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

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 04, April 2023, pp : 10-15

e-ISSN : 2583-1062

Impact Factor : 2.625

SPEECH EMOTION RECOGNITION USING CONVOLUTIONAL NEURAL NETWORK

Karthick M[1], Prasanna Venkatesan K J[2], Ajeethkumar S[3], Naveen Kumar V[4]

[1]Assistant Professor, Department of Computer Science and Engineering, Nandha College of Technology, Perundurai 638 052, Tamilnadu, India.

[2,3,4]UG Scholars, Department of Computer Science and Engineering, Nandha College of Technology, Perundurai 638 052, Tamilnadu, India.

ABSTRACT

Convolutional neural network (CNN)-based speech emotion recognition is a new area of study that aims to automatically identify emotions in speech signals. This is a moving undertaking because of the perplexing idea of discourse signals, which contain different acoustic highlights that convey feelings like pitch, volume, and tone. In recent years, deep learning approaches, particularly CNNs, have outperformed conventional machine learning methods in speech and emotion recognition. CNNs have been extensively utilized in image and speech processing tasks due to their capability of automatically learning complex features from raw speech signals. CNN-based speech emotion recognition has numerous potential applications, including in the entertainment industry and healthcare, where it can be used to improve users' emotional experiences and to monitor patients' emotional states. In general, CNN-based speech emotion recognition is a rapidly developing field that has the potential to have a significant impact on a variety of fields. Further research in this field will result in the creation of more accurate and robust models for identifying emotions in speech signals.

Keywords: Speech emotion recognition, deep learning, deep neural network, recurrent neural network, convolutional neural network

1. INTRODUCTION

1.1 Speech Emotion Recognition

The process of determining a speaker's emotional state from their speech signal is referred to as speech emotion recognition (SER). It is a branch of speech processing and natural language processing with applications in fields like affective computing, speech therapy, and human-computer interaction. In most cases, SER systems make use of machine learning algorithms to analyze the acoustic characteristics of speech signals and divide them into various emotional categories like happy, sad, angry, or neutral. Pitch, loudness, tempo, spectral energy, and speech rate are a few of the most frequently utilized acoustic features. Using supervised learning algorithms, a dataset of labeled speech samples is used to train a model to recognize various emotions, which is one approach to SER. The model is then tried on a different dataset to assess its exhibition. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two examples of deep learning models that have produced promising outcomes in SER.

1.2 Deep Learning

Building and training neural networks with multiple layers for complex tasks like image recognition, natural language processing, and speech recognition is deep learning, a subfield of machine learning. With each layer of the network extracting more complex and meaningful features from the input data, deep learning models learn to represent data in increasingly abstract and hierarchical ways.

Convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models are all examples of popular deep learning architecture. On a wide range of tasks, such as speech recognition, machine translation, image classification, and object detection, these models have performed at the highest possible level. The availability of large datasets, powerful computing hardware, and advancements in algorithms like stochastic gradient descent and back propagation, which make it possible to train deep neural networks effectively, are all to blame for the success of deep learning.

1.3 Emotion Communication

The act of conveying one's emotional state to others through verbal and nonverbal means is known as emotion communication. [3] It entails expressing feelings like joy, sadness, rage, fear, and surprise in a way that other people can understand. Individuals can benefit from improved relationships, conflict resolution, and smoother social situations through effective emotion communication. Emotional communication verbally involves expressing one's feelings through language. Statements like "I feel happy" or "I'm feeling sad today" can accomplish this. Emotional expressions such as "I'm excited about my upcoming vacation" or "I'm frustrated because I missed my train" can also

be part of the process. Using facial expressions, voice tone, body language, and other nonverbal cues to convey one's emotional state is known as nonverbal emotion communication. The success of deep learning can be attributed to the availability of large datasets, powerful computing hardware, and advancements in algorithms such as back propagation and stochastic gradient descent, which enable efficient training of deep neural networks.

1.4 CNN

A type of artificial neural network (ANN) known as a Convolutional Neural Network (CNN) is frequently utilized in computer vision and deep learning tasks. CNNs, in contrast to fully connected traditional neural networks that treat input data as a flat vector, are made to process and analyze images and other multidimensional data by making use of the structure and spatial dependencies of the data. Convolutional, pooling, and fully connected layers are among the many layers that make up a CNN. When filters are applied to the input data by convolutional layers, various features like edges, textures, and shapes are identified. [4] Pooling layers down samples the Convolutional layers' output, helping to prevent overfitting and reducing the feature maps' dimensionality. Final classification or regression on the output of the preceding layers is carried out by fully connected layers.

