite impulse. Both classes of networks exhibit temporal dynamic

behavior. A finite impulse recurrent network is a directed acyclic graph that can be unrolled and replaced with a strictly feedforward neural network, while an infinite impulse recurrent network is a directed cyclic graph that cannot be unrolled.

Both finite impulse and infinite impulse recurrent networks can have additional stored states, and the storage can be under direct control by the neural network. The storage can also be replaced by another network or graph, if that incorporates time delays

or has feedback loops. Such controlled states are referred to as gated state or gated memory, and are part of long short-term memory networks (LSTMs) and gated recurrent units. This is also called Feedback Neural Network (FNN).

G. Existing System

In the existing system various techniques are proposed to analyze voice data. These techniques include Dynamic Time Warping (DTW), Hidden Markov Model(HMM). In the analysis and prediction of the audio file. The major drawback found in the existing work is less accuracy.

H. Proposed System

This section explains the proposed methodology, emotion database used for research, Inception model.

1) Emotion database

IEMOCAP corpus Database is prepared by the Speech Analysis and Interpretation Laboratory (SAIL), at the University of Southern California (USC) a new

corpus named "Interactive Emotional Dyadic Motion Capture Database" (IEMOCAP) issued in this paper. Since this data is rarely used, so this project explores more on this dataset. Corpus Data consists of ten actors in dyadic sessions with markers on the face, head, and hands, which provide detailed information about their facial expression and hand movements during scripted and spontaneous spoken communication scenarios.

The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions (happy, anger, sadness, fear and enthusiast state). Database consist of twelve hours of audio-visual data. We have chosen audio clips of various sessions.

Based on certain annotators, these audio clips of 10 seconds(approx.) are classified into one of the emotions classes. All the audio-visual data is divided into five sessions audio data in .wav format and video data in .mp4 format. During the sessions of capturing the data, actor's emotions is evaluated by various annotators into seven range of emotions. All the data are given along with the database

2) Transfer learning

Transfer learning is one of the machine learning models which uses the knowledge gained from solving one problem is incorporated to solve another problem. Itis evident that Transfer learning solves many problems within short interval of time. Transfer Learning is incorporated whenever there is any need to reduce computation cost, achieve accuracy with less training

3) Inception Net v3 model

Inception Net v3 Model is used to build an emotion recognition model. Inceptions evolved from GoogleNet Architecture with some enhancements. Inception model issued for automatic image classification and image labelling according to the image. Inception-v3 is used for image classification in Google Image Search. Inception-v3achieved top 5.6% error rate in ILSVRC 2012 classification challenge validation. Inception net consist of network in a network in a network which consist of three inception modules that are embed inside the inception architecture which helps in reduction of numerical array.

4) System setup

For performing the experiment, we have used system setup consist of Core i7 6thGeneration 3.7 GHz Processor, Samsung SSB of 512 GB memory space, NVIDIA GeForce GT 730 2GB GPU Card with Ubuntu 16.04 installed. For deep learning We have used Tensor Flow 1.5 for implementing the Inception net model and Tensor Board for visualizing the learning, graphs, histograms and so on.

5) Training method

The spectrograms were generated from the IEMOCAP are resized to 500 x 300.

More than 400 spectrograms were generated from all the audio files in the dataset. For each emotion, the training process was run for 20 epochs with a batch size set to 100.Initial learning rate was set to 0.01 with a decay of 0.1 after each 10 epochs. Training data model was performed on a single NVidia GeForce GT 730 with 2GB onboard memory. The training took around 35 minutes and the best accuracy was achieved after 28 epochs. On the training set, a loss of 0.71 was achieved, whereas 0.95loss was recorded on the test set. An accuracy of 92% was achieved per spectrogram. It is important to notice here that the overall accuracy is very low. These may be due to transfer learning used and less dataset for each class of emotion.

2. System Design

A. Speech Emotion Recognition (SER) System

In fig. 2 The idea behind speech recognition is to provide a means to transcribe spoken words into written text. There exist many approaches to achieve this goal. The simplest technique is to build a model for every word that needs to be recognized. Speech signal primarily conveys the words or message being spoken. Area of speech recognition is concerned with determining.

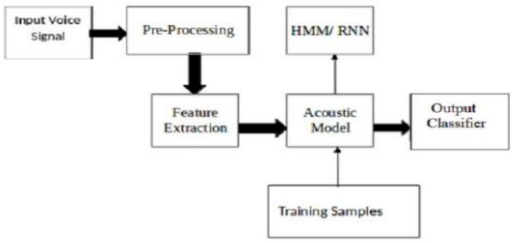

Fig. 2. Architecture diagram of speech recognition system

B. Pre-processing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

It involves below steps:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

C. Feature Extraction

The speech signal contains a large number of parameters that reflect the emotional characteristics. One of the sticking points in emotion recognition is what features should be used. In recent research, many common features are extracted, such as energy, pitch, formant, and some spectrum features such as linear prediction coefficients(LPC), MFCC, and modulation spectral features. In this work, we have selected modulation spectral features and MFCC, to extract the emotional features. MFCC is the most used representation of the spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. For each frame, the Fourier transform and the energy spectrum were estimated and mapped into the Mel-frequency scale. The Discrete Cosine Transform(DCT) of the Mel log energies was estimated, and the first 12 DCT coefficients provided the MFCC values used in the classification process.

Usually, the process of calculating MFCC. In our research, we extract the first 12order of the MFCC coefficients where the speech signals are sampled at 16 KHz. For each order coefficients, we calculate the mean, variance, standard deviation, kurtosis, and skewness, and this is for the other all the frames of an utterance. Each MFCC feature vector is 60-dimensional. Modulation spectral features (MSFs) are extracted from an auditory-inspired long-term spectro-temporal representation.

