© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 13 Issue: 05 | October - December 2022
Paper is available at http://www.jrps.in | Email : info@jrps.in
DOI : 10.5281/zenodo.7335435 Refereed & Peer Reviewed

EMOTION DETECTION WITH SPEECH ASINPUT USING CONVOLUTION NEURAL NETWORKS

KARTIK SHARMA
Dept. of Computer Engineering
MPSTME, NMIMS University
Mumbai,India
sharma.kartik546@gmail.com

JEET SHAH
Dept. of Computer Engineering
MPSTME, NMIMS University
Mumbai,India
jeetpareshshah@gmail.com

SHREYANSH KUMAR
Dept. of Computer Engineering
Bennett University
Greater Noida,India
shreyanshkumar2304@gmail.com

***Abstract*—With stress rising exponentially in day to day life, sudden emotions changes are triggering serious health parameters leading to surgeries. Even though it is challenging to identify, recognise emotions in spontaneous speech, it is a principal part of interactions between humans and computers. The challenges in identifying emotions during spontaneous speech are mainly attributed to the emotions expressed by the speaker not standing out as they do in acted speech. This paper proposes a framework for automatic speech stimulation that uses relevant information. The framework is moved to see that there is a great deal of inconsistency between human annotations when describing spontaneous speech; disagreement is greatly reduced when additional information is provided. The suggested framework employs the known emotions shown by the sound bytes of the RAVDESS dataset and the understanding of how spoken words change over time during an audio call in order to accurately identifythe speaker's current emotional state during a conversation. We use convolution neural networks as well as the softmax activation function to classify the data into multiple classes once the features of the audio have been extracted. Our experimental results demonstrate that the accuracy of detection of emotion using speech as input is 68.7% and is feasible to deploy in a real-life scenario.**

***Index Terms*—**
Emotion recognition
Feature Selection
Feature Extraction
Emotion Analysis
Anger Detection
Deep Learning
Speech Input
Artificial Intelligence.

I. INTRODUCTION

SPEECH is one of the most basic and natural forms of communication between human beings. With the advent of human machine communication technology, an easy-to-use interface is becoming increasingly important for speech-focused programs. Humans have a natural ability to use all their senses to learn more about the message you are receiving. With all available sensors people can easily see the emotional state of their communication partner. Emotional recognition is natural for humans but it is the most difficult task of a machine. The purpose of the sensory system is therefore to improve machine-to-human communication using information-related emotions in such a way that the human machine interface works better. Emotional awareness of speech has a variety of processes in everyday life. Some of the Speech Recognition program include:

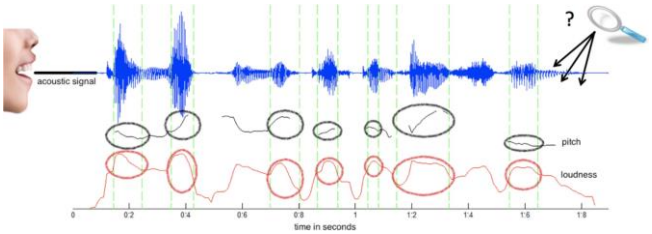

Fig. 1: Wave-plot

1. In psychiatric diagnosis and lie detection
2. In a call center conversation it can be used to analyze the ethics of telephone assistants and clients who help improve the quality of service of the telephone operator.
3. In the cockpit of aircraft, speech recognition systems that are trained to detect focused speech are used for better performance.
4. An Analysis of the feelings of telephone conversations between criminals can help the crime investigation department.
5. An interactive movie, discussing E-tutoring applications can be very helpful, if they are able to adapt to the audience or students of emotional states.
6. Conversing with humanoid partners can be realistic and enjoyable, if they are able to understand and express human-like emotions. In speech recognition, emotions are found in the speech of male or female speakers .

