Speech Emotion Recognition Using Convolutional Neural Networks

Aishwarya V[1], Faseeha Fathima J[1], Jagadale Rutuja T[1], Jaganathan K[1], Sangeetha Priya R[1]

[1]Student, Department Of Information Technology, Sona College of Technology, Salem, Tamil Nadu, India.

Corresponding Author: aishwarya1972000@gmail.com

Abstract: - Speech is the most natural and convenient ways by which humans communicate, and understanding speech is one of the most intricate processes that human brain performs. Speech Emotion Recognition (SER) aims to recognize human emotion from speech. This is on the fact that voice often reflects underlying emotions through tone and pitch. The libraries used are Librosa for analyzing audio and music, sound file for reading and writing sampled sound file formats, sklearn for building the model. In the current study, the efficacy of Convolutional Neural Network (CNN) in recognition of speech emotions has been investigated. Spectrograms of the speech signals are used as the input features of the networks. Mel-Frequency Cepstral Coefficients (MFCC) is used to extract features from audio. Our own speech dataset is used to train and evaluate our models. Based on the evaluation, the emotions (happy, sad, angry, neutral, surprised, disgust) of the speech will be detected.

Key Words: — ***Speech emotion, Energy, Pitch, Librosa, Sklearn, Sound file, CNN, Spectrogram, MFCC.***

I. INTRODUCTION

Speech emotion recognition (SER) was a technology that extracts emotional feature from speech by analysis the characteristic parameters and the emotional change acquired. At present, speech emotion recognition is an emerging crossing field of artificial intelligence [1]. Speech emotion processing and recognition system is generally composed of three parts which are speech signal acquisition, feature extraction, and emotion recognition. In this system, the quality of extraction directly affects the accuracy of speech emotion recognition. In feature extraction, it usually took the whole emotion sentence as units for feature extracting, and extraction contents. The neural networks of human brain are strongly competent to learn high-level abstract concepts from experiencing low-level information processed by sensory periphery. Speech is humans communicate, and understanding speech is one of the most intricate processes that human brain performs. It has been argued that children who are not able to understand the emotional states of the speakers developed poor social skills and in some cases they show psychopathological symptoms [2, 3].

This highlights the importance of recognizing the emotional states of speech in effective communication.

Manuscript revised March 26, 2021; accepted March 27, 2021. Date of publication March 28, 2021.
This paper available online at www.ijprse.com
ISSN (Online): 2582-7898

Detection of emotion from facial expressions and biological measurements such as heart beats or skin resistance formed the preliminary framework of research in emotion recognition [4].

More recently, emotion recognition from speech signal has received growing attention. The traditional approach toward this problem was based on the fact that there are relationships between acoustic features and emotion. In other words, the emotion is encoded by acoustic and prosodic correlates of speech signals such as speaking rate, intonation, energy, formant frequencies, fundamental frequency (pitch), intensity (loudness), duration (length), and spectral characteristic (timbre) [5, 6]. There are a variety of machine learning algorithms that have been examined to classify emotions based on their acoustic correlates in speech utterances. In the current study, we investigated the capability of convolutional neural networks in classifying speech emotions using our own dataset. There are a variety of machine learning algorithms that have been examined to classify emotions based on their acoustic correlates in speech utterances. In the current study, we investigated the capability of convolutional neural networks in classifying speech emotions using our own dataset. The specific contribution of this study is using wide-band spectrograms instead of narrow-band spectrograms as well as assessing the effect of data augmentation on the accuracy of models. Our results revealed that wide-band spectrograms and data augmentation equipped CNNs to achieve the state-of-the art accuracy and surpass human performance.

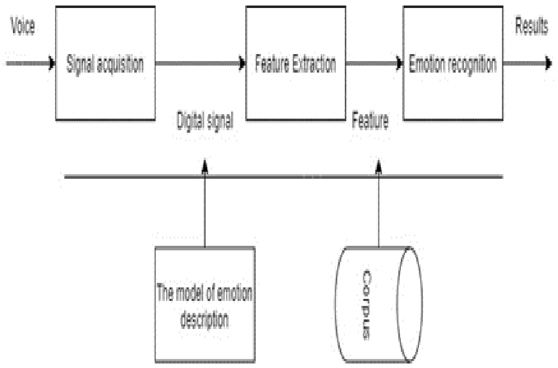

Fig.1. Speech emotion recognition block diagram

II. Related Work

Most of the papers published in last decade use spectral and prosodic features extracted from raw audio signals. The process of emotion recognition from speech involves extracting the characteristics from a corpus of emotional speech selected or implemented, and after that, the classification of emotions is done on the basis of the extracted characteristics. The performance of the classification of emotions strongly depends on the good extraction of the characteristics (such as combination of MFCC acoustic feature with the energy prosodic feature [7]. Yixiong Pan in [8] used SVM for three class emotion classification on Berlin Database of Emotional Speech [9] and achieved 95.1% accuracy.

