Effect Of Activation Function In Speech Emotion Recognition On The Ravdess Dataset

Komal D. Anadkat[1], Dr.Hiteishi M. Diwanji[2]

Information Technology department[12]
Government Engineering College[1], L.D.College of Engineering[2],
Gandhinagar, India[1], Ahmedabad, India[2]
komalanadkat@gecg28.ac.in[1], hiteishi.diwanji@gmail.com[2]

Abstract

Since last decade, Speech Emotion recognition has attracted extensive research attention to identify emotions by user's pitch and voice. Many research has been done in this field to recognize emotions using different machine learning as well as deep learning approaches. In this paper ,we tried three different machine learning algorithms named SVM, Logistic regression and Random forest which take four different features named MFCC,Chroma,Mel-scale spectrogram and tonnetz as an input on RAVDESS dataset where SVM is more accurate than others. As deep learning approaches are more capable to identify hidden patterns and classify the data more accurately, we tried popular algorithm like MLP,CNN and LSTM. In deep learning approach, activation function is one of the most dominant parameters which a designer can choose to make classification more accurate. In this paper, we tried to show the effect of different activation functions on the overall accuracy of the model and analyzed the results.

Keywords: CNN; MLP; activation function; neural network; deep learning; SER;

I. Introduction

Through all the inherently available senses, people can identify the emotional condition of their correspondence accomplice. This emotion detection is normal for people, yet it is troublesome undertaking for computers; Emotion is a way to express how an individual feels. Emotion recognition play a significant factor in many important areas, like health-care, education, and human resources. Emotion recognition is an intense errand as each individual has an alternate tone and pitch of voice [1].SER is the demonstration of endeavoring to perceive human feeling and emotional states from speech. This is gaining by the way that voice frequently reflects fundamental feeling through tone and pitch. Speech Emotion recognition (SER) is the common and quickest method of correspondence among people and PCs and assumes a significant part progressively uses of human-machine association. To recognize the emotion of the individual from voice is one of the interesting research area for various researcher. In SER, the robust and discriminative features selection and extraction is a most difficult task [2]. In this paper, we compare different machine learning algorithms like Linear SVM, Logistic regression and Random forest, with different neural network algorithms for classifying emotions from speech.

Along with all major problems in machine learning, SER has started to gain an advantage from the tools made available by deep learning. Feed-forward Neural Network takes input, processes input with 1 or more hidden layers and then outputs with the help of output layer at the end. In between each layer, no matter which type of layer it be, there is always an activation function which

decides what amount of input should be carried forward. Different neural networks serve different and specific purposes. For a case to consider, for image classification task, convolutional neural network (CNN) has been successful in achieving great results in various applications in machine learning domain in recent years [4] [5]. Whereas an artificial neural network with the capability of back-propagation learning helps in approximating multi-variate and non-linear kind of relationship between input(s) and output(s)[6].

One of the most important part of any neural network is the activation function [7] that helps decide if the current input and its setting allow what amount of input is to be carried forward in the neural network setting. Deep learning algorithms work just like our brains work and the activation functions act as if what is the limit to which a neuron can allow certain inputs to be tolerated, which is called threshold [8]. Activation function helps classifying input data into binary or multi-class based on certain threshold value, which is the basic use of an activation function of any neuron [9]. There are various types of activation functions, some of which are depicted below. It should be noted that this list and its details are not exhaustive.

II. Related Work

Many Researchers have previously done work on speech emotion recognition using different datasets and algorithms. Main two tasks of Speech emotion recognition are :- (i) Extracting and selecting the important and salient features from dataset and (ii) The selection of appropriate classifier which can classify the accurate emotions from the given speech. Many authors have used handcrafted features like MFCC, Contrast, Mels Spectrograph Frequency, Chroma and Tonnetz, while some authors have used deep learning approaches to improve the recognition accuracy.

I. Hand-Crafted Feature-Based Speech Emotion Recognition (SER)

Many researchers tried to efficiently recognize emotions from person's voice using handcrafted features. Audio signal have many features and so that it is major concern of researcher to select the appropriate features in the SER task. One researcher , Dave et al. [10] analyzed various speech emotions features and conclude that Mel frequency cepstral coefficient (MFCC) [11] features are better to use for SER than other low-level features like linear productivity code (LPC) [12], like formant, loudness,etc. On the other side Liu et al. [13] analyzed that to increase the Unweighted accuracy upto 3.6 %, it is better to use gamma tone frequency cepstral coefficient (GFCC) features using additional voice features like jitter and shimmer for SER. Liu et al. [14] proposed a novel method to extract the features from Chinese speech dataset (CASIA), using correlation and an extreme learning machine (ELM)-based decision tree for classification. Fahad et al. [15] proposed a technique which select glottal and MFCC features. These features fed to train a model based on deep neural network for SER. Wei and Zhao [16] proposed a novel method which used sparse classifier and auto encoders on a Chinese speech emotion dataset. To train support vector machine (SVM) classifier to recognize the emotion of a person from speech, author used auto encoder which extracts the large dimensions features and sparse network which extract small dimensions sparse features.

