International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 87-91

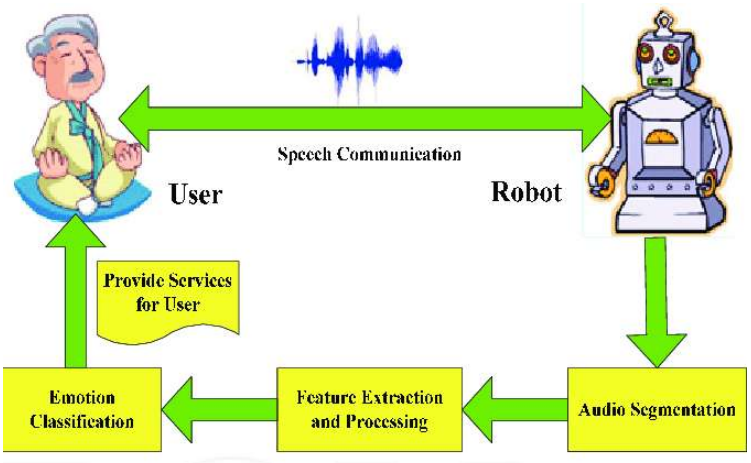
Speech Emotion Recognition and Implementation: A Survey

Varshitha M S[1], Mamatha G[2]
[1] *(Student, Department of ISE, JSS Academy of Technical Education, Bengaluru, India)*
[2] *(Assistant Professor, Department of ISE, JSS Academy of Technical Education, Bengaluru, India)*

Abstract: *The last few decades have seen a wide range of research projects focusing on automatic emotion recognition based on speech for human-machine communication, Speech is the most fundamental and natural means of communication while communicating with others, speech signal can be one of the fastest technique of communication between living beings, therefore it can serve as an efficient and fast method of communication. As the technology establishes a seamless communication between Man and Machine, speech is progressively becoming the key element of the Man-Machine interface in the IT area, which includes Computer science, signal processing, psychology, linguistics, and more. Speech Recognition technology is increasingly becoming the essential module of the Man-Machine communication system, and Speech Recognition is even correlated to an individual's body language. A study of Speech Recognition technology, its fundamental principles, methods, voice recognition process, classifications of SR systems, and datasets used is presented in this paper.*
Keywords: *Speech recognition, Speech Emotion Recognition (SER), Artificial Intelligence, CNN algorithm, dataset*

I. Introduction

Speech is the most common form of communication which is rich in paralinguistic information, to convey emotion, age, gender and other attributes in real-time. From the past few decades, Speech Emotion Recognition (SER) has developed intoan interesting research area of Computer Science related to smart home automation, social media, education, health care, and a variety of other Artificial Intelligence (AI) based applications. Fig. 1 shows a simple setup for automated SER. One of the most challenging steps of SER is the feature generation for emotions, because the features derived from raw speech signals will be able to effectively distinguish emotion states [6].



Fig, 1 Automated Speech emotion detection

Speech emotion recognition is a process of identifying human emotions from a recorded speech or in real time through theuse of advanced technologies, algorithms, and accurate datasets to train the machine or system to detect and classify these emotions based on the words used or tone of the voice. Fig. 2 shows the Architectural components of an ideal SER system. Due to the gap or disparity amongst Acoustic characteristics (intensity and frequency pattern of sound) and Human emotions (happy, sad, etc), automated Speech Emotion Recognition is a challenging procedure, which depends greatly on the distinguishable acoustic characteristics captured from a specified recognition task.