In the last century studied certain speech features including basic waves, Spectral features have been widely used in sensory perception, such as Mel frequency cepstrum coeffi-

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 13 Issue: 05 | October - December 2022
Paper is available at http://www.jrps.in | Email : info@jrps.in
DOI : 10.5281/zenodo.7335435 Refereed & Peer Reviewed

cient (MFCC), coefficient coepffum (LPCC)[2,4,8]. Another effective group of features are Prosodic features for describing emotional states such as energy , pitch , intensity , formants etc. Most traditional speech culture cultures have focused on prosodic or spectral aspects that form the basis of modern speech processing. In one of the studies, a correlation between speech and emotion was present. The levels of human emotional awareness were compared, when both identical levelsof recognition were established. In other studies of physical and emotional expression the method of spectrograms was studied and received the same amount of recognition for both of us, which is recommended for use in the speech recognitionsystem Activity recognition in the Speaker is a major challengefor the following reasons: In this case what specific aspects of speech are needed to distinguish between different emotions are unclear. Due to the presence of different speaking styles, speakers, sentences, languages and speaking values, aspects of speech are directly affected. Another problem is that the expression of emotions is based on the speaker and their culture and environment. As there is a change in the culture and environment of different speakers their style of speaking also undergoes a change. This is another challenge in frontof a speech recognition program.In this paper we propose a emotion detection system using convolution neural networks that uses manually annotated audio clips with 5 main emotionsnamely ;happy , sad , fearful, calm and angry. We tried 3 different models to achieve our output , i.e. LSTM , MLP and CNN. We discuss the merits and demerits of each model.[3]

II. Literature Survey

A. Spontaneous speech emotion recognition using prior knowledge

Automatic and spontaneous speech emotion recognition is an important part of a human-computer interactive system. However, emotion identification in spontaneous speech is difficult because most often the emotion expressed by the speaker are not necessarily as prominent as in acted speech. The theory propose a spontaneous speech emotion recognition framework that makes use of the associated knowledge.Recognizing emotion in spontaneous speech is difficult because it does not carry sufficient intensity to distinguish one emotion from the other.

B. Detection of negative emotions in speech signals usingbags-of-audio-words

Robust, low-complex classification system for the detection of negative emotions in speech signals was implemented on the basis of a spontaneous, strongly emotionally colored speech corpus.The prediction is made on whether or not the emotion is negative using a very broad meaning of the word 'negative'. This is done using a bag of words approach.

C. Deep neural networks for anger detection from real life speech data

Three state-of-the-art, deep neural networks for anger detection have been thoroughly investigated and the results were compared with each other to see which one was the best.The proposed neural networks significantly outperform traditional modelling algorithms for speech anger detection.

D. Anger recognition in speech using acoustic and linguistic cues

The present study focuses on human computer interactions by studying methods to automatically detect whether utterances spoken by the speaker are emotional or non-emotional.It focuses on all kinds of excitations or abnormalities from the normal speech pattern using acoustic features and so it broadens the domain to all kinds of emotions other than anger.

