INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 544-548

e-ISSN : 2583-1062

Impact Factor: 5.725

AUTOMATIC SPEECH EMOTION RECOGNITION USING DEEP LEARNING TECHNIQUES

Prajkta zanje[1], Mohini Bhosale[2], Dhas Payal[3], Thatar Pooja[4], Unde Suvarna[5]

[1,2,3,4]Student, Computer Engg, RGCOE Ahmednagar, India.

[5]Asst. Prof, Computer Engg, RGCOE Ahmednagar, India.

ABSTRACT

The speech emotion recognition system as a set of methodologies that process and classify speech signals to detect the emotions embedded in them. In this study, we set out to detect basic emotions in recorded speech by analysing the acoustic properties of the audio data of the recordings. Emotions are an integral part of human behaviour and inherited characteristics as a way of communication. We humans are well trained because your reading experience recognizes different emotions which make us more reasonable and understandable. But only in the case of a machine, it can easily understand content-based information such as information in text, audio or video, but it is still far behind in accessing the depth of content. There are three classes of features during speech, namely lexical features (vocabulary used), visual features (expressions the speaker makes) and thus acoustic features.

Key Words: Emotion recognition, Emotion detection, lexical features, visual features, acoustic features, Image recognition; Signal processing; Image classification**.**

1. INTRODUCTION

As human beings, speech is among the most natural ways to express ourselves. We are so dependent on it that we realize its importance when we resort to other forms of communication, such as emails and text messages, where we often use emoticons to express the emotions associated with messages. Since emotions play a vital role in communication, their detection and analysis 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 set of methodologies that process and classify speech signals to detect emotions embedded in them. Such a system can find use in a wide range of application areas, such as interactive analysis of caller-agent conversation. In this study, we attempt to reveal the underlying emotions in recorded speech by analysing the acoustic properties of the audio data of the recordings. Emotions play a significant role in everyday human interactions. This is necessary for our rational and intelligent decisions. It helps us to tune in and understand the feelings of others by communicating our feelings and giving feedback to others. Research has revealed the powerful role that emotions play in shaping a person social interaction. Emotional expressions provide considerable information about an individual's mental state.

2. LITERATURE SURVEY

Bjorn Schuller et.al., This system uses two approaches. Gaussian mixture models are used to classify the global statistical framework of the first speech method using inferred raw pitch and energy contour features of the speech signal. The second technique used continuous hidden Markov models to make time more complex. They extracted 20 features from the background and introduced rough outlines using the first method. Pitch-related properties, energy-related properties and processing of the resulting properties.

Li Zheng et.al., A CNN model was used as a feature extractor to extract high-order features from the spectrogram. A voice emotion recognition system was designed and implemented using RF as a classifier. The speech signals in this case are divided into frames and the spectrogram was generated from the emotional speech samples using framing, windowing, short-time Fourier transform (STFT) and power spectral density (PSD).

The normalized spectrogram was then fed into the CNN model as input. The CNN was used to extract the speech emotion features and the output of the CNN Flatten layer was used to feed the speech emotion eigenvectors to the RF classifier. The test voice signals were converted into spectrograms for the recognition phase, where they were then fed into a CNN-RF model classifier to identify different speech moods. In this study, the RF classifier is integrated with the CNN model as a feature extractor.

III. PROPOSED MODEL

Human speech consists of many parameters that show what emotions are contained in it. Since there is a change in emotion, this right vector is intended to identify the emotion. The features are categorized as stimulus source features, spectral features, and prosodic features. The excitation source characteristics are achieved by suppressing the vocal tract (VT) characteristics. Spectral properties used for emotion recognition are linear prediction coefficients (LPC),

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

e-ISSN : 2583-1062

Impact Factor: 5.725

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 544-548

perceptual linear prediction coefficients (PLPC), Mel-frequency spectrum coefficients (MFCC), cestrum linear prediction coefficients (LPCC), perceptual linear prediction (PLP). Prosodic properties used for emotion recognition are pitch, energy, intensity. Statistical measures are also used to distinguish emotions such as minimum, maximum, standard deviation, range, mean, median, variance, skewness, kurtosis, etc. properties.

Categorization of emotions

In this section, People can express their emotions through many different types of non-verbal communication including facial expressions, quality of speech produced and physiological signals of the human body. we discuss each of these categories.

1.Facial expressions- The human face is extremely expressive, capable of expressing countless emotions without words [31]. And unlike some forms of non-verbal communication, facial expressions are universal. Facial expressions for happiness, sadness, anger, surprise, fear, and disgust are the same across cultures.