Noroozi et.al. Proposed a versatile emotion recognition system based on the analysis of visual and auditory signals. He used 88 features (Mel frequency cepstral coefficients (MFCC), filter bank energies (FBEs)) using the Principal Component Analysis (PCA) infeature extraction to reduce the dimension of features previously extracted revealed that wide-band spectrograms and data augmentation equipped CNNs to achieve the state-of-the art accuracy and surpass human performance.

The performance of the classification of emotions strongly depends on the good extraction of the characteristics (such as combination of MFCC acoustic feature with the energy prosodic feature [7]. Yixiong Pan in [8] used SVM for three class emotion classification on Berlin Database of Emotional Speech [9] and achieved 95.1% accuracy.

Noroozi et.al. proposed a versatile emotion recognition system based on the analysis of visual and auditory signals. He used 88 features (Mel frequency cepstral coefficients (MFCC), filter bank energies (FBEs)) using the Principal Component Analysis (PCA) in feature extraction to reduce the dimension of features previously extracted [10]. S. Lalitha in [11] used pitch and prosody features and SVM classifier reporting 81.1% accuracy on 7 classes of the whole Berlin Database of Emotional Speech. Zamil et al also used the spectral characteristics which is the 13 MFCC obtained from the audio data in their proposed system to classify the 7 emotions with the Logistic Model Tree (LMT) algorithm with an accuracy rate 70% [12]. Yu zhou in [13] combined prosodic and spectral features and used Gaussian mixture model super vector based SVM and reported 88.35% accuracy on 5 classes of Chinese-LDC corpus.

H.M Fayek in [14] explored various DNN architecture and reported accuracy around 60% on two different database eNTERFACE [15] and SAVEE [16] with 6 and 7 classes respectively. Fei Wang used combination of Deep Auto Encoder, various features and SVM in [17] and reported 83.5% accuracy on 6 classes of Chinese emotion corpus CASIA. In contrast to these traditional approaches more novel papers have been published recently employing Deep Neural Networks into their experiments with the promising results. Many authors agree that the most important audio characteristics to recognize emotions are spectral energy distribution, Teager Energy Operator (TEO) [18], MFCC, Zero Crossing Rate (ZCR), and the energy parameters of the filter bank energies (FBEs) [19].

III. Traditional System

The traditional system was based on the analysis and comparison of all kinds of emotional characteristic parameters, selecting emotional characteristics with high emotional resolution for feature extraction. In general, the traditional emotional feature extraction concentrates on the analysis of the emotional features in the speech from time construction, amplitude construction, and fundamental frequency construction and signal feature [28].

IV. Proposed Method

Convolutional Neural Network (CNN) is used to classify the emotions (happy, sad, angry, neutral, surprised, disgust) and to predict the output by showing its accuracy.

The given speech is plotted as spectrogram by using matplot library and this is used as input for CNN to build the model.

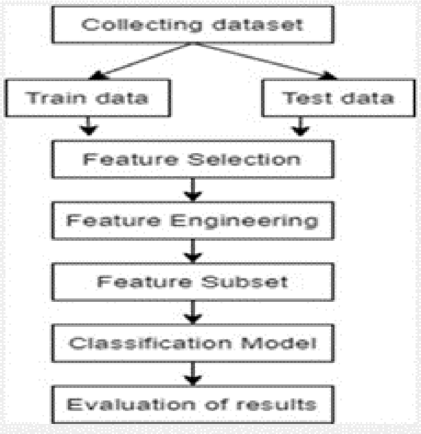

Fig.2. Flow diagram of proposed system.

A. Data Set Collection

The first step is to create an empty dataset that will hold the training data for the model. After creating an empty dataset, the data's (audio) have to be recorded and labeled in different classes. Once the labeling is done, the data's have to be pre-processed which will produce the clear pitch of the data by removing its unwanted background noise. After pre-processing the data's are classified into train dataset and test dataset, where the train dataset hold 75% of the data and the test dataset holds 25% of the data.

B. Feature Extraction of Speech Emotion

Human speech consists of many parameters which show the emotions compromise in it. As there is change in emotions these parameters also gets changed. Hence it's necessary to select proper feature vector to identify the emotions. Features are categorized as excitation source features, spectral features, and prosodic features. Excitation source features are achieved by suppressing characteristics of vocal tract (VT). Spectral features used for emotion recognition are linear prediction coefficients (LPC), Perceptual Linear prediction coefficients (PLPCs), Mel-frequency cepstral coefficients (MFCC), linear prediction cepstrum coefficients (LPCC), and perceptual linear prediction (PLP). The accuracy of differentiating different emotions can be achieved by using MFCC, LFPC [20, 21].