III. Architecture

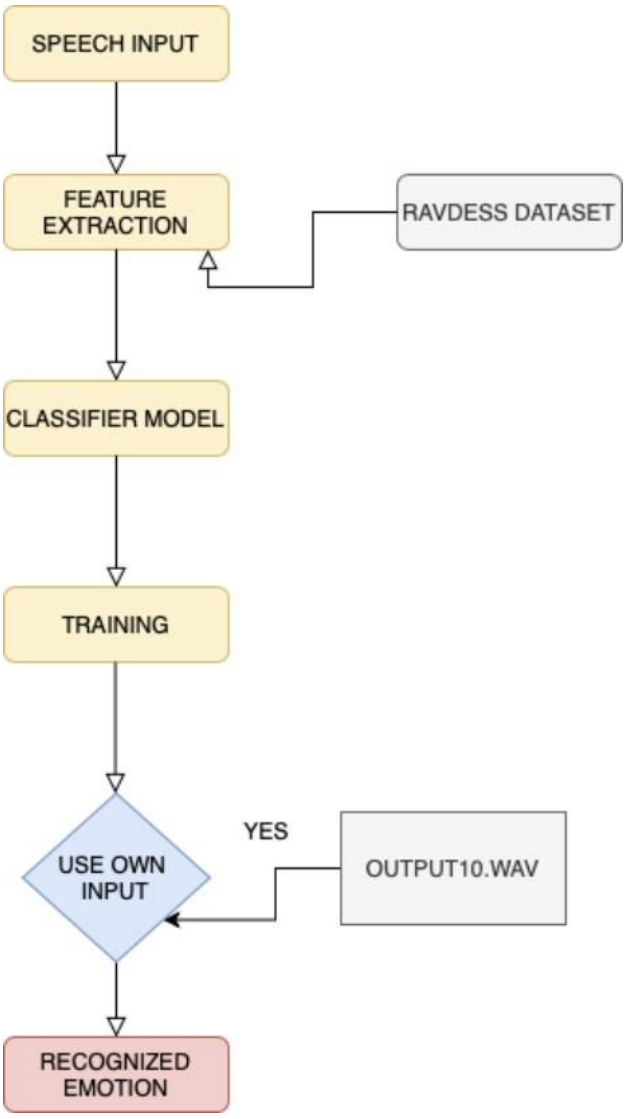

Fig. 2: Architecture

Speech emotion recognition system is typical pattern recognition system. This indicates that the stages involved in the

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 13 Issue: 05 | October - December 2022
Paper is available at http://www.jrps.in | Email : info@jrps.in
DOI : 10.5281/zenodo.7335435 Refereed & Peer Reviewed

pattern recognition process are also present in the emotional recognition system. There are five key modules in the speech recognition system fig.[2] that consist emotional speech input, feature extraction, feature selection, classification and recognized emotional output. This is because there are more than 300 types of emotions present and with our small database classification can be very difficult, The input as database feed to the speech emotion recognition system may contain the real world emotions or artificial feelings.[10] In the discovery of speech recognition data speech is very important to consider. The evaluation of emotion recognition system is depends on the naturalness and efficiency of the database which is usedas an input to the speech emotion recognition system. If an incorrect database is provided as an inclusion in the system then the incorrect conclusion will be drawn. We have elabo- rated more about our database further. For the classifier modelwe have used a convolution neural network. The summary is shown in Fig[2].

```
In [51]: model.summary()
```

Layer (type)	Output Shape	Param #
conv1d_1 (Conv1D)	(None, 216, 128)	768
activation_1 (Activation)	(None, 216, 128)	0
conv1d_2 (Conv1D)	(None, 216, 128)	82048
activation_2 (Activation)	(None, 216, 128)	0
dropout_1 (Dropout)	(None, 216, 128)	0
max_pooling1d_1 (MaxPooling1	(None, 27, 128)	0
conv1d_3 (Conv1D)	(None, 27, 128)	82048
activation_3 (Activation)	(None, 27, 128)	0
conv1d_4 (Conv1D)	(None, 27, 128)	82048
activation_4 (Activation)	(None, 27, 128)	0
conv1d_5 (Conv1D)	(None, 27, 128)	82048
activation_5 (Activation)	(None, 27, 128)	0
dropout_2 (Dropout)	(None, 27, 128)	0
conv1d_6 (Conv1D)	(None, 27, 128)	82048
activation_6 (Activation)	(None, 27, 128)	0
flatten_1 (Flatten)	(None, 3456)	0
dense_1 (Dense)	(None, 10)	34570
activation_7 (Activation)	(None, 10)	0

Total params: 445,578
Trainable params: 445,578
Non-trainable params: 0

Fig. 3: Model Summary

We have used multiple convolution layers fig[3] to get a better classification, the dropout layers are to prevent over- fitting of data . In the model while training we have used a ReLu activation layer since the sigmoid and tanh functions very sensitive to change and we got better classification using ReLu. We have used pooling to reduce the size of the matrix tothe next levels in the neural network. Lastly, we use a softmaxactivation function to get multiclass classification. One thing to note is that we are classifying using gender as well as the features of emotions for both genders were significantly different and we got a higher accuracy using this model.

IV. Dataset

The database we used was the RAVDESS Dataset: Ry-erson Audio-Visual Database Emotional Speech and songs (RAVDESS) contains 7356 files (total size: 24.8 GB). The database estimates gender consisting of 24 professional ac- tors, uttering language-related statements differently in North America.[7] The speech includes calm, joy, sadness, anger, fear, surprise, and disgust speech, and the songs contains calm, happy, sad, angry, and terrifying emotions. Each speech is produced in two levels of emotional strength, with additional neutral speech. All modes are available with face and voice, face only, and voice structures only. Each set of 7356 records was rated 10 times for emotional, dynamic, and factual per- formance. The ratings were presented by 247 people whowere features of untrained research participants from North America. An additional set of 72 participants provided details of the re- examination. High levels of emotional performance and reliability of test-retest intrarater were reported. Corrected accuracy and integrated “aesthetic” measures are introduced to assist researchers in selecting stimulants.

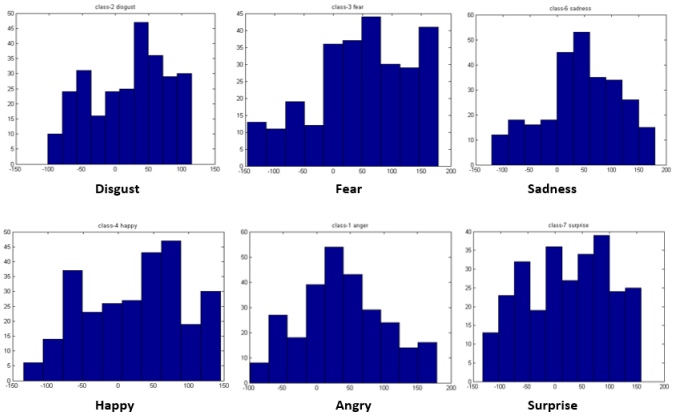

Fig. 4: Bucket Wise Distribution of Results

V. Feature Extraction

Extraction/Exclusion is a very important part of analysing and finding relationships between different objects. The in- formation provided by the audio cannot be understood bythe models directly. So, we convert it into an understandable format using feature extraction. The important issue in feature extraction is the region of analysis of the speech signal used which is to be considered in the feature extraction. The speech signal is non-stationary hence divided into the small intervals which are called as frames. The Spectral and prosodic features are known to be the main indicators of emotional state speak- ers. Sensory cognitive studies have shown important factors in speech rate, intensity, tone, energy, duration, formants, linear prediction cepstrum coefficient (LPCC) and Mel frequency cepstrum coefficient (MFCC)[1]. In a different emotional

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 13 Issue: 05 | October - December 2022
Paper is available at http://www.jrps.in | Email : info@jrps.in
DOI : 10.5281/zenodo.7335435 Refereed & Peer Reviewed

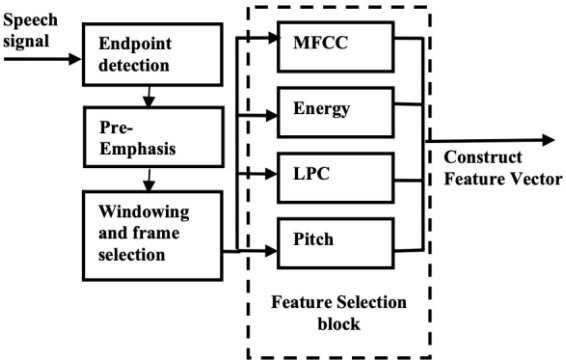

Fig. 5: Feature Extraction

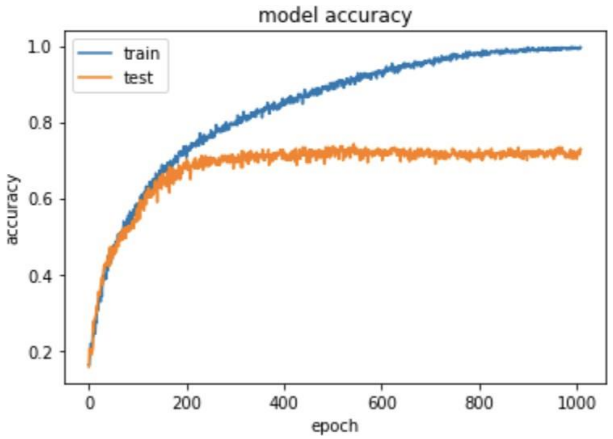

Fig. 6: Accuracy vs Iterations

state, a corresponding change occurs in pitch, energy, speech rate and spectrum. Anger often has a high level of energy, mean and variation of the pitch, with high frequency levels. Sadness, on the other hand, has decrease in the mean value, variance of pitch, the rate of speech is slow, high frequency components decreases and the energy is weak. So, visualizing emotions from speech, statistics of energy, formants, pitch,and some aspects of the spectrum can be extracted. We usethe librosa library in python to extract MFCC - Mel-FrequencyCepstral Coefficients[7] fig.[5]. This feature is one of the mostimportant ways to remove the audio signal feature and is used mainly when working with audio signals. The mel frequency cepstral coefficients help us to briefly describe the complete structure of the envelope. When extracting files we keep the time stamps of all the files for 3 seconds so that we can getthe same characteristics of all sound bytes and it is evenly measured. The sample size of each file is doubled to keep the sample frequency to find other features that will help separate the audio file because the size of the database is small.

VI. Experimental Results

After building numerous models, we have found CNN model as a best fit for our emotion classification problem.We achieved a validation accuracy of 68.7% with our ex- isting model. Our model could perform better if we had more data to work on. What is more surprising is that the model performed excellent when distinguishing between maleand female voices. We can also see fig.[6] the way model predicted against the actual values. In the future we couldbuild a sequence to sequence model to generate voice based on different emotions. E.g. A happy voice, A surprised oneetc.

TABLE I: ACCURACIES USING DIFFERENT MODELS.

Model	Layers	Activation F(n)	Epochs	Accuracy
MLP	8	softmax	550	24.76
LSTM	5	tan h	50	15.59
CNN	18	softmax,rmsprop	1000	68.7

An important function in the speech recognition systemis to select distinctions. after calculating the features of the Speech, the appropriate elements are given to the separator. The separator gets the feel from the speech of the various speakers. Creating emotional recognition in speech Various types of planning have been suggested. We used an in-depth learning model to find the components divided into five main components, namely:

0 - Female angry
1 - Female calm
2 - Female fearful
3 - Female Happy
4 - Female sad
5 - Male angry
6 - Male calm
7 -Male fearful
8 - Male happy
9 - Male sad

It was important that we distinguished using gender and the characteristics of men and women that we were very different from and gave us better accuracy in this way. Since the project is a problem of classification, the Convolution Neural Network seems to be a popular option. We also developed multilayer perceptrons and models of Long Short Term Memory but they donot work well with very low data that could not pass the test while predicting the right emotions. Building and repairing a model is a time-consuming process. The idea is to always start first without adding too many layers to make it more complex. After layout testing, the model that provided the maximum amount of complete validation against test data was 68.7%.

MLP Model: The MLP model we have developed has a very low level around 24.76% authentication with 8 layers, softmax output performance, batch size of 32 and 550 epochs.

LSTM: The LSTM model had a very low training accuracy of about 15.59% in 5 layers, tan h activation function, 32 batch size and 50 epochs.

CNN: The CNN model was at the forefront of our editing

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 13 Issue: 05 | October - December 2022
Paper is available at http://www.jrps.in | Email : info@jrps.in
DOI : 10.5281/zenodo.7335435 Refereed & Peer Reviewed

problem. After training many models we found 68.7% best verification accuracy in 18 layers, softmax activation function, rmsprop activation function, 32 batch size and 1000 epochs.

VII. CONCLUSION

Recognizing emotion in spontaneous speech is difficult because it does not carry sufficient intensity differences to distinguish one emotion from the other. In this paper, we assume that the absence of discriminatory features in noise can be managed by using prior knowledge about the status of language content and the timing of words. A major contribution of this paper is the development of an automatic speech recognition framework. Such a framework is general undermines the general mode of emotional recognition. Test results validate the use of different content-dependent information according to the end-of-speech audio and language content. The inclusion of prior knowledge and integrated integration to improve emotional recognition in spontaneous speech, or in the case of an incomparable train test, clearly sets out the benefits of the proposed framework.

REFERENCES

[1] Chakraborty, R., Pandharipande, M. and Kopparapu, S.K., 2016, December. Spontaneous speech emotion recognition using prior knowledge. In 2016 23rd International Conference on Pattern Recognition (ICPR) (pp. 2866-2871). IEEE.

[2] Aggarwal, A., Srivastava, A., Agarwal, A., Chahal, N., Singh, D., Alnuaim, A.A., Alhadlaq, A. and Lee, H.N., 2022. Two-way feature extraction for speech emotion recognition using deep learning. Sensors, 22(6), p.2378.

[3] Pappas, D., Androutsopoulos, I. and Papageorgiou, H., 2015, October. Anger detection in call center dialogues. In 2015 6th IEEE international conference on cognitive infocommunications (CogInfoCom) (pp. 139-144). IEEE.

[4] Pokorny, F.B., Graf, F., Pernkopf, F. and Schuller, B.W., 2015, September. Detection of negative emotions in speech signals using bags-of-audio-words. In 2015 International Conference on Affective Computing and Intelligent Interaction (ACII) (pp. 879-884). IEEE.

[5] Yun-Maw, C., Yue-Sun, K., Jun-Heng, Y., Yu-Te, C. and Tsang-Long, P., 2006. Using Recognition of Emotions in Speech to Better Understand Brand Slogan. In Proceedings of the International Workshop on Multimedia Signal Processing.

[6] Burkhardt, F., Van Ballegooy, M., Engelbrecht, K.P., Polzehl, T. and Stegmann, J., 2009, September. Emotion detection in dialog systems: Applications, strategies and challenges. In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (pp. 1-6). IEEE.

[7] Atila, O. and Şengür, A., 2021. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. Applied Acoustics, 182, p.108260.

[8] Mohamoud, A.A. and Maris, M., 2008. An implementation of anger detection in speech signals.

[9] Proux, D., Marchal, P., Segond, F., Kergourlay, I., Darmoni, S., Pereira, S., Gicquel, Q. and Metzger, M.H., 2009, September. Natural language processing to detect risk patterns related to hospital acquired infections. In Proceedings of the Workshop on Biomedical Information Extraction (pp. 35-41).

[10] Pawar, M.D. and Kokate, R.D., 2021. Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients. Multimedia Tools and Applications, 80(10), pp.15563-15587.

[11] Abdelhamid, A.A., El-Kenawy, E.S.M., Alotaibi, B., Amer, G.M., Abdelkader, M.Y., Ibrahim, A. and Eid, M.M., 2022. Robust Speech Emotion Recognition Using CNN+ LSTM Based on Stochastic Fractal Search Optimization Algorithm. IEEE Access, 10, pp.49265-49284.

[12] Polzehl, T., Schmitt, A., Metze, F. and Wagner, M., 2011. Anger recognition in speech using acoustic and linguistic cues. Speech Communication, 53(9-10), pp.1198-1209.

[13] Damiano, R., Lombardo, V., Monticone, G. and Pizzo, A., 2019, September. All about face. An experiment in face emotion recognition in interactive dramatic performance. In 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-7). IEEE.

[14] Kadiri, S.R. and Alku, P., 2020. Excitation features of speech for speaker-specific emotion detection. IEEE Access, 8, pp.60382-60391.

[15] Polzehl, T., Schmitt, A., Metze, F. and Wagner, M., 2011. Anger recognition in speech using acoustic and linguistic cues. Speech Communication, 53(9-10), pp.1198-1209.

[16] Schuller, B., Batliner, A., Steidl, S. and Seppi, D., 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech communication, 53(9-10), pp.1062-1087.

[17] Ververidis, D. and Kotropoulos, C., 2006. Emotional speech recognition: Resources, features, and methods. Speech communication, 48(9), pp.1162-1181.