2. Speech- In addition to faces, voices are an important modality for emotional expression. Speech is a relevant communication channel enriched with emotions: the voice in speech conveys not only a semantic message, but also information about the speaker's emotional state. Some important voice vectors have were selected for research such as fundamental frequency, mel-frequency cepstral coefficient (MFCC), predictive cepstral coefficient (LPCC), etc.

3. Physiological signals- Physiological signals related to the autonomic nervous system make it possible to objectively assess emotions. These include electroencephalogram (EEG), heart rate (HR), social media and machine learning electrocardiogram (ECG), respiration (RSP), blood pressure (BP), electromyogram (EMG), skin conductance (SC), blood volume pulse (BVP) and skin temperature (ST). Using physiological signals to recognize emotions is also useful for people who suffer from physical or mental illness and therefore have problems with facial expressions or tone of voice.

3. SPEECH EMOTION RECOGNITION (SER) SYSTEM

Our SER system consists of four main steps. The first is a collection of voice samples. The second feature vector, which is formed by feature extraction. As a next step, we tried to determine which properties are most relevant for differentiating individual emotions. These features are fed into a machine learning classifier for recognition.

The speech signal contains a large number of parameters that reflect emotional characteristics. One of the problems in emotion recognition is what features should be used. In recent research, many common features such as energy, pitch, formant and some spectral features such as linear prediction coefficients (LPC), mel-frequency cepstrum coefficients (MFCC) and modulation spectral features have been extracted. In this work, we selected modulation spectral features and MFCC to extract emotional features. Mel-frequency cepstrum coefficient (MFCC) is the most widely used expression of the spectral properties of voice signals. These are best for speech recognition because it takes into account the sensitivity of human perception with respect to frequencies.

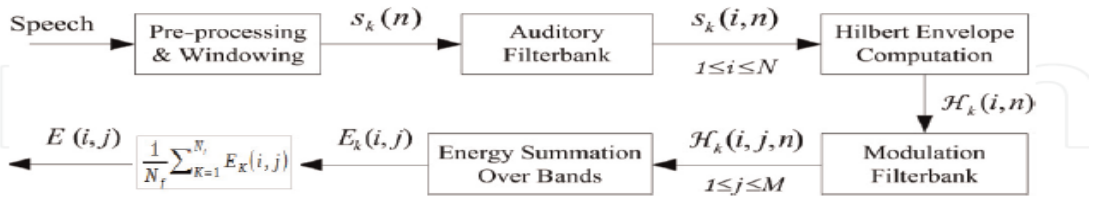

Fig 1. Process for computing the ST representation

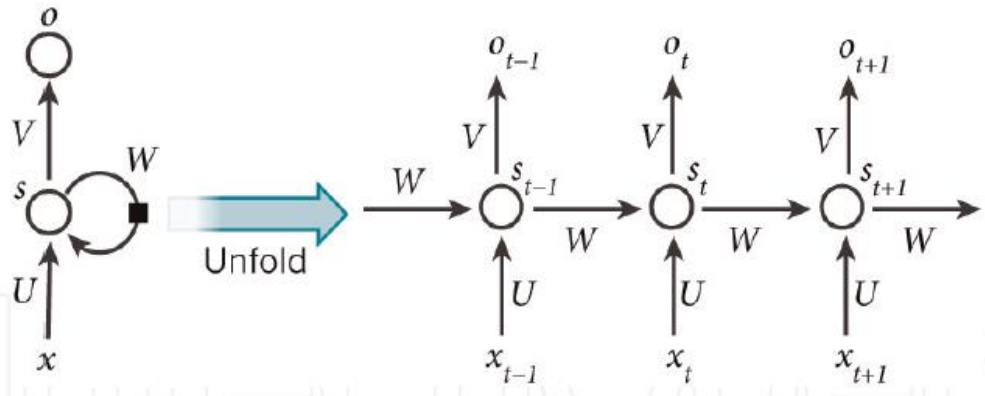

Fig. 2 A basic concept of RNN

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)

www.ijprems.com
editor@ijprems.com

Vol. 04, Issue 03, March 2024, pp: 544-548

e-ISSN : 2583-1062

Impact Factor: 5.725

TABLE 1. Confusion Matrix For Feature Combination After Selection Based On Spanish Database.