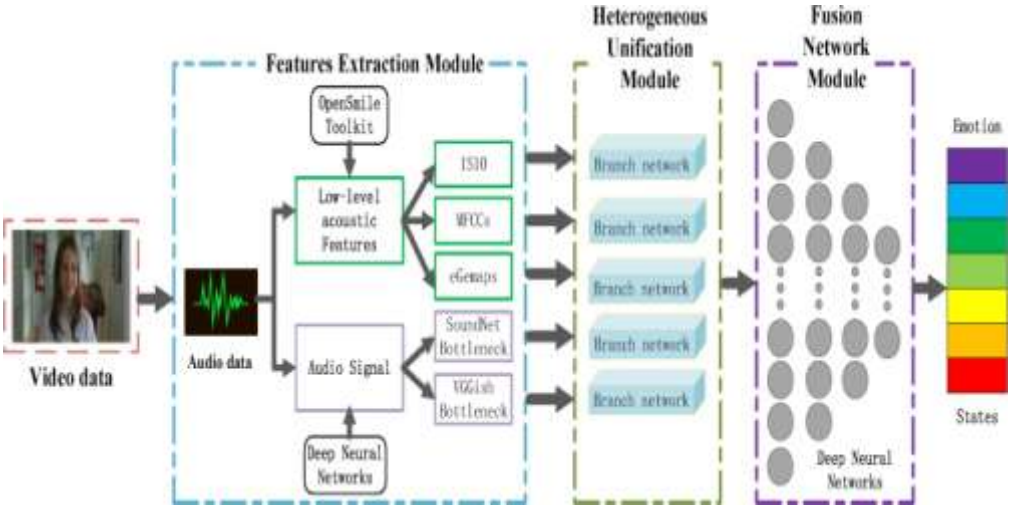


Fig. 2 Architectural components of a Speech Emotion Recognition system

Diverse people have different forms of emotion and they are expressed in different ways. Speech emotions vary in terms of the sound energy and pitch variations. Speech emotion detection is a challenging task at some stage in Computer Vision, herethe speech emotion detection is based on algorithms that utilize different modules to recognize different emotions, and classifiers are used to distinguish emotions like happiness, anger, neutrality, surprise, sadness, etc. SER system uses speech samples to identify emotions, and the characteristics are extracted from the samples using LIBROSA. The classification is based on the extracted characteristics, and we can then deduce the emotions from the speech signals.

Convolutional Neural Network (CNN) consists of convolution layers, pooling layers, fully connected layers, and a SoftMax unit; this sequential network forms a feature extraction. Initially, input spectrograms are convolved with different filters during the training phase and feature maps are obtained. Polling layers accumulate maximum activation functions from the feature maps, to reduce their dimensionality. Lastly, SoftMax unit performs the task of classification.

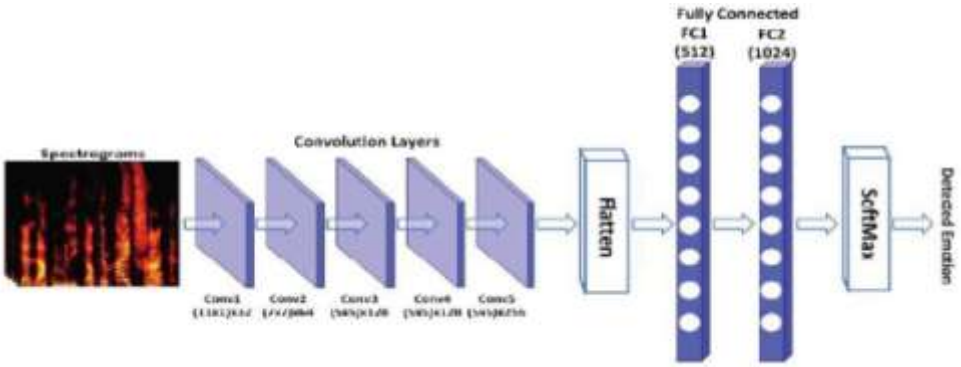


Fig. 3 Deep Stride CNN design for speech emotion recognition

Deep Stride Convolutional neural network (DSCNN) design for SER is as shown in Fig. 3, in contrast to CNN model, DSCNN eliminate the pooling layer and as a substitute uses special Strides for down-sampling or dimension reduction. Strides signify the number of pixels shifted over the input matrix and the reduction in the number of computations that are to be completed by the successive layers in the network.

DSCNN for recognizing emotions from the speech samples consists of convolutional layers, a flatten layer, fully connected layers, and a SoftMax unit. for learning and deep feature-extraction at convolutional layers DSCNN network use similar filter size (5×5) as CNN model, stride setting of 2×2 pixels is used to decrease the resolution of feature map size, spectrograms produced from SAVEE database are taken as input.

II. Literature Survey

Rohit Raj Sehgal [1] propose a technology that uses a set of unique frequencies, which is transmitted via audio channels sothe Computer can understand it quickly, at Call centers Artificial Intelligence can be used to achieve Automatic Speech Recognition (ASR), and Sentiment analysis can be used to determine whether customers are satisfied with their performance through ASR system. [1] discusses how Call centers can use ASR, virtual calling centers, Cloud computing, DTMF and PBX machines for sentiment analysis to identify customers' emotions.

Kun-Yi Huang [2] proposes a method that uses a CNN-based algorithm with Audio Word Embedding for Speech emotion recognition. When compared to an end-to-end approach, AWE provides a more comprehensible representation. In addition, CNN-based methods would be able to deal with long sequences, as opposed to LSTM methods. On the NCKU-ES database, experimental results showed that proposed method [2] outperformed the LSTM-based method by 82.34% in Emotionrecognition accuracy. Further, EMODB database was also evaluated for the portability of [2] method to other Emotion databases. The result of the evaluation was that the Audio word-based embedding was found to be beneficial for Speech emotion recognition.