II. Deep Learning

Many researchers have worked to improve the efficiency of Speech emotion recognition using many different methods and using various datasets. Many of them used different neural network architecture and different pre trained networks to classify the emotions. Navya Damodar[17] have proposed novel approach to classify the emotions using CNN and decision tree. They used MFCC to extract the useful features from the Pre-processed audio files and applied these features with Decision tree as well as CNN both to compare the accuracy .They concluded that CNN outperforms Decision tree with 72% accuracy ,where the other achieved 63%. Jianfeng Zhao[18] have proposed

an approach which used Merged Deep CNN to learn deep features for speech emotion recognition. The merged deep CNN is the combination of 1D and 2D CNN models. These two models designed and evaluated individually and them merged. 1 D and 2D CNNs are designed to learn deep features from audio clips and from log-mel spectrograms respectively. Authors have used transfer learning to train the model which makes training faster. Fei and Liu [19] have proposed a method which uses advance long short-term memory (A-LSTM) to learn the sequences using pooling recurrent neural network (RNN) scheme in SER. Then they compare the results of A-LSTM with simple LSTM and found that former is better. Saurabh et al. [20] used IEMOCAP dataset and autoencoder approach for Speech emotion recognition. Hajar and Hasan [21] proposed a novel technique in which they split the audio signal into frames and extract MFCCs features. They used K-means clustering algorithm to select key spectrograms and used 3D CNN to predict speech emotions.

III. Proposed Methodology

Every individual express his/her emotion by face, speech, and text etc. The Speech emotion can be captured by tone and pitch of a person. In this paper we tried to recognize the different types of emotions of a person like sad, happy, angry , neutral ,disgust and surprised. In this paper the emotions in the speech are predicted using different machine learning algorithm along with neural networks. We have also enlighten the effect of different activation functions on accuracy. For analysis purpose we have used RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song dataset) dataset.

Fig.1 shows the work flow of speech emotion recognition. First, we load and read the RavDess dataset. In the next step, important features like MFCC, Mel, Chroma and Tonnetz are extracted from the data. In the third step, these extracted features become inputs for the training algorithm. Here we have used three different machine learning algorithms, Multilayer Perceptron, CNN and CNN-LSTM for the analysis purpose. When we used neural network algorithms, we tried to show the effect of different activation functions on the classification accuracy. Hence we compared each neural network on the bases of accuracy they achieved with different activation functions.

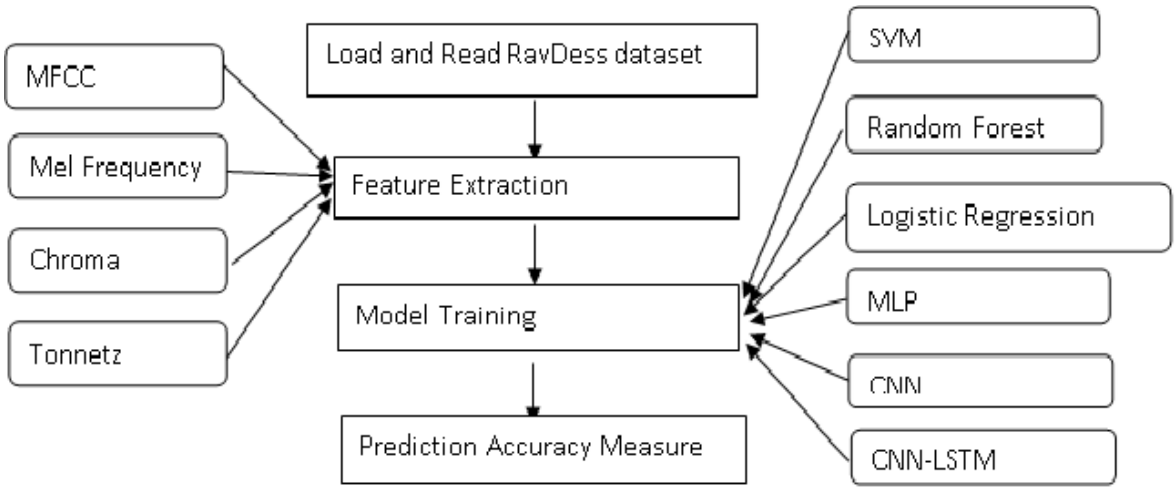

Figure 1: *Work Flow of Speech emotion recognition*

I. Traditional Methods

Before the era of deep neural networks, emotion recognition from speech has been mostly done

using machine learning and signal processing techniques. Machine learning algorithms like Decision trees, Random forest, SVM and logistic regression were used to classify the emotions automatically from speech.