Emotion	Anger	Disgust	Fear	Joy	Neutral	Surprise	Sadness	Rate (%)
Anger	79	1	0	1	2	3	0	91.86
Disgust	0	67	3	0	1	0	1	93.05
Fear	0	3	70	0	1	0	2	93.33
Joy	3	1	1	71	0	0	0	93.42
Neutral	2	0	1	0	156	0	1	97.50
surprise	2	1	0	3	0	60	0	92.30
Sadness	0	0	1	0	2	0	66	95.65
Precision (%)	91.86	91.78	92.10	94.66	96.29	95.23	94.28	

4. CLASSIFIER SELECTION

In a speech emotion recognition system, after feature calculation, the best features are provided to the classifier. The classifier recognizes the emotion in the speaker's speech. Different types of classifiers have been proposed for the speech emotion recognition task. Gaussian mixtures model (gmm), k-nearest neighbors (knn), hidden markov model (hmm) and support vector machine (svm), artificial neural network (ann) etc. are classifiers used in speech emotion recognition system. Each classifier has certain advantages and limitations over the others.

Another classifier used for emotion classification is the Artificial Neural Network (ANN), which is used due to its ability to find non-linear boundaries separating emotional states. Among the many types, the feedforward neural network is most commonly used in speech emotion recognition [7]. Perceptron-layer multilayer neural networks are relatively common in speech emotion recognition because they are easy to implement and have a well-defined training algorithm [1]. ANN-based classifiers can achieve a correct classification of 51.19% in speaker-dependent recognition and 52.87% for speaker-independent recognition. According to the emotional state of utterances k, the k-nearest neighbor (k-NN) classifier assigns the utterance to an emotional state. The classifier can correctly classify all utterances in the design set if "k" is equal to 1, but its performance on the test set will decrease. Using pitch and energy contour information, the K-NN classifier achieves an accurate classification of 64% for the four emotional states.

5. CNN ARCHITECTURE

A convolutional neural network, also known as CNN or ConvNet, is a class of neural network that specializes in processing data that has a grid-like topology, such as an image. A digital image is a binary representation of visual data. It contains an array of pixels arranged in a grid that contains pixel values that determine how bright and what color each pixel should be.

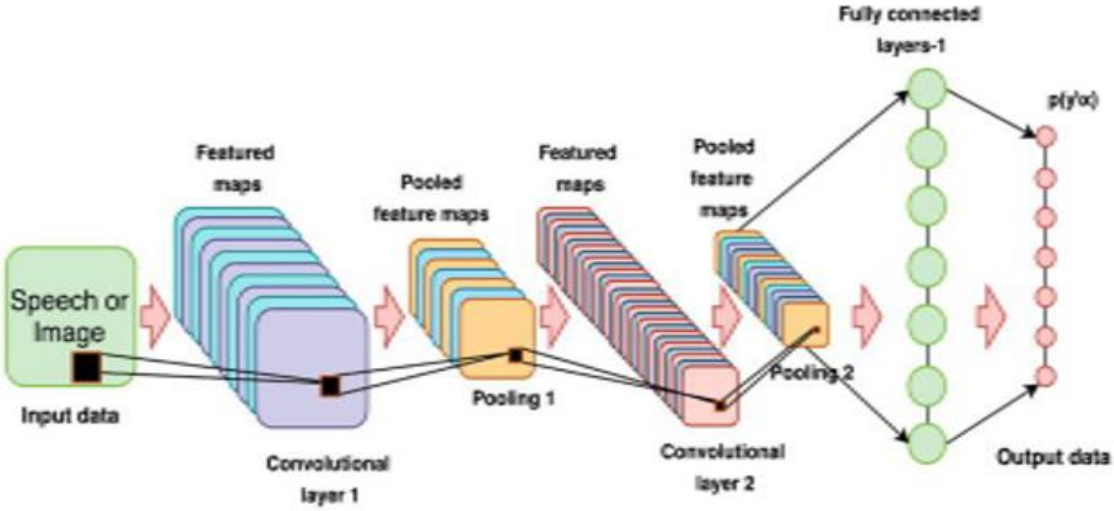

Fig 3. CNN Architecture

Convolutional Neural Networks, sometimes referred to as CNNs or ConvNets, are a subclass of neural networks that are particularly adept at processing inputs with a lattice-like architecture. Binary visual data is represented as a digital image. It has a number of pixels that are arranged like a grid and each one is assigned a value that indicates how bright and what color each pixel should be.

1.Convolution layer: The basic component of CNN is the convolutional layer. It carries most of the network's computing load. The kernel traverses the entire image in several iterations. At the end of each iteration, the dot product between the input pixels and the filter is calculated. A feature map or convolved feature is the result of connecting dots in a certain pattern. In this layer, the image is finally transformed into numerical values that the CNN can understand and extract relevant patterns from.