2. LITERATURE REVIEW

2.1 Speech Emotion Recognition Two Decades in a Nutshell, Benchmarks, and Ongoing Trends

The role of human emotion's acoustics was previously the subject of a series of psychological rather than computer science-based studies (see, for instance, references). Blanton, for instance, composed that "the impact of feelings upon the voice is perceived by all individuals.[5] The tones of love, fear, and rage can be recognized by even the most primitive individuals; the animals also share this knowledge. The meaning of the human voice can be understood by the dog, horse, and numerous other animals. The oldest and most widely used form of human communication is the tonal language.

This holds true for the entire field of affective computing. Picard's field-coining book by the same name appeared around the same time as SER29, describing the broader concept of lending machines emotional intelligence that can recognize human emotion and to synthesize emotion and emotional behavior. It appears that the time has come for computing machinery to understand it as well.

2.2 Emotion Recognition using Deep Learning Approach from Audio-Visual Emotional Big Data

The role of human emotion's acoustics was previously the subject of a series of psychological rather than computer science-based studies (see, for instance, references). Blanton, for instance, composed that "the impact of feelings upon the voice is perceived by all individuals. The tones of love, fear, and rage can be recognized by even the most primitive individuals; the animals also share this knowledge. [1] The meaning of the human voice can be understood by the dog, horse, and numerous other animals. The oldest and most widely used form of human communication is the tonal language. This holds true for the entire field of affective computing. Picard's field-coining book by the same name appeared around the same time as SER29, describing the broader concept of lending machines emotional intelligence that can recognize human emotion and to synthesize emotion and emotional behavior. It appears that the time has come for computing machinery to understand it as well.

2.3 Emotion Communication System

People are shifting their attention away from the material world and toward the spiritual world in this increasingly materialistic world. Systems for human-machine interaction have been developed to identify and treat people's emotions. Most communications in terms of human-to-human and human-to-machine are nonlife-of-sight (NLOS), but the currently available human-machine interaction systems frequently support interaction between humans and robots in an environment with line-of-sight (LOS) propagation. We propose a NLOS-based emotion communication system to overcome the limitations of the conventional human-machine interaction system. We initially characterize the feeling as a sort of media which is like voice and video. Emotional information can be both recognized and conveyed over a significant distance. Then, considering the need for real-time communications between the parties involved, we suggest an emotion communication protocol that offers dependable support for the implementation of emotion communications. We plan a pad robot discourse feeling correspondence framework, wherein the cushion robot goes about as a model for client feeling planning.

2.4 Can Deep Learning Revolutionize Mobile Sensing?

Sensor-prepared PDAs and wearables are changing an assortment of portable applications going from wellbeing checking to computerized colleagues. However, a major obstacle to the development of sensor apps is the inability to reliably infer user behavior and context from complex and noisy sensor data collected under mobile device constraints. In related inference tasks like speech and object recognition, advances in the field of deep learning have led to gains

that are nearly unprecedented. Deep learning [1] has not yet been thoroughly investigated in the sensing field, even though mobile sensing faces many of the same issues with data modeling. By rapidly expanding the number of sensor apps that are ready for widespread use, deep learning could result in mobile sensor inference that is significantly more robust and effective.

By prototyping a low-power Deep Neural Network (DNN) inference engine that takes advantage of both the CPU and DSP of a mobile device SoC, we study typical mobile sensing tasks like activity recognition using DNNs and compare results to learning techniques that are more commonly used. This paper provides preliminary answers to this potentially game-changing question. Our initial findings demonstrate how DNNs can improve inference accuracy while also illustrating how to use them in ways that do not overburden current mobile hardware.

2.5 Speech Emotion Recognition in Emotional Feedback for Human-Robot Interaction

Recognizing human emotions is essential if robots are to plan their actions autonomously and interact with people. Emotions can be effectively conveyed through nonverbal cues like pitch, loudness, spectrum, and speech rate for most people. A machine might be able to recognize emotions by using the features of a spoken voice's sound, which probably contain important information about the speaker's emotional state. Six distinct types of classifiers were tested in this study to predict six fundamental universal emotions based on nonverbal characteristics of human speech. The arrangement methods utilized data from six sound documents separated from the eNTERFACE05 varying media feeling information base.

