International Journal for Modern Trends in Science and Technology, 9(SI01): 38-43, 2023

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST09SI0107
Available online at: http://www.ijmtst.com/vol9si01.html

Deep learning for the purpose of speech and motion recognition

Dr.G.Chenchamma, P.N.V Siva Kumar, P.Silpa, E.Ravi Kumar

Department of Electronics and Communication Engineering, Vijaya Institute of Technology for Women, Enikepadu, Vijayawada.

To Cite this Article
Dr.G.Chenchamma, P.N.V Siva Kumar, P.Silpa and E.Ravi Kumar. Deep learning for the purpose of speech and motion recognition. International Journal for Modern Trends in Science and Technology 2022, 9(SI01), pp. 38-43. https://doi.org/10.46501/IJMTST09SI0107

Article Info
Received: 26 January 2023; Accepted: 22 February 2023; Published: 26 February 2023

ABSTRACT

Speech emotion recognition has progressed from a specialty to a critical component of Human-Computer Interaction(HCI). These systems strive to make natural human-machine contact easier by using direct voice interaction rather than standard devices as input to understand verbal information and make it simple for human listeners to react. Dialogue systems for spoken languages, such as cell center discussions, onboard car driving systems, and the use of emotion patterns from speech in medical applications are just a few examples. Numerous strategies have been used to extract emotions from signals in the literature of speech emotion recognition (SER), including many well-established speech analysis and classification techniques.The feature extraction and feature classification phases are the most important parts of the speech emotion recognition(SER) process. Researchers have derived several features such as prosodic features, vocal traction features and other hybrid features for speech processing. The second phase includes feature classification using deep learning techniques. These techniques are recently proposed as on alternative to traditional techniques in SER.

KEY WORDS: Feature extraction, prosodic features, vocal traction features, Feature classification, Deep Learning techniques.

1. INTRODUCTION

1.1 Brief information about the project:

The ability to notice, interrupt and respond tosocialinteractions, which is usually assessed through effective expressions is one of the corner stones of human communication understanding emotion expressions can increase the efficiency and complexity of human-machine interaction by improving the processing and responsiveness of automatic emotion detection system, such as robots or experted systems to natural human behavior. If a robot can recognize human emotion expressions, it can change how it interacts with its surroundings. It can increase its problem-solving abilities by incorporating these expressions into a decision-making process. Emotional expression recognition has been a frustratingly tough topic that has drawn a lot of attention in recent years. There is no unanimity intheliterature on how to define emotions, yet features, prosodic features, vocal traction factors and other hybrid features are all examples of hybrid features using linear and non-linear classifiers.

The second phase involves feature classification Bayesian networks (BN) or the maximum likelihood principle (MLP) and support vector machine (SVM) are two of the most often used linear classifiers for emotion reorganization. The voice signal is usually considered

non- stationary. As a result, non-linear classifiers are thought to be successful for SER. Deep Learning is a new research subject in Machine Learning that has gotten a lot of interest in recent years. Deep Learning techniques for SER have several advantages over traditional methods, including the ability to detect complex structure and features without the need for manual feature extraction and tuning the tendency to extract low-level features from the given data and the ability to detect complex structure and features without the need for manual feature extraction and tuning.

1.2. Motivation and Contribution of the project:

The inspiration of the venture is, discourse feeling acknowledgement enhances cutting edge man-made reasoning capacities by getting a handle on the feeling from voice. After over 20 years of exploration, the field has developed to where it very well may be the following enormous thing in discourse UIs, communicates in language handling and examination for well-being, recovery, advanced mechanics, security and a few further applications. The impact of feelings on the voice is perceived by all individuals. The language of the tones is the most seasoned and generally wide spread of everyone of our method correspondence. It seems like the opportunity has arrived for registering apparatus to get it too. This holds for the whole field of emotional figuring. As SER depicts the more extensive thought of cloning machines the ability to appreciate people at their core ready to pursue human inclination and orchestrate feeling and enthusiastic way of behaving.

1.3. Objective of the project:

In this undertaking, we portray our Discourse feeling acknowledgment utilizing profound learning methods: a survey, through the contribution of crude discourse information, empowered us to perceive different discourse feelings in the discourse. This framework is valuable in call places, locallyavailable vehicle driving frameworks, and clinicalapplications.