2. Pooling Layer: A pooling layer, like a convolutional layer, spreads a kernel or filter over the input image. Unlike the convolutional layer, the pooling layer has fewer input parameters, but it also causes some information to be lost. Positively, this layer simplifies the CNN and increases its efficiency. By calculating summary statistics from surrounding outputs, the pooling layer replaces the output of the network at specific locations.

3. Fully connected layer: Based on the features extracted in the preceding layers, picture categorization in the CNN takes place in the FC layer. Fully connected in this context means that every activation unit or node of the subsequent layer is connected to every input or node from the preceding layer. The CNN does not have all of its layers fully connected because that would create an excessively dense network. It would cost a lot to compute, increase losses, and have an impact on output quality.

4. Working of CNN: Multiple layers of a CNN are possible, and each layer trains the CNN to recognize the various elements of an input image. Each image is given a filter or kernel to create an output that gets better and more detailed with each layer. The filters may begin as basic characteristics in the lower layers. In order to check and identify features that specifically reflect the input item, the complexity of the filters increases with each additional layer. As a result, the partially recognized image from each layer's output, or convolved image, serves as the input for the subsequent layer. The CNN recognizes the image or object it represents in the final layer, which is an FC layer.

6. CONCLUSION

The proposed scheme represented an approach to recognizing emotions from human speech. This approach was implemented by a neural network. This dissertation focuses on a feature extraction method that is useful in emotion recognition through a speech signal. Mel Frequency Cepstrum Coefficient (MFCC) is used for feature extraction purposes. A high pass function is designed to achieve good extraction. In this current study, we presented an automatic speech emotion recognition (SER) system using three machine learning algorithms (MLR, SVM, and RNN) to classify seven emotions. So there were two types of functions (MFCC and MS).

extracted from two different controlled databases (Berlin and Spanish databases) and a combination of these properties was presented.

7. REFERENCES

[1] Pan Y, Shen P, Shen L. Speech emotion recognition using support vector machine. International Journal of Smart Home. 2012; 6:101-108

[2] Schirmer A, Adolphs R. Emotion perception from face, voice, and touch: Comparisons and convergence. Trends in Cognitive Sciences. 2017;21(3): 216-228

[3] Ekman P. An argument for basic emotions. Cognition & Emotion. 1992; (3–4):169-200

[4] A.P. Wanare, S.N. Dandare, "Human Emotion from Speech", Int. Journal of Research and Applications,vol. 4, no. 7, pp. 74-78, July 2014

[5] Gunn SR. Support vector machines for classification and regression [PhD thesis]. 1998

[6] Y. Kim, H. Lee, and E. M. Provost, "Machine learning for robust feature generation in audio emotion recognition," in Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '13), Vancouver, Canada, 2013.

[7] E. Bozkurt, E, Erzin, C. E. Erdem, A. Tanju Erdem, "Formant Position Based Weighted Spectral Features for Emotion Recognition", Science Direct Speech Communication, 2011.

[8] A. Nogueiras, A. Moreno, A. Bonafonte, Jose B. Marino, "Speech Emotion Recognition Using Hidden Markov Model", Eurospeech, 2001.

[9] M. E. Ayadi, M. S. Kamel, F. Karray, "Survey Speech Emotion Recognition: Features, Classification Schemes,and Databases", Pattern Recognition 44,PP.572-587, 2011.

[10] T. L. Nwe, S. W. Foo, and L. C. de Silva, "Speech emotion recognition using hidden Markov models," Speech Communication, vol. 41, no. 4, pp. 603–623, 2003.

[11] 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.

[12] Kaur, Jasmeet , and Anil Kumar. "Speech Emotion Recognition Using CNN, k-NN, MLP and Random Forest." In Computer Networks and Inventive Communication Technologies, pp. 499- 509. Springer, Singapore, 2021.

[13] Cabanac M. What is emotion Behavioral Processes. 2002;60(2):69-83

[14] Balti, H.; Elmaghraby, A.S. Emotion analysis from speech using temporal contextual trajectories. In Proceedings of the IEEE Symposium on Computers and Communications (ISCC), Funchal,Portugal, 23–26 June 2014.

[15] Balti, H.; Elaraby, A.S. Speech emotion detection using time dependent self-organizing maps. In Proceedings of IEEE International Symposium on Signal Processing and Information Technology, Athens, Greece, 12–15 December 2013.

[16] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3):572–587, 2011.

[17] R. Elbarougy and M. Akagi. Cross-lingual speech emotion recognition system based on a thre layer model for human perception. 2013 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pages 1–10, 2013.