Journal of Advanced Research in Technology and Management Sciences
Volume: 04 Issue: 03 ISSN: 2582-3078 June 2022
Available online at: http://www.jartms.org

Speech Emotion Recognition (SER) & Gender Detection Using Deep Learning

N. HARSHAVARDHAN, Dept. of CSE, Kallam Haranadhareddy Institute of Technology, Guntur.
G. VASISHTA Dept. of CSE, Kallam Haranadhareddy Institute of Technology, Guntur.
B. BALAJI Dept. of CSE, Kallam Haranadhareddy Institute of Technology, Guntur.
SK. HASEENA Dept. of CSE, Kallam Haranadhareddy Institute of Technology, Guntur.
M. DEVA KUMAR Dept. of CSE, Kallam Haranadhareddy Institute of Technology, Guntur.

Abstract: Emotion recognition from speech signals is an important but challenging component of Human-Computer Interaction (HCI). In the literature of speech emotion recognition (SER), many techniques have been utilized to extract emotions from signals, including many well-established speech analysis and classification techniques. Deep Learning techniques have been recently proposed as an alternative to traditional techniques in SER. This paper presents an overview of Deep Learning techniques and discusses some recent literature where these methods are utilized for speech-based emotion recognition. The review covers databases used, emotions extracted, contributions made toward speech emotion recognition and limitations related to it. A Multilayer Perceptron (MLP) deep learning model has been described to recognize voice gender. The data set have 3,168 recorded samples of male and female voices. The samples are produced by using acoustic analysis. An MLP deep learning algorithm has been applied to detect gender specific traits.

Introduction: As human beings speech is amongst the most natural way to express ourselves. We depend so much on it that we recognize its importance when resorting to other communication forms like emails and text messages where we often use emojis to express the emotions associated with the messages. As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today's digital world of remote communication. Emotion detection is a challenging task, because emotions are subjective. There is no common consensus on how to measure or categorize them. We define a SER system as a collection of methodologies that process and classify speech signals to detect emotions embedded in them. Such a system can find use in a wide variety of application areas like interactive voice based-assistant or caller-agent conversation analysis.
Acoustic analysis of the voice depends upon parameter settings specific to sample characteristics such as intensity, duration, frequency and filtering. The acoustic properties of the voice and speech can be used to detect gender of speaker. The data set which has acoustic parameters can be obtained with this analysis. The data set can be trained with different machine learning algorithms. In this paper, MLP has been used to obtain model. The results have been compared with related work. A algorithm has been designed to detect the gender of voice by using obtained model.

Literature Survey: For SER, many deep learning algorithms have been developed. However, there exist meaningful prospects and fertile ground for future research opportunities not only in SER but many other domains. The layer-wise structure of neural

networks adaptively learns features from available raw data hierarchically. The remainder of this section summarizes the literature on deep layer architectures, learning and regularization methodologies discussed in the context of SER. The pattern of identification is the main object for which, the basic mathematical tool is utilized. On verification, it is observed that no model is proved consistently and effectively to be predicted in its classification. This paper is, therefore, introduces a procedure for Raaga Identification with the help of Hidden Markov Models (HMM) which is rather an appropriate approach in identifying Melakarta Raagas. This proposed approach is based on the standard speech recognition technology by using Hidden continuous Markov Model. Data is collected from the existing data base for training and testing of the method with due design process relating to Melakarta Raagas. Similarly, to solve the problem of automatic identification of raagas, a suitable approach from the existing database is presented. The system, particularly, this model is based on a Hidden Markov Model enhanced with Pakad string matching algorithm. The entire system is built on top of an automatic note transcriptor. At the end, detailed elucidations of the experiments are given. It clearly indicates the effectiveness and applicability of this method with its intrinsic value and significance.The main problem that affects the overall performance of the RNN is its sensitivity towards the disappearance of gradients. An adaptive SER system based on deep learning technique known as DRNN is used for SER. The learning stage of the model includes both frame-level and short-time acoustic features, due to their similar structure. Another multi-tasking deep neural network with shared hidden layers named MT-SHL-DNN is utilized, where the transformation of features is shared. Here, the output layers have an association separately with each data set used. The DNN also helps in measuring the SER based on the nature of the speaker and gender. When the DNNs are used for encoding of segments into length vectors that are fixed in nature, this is done by using pooling of various hidden layer over the specified time. The design of the feature encoding procedure is done in such a manner that it can be used jointly with segmental level classifier for efficient classification.