3. EXISTING SYSTEM

In Human-Computer Interaction (HCI), emotion recognition from speech signals is a crucial but challenging component. Many well-known speech analysis and classification techniques have been used to extract emotions from signals in the literature on speech emotion recognition (SER). In SER, deep learning methods have recently been proposed as an alternative to traditional methods. The speech-based emotion recognition applications of Deep Learning are the subject of some recent research, and this paper provides an overview of these techniques. The database used, the emotions extracted, the contributions made to speech emotion recognition, and the limitations associated with it are all covered in the review. In recent years, more attention has been paid to the emerging machine learning research field of deep learning. The ability of Deep Learning techniques [3] for SER to detect complex structure and features without the need for manual feature extraction and tuning is one of their many advantages over traditional methods. Ability to deal with unlabeled data and inclination to extract low-level features from the raw data. Feed-forward structures with one or more underlying hidden layers between inputs and outputs serve as the foundation For Deep Neural Networks (DNNs). The feed-forward structures like Profound Brain Organizations (DNNs) and give productive outcomes to picture and video handling.

4. PROPOSED SYSTEM

Due to the gap between acoustic characteristics and human emotions, Speech Emotion Recognition is a challenging process that heavily relies on the discriminative acoustic characteristics extracted for a given recognition task. Emotions and ways of expressing them vary greatly from person to person. Emotional speech does have different energies, and when considering various subjects, pitch variations are emphasized. As a result, speech emotion detection is a challenging computing vision task. Here, the Convolutional Neural Network (CNN) algorithm is used for speech emotion recognition. Different modules are used for emotion recognition, and classifiers are used to distinguish between happy, surprised, angry, neutral, sad, and other emotions. A novel speech-emotion recognition system based on a Convolutional Neural Network (CNN) was proposed by us. Speech is regarded as the most universal and natural form of communication.

A person's mental, behavioral, and emotional traits can be communicated in a variety of ways through speech. Additionally, work related to speech-emotion recognition can assist in preventing cybercrimes. Emotion communication refers to the process of conveying one's emotional state to others through verbal and nonverbal means.

4.1 Audio Feature Extraction and Visualizations

Classification and depiction necessitate the extraction of characteristics. The audio signal is a three-dimensional signal with time, amplitude, and frequency indicated by three axes. Any audio signal can be analyzed and its characteristics extracted using librosa. load) capability pulls a sound document and decodes it into a 1D cluster which is of time series x, and SR is really testing pace of x. As a matter of course SR is 22 kHz. I will demonstrate one audio file display using the (IPython.display) function in this example. Librosa.display is essential for displaying audio files in wave plot, spectrogram, and colormap formats. Wave plots make use of the volume of the audio at a specific time. A spectrum graph shows the amplitude of various frequencies for a specific period.