Machine learning algorithm based solutions of SER problems need deep understanding of feature extraction and selection methods. Speech signals have many features available like Mel,Chroma, mel-frequency cepstrum coefficients (MFCC), modulation spectral features ,pitch,, energy, linear prediction coefficients (LPC) and Tonnetz .The main challenge here is to select the relevant and useful features for the classification task. After feature extraction, the next stage is classification. In the classification step, emotions will be classify in different class. Many classification algorithms are available and they can't be compare with each other as each of them has its own pros and cons. In this paper, we chose Linear SVM, logistic regression and Random Forest to compare the efficiency of three classifiers.

II. Multi-Layer Perceptron Classifier

Multi-layer Perceptron Classifier (MLP Classifier) is fall in the category of neural network used for the classification task. It uses MLP algorithm and trains model which uses Back-propagation to update the weights of neurons. Below are the steps to build MLP classifier:-

- In the first step, We need to initialize the parameters and define the classifier. Here we experimented different activation functions to find the effect on accuracy of the model.
- Then in the next step data is feed to neural network to train the model
- Now model trained in the step -2 is used to predict the output of the test data.
- Now Measure the accuracy of the prediction.

III. Convolutional Neural Networks

Convolutional neural networks (CNNs) are one of the most popular deep learning models that have manifested remarkable success in the research areas such as 14 object recognition , face recognition , handwriting recognition , speech recognition , and natural language processing . The term convolution comes from the fact that convolution—the mathematical operation—is employed in these networks. Generally, CNNs have three fundamental building blocks: the convolutional layer, the pooling layer, and the fully connected layer.

IV. Long Short Term Memory Networks

LSTM Networks are Recurrent neural network, which can learn long-term dependencies. These networks are used to solve many complex real world problems like speech recognition, object detection, facial expression recognition etc. the architecture of LSTM contains one input layer, one hidden layer and one output layer. The hidden layers have memory cells with gate units named the input, output and forget gates.

IV. Results and Discussion

I. Traditional methods Experiments

As A traditional methods, we have used SVM, Logistic regression and Random Forest. Table:-1 shows the comparison of three machine learning algorithm and it shows that SVM has achieved highest accuracy of 69.17% with very less training time. Other two algorithms have nearly performed well in terms of accuracy but random forest take more time in training and testing phase.

Table 1: *Performance Comparison of Tradition methodS*

Algorithm	Accuracy	Training Time(s)	Testing Time(s)
SVM	69.17%	0.024	0.001
Logistic Regression	67.10%	0.549	0.001
Random Forest	67.19%	3.413	0.155

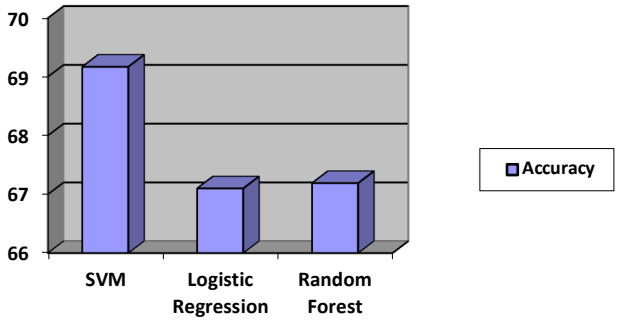

Figure 2: *Accuracy analysis of Tradition method*

II. MLP classifier Experiments

For MLP classifier, we use categorical cross entropy as our loss function. We use the Adam optimizer for optimizing our model. It combines various improvements to traditional stochastic gradient descent . Here highest accuracy achieved by Tanh activation function which was 81.77%, where Logistic also performs almost similar with 81.27% but it takes much time to train the model.