2. LITERATURE SURVEY

M. S. Hossain and G. Muhammad proposed "Feeling acknowledgment utilizing profound gaining come closer from general media enthusiastic enormous information" [19]. A feeling acknowledgment framework utilizing a profound gaining come nearer from enthusiastic Large Information. Enormous Information involves discourse and video. In that proposed framework, a discourse signal is first handled in the recurrence space to acquire a Mel-spectrogram, which can be treated as a picture. Then, at that point, Mel-spectrogram is taken care of to a convolutional brain organization (CNN). For video flags, a few agent outlines from a video fragment are separated and taken care of the CNN. The results of the twoCNNs are intertwined utilizing two sequential outrageous learning machines (ELMs). The result of the combination is given to a helpservice vector machine (SVM) for the last arrangement of the feelings. That proposed framework is assessed utilizing two general media passionate information bases, one of which is Enormous Information. Exploratory outcomes affirm the adequacy of the proposed framework including the CNNs and the ELMs.

Hrithik Patni, Akash Jagtap, Vaishali Bhoyar, Dr. Aditya Gupta, proposed the "Speech Emotion Recognition using MFCC, Chromagram and RMSE features"[16]. The model distinguishes between emotions such as neutral, happy, sadness and rage. The classification system's performance is based on characteristics gathered and models developed. Energy, pitch, chromagram, MFCC and Gammatone frequency spectrum coefficients are some of the features used in this method. A two-dimensional Convolutional Neural Network (CNN) is used to classify the emotions.

D. Le and E.M.Provost proposed "Feeling acknowledgment from unconstrained discourse utilizing stowed away Markov models with profound conviction organizations" [17] Programmed feeling acknowledgment from the unconstrained discourse is trying because of non-ideal recording conditions and exceptionally questionable ground truth names. Further, feeling acknowledgment frameworks regularly work with uproarious high-layered information, delivering it hard to track down agent elements and train a viable classifier. He handles this issue by utilizing Profound Conviction Organizations, which can display complex and non-direct undeniable level connections between low-level elements. He proposes and assesses a set-up of crossover classifiers in view of Stowed away Markov Models and Profound Conviction Organizations. He accomplishes cutting edge outcomes on FAU Aibo, a benchmark dataset in feeling acknowledgment. His work gives experiences into significant likenesses and contrasts among discourse and feeling.

S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh proposed "Discourse feeling acknowledgment" Discourse is the most normal and helpful way by which people impart, and understanding discourse is one of the most many-sided processes that the human mind performs. Discourse Feeling Acknowledgment (SER) expects to perceive human feeling from discourse. This is on the way that voice frequently reflects hidden feelings through tone and pitch. The libraries utilized are Librosa for dissecting sound and music, sound documents for perusing and composing inspected sound record designs, and sklearn for building the model. The adequacy of the Convolutional Brain Organization (CNN) in the acknowledgment of discourse feelings has been examined. Spectrograms of the discourse signals are utilized as the info elements of the organizations. Mel-Recurrence Cepstral Coefficients (MFCC) are utilized to extricate highlights from sound. Her discourse dataset is utilized to prepare and assess our models. In view of the assessment, the feelings (blissful, miserable, irate, impartial, astounded, disdain) of the discourse will be recognized.

3. DATASET

RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [1] is a collection of 24 professional voices (i.e., 12 males, 12 female). All professionals speak the same two sentences in a North-American accent. The many spoken emotion expressions utilized include angry, pleased, calm, afraid, sad, disgust, and surprise. Every expression has two emotional intensity levels (light and bold), as well as a neutral expression. Every one of the 1440 files have a different manner of naming it. The filename is made up of seven components. The RAVDESS dataset was chosen because each file is graded ten times for emotional sincerity and intensity. Interrater reliability, emotional validity, and test-retest reliability are also excellent.

SAVEE stands for Surrey Audio-Visual Expressed Emotion. A requirement for the creation of an automatic emotion identification system was recorded. The database contains 480 British English utterances recorded by four male actors in seven different emotions. For each emotion, sentences were taken from the normal TIMIT corpus and phonetically balanced. The data was recorded, analyzed, and labelled at a visual media lab with high-quality audio-visual equipment. The recordings were assessed by ten individuals under audio, visual, and audio-visual circumstances to ensure that they were of high quality. For each of the auditory, visual, and audio-visual modalities, classification systems were created using standard features and classifiers, and speaker independent identification rates of 61 percent, 65 percent, and 84 percent were attained, respectively.