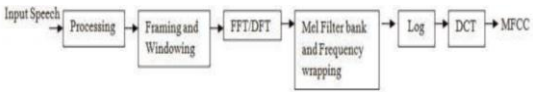

Fig. 3. Schema of MFCC extractions

These features are obtained by emulating the spectro-temporal (ST) processing performed in the human auditory system and consider regular acoustic frequency jointly with modulation frequency. The steps for computing the ST representation are illustrated in figure. In order to obtain the ST representation, the speech signal is first decomposed by an auditory filter bank (19 filters in total). The Hilbert envelopes of the critical-band outputs are computed to form the modulation signals. A modulation filter bank is further applied to the Hilbert envelopes to perform frequency analysis. The spectral contents of the modulation signals are referred to as modulation spectra, and the proposed features are thereby named Modulation Spectral Features (MSFs). Lastly, the ST representation is formed by measuring the energy of the decomposed envelope signals, as a function of regular acoustic frequency and modulation frequency. The energy, taken overall frames in every spectral band, provides a feature. In our experiment, an auditory filter bank with N ¼ 19 filters and a modulation filter bank with M ¼ 5 filters are used. In total,95 19 ð Þ 5 MSFs are calculated in this work from the ST representation.

D. Categorization of Emotions

The categorization of emotions has long been a hot subject of debate in different fields of psychology, affective science, and emotion research. It is mainly based on two popular approaches: categorical (termed discrete) and dimensional (termed continuous). In the first approach, emotions are described with a discrete number of classes. Many theorists have conducted studies to determine which emotions are basic. A most popular example is Ekman who proposed a list of six basic emotions, which are anger, disgust, fear, happiness, sadness, and surprise. He explains that each emotion acts as a discrete category rather than an individual emotional state. In the second approach, emotions are a combination of several psychological dimensions and identified by axes.

Other researchers define emotions according to one or more dimensions. WilhelmMax Wundt proposed in 1897 that emotions can be described by three dimensions: strain

versus relaxation, pleasurable versus un pleasurable, and arousing versus subduing. PAD emotional state model is another three-dimensional approach by Albert Mehrabian and

James Russell where PAD stands for pleasure, arousal, and dominance. Another popular dimensional model was proposed by James Russell in 1977.Unlike the earlier three-dimensional models, Russell's model features only two dimensions which include arousal (or activation) and valence (or evaluation).

The categorical approach is commonly used in SER. It characterizes emotions used in everyday emotion words such as joy and anger. In this work, a set of six basic emotions (anger, disgust, fear, joy, sadness, and surprise) plus neutral, corresponding to the six emotions of Ekman's model, were used for the recognition of emotion from speech using the categorical approach.

3. Algorithm

A. Long Short Term Algorithm

Long short-term memory (LSTM) is an artificial recurrent network (RNN) architecture used in the field of deep learning. Unlike standard feed forward neural network, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video).

LSTM networks are well-suited to classifying, processing and making prediction based on time series data, since there can be lags of unknown duration between important events in a time series. LSTM were developed to deal with the vanishing gradient problem that can been countered when training traditional RNN. Relative insensitivity to gap length is an advantage of LSTM over RNN.

4. Conclusion

This paper presented an overview on face emotion and audio analysis using Machine Learning.

References

[1] J. Akshay, S. L. and Utane, S. (2013) 'Emotion Recognition through Speech UsingGaussian Mixture Model and Support Vector Machine', International Journal of Scientific & Engineering Research, no. 5, pp. 1439-1443.

[2] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," IEEE Trans. Electron Devices, vol. ED-11, no. 1, pp. 34–39, Jan. 1959.

[3] Ankur Sapra, S. and Nikhil Panwar, A. (2013) 'Emotion Recognition from Speech', International Journal of Emerging Technology and Advanced Engineering, Vol. 3, pp. 341-345.

[4] J. Bertero, D. and Fung, P. (2017) 'A first look into a Convolutional Neural Network for speech emotion detection', IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, vol.12, no.4, pp. 5115-5119.

[5] D. B. Payne and J. R. Stern, "Wavelength-switched pas- sively coupled single-mode optical network," in Proc. IOOC-ECOC, 1985, pp. 585–590.

[6] BjörnSchuller, J. and Manfred Lang, L. (2013) 'Automatic Emotion Recognitionby the Speech Signal', National Journal, Vol. 3, no. 3, pp. 342-347.

[7] Dalal, H. and Triggs, B. (2005) 'Histograms of oriented gradients for human detection', in Computer Vision and Pattern Recognition. IEEE Computer Society Conference on, volume 1, pp. 886–893

[8] Furui, T. and Hori, F. (2004) 'Speech-to-Text and Speech-to-Speech Summarization', vol. 12, no. 4, pp. 401–408.

[9] Koolagudi, D. and. Krothapalli, S. (2012) 'Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features', Int. J. Speech Technol, vol. 15, no. 4, pp. 495–511.

[10] Lingli Yu, T. (2013) 'A hierarchical support vector machine based on feature-driven method for speech emotion recognition', Artificial Immune Systems -ICARIS, pp. 901-907.

[11] Krizhevsky, A. and Hinton, G. (2012) 'Image net classification with deep

[12] convolutional neural networks', in Advances in neural information processing systems, vol.12, no. 7, pp. 1097–1105.

[13] Kamel, L. S and Karray, F. (2011) 'Survey on speech emotion recognition: Features, classification schemes, and databases', Pattern Recognit., vol. 44, no. 3, pp. 572–587.