Table 2: *Effect Of Activation Functions On Accuracy Of MLP Classifier*

Activation Function	Accuracy	Training Time(s)	Testing Time(s)
Relu	71.35%	2.408	0.002
Tanh	81.77%	9.681	0.002
Logistic	81.27%	16.41	0.003
Identity	52.08%	1.172	0.002

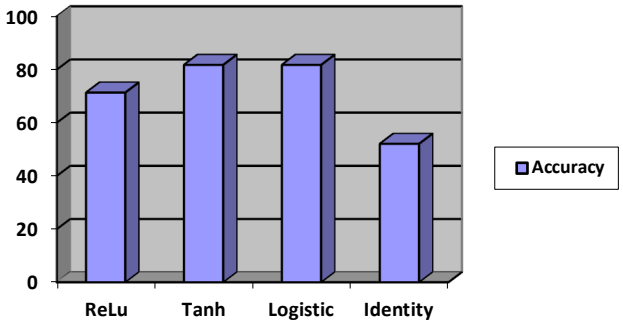

Figure 3: *Accuracy analysis of MLP classifier with different activation function*

III. Convolutional Neural Networks Experiments

The architecture followed here is 3 convolution layers followed by Max pooling layer, a fully connected layer and softmax layer respectively with 75 epochs. Multiple filters are used at each convolution layer, for different types of feature extraction. Here highest accuracy achieved by ReLu activation function which was 87.46%, where Tanh also performs almost similar with 85.89%.

Table 3: *Effect Of Activation Functions On Accuracy Of CNN Classifier*

Activation Function	Testing Accuracy	Training Accuracy
Relu	87.46%	96.50%
Tanh	85.89%	98.11%
Linear	75.61%	81.11%
Sigmoid	70.19%	75.40%

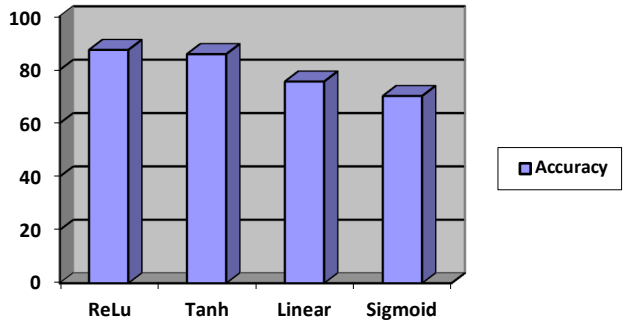

Figure 4: *Accuracy analysis of CNN classifier with different activation function*

IV. CNN-LSTM Experiments

Here we have created a 2D convolutional network as comprised of Conv2D and MaxPooling2D layers ordered into a stack of the required depth.We have defined a CNN LSTM model in Keras by first defining the CNN layer or layers, wrapping them in a TimeDistributed layer and then defining the LSTM and output layers. We have defined the CNN model first, then add it to the LSTM model by wrapping the entire sequence of CNN layers in a TimeDistributed layer. Here highest accuracy was achieved by ReLu activation function which was 84.97%.

Table 3: *Effect Of Activation Functions On Accuracy Of LSTM Classifier*

Activation Function	Testing Accuracy	Training Accuracy
elu	74.31%	90.81%
Tanh	54.67%	69.99%
Relu	84.97%	88.05%
Sigmoid	40.14%	50.95%

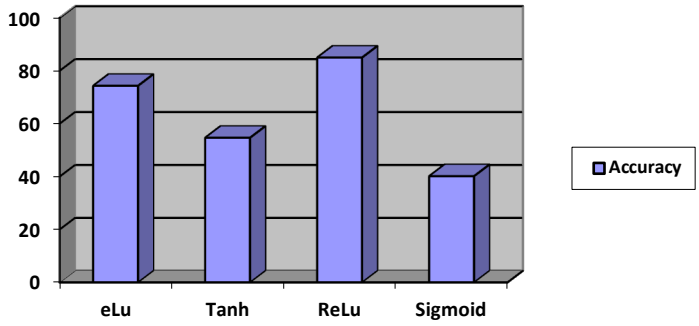

Figure 5: *Accuracy analysis of LSTM classifier with different activation function*

V. Conclusion & Future Work

Even though many works have been done for the speech emotion recognition area, still it faces many challenges. We applied three different machine learning algorithm named SVM, Logistic regression and Random forest on RevDess dataset where SVM outperforms with 69.17% accuracy and it takes less time to train the model. Then we applied ReLu, Tanh, Identity Logistic activation functions on MLP classifier where Tanh outperform with 81.77% of accuracy. After having ReLu, Tanh, Linear and sigmoid on CNN classifier, ReLu outperform with 87.46% accuracy. At Last we applied eLu, Tanh, Relu and sigmoid activation functions on CNN-LSTM model where ReLu outperform with 84.97% accuracy. Deep learning algorithms give better performance than traditional machine learning algorithms. These results of machine learning algorithm, MLP, CNN, LSTM are for Speech emotion recognition task on Revdess dataset. In future these results can be cross verified on other wide and real datasets of speech emotion recognition. Speech signals have

many features, here we have selected Mel, MFCC, Chroma and Tonnetz. In future, other features can also be used as a feature extraction methods and check whether it can improve the accuracy.