The tendency of DNNs to learning the specific features from various auditory emotion recognition systems is analyzed. These features include voice and music-based recognition. Further, the utilization of cross-channel architecture can improve the general performance in a complex environment. The model provided some good results for human speech signal and music signal; however, the results for generalized auditory emotion recognition are not optimal. The purpose of this cross-channel hierarchy is to extract specific features and combine them into a much more generalized scenario. Also, these models can be coupled with visual-based DNNs to improve automatic SER. RNNs utilization in such a scenario can further boost the performance for the input data with time-dependent constraints.A hybrid deep learning modality may inherit the underlying properties of RNN with CNN, with convolutional levels implanted with RNN. This enables the model to obtain both frequency and temporal dependency in a given speech signal. Sometimes, a memory enhanced reconstruction-error-based RNN for continuous speech emotion recognition can also be used. This RNN model uses two components, first an auto-encoder for the reconstruction of features, and second for the prediction of emotions. It can also be used to obtain further insights into the behavior of BLSTM-based RNN using regression models such as SVR.As a

final remark, deep learning is rapidly becoming a method of choice over traditional techniques for SER. Also, most of the research is evolving towards multimodal and unsupervised SER, speech recognition and NLP. Multimodal emotion recognition can use input data such as audio-visual at the same time, in an efficient way.The biggest hurdle is the noise factor which added with speech in real life environment. The noise can interfere with the actual speech and this can lead to wrong classification. Noise can be anything like, babble noise, street noise, suburban train noise, car noise, restaurant noise or similar kind. So in order to have a reliable gender recognition system, some pre-processing techniques and robust features need to be applied. In this thesis, the focus is mainly on three speech features: pitch and MFCC. This thesis can be differentiated from other works in the way that here a new algorithm is proposed for extracting appropriate frames of speech which are required for gender identification.

Proposed System: Our goal is to implement machine learning model in order to classify, to the highest possible degree of accuracy, ravdess-emotion-speech-audio is a dataset gathered from Kaggle. After initial data exploration, for the best accuracy reports Deep learning model is implemented.

Neural networks are a set of algorithms, modeled loosely after the human brain, that are designed to recognize patterns. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be- it images, sound, text or time series, must be translated. It helps us cluster and classify. You can think of them as a clustering and classification layer on top of the raw data you store and manage. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labeled dataset to train on.

A deep feed-forward neural network is an artificial neural network with multiple hidden layers of units between the input and output layers.

The Mel-frequency cepstral coefficients (MFCC) is one of the most popular audio feature. It is a representation of the speech signals where a feature called the cepstrum of a windowed short-time signal is derived from the FFT of that signal.