4. PROPOSED METHOD

As we probably are aware, discourse feeling acknowledgment is basically involving two stages known as element extraction and component order. The sound dataset should be first changed over to a machine understanding language and this can be accomplished by utilizing librosa instrument then to make feature extraction feasible, for this proposed work we use python 3.7 language to compose the code since it is powerful and less mind boggling. Before feature extraction, the sound signal is changed to an analog signal and goes through sampling at a sampling rate of 20KHz per second. The absolute dataset tests are 2556 samples. Fig1 gives a block representation of this proposed.

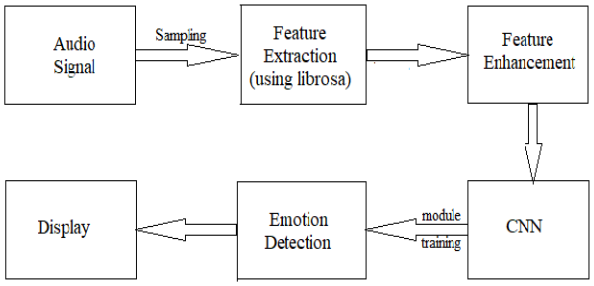

Fig1: Block Diagram representation.

4.1 Feature Extraction:

SER is a process involving the conversion of speech input into digital signals and then processing it to extract related features suitable for training the model. Here are the important extracted features in the proposed system:

Zero-Crossing Rate:

An audio frame's Zero-Crossing Rate (ZCR) is the rate at which the signal's sign changes during the frame. In other words, it's the number of times the signal's value changes from positive to negative and back, divided by the frame's length. The ZCR is calculated using the following formula:

$$\boldsymbol{Z(\vec{\imath}) = \frac{1}{2w_L}\sum_{n=1}^{w_L}|sgn[x_i(n)] - sgn[x_i(n-1)]|\,,}$$

Where sgn(.) is the sign function, i.e.

$$sgn[x_i(n)] = \begin{cases} 1; x_i(n) \geq 0 \\ -1; x_i(n) < 0 \end{cases}$$

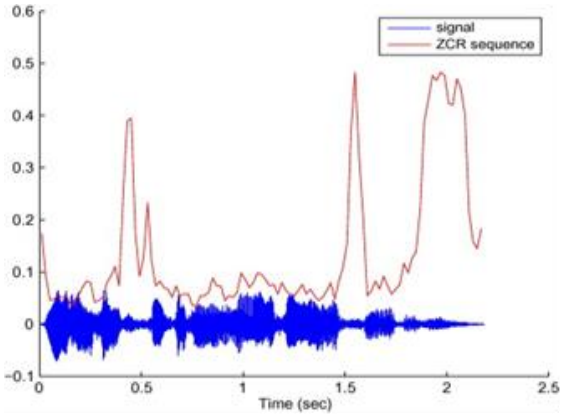

Fig2: Example of a speech segment and the respective sequence of ZCR values.

Mel frequency cepstrum coefficients (MFCC):

MFCC is broadly utilized in voice acknowledgment and SER based applications [2]. MFC is delineating the transient power range of sound.

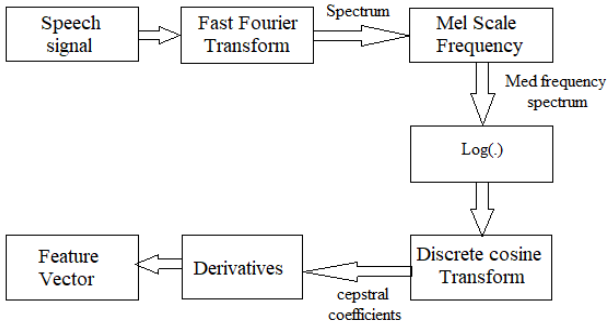

Fig3: MFCC feature extraction procedure.

A high feeling acknowledgment rate was accomplished utilizing MFCC. Because of MFCC channel bank attributes, it accomplishes better recurrence goal and heartiness to commotion in the low-recurrence area in contrast with the high-recurrence locale. MFCC feature extraction procedure is shown in Figure 2 [3].