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

www.ijprems.com
editor@ijprems.com

Vol. 03, Issue 04, April 2023, pp : 10-15

e-ISSN : 2583-1062

Impact Factor : 2.625

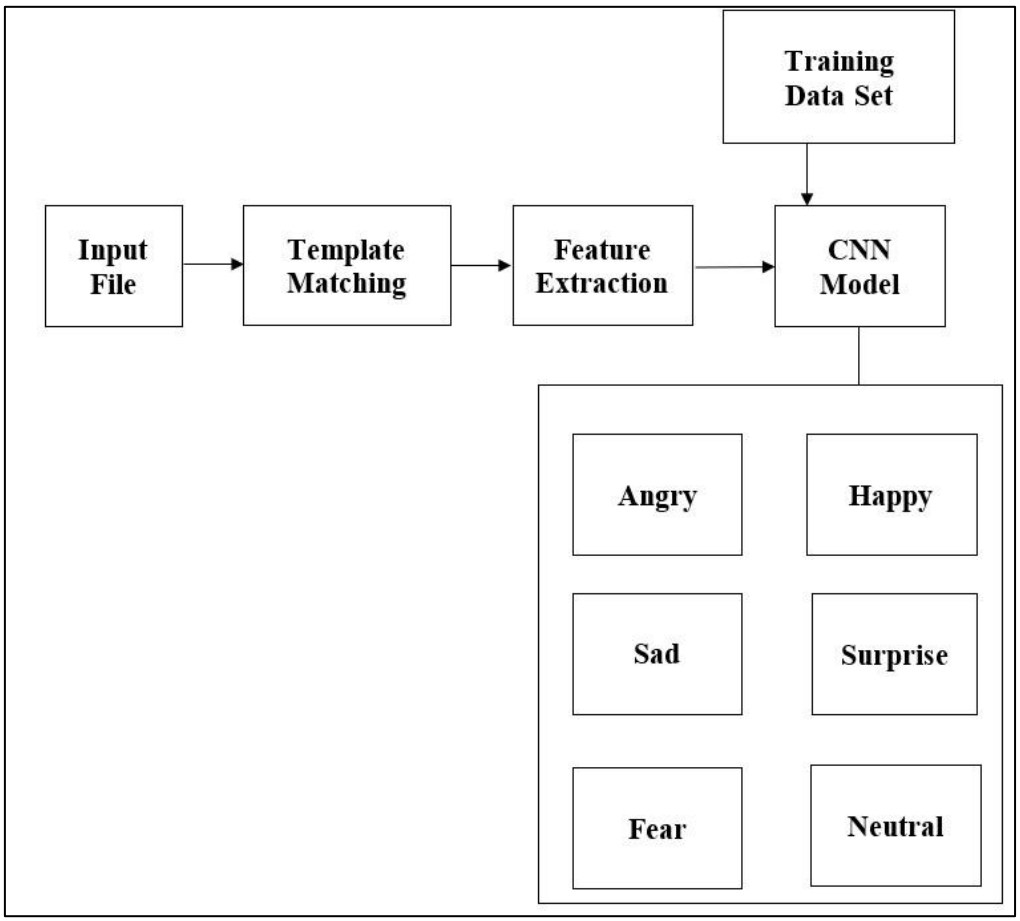

Figure. 1 Flow diagram for audio feature extraction and visualization

4.2 To Train the Model for Accuracy Calculation

Make a dataset of speech signals with emotional states labeled on them. In order to evaluate the CNN model's performance, preprocess the raw signals to extract relevant features and divide the dataset into training and testing sets. Using back propagation and gradient descent algorithms, train the CNN model on the training set. The system receives the expression label as training data, and weight training is also provided for that network. As an input, an audio is used. The audio is then subjected to intensity normalization. The Convolutional Network is trained on normalized audio to ensure that the presentation sequence of the examples has no effect on training performance. This training process results in the collection of weights, which achieve the best results with this learning data. During testing, the dataset retrieves the system's pitch and energy, and the determined emotion is provided by the system based on the final network weights trained.

4.3 Implementation Process of CNN Model

For the CNN model, specify the activation functions, the number of layers, and the type of layers (such as pooling, fully connected, or Convolutional). Create a model that is suitable for the job of recognizing emotions in speech. On the testing set, test the CNN model's performance. Measure the model's performance using evaluation metrics like accuracy, precision, recall, and F1 score. An image with three layers depicts speech. When using CNN, keep in mind the first and second derivatives of the speech image in terms of frequency and time. CNN can make predictions, look at speech data, learn from speeches, and figure out words or phrases.

4.4 Classification of Speech Emotions

While testing we give the sound info. After that, we use the IPython.display packages to run the audio so that we can hear. After that, use the librosa. display. waveplot packages to plot the audio features. Using librosa, extract the Characteristics. Load. One data frame is transformed into a structured form for display. It also compares the loaded model using a predict function with a batch size of 32. At last, it shows the result from the sound document what kind of articulation/feeling that sound record.

5. RESULTS AND DISCUSSIONS

Table. 1 summarizes the performance of the CNN model on the test dataset. From the table it can be concluded that the model performs almost uniformly on all the classes of emotions. The normalized confusion matrix for the CNN

model applied on the test dataset.

Table. 1 CNN Performance table

Class	Precision	Recall	F1 Score
Disgust	0.82	0.81	0.82
Fear	0.85	0.8	0.82
Happiness	0.83	0.87	0.85
Neutral	0.83	0.86	0.85
Sadness	0.83	0.9	0.86
Surprise	0.84	0.87	0.86