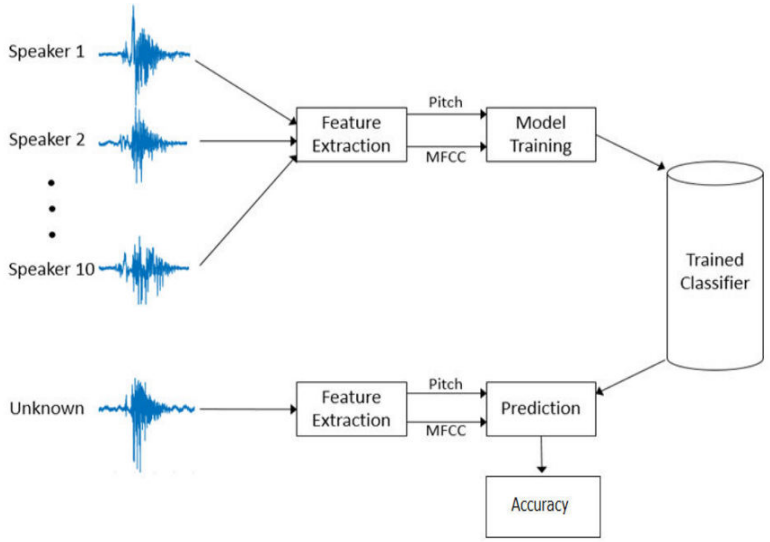

Fig: Speech Emotion Recognition and Gender Detection System

ADVANTAGES OF PROPOSED SYSTEM

- Provides the flexibility to work with nonlinear values
- Less number of parameters required.
- Higher performance compared to previous systems.
- Better classification of parameters is shown
- Can handle missing values, model complex relationships and support multiple inputs.

SYSTEM DESIGN: The purpose of the design phase is to plan a solution of the problem specified by the requirement document. This phase is the first step in moving from problem domain to the solution domain. The design of a system is perhaps the most critical factor affecting the quality of the software, and has a major impact on the later phases, particularly testing and maintenance. A design methodology is a systematic approach to creating a design by application f set of techniques and guidelines. Most methodologies focus on system design. The two basic principles used in any design methodology are problem partitioning and abstraction. A large system cannot be handled as a whole, and so for design it is partitioned into smaller systems. Abstraction is a concept related to problem partitioning. When partitioning is used during design, the design activity focuses on one part of the system at a time. Since the part being designed interacts with other parts of the system, a clear understanding of the interaction is essential for properly designing the part. For this, abstraction is used.

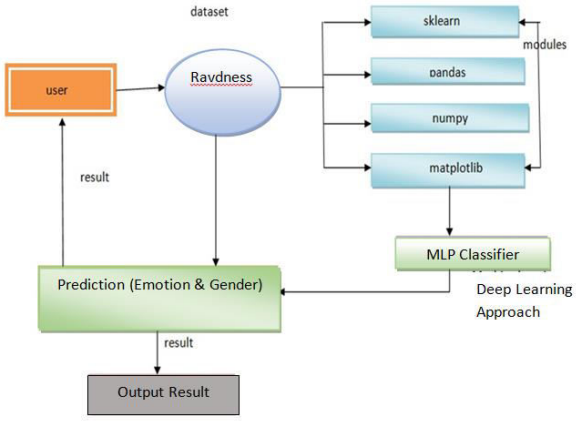

Fig: System Design for Speech Emotion and Gender Detection.

The design of the project is described as , the user has the predefined data set which contains all the information related to credit card holders. Later, Python libraries are imported for th data set. The libraries include NumPy, Panda, Sklearn, TQDM, Librosa.

Implementation: This chapter includes the implementation of the design and source code. In this phase the design is translated into code. Computer programs are written using a conventional programming language or an application generator. Programming tools like Compilers, Interpreters, and Debuggers are used to generate the code. Different high-level programming languages like C, C++, Pascal, Java, .Net are used for coding. With respect to the type of application, the right programming language is chosen. This dataset is about audio files. The ravdess-emotional-speech-audio is a dataset that consists of different actors which had male and female with different emotions,it includes the following.

Journal of Advanced Research in Technology and Management Sciences
Volume: 04 Issue: 03 ISSN: 2582-3078 June 2022
Available online at: http://www.jartms.org