The signal is separated into frames during the MFCC Extraction procedure. The energy spectrum and Fourier transform are calculated for each frame and mapped on the Mel-frequency scale. The Mello energies' discrete cosine transform (DCT) is evaluated. After the DCT procedure, a vector comprising 40 MFCC coefficients is extracted. Only the first 16 coefficients of the 40 retrieved coefficients are used as features in emotion recognition because the subsequent coefficients are redundant and lower the system's accuracy performance. The FFT of the signal x[n] is represented by the equation below.

$$\boldsymbol{X(k) = \sum_{n=0}^{N-1} \left(x[n] W_N^{nk}\right)}$$

Pitch and Chromagram:

Pitch is a significant attribute of the discourse signal. It is otherwise called the glottal waveform. Feeling relies upon the subglottal pneumatic force and strain in vocal folds and this data can be removed from the pitch signal. Chroma highlights [4] address the symphonious substance of a brief time frame window of the info discourse signal. The component vector is separated from the extent range by utilizing a brief time frame Fourier transform (STFT), Consistent Q changes (CQT), Chroma Energy Standardized (CENS). The mean worth of fluctuation, variety ranges, pitch and forms are different for eight feelings. 13 pitch and chromogram coefficients are extricated from the discourse signal. STFT of the sign is given in condition 4 where x[n] is the information discourse signal in discrete structure and w[n] is the window work.

$$\boldsymbol{F[\tau, \omega] = x \int_{-\infty}^{\infty} x[n]^* w(n-\omega)^* e^{-i\omega n}}$$

After feature extraction, in view of the upsides of mfcc, pitch, size, and chromogram; every inclination is left with various number of tests which are Outrage: 355 examples, disdain: 192 examples, dread: 353 examples, cheerful: 351 examples, nonpartisan: 255 examples, miserable: 339 examples, shock: 199 examples.

Classifier Training:

The whole dataset samples were divided into training and testing samples with a ratio of 0.8. This dataset is in csv format, and it walks over the existing dataset, attaching pathways to emotions and labelling them. The following is the labelling:

0: Anger (355 samples)
1: disgust (192 samples)
2: fear (353 samples)
3: happy (351 samples)
4: neutral (255 samples)
5: sad (339 samples)
6: surprise (199 samples)

For detecting the emotion based on extracted values, we develop a convolution neural network.

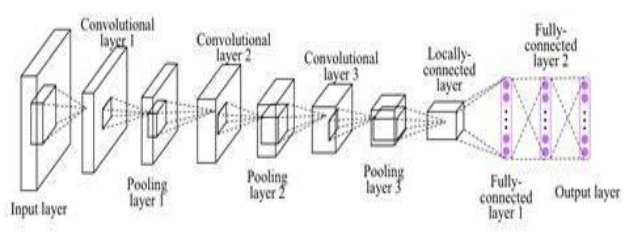

Fig 4: General Representation of CNN model.

The first convolutional layer receives a picture or audio as input. An activation map is created from the convoluted output. The convolution layer's filters collect relevant features from the input image that are then passed on. Each filter should have a unique feature to aid in accurate class prediction. We utilize same padding (zero padding) if we need to keep the image's size; otherwise, valid padding is used because it helps to reduce the number of features. The number of parameters is then further reduced by adding pooling layers. Before making a prediction, several convolution and pooling layers are included. The use of a convolutional layer aids in the extraction of features. In comparison to going deeper into the network, more particular features are recovered as we go further into the network. A shallow network in which the retrieved features are more generic. As previously stated, the output layer in a CNN is a fully linked layer that flattens and sends the input from the other layers in order to change the output into the number of classes requested by the network. After that, the output layer generates the output, which is then compared to the output layer for error creation. To compute the mean square loss, a loss function is established in the fully linked output layer. The error gradient is then determined. The error is then backpropagated to update the bias and filter(weights). A single forward and backward pass complete one training cycle.

5. EXPERIMENTAL ANALYSIS

The below figure shows the training and testing loss on our dataset. As we can see from the graph that both training and testing errors reduces as number of epochs to the training model increases.