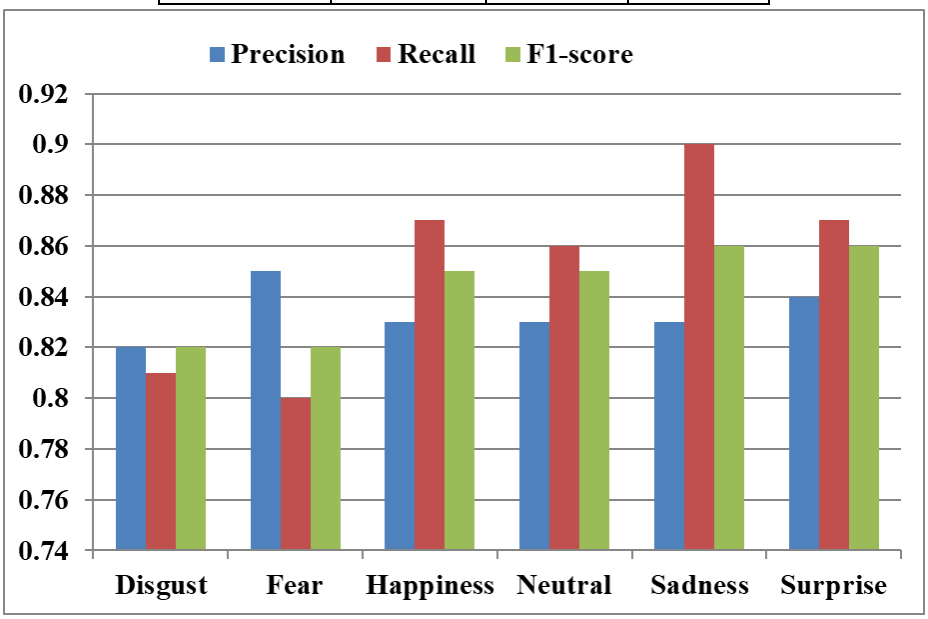

Figure 2. Performance variations

6. CONCLUSIONS

For the emotion distinction task, we developed the superior CNN model after constructing various models. Compared to the model that was previously available, our accuracy was 71%. With more data, our model would have performed better. Additionally, our model performed exceptionally well when separating a feminine voice from a masculine one. Our project can be extended to work with the robot to help it understand the mood of the person it is talking to better. This will help the robot have a better conversation with the person. It can also work with music apps to suggest songs to users based on how they feel. It can also be used in online shopping apps like Amazon to help users find better products. In addition, in the coming years, we will be able to develop a sequence-to-sequence model that will enable us to produce voices with varying degrees of emotion. As an advertisement voice, an excited voice, etc.

7. REFERENCES

[1] Speech and emotion recognition: In a nutshell: benchmarks, ongoing trends, and two decades," Communications of the ACM, vol. 61, no. 5, pp. 90–99, 2018.

[2] "Emotion recognition using deep learning approach from audio–visual emotional big data," by M. S. Hossain and G. Muhammad, Information Fusion, vol. 49, pp. 69–78, 2019.

[3] M. Chen, P. Zhou, and G. Fortino, "Feeling correspondence framework," IEEE Access, vol. 5, pp. 326–337, 2017.

[4] "Can deep learning revolutionize mobile sensing?" by N. D. Lane and P. Georgiev. in the 16th International Workshop on Mobile Computing Systems and Applications's Proceedings. ACM, 2015, pp. 117–122.

[5] "Speech emotion recognition in emotional feedback for human-robot interaction," International Journal of Advanced Research in Artificial Intelligence (IJARAI), vol. 5, J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson 4, no. 2, pp. 20–27, 2015.

[6] CNN and DBN with four LFLBs and one LSTM Deep 1D and 2D CNN LSTM to achieve 91.6% and 92.9%

accuracy J. Zhao et. al (2019).

[7] LSTM based RNN with 3 layers Model is tested on MoCap data with overall accuracy of 71.04%S. Tripathi et. al (2018).

[8] Deep Convolutional Neural Network (DCNN) Merged Deep 1D and 2D CNN for high level learning of features from input audio and log-mel spectrograms with 92.71% accuracy J. Zhao et. al (2018).

[9] Combined SVM and DBN for SER The combined model achieves 94.6% accuracy Z. Lianzhang et. al (2017).

[10] CAS emotional speech data base W. Zhang et. al (2017).

[11] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, "Emotion analysis in man-machine interaction systems," in International Workshop on Machine Learning for Multimodal Interaction. Springer, 2004, pp. 318–328.

[12] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal processing magazine, vol. 18, no. 1, pp. 32–80, 2001.

[13] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in Eighth European Conference on Speech Communication and Technology, 2003.

[14] R. W. Picard, Affective computing. Perceptual Computing Section, Media Laboratory, Massachusetts Institute of Technology, 1995.

[15] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," International journal of speech technology, vol. 15, no. 2, pp. 99–117, 2012.

[16] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition, vol. 44, no. 3, pp. 572–587, 2011.

[17] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in Eighth European Conference on Speech Communication and Technology, 2003.

[18] Dileep A D and C. C. Sekhar, "Gmm-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," IEEE transactions on neural networks and learning systems, vol. 25, no. 8, pp. 1421–1432, 2014.

[19] L. Deng, D. Yu et al., "Deep learning: methods and applications," Foundations and Trends® in Signal Processing, vol. 7, no. 3-4, pp. 197– 387, 2014.

[20] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural networks, vol. 61, pp. 85–117, 2015.