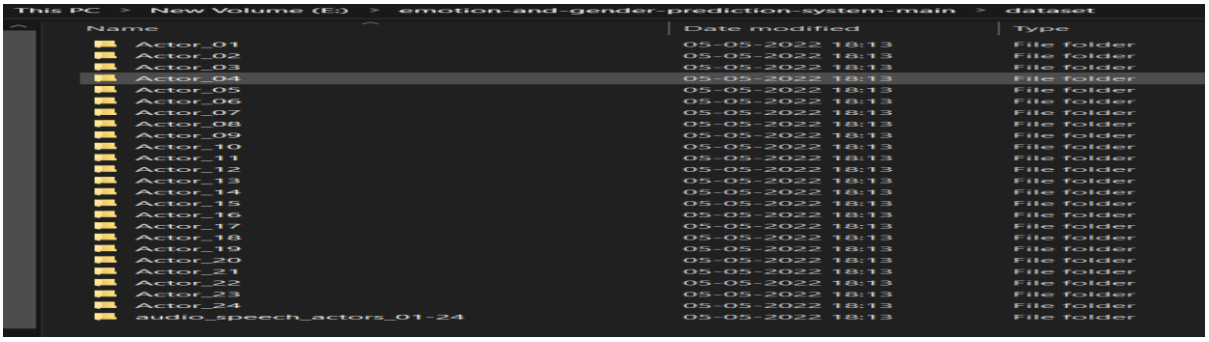

Fig: Dataset

Above Fig shows dataset in the local repository/local file System.

Step 2: Load Data from WAV File

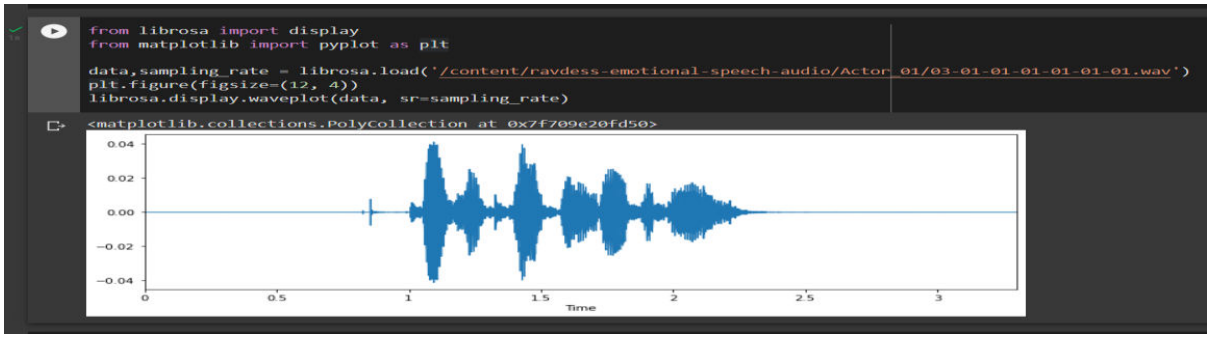

Fig: Actor speech signals are visualized as a wave plot

Step 3: Split dataset for training and validating

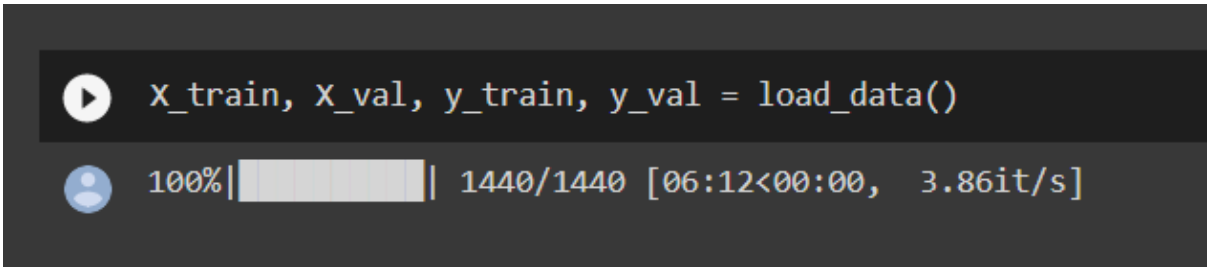

Fig: Splitting data

Splitting the dataset into 2 categories named Gender and emotion based on the naming convensions in our dataset.

