International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

Audio Emotion Analysis

Ashok K[1], Dhanasekar M[2], Dilipwaran R[3], Kailesh M K[4], Maithili K[5]

[1,2,3,4]UG scholar, Department of Computer Science Engineering, Kingston Engineering College, Tamilnadu, India.
[5]Assistant Professor, Department of Computer Science and Engineering, Kingston Engineering College, Tamilnadu, India.

Abstract - *While tightening and expansion of our facial muscles cause some changes called facial expressions as a reaction to the different kinds of emotional situations of our brain, similarly there are some physiological changes like tone, loudness, rhythm and intonation in our voice, too. These visual and auditory changes have a great importance for human-human interaction human-machine interaction and human-computer interaction as they include critical information about humans' emotional situations. Automatic emotion recognition systems are defined as systems that can analyze individual's emotional situation by using this distinctive information. In this study, an automatic emotion recognition system in which auditory information is analyzed and classified in order to recognize human emotions is proposed. In the study spectral features and MFCC coefficients which are commonly used for feature extraction from voice signals are firstly used, and then deep learning-based LSTM algorithm is used for classification. Suggested algorithm is evaluated by using three different audio data sets (SAVEE, RAVADES and RML).*

Key Words: Speech, Dataset, training data, CNN, LSTM.

1. INTRODUCTION

Human machine interaction is widely used nowadays in many applications. One of the medium of interaction is speech. The main challenges in human machine interaction is detection of emotion from speech. When two persons interact to each other they can easily recognize the underlying emotion in the speech spoken by the other person. The objective of emotion recognition system is to mimic the human perception mechanisms. There are several applications in speech emotion recognition. Emotion can play an important role in decision making. Emotion can be detected from different physiological signal also. If emotion can be recognized properly from speech then a system can act accordingly. CNN has been widely used in computer vision. Intuitively, it is also applicable to speech recognition since the audio signal can be converted via short-time fourier transform (STFT) into a spectrogram which can be viewed as a 2-dimension image indexed by the time-axis and frequency-axis. Despite some positive results, it has long been argued that CNNs overkill the variation along time-scale by pooling within a temporal window, resulting in deep fully-connected neural network's dominance in modeling time variation. Introduced a limited-weight-sharing convolutional scheme and found that using convolution along the frequency axis or time axis increased recognition accuracy, but the improvement was less significant along the time axis. To alleviate the problem, a bottleneck network was constructed in place of the pooling layer. Furthermore, T6th in proposed treating time-domain and frequency-domain separately and achieved the best performance on the TIMIT dataset by constructing such a hierarchical convolutional network. Inspired by the temporal characteristics of speech, RNN, which tries to predict the current frame based on feature information collected from previous frames, has long been used in speech recognition tasks. Due to its capability of modeling sequential data, RNN can be combined with HMM, or even replace HMM. In the latter case, the model can be trained "end-to-end", and for this purpose, the connectionist temporal classification (CTC) and RNN Transducer were proposed to deal with the specific evaluation metric for sequence labeling. Two special RNN models, Long Short-Term Memory (LSTM) and Gated Recurrent Unit are now widely used in speech recognition. These methods have showed good results in many tasks. One of their limitations refers to the difficulty in training, and in practice, their performance relies heavily on pre-training. Several recent works attempted to combine CNN and RNN for speech recognition. The CNN-RNN hybrid model for Large Vocabulary Continuous Speech Recognition. Here proposed an architecture, which unit files CNN, LSTM. In the two models, however, the CNN module and RNN module are separated. A similar combination method was proposed for text classification. Recently, Liang et al. proposed a deep learning model in which RNN and CNN were tightly coupled. The hallmark of the model is that there exist intra layer recurrent connections among units in the convolutional layer of CNN. This model was used in experiments on static images, but has not been tested on speech data.

1.1 Existing System

We use four types of speech emotions, neutral style, anger, happiness and sadness. The data size of each emotion is respectively two hours, one hour, half an hour and two hours. The speech are all sentences uttered by a recorder. Sentences were extracted from each emotion as testing sentences. The estimated clean speech can be obtained from the observed signal through a filtering process. Diverse classification of audio expressions might be used in numerous applications like; Human Behavior Predictor, Surveillance System and Medical Rehabilitation. Seven elementary categories of human emotions are unanimously predictable across different cultures and by numerous people are: anger, disgust, fear, happiness, sadness, surprise and neutral. Numerous scholars have used dissimilar methods for classifying audio expression.

DISADVANTAGE:

- Accuracy is low when compared to new algorithms.
- It also requires some computational devices.
- Implementation cost is high

1.2 Proposed System:

We propose to use a recently developed deep learning model, recurrent convolutional neural network (RNN), for speech processing, which inherits some merits of recurrent neural network (RNN) and convolutional neural network (CNN). The core module can be viewed as a convolutional layer embedded with an RNN, which enables the model to capture both temporal and frequency dependence in the spectrogram of the speech in an efficient way. If speech samples are collected under real life condition then speech signal corrupted with several noise. To recover from this problem, a noise reduction phase is performed. We compared the classification accuracy using segment-level features from different layers speech recognition applications, by adding several fully connected layers on the top will boost the performance of CNN. The emotion having the maximum parentages is projected as its resulting emotion on a specified audio. Likewise, founded on experimental outcomes, training and examination of various emotional phases has also inspired us to develop a real-time audio expression recognition system.

ADVANTAGE:

1) Accuracy is high.

2) High Computational processing.

3) Independent of ethnicity.

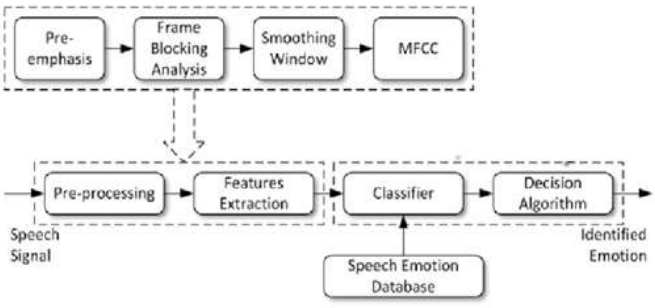

Fig 1: Architecture design of audio emotion analysis