References

[1] Jerry Joy1, Aparna Kannan2, Shreya Ram3, S. Rama4," Speech Emotion Recognition using Neural Network and MLPClassifier ", IJESC,April-2020.

[2] Mustaqeem and Soonil Kwon * ,"A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition", Sensors 2020, 20, 183; doi:10.3390/s20010183.

[3] Dave, N. Feature extraction methods LPC, PLP and MFCC in speech recognition. Int. J. Adv. Res. Eng.Technol. 2013, 1, 1–4.

[4] Y. Yin, L. Want, E. Gelenbe, "Multi-Layer Neural Networks for Quality of Service oriented Server-State Classification in Cloud Servers", 2017, ISBN No. 978-1-5090-6182-2/17, pp. 1623-1627

[5] Y. Guo, L. Sun, Z. Zhang, H. He, "Algorithm Research on Improving Activation Function of Convolutional Neural Networks", 2019, ISBN No. 978-1-7281-0106-4/19, pp. 3582-3586

[6] Y. Zhang, Q. Hua, D. Xu, H. Li, Y. Bu, P. Zhao, "A complex valued CNN for different activation functions in polarsar image classification", 2019, IGARSS 2019, pp. 10023-100

[7] J. Demby, Y. Gao, G. N. DeSouze, "A study on Solving the Inverse Kinematics of Serial Robots using Aritifical Neural Network and Fuzzy Neural Network", 2019, ISBN No. 978-1-5386-1728-1/19

[8] S. Baraha, P. K. Biswal, "Implementation of Activation Functions for ELM based Classifiers", 2017, WiSPNET 2017, pp. 1038-1042

[9] S. Jeyanthi, M. Subadra, "Implementation of Single Neuron Using Various Activation Functions with FPGA", 2014, "International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), pp. 1126-1131

[10] Luque Sendra, A.; Gómez-Bellido, J.; Carrasco Muñoz, A.; Barbancho Concejero, J. Optimal Representation of Anuran Call Spectrum in Environmental Monitoring Systems UsingWireless Sensor Networks. Sensors 2018, 18, 1803. [CrossRef]

[11] Erol, B.; Seyfioglu, M.S.; Gurbuz, S.Z.; Amin, M. Data-driven cepstral and neural learning of features for robust micro-Doppler classification. In Proceedings of the Radar Sensor Technology XXII, Orlando, FL, USA, 16–18 April 2018; p. 106330J.

[12] Liu, G.K. Evaluating Gammatone Frequency Cepstral Coe_cients with Neural Networks for Emotion Recognition from Speech. arXiv 2018, arXiv:1806.09010.

[13] Liu, C.-L.; Yin, F.; Wang, D.-H.; Wang, Q.-F. CASIA online and o_ine Chinese handwriting databases.In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 37–41.

[14] Liu, Z.-T.; Wu, M.; Cao, W.-H.; Mao, J.-W.; Xu, J.-P.; Tan, G.-Z. Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing 2018, 273, 271–280. [CrossRef]

[15] Fahad, M.; Yadav, J.; Pradhan, G.; Deepak, A. DNN-HMM based Speaker Adaptive Emotion Recognition using Proposed Epoch and MFCC Features. arXiv 2018, arXiv:1806.00984.

[16] Zhu, L.; Chen, L.; Zhao, D.; Zhou, J.; Zhang, W. Emotion recognition from Chinese speech for smart services using a combination of SVM and DBN. Sensors. 2017, 17, 1694.

[17] Navya Damodar, Vani H Y, Anusuya M A. Voice Emotion Recognition using CNN and Decision Tree. International Journal of Innovative Technology and Exploring Engineering (IJITEE), October 2019.

[18] Jianfeng Zhao, Xia Mao, Lijiang Chen. Learning Deep features to Recognise Speech Emotion using Merged Deep CNN. IET Signal Process., 2018

[19] Tao, F.; Liu, G. Advanced LSTM: A study about better time dependency modeling in emotion recognition.In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2906–2910.

[20] Sahu, S.; Gupta, R.; Sivaraman, G.; AbdAlmageed,W.; Espy-Wilson, C. Adversarial auto-encoders for speech based emotion recognition. arXiv 2018, arXiv:1806.02146.

[21] Hajarolasvadi, N.; Demirel, H. 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms. Entropy 2019, 21, 479.