Step 4: Model Loss

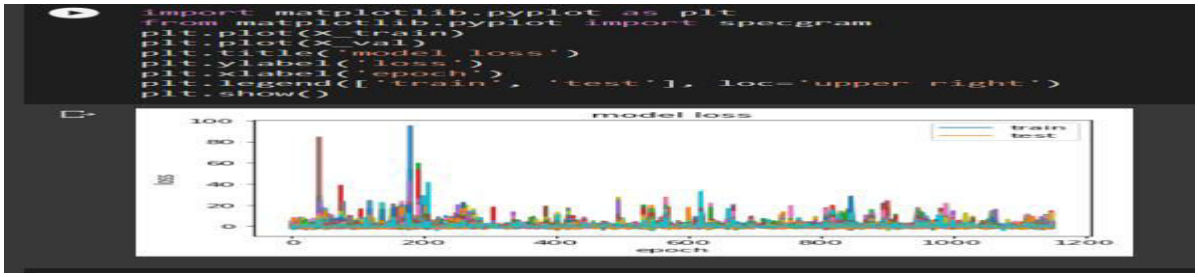

Fig: Model of Loss

Step 5: Make predictions on validation set and accuracy of model

```
y_pred=model.predict(X_val)
print(model.score(X_val, y_val))
```

```
0.7118055555555556
```

Fig. Accuracy of model

Step 6: Visualizing of Actual and predicted data using DataFrame.

```
df=pd.DataFrame({'Actual': y_val, 'Predicted':y_pred})
df.head(15)
```

	Actual	Predicted
0	neutral_male	sad_male
1	fearful_male	sad_male
2	surprised_male	surprised_male
3	surprised_female	surprised_female
4	disgust_female	disgust_female
5	neutral_female	neutral_female
6	sad_male	sad_male
7	neutral_female	sad_female
8	surprised_male	surprised_male
9	fearful_male	fearful_male
10	happy_female	happy_female
11	angry_female	surprised_female
12	surprised_female	surprised_female
13	calm_male	calm_male
14	fearful_female	fearful_female

Fig. Actual vs Predict

Step 7: Record a Speech signals from the User

```
import IPython.display as ipd

def synth():
  print("Now recording for 10 seconds, say what you will...")
  record(5)
  print("Audio recording complete")
  in_fpath = Path("audio.wav")
InvokeButton('Start recording', synth)
```

Start recording

```
Now recording for 10 seconds, say what you will...
Audio recording complete
```

```
[31] ipd.Audio("audio.wav")
```

0:04 / 0:04

Fig. Record a speech signals

Step 8: Extracting Features from the audio speech

```
np.array(x_audio)

array([[-4.25651093e+02,  1.12727928e+02, -3.90526619e+01,
         1.15827732e+01,  5.17359400e+00, -6.54628897e+00,
         7.31685829e+00, -4.20395708e+00, -8.19178009e+00,
        -6.65955973e+00,  6.42200530e-01, -1.03814316e+01,
        -4.15901232e+00, -3.66859603e+00, -4.33219481e+00,
         2.75568485e-01, -7.65198708e+00,  1.94887781e+00,
        -3.83501720e+00, -3.47239852e+00, -4.84950209e+00,
         1.34823895e+00, -5.34854555e+00, -1.84901997e-01,
        -5.45019627e+00, -3.78209233e-01, -5.34057140e+00,
        -5.41756153e+00, -4.70523787e+00, -5.16137791e+00,
        -3.64355755e+00, -1.89193380e+00, -1.31904209e+00,
         1.68172967e+00, -1.49407351e+00,  2.18041801e+00,
        -2.39506531e+00, -1.76568782e+00, -3.26771355e+00,
         1.81378961e-01,  6.05872035e-01,  6.26412272e-01,
         7.32697725e-01,  7.31683373e-01,  7.20146418e-01,
         7.00939655e-01,  7.26864278e-01,  7.31238782e-01,
         7.89489448e-01,  7.01565504e-01,  6.45269632e-01,
         6.38535619e-01,  3.29855364e-04,  7.65837322e-04,
         1.25186564e-03,  2.71665049e-03,  1.81659982e-02,
         6.12572441e-03,  7.01479008e-03,  6.66460674e-03,
         4.37895022e-03,  2.92440038e-03,  5.94054209e-03,
         9.43256449e-03,  1.01492126e-02,  8.77916440e-03,
         1.23901710e-01,  4.12039191e-01,  1.08784653e-01,
         8.27019569e-03,  6.88834628e-03,  5.98414103e-03,
         3.61182401e-03,  9.59070586e-03,  2.69957595e-02,
         2.24997662e-02,  9.80041455e-03,  2.01640720e-03,
         1.38652313e-03,  1.02607603e-03,  2.97920429e-03,
         6.56045927e-03,  4.28216299e-03,  4.93356120e-03,
```

Fig. Extracting Features

Step 9: Prediction of Emotion and Gender:

```
y_pred_audio=model.predict(X_audio_val)
hasil = str(y_pred_audio[0])
emotion_value = hasil.split("_")[0]
gender_value = hasil.split("_")[1]
print("The system detects that you are a",gender_value, "and your current emotions is", emotion_value)

The system detects that you are a female and your current emotions is neutral
```

Fig. Prediction of Emotion and Gender

Conclusion: This Project shows that MLPs are very powerful in classifying speech signals. Even with simplified models, a limited set of characters can be easily identified. We have obtained higher accuracies as compared to other approaches for individual emotions along with that we can able to recognize Gender. The performance of a module is highly dependent on the quality of pre-processing. Mel Frequency Cepstrum Coefficients are very dependable. Every human emotion has been thoroughly studied, analyzed and the accuracy has been checked The results obtained in this study demonstrate that speech recognition is feasible, and that MLPs can be used for any task concerning recognizing of speech and demonstrating the accuracy of each emotion present in the speech. One widespread limitation in almost all the related works examined was the fact that they were only reporting the accuracy of the recognition as their performance measure, but statistically, accuracy by itself is not a comprehensive measure of the performance of a system.

BIBLIOGRAPHY

1) *B.Tarakeswara Rao, P.Lakshmikanth, E.Ramesh, "Some Studies on Raaga Emotions of Singers Using Gaussian Mixture Model", International Journal for Modern Trends in Science and Technology,Vol. 03, Special Issue 01, 2017, pp. 149-153, http://www.ijmtst.com/NCRACSE2017/29NCRACSE95.pdf*
2) *Prasad Reddy PVGD, B Tarakeswara Rao, Dr. K.R Sudha, Hari CH.V.M.K, "Automatic Raaga Identification System for Carnatic Music using Hidden Markov Model", Volume 11,Issue 22 ,December 2011, Global Journal Of Computer Science And Technology, Online ISSN: 0975-4172 & Print ISSN: 0975-4350, IF=1.02*

3) *Tarakeswara Rao B, Dr. Prasad Reddy PVGD, "A Novel Process for Melakartha Raaga Recognition using Hidden Markov Models (HMM),Vol. 2, No. 2, April 2011, International Journal of Research and Reviews in Computer Science (IJRRCS), Page No. 508-513*
4) *Tarakeswara Rao B, Dr. Prasad Reddy PVGD, "Raaga Recognition System using Gaussian Mixture Model for Gender Dependence and Independence:Melakartha Raagas", Volume 1, Number 4 Nov - Dec 2010, International Journal of Advanced Research in Computer Science, 0976-5697, IF=2.5, DOI: https://doi.org/10.26483/ijarcs.v1i4.210*
5) *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.*
6) *S. Brave and C. Nass, "Emotion in human–computer interaction," in Human-computer interaction fundamentals, vol. 20094635: CRC Press Boca Raton, FL, USA, 2009, pp. 53-68.*
7) *G. Trigeorgis et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2016: IEEE, pp. 5200-5204.*
8) *K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Fifteenth annual conference of the international speech communication association, 2014.*
9) *W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), 2016: IEEE, pp. 1- 4.*
10) *M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," Proceedings of the IEEE, vol. 91, no. 9, pp. 1370-1390, 2003.*
11) *S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PloS one, vol. 13, no. 5, p. e0196391, 2018.*
12) *D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," Biomedical Signal Processing and Control, vol. 59, p. 101894, 2020.*
13) *L. Auria and R. A. Moro, "Support vector machines (SVM) as a technique for solvency analysis," 2008.*