Cornejo [3] explains, in order to recognize emotions in Videos based on audio and facial parts, a Two-dimensional Convolution Neural Network is used to process both audio and visual information. Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are used in conjunction with the CNN to reduce both audio and visual information by converting Audio signal into an image, then feeding it to the CNN to extract the audio and visual information. There are a number of classifiers available to recognize emotions, such as K-Nearest Neighbor (K-NN), Support Vector Machine (SVM),Gaussian Naive Bayes (GNB), Logistic Regression (LR). Datasets used in [3] are RML, BAUM- 1 and ENTERFACE 05, and they found that competitive recognition rates were 5% higher than previously reported approaches.

Girija Deshmukh [4] paper explains the use of three feature vectors - pitch, short-term energy, and Mel frequency cepstral coefficients (mfcc) – they classified three different emotions from audio signals as: anger, sadness, and happiness. Training was divided into the acted and natural corpora in a 4:1 ratio. Unlike the Mode method, which only considers the highest frequency values, the Mean method considered every value present, so it was more accurate. It was also found that accuracy was influenced by the speaker's natural tone. Their future study was to examine how emotions are classified in Indian regional languages and how they differ with respect to age.

Apoorv Singh [5] in this paper the SER system relies on the CNN algorithm which is composed of modules for recognizing emotions, and classifiers that distinguish between emotions like happiness, surprise, anger, neutrality, and sadness. Speech samples are used as a dataset for SER, and LIBROSA is used to extract the characteristic features from them. [5] They have extended their work, to integrate with the robot so that the SER system will better understand the mood of the corresponding human, helping the system to converse with the human more effectively. In addition, the proposed system can be integrated with Music apps, so as to recommend songs to users based on their mood/ emotions. For emotion distinction tasks, they created many models and identified the best CNN model. The new model was 71% accurate, but could have done better if it could distinguish voices of male and female.

Huihui [6] this paper proposes an ICNN that improves the traditional CNNs performance used for SER by making the representations of Emotion-related features more discriminate. To achieve this, they employed Convolution processes that are interactive, for feature maps of different scales. In particular, the chosen MFCC are split into parallel channels, i.e. H-MFCC and L-MFCC, by means of an ICNN process, this maximizes the benefit of the complementary advantage of each channel.To differentiate between different emotional states the derived high abstract representations are used. Through their extensive experiments, they have shown the influence of branches interactively involved during the deep feature-extraction process. ICNN have provided new guidelines for SER's fusion work, primarily with respect to multi-model fusion.

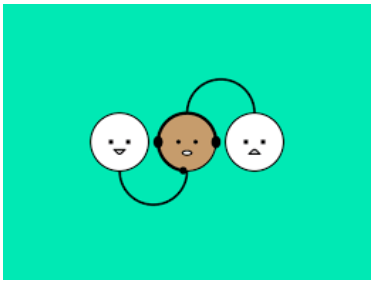
Taiba [7] in this paper, they focus on identifying emotions and designing robust methods to detect them. By eliminating Pooling layers, [7] modified the recently proposed Deep Stride Convolution Neural Networks (DSCNN), as a substitute they used Special turns to reduce the dimensionality of feature- maps. The paper proposed two experiments, to evaluate the efficiency of the current CNN and DSCNN model. Experiments, considered the emotion spectrograms generated from the speech database for training epochs: 500, 1200, and 1500. In both experiments, the performances of the models were improved as more epochs were added to the training process. DSCNN performance was better than the state-of-the-art CNN model, with 87.8% accuracy obtained by DSCNN compared to 79.4% by CNN. This infers that for convincingly detecting emotions, CNN requires further enhancement in the architecture.

III. Datasets for Speech Emotion Recognition

Datasets are the collections of different data that helps to train the System. A number of datasets are available which has their own unique characteristics and accuracy levels to train the system or robot, to detect human emotion of humans from their speech or voice. Description of few datasets is as follows:

- The RAVDESS dataset can be downloaded from Kaggle.com , it contains 1440 files (voice samples of 24 actors, 60 trails each), 24 professional voices (12 females, 12 males), actors speaking in English with North-American accent, and expressing emotions such as happy, angry, sad, fearful, calm, disgust and surprise. The expressions aregenerated in two intensity levels (light and bold), and with a neutral expression; every file out of 1440 files, has a unique filename and the filename contains a unique 7-part numerical identification.
- Chinese Academy of Sciences' Institute of Automation (CASIA) has compiled a database of Chinese emotional corpora which include four pronouncers, 9600 different pronunciations, 300 texts that are same and 100 others that are different, and six emotion labels: angry, happy, fearful, shocked, and neutral.
- Surrey Audio-Visual Expressed Emotion (SAVEE), has 480 English utterances, recorded by DC, JE, JK, and KL, consisting of 15 sentences for each of the seven different emotions like- happiness, disgust, sadness, anger, fear, surprise, and neutral. Each emotion has 60 sentences, and neutral has 120 sentences.
- AESDD dataset is an audio file with a file size of 0.392 GB, having around 500 utterances by five different actors, simulated to represent a range of emotions like- fear, happiness, anger, disgust, sadness.
- Dataset developed by CMU-MOSEI in 2018, contains 65 hours of annotated video from more than 1000 speakers on250 topics, with 6 Emotions (anger, fear, disgust, happiness, sadness, surprise) and dataset is available in both Audio and Video format to train a system.

IV. Applications of Speech Emotion Recognition



(a) BPO call centers

(b) Automated Health diagnosis

(c) Automated Helpline services

Fig. 4 Applications of Speech emotion recognition

In our daily life, Speech Emotion Recognition are used in a number of applications some are shown in fig. 4,

1. To analyze the behavior of Call center representatives with Customers, and helping representatives to improve thequality of customer service.
2. SER is useful for enhancing the naturalness of Human-computer interaction in speech recognition systems.
3. Diagnosis of Mental disorders,

4. Lie detection, in crime investigation
5. Emotion analysis of Telephone conversations between Criminals would assist the Law enforcement department.
6. In Aircraft cockpit, for better performance, Speech recognition systems are equipped with speech detectors andSER.
7. Interesting and more practical application of SER is in E-Tutoring, Interactive movies, and Story-telling, asapplications are expected to adapt to Listener or Student emotional state.
8. Conversations with Robotic pets and Humanoids would be more realistic and entertaining, if they could understand, and respond to Human emotions.

V. Conclusion

In this paper we surveyed a few existing research works for finding different ways of human emotion recognition through speech or voice signals. Emotion recognition is a technique that allows for reading of emotions of a human through different forms of human actions such as facial expression, vocal-speech, hand writing/ textual meaning, hand signals or body gestures, etc. In this paper we are mainly studying Speech based emotion detection using advanced technologies. After studying some research papers, we are able to list the possible AI/ ML algorithms used for speech emotion detection, some are listed as follows, CNN, LSTM, DCNN, VSM and hybrid approaches. It has been observed that for Speech emotion detection and feature-extraction, Convolution Neural Network (CNN) algorithm and its variants DCNN gives greater accuracy levels compared to any other methods. Good number of Datasets with a variety of speech samples was studied, as these classifiers play a very important role to differentiate discrete Human emotions such as happiness, anger, neutral, etc based on the range of voice or speech waves. After the study of essential features of the basic components required for building SER system and knowing the versatile applications of SER system in real world applications in the field of Education, health, BPO, crime detection, etc. we are interested to propose a conversing Human- interaction model based on the mood of the user.

Reference

[1] R. R. Sehgal, S. Agarwal and G. Raj, "Interactive Voice Response using Sentiment Analysis in Automatic Speech Recognition Systems," 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2018, pp. 213-218, doi: 10.1109/ICACCE.2018.8441741.

[2] K. Huang, C. Wu, Q. Hong, M. Su and Y. Zeng, "Speech Emotion Recognition using Convolutional Neural Network with Audio Word- based Embedding," 2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2018, pp. 265-269, doi: 10.1109/ISCSLP.2018.8706610.

[3] J. Cornejo and H. Pedrini, "Bimodal Emotion Recognition Based on Audio and Facial Parts Using Deep Convolutional Neural Networks," 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 2019, pp. 111-117, doi: 10.1109/ICMLA.2019.00026.

[4] Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni, "Speech based Emotion Recognition using Machine Learning", 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)

[5] Apoorv Singh, Kshitij Kumar Srivastava, Harini Murugan, "Speech Emotion Recognition Using Convolutional Neural Network (CNN)", International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08 2020.

[6] H. Cheng and X. Tang, "Speech Emotion Recognition based on Interactive Convolutional Neural Network," 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), 2020, pp. 163-167, doi: 10.1109/ICICSP50920.2020.9232071.

[7] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi and N. Ismail, "Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional Neural Networks," 2020 6th International Conference on Wireless and Telematics (ICWT), 2020, pp. 1-6, doi: 10.1109/ICWT50448.2020.9243622.