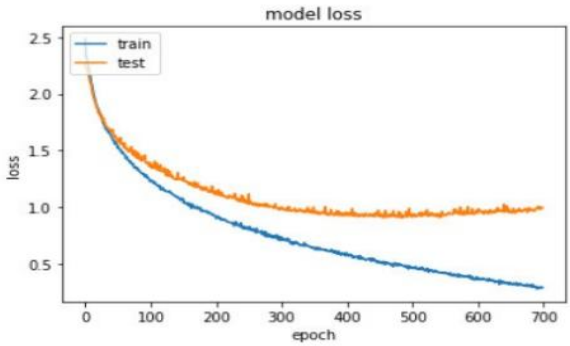

Fig5: Result training Vs Testing accuracy.

We got the accuracy of the training model about 92% but the accuracy of cross validation about 47%. Dataset could be modified to get better accuracy.

6. CONCLUSION

Henceforth our venture presents a better approach to provide the capacity to machine to decide the feeling with the assistance of the human voice. It will enable the machine to have a superior methodology towards having a superior discussion and consistent discussion like human does.

Our undertaking means to decide the feeling with the discourse of a human. Our task can be reached out to incorporate with the robot to assist it with having a superior comprehension of the state of mind the comparing human is in, which will assist it with having a superior discussion.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Kerikeri, Leila, Youssef, Kosaiand,Catherine, Mahjoub, Mohamed, "Automatic Speech Emotion Recognition Using Machine Learning." March 2019.

[2] Issa, Dias, M. Fatih Demirci, and Adnan Yazici, "Speech emotion recognition with deep convolutional neural networks." Biomedical Signal Processing and Control 59 (2020): 101894.

[3] M.S. Likitha, Sri Raksha R. Gupta, K. Hasitha and A. Upendra Raju, "Speech Based Human Emotion Recognition Using MFCC." IEEE WiSP- NET 2017 conference, 2017.

[4] M. Kattel, A. Nepal, A. K. Shah, D.Shrestha," Chroma Feature Extraction" Department ofComputer Science and Engineering school of Engineering Kathmandu University, Nepal.

[5] A.K. Verma, A.R. Verma, Manoj Kumar, "2-D Speech Enhancement based on Curvelet Transform using Different Window Functions." International Journal of Computer Applications (Volume 81 – No.13).

[6] Hrithik Patni, Akash Jagtap, Vaishali Bhoyar, Dr Aditya Gupta, "Speech Emotion Recognition using MFCC, GFCC, Chromagram and RMSE features." IEEE February 2021.
[7] Harshit Dolka, Arul Xavier V M, Sujitha Juliet, "Speech Emotion Recognition Using ANN on MFCC Features." IEEE May 2021.
[8] W.Q. Zheng, J.S. Yu, Y.X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks." in International Conference on Affective Computing and Intelligent Interaction IEEE, 2015, pp. 827–831
[9] Liberman, Mark, et al. Emotional Prosody Speech and Transcripts LDC2002S28. Web Download. Philadelphia: Linguistic Data Consortium, 2002.
[10] M.M.H.E. Ayadi, M.S. Kamel, F. Karray, "Survey on speech emotion recognition: Features classification schemes and databases", Pattern Recognition, pp. 572-587, 2011.
[11] Badshah, Abdul Malik, et al, "Speech emotion recognition from spectrograms with a deep convolutional neural network." 2017 international conference on platform technology and service (Plat Con). IEEE, 2017.
[12] M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases." Pattern Recognition 44, PP.572-587, 2011.
[13] S. An, Z. Ling and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs." 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, 2017, pp. 1613-1616.
[14] M. Kattel, A. Nepal, A. K. Shah, D. Shrestha, "Chroma Feature Extraction" Department of Computer Science and Engineering, School of Engineering Kathmandu University, Nepal.
[15] S. Yildirim, M. Bulut, and C. Lee, "An acoustic study of emotions expressed in speech." Proceedings of Inter Speech, pages 2193–2196, 2004.
[16] Hrithik Patni, Akash Jagtap, Yaishalu Bhoyar, Dr. Aditya Gupta, "speech emotion recognition using Mfcc, chromagram and RMSE features." IEEE 2021.
[17] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks," 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, 2013, pp. 216-221.
[18] Min Chen, Ping Zhou, Glancario Fortino, "Emotion Communication System." In IEEE access, vol. 5, pp. 326-337, 2017.
[19] M.S. Hossain, G. Muhammad, "Emotion Recognition using Deep Learning approach from audio- visual emotional big data." 2019 Information fusion, pages. 69-78.