C. Mel-Frequency Cepstral Coefficients

The Mel-Frequency Cepstral Coefficients (MFCC) feature extraction method is a leading approach for speech feature extraction. The various steps involved in MFCC feature extraction are:

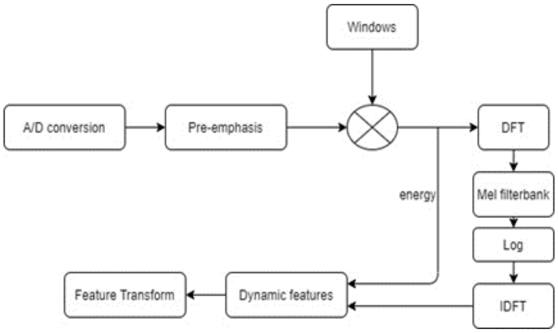

Fig.3. Flow of MFCC

*A/D conversion***:**

This converts the analog signal into discrete space.

*Pre-emphasis***:**

This boosts the amount of energy in the high frequencies.

Windowing:

Windowing involves the slicing of audio waveform into sliding frames.

Discrete Fourier Transform:

DFT is used to extract information in the frequency domain [22, 23].

D. Classifiers

After extracting features of speech, it is essential to select a proper classifier. Classifiers are used to classify emotions. In the current study, we use Convolutional Neural Network (CNN). The term Convolutional comes from the fact that Convolution-the mathematical operation is employed in these networks. Convolutional Neural Networks is one of the most popular Deep Learning Models that have manifested remarkable success in the research areas. CNN is a deep learning algorithm that takes image as an input, assign importance to various aspects in the image and will be able to differentiate from other. Generally CNNs have three building blocks: the convolutional layer, the pooling layer, and the fully connected layer. Following, we describe these building blocks along with some basic concept such as soft max unit, rectified linear unit, and drop out.

- *Input layer:* This layer holds the raw input image.
- *Convolution Layer:* This layer computes the output volume by computing dot product between all filters and image patch.
- *Activation Function Layer:* This layer will apply element wise activation function to the output of convolution layer.

- *Pool Layer:* This layer is periodically inserted in CNN and its main function is to reduce the size of volume which makes computation fast and reduces memory. The two types are Maxpooling and average pooling.
- *Fully-Connected Layer:* This layer takes input from the previous layer and computes the class scores and outputs the 1-D array of size equal to the number of classes [24, 25].

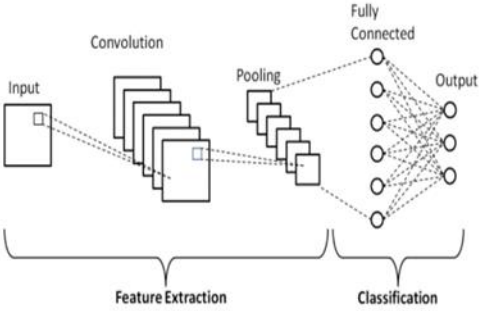

Fig.4. CNN Algorithm.

V. Application

The applications of speech emotion recognition system are, psychiatric diagnosis, conversation with robots, intelligent toys, mobile based emotion recognition, emotion recognition in call centre where emotions of customer can be identified and can help to get better service quality, intelligent tutoring system, lie detection, games[26,27]. It is also used in healthcare, Psychology, cognitive science and marketing, voice-based virtual assistants.

VI. Conclusion

In this paper, we proposed a method that used the CNN algorithm, one of the Deep Learning algorithms, to extract the emotional characteristic parameter from emotional speech signal. Previous literature mostly used narrow-band spectrograms, which have higher frequency resolution than wide-band spectrograms and resolved individual harmonics. On the other hand, wide-band spectrograms have higher time resolution than narrow-band spectrograms and show individual glottal pulses, which are associated with fundamental frequency and pitch. The CNNs perform well on training data. The results of current study manifested the competency of CNNs in learning the underlying emotional features of speech signals from their low-level representation using wide-band spectrums.

VII. Future Scope

For future work, we suggest to use audio-visual database or audio-visual-linguistic databases to train Deep Learning models where facial expressions and semantic information are taken into account as well as speech signals, which allows improving the recognition rate of each emotion. In future, we can think about using other types of features and apply our system on other bases that are larger and used other method for feature extraction.

References

[1]. Z. Yongzhao and C. Peng, “Research and implementation of emotional feature extraction and recognition in speech signal,” Journal of Jiangsu University, volume. 26, No. 1, pp.72-75, 2005.

[2]. Monita Chatterjee, Danielle J Zion, Mickael L Deroche, Brooke A Burianek, Charles J Limb, Alison P Goren, Aditya M Kulkarni, and Julie A Christensen. Voice emotion recognition by Cochlear-implanted children and their normally-hearing peers. Hearing research, 322:151-162, 2015.

[3]. Nancy Eisenberg, Tracy L Spinrad, and Natalie D Eggum. Emotional-related self-regulation and its relation to children’s maladjustment. Annual review of clinical pschycology, 6:495-525, 2010.

[4]. Harold Schlosberg. Three dimensions of emotion. Psychological review, 61(2):81, 1954.

[5]. Louis Ten Bosch. Emotions, speech and the asr framework. Speech communication, 40(1-2):213-225, 2003.

[6]. Thomas S Polzin and Alex Waibel. Detecting emotions in speech. In proceedings of the CMC, volume 16. Citeseer, 1998.

[7]. Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag Singh, Promod Yenigalla,“Speech emotion recognition using kernel sparse representation based classifier,” in 2016 24th European Signal Processing Conference(EUSIPCO), pp.374-377, 2016.

[8]. Pan, Y., Shen, P. and Shen, L., 2012. Speech emotion recognition using support vector machine. International Journal of Smart Home, 6(2), pp.101-108.

[9]. Burkhardt, F.,Taeschke, A.,Rolfes, M.,Sendlmeier, W.F.and Weiss,B., 2015, September. A database of German emotional speech. In Interspeech (vol.5,pp.1517-1520).

[10]. Noroozi, F., Marjavonic, M., Njegus, A., Escalera, S., &Anbarjafari, G. Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing, 2017.

[11]. Lalitha , S., Madhavan, A., Bhushan, B, and Saketh, S., 2014, October. Speech emotion recognition. In Advances in Electronics, Computer characteristics. The performance of the classification of emotions strongly depends on the good

extraction of the characteristics (such as combination of MFCC acoustic feature with the energy prosodic feature [7]. Yixiong Pan in [8] used SVM for three class emotion classification on Berlin Database of Emotional Speech [9] and achieved 95.1% accuracy. Noroozi et al proposed a versatile and Communications (ICAECC), 2014 International Conference on (pp. 1-4). IEEE.

[12]. Zamil, Adib Ashfaq A., et al. "Emotion Detection from Speech Signals using Voting Machanism on Classified Frames." 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, 2019.

[13]. Zhou, Y., Sun, Y., Zhang, J., and Yan, Y., 2009, December. Speech emotion recognition using both spectral and prosodic features. In 2009 International Conference on Information Engineering and Computer Science (pp. 1-4). IEEE.

[14]. Fayek, H. M., M. Lech, and L. Cavedon. "Towards real-time speech emotion recognition using Deep Neural Networks." Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on IEEE, 2015.

[15]. Martin, O., Kotsia, L., Macq, B., and Pitas, I., 2006, April. The eNtERFACE'05 audio-visual database. In 22nd International Conference on Data Engineering Workshops (ICDEW'06) (pp 8-8). IEEE.

[16]. Sanjitha. B. R, Nipunika. A, Rohita Desai. "Speech Emotion Recognition using MLP", IJESC.

[17]. Ray Kurzweil. The singularity is near. Gerald Duckworth & Co, 2010.

[18]. Hadhami Aouani and Yassine Ben Ayed, "Speech Emotion Recognition with Deep Learning", 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems.

[19]. Pavol Harar, Radim Burget and Malay Kishore Dutta "Efficiency of chosen ` speech descriptors in relation to emotion recognition," EURASIP Journal on Audio, Speech, and Music Processing, 2017.

[20]. Idris I., Salam M.S. "Improved Speech Emotion Classification from Spectral Coefficient Optimization". Lecture Notes in Electrical Engineering, vol 387. Springer, 2016.

[21]. Pao TL., Chen YT., Yeh JH., Cheng YM., Chien C.S. "Feature Combination for Better Differentiating Anger from Neutral in Mandarin Emotional Speech", LNCS: Vol. 4738 Berlin: Springer 2007.

[22]. J&M: Daniel Jurafsky and James H. Martin (2008). Speech and Language Processing, Pearson Education (2nd edition).

[23]. Hyenk Hermansky, "Perceptual linear Predictive (PLP) analysis of speech," The Journal of the Acoustical Society of America, Vol.87, No.4, pp.1737-1752, 1980.

[24]. A. Berg, J. Deng, and L. Fei-Fei. Large scale visual recognition challenge 2010. www.upgrad.com/blog/.

[25]. L.Breiman, Random forests, Machine Learning, 45(1):5-32, 2001.

[26]. A. B Ingale and D. S. Chaudhari, "Speech Emotion Recognition", International Journal of Soft Computing and Engineering (IJSCE), March 2012,pp.235-238.

[27]. A. S. Utrane and S.L. Nalblwar, "Emotion Recognition through Speech", International Journal of Applied Information Systems (IJAIS), 2013, pp.5-8.

[28]. Z. Li, "A study on emotional feature analysis and recognition in speech signal, " Journal of China Institute of Communications, vol. 21, no.10,pp. 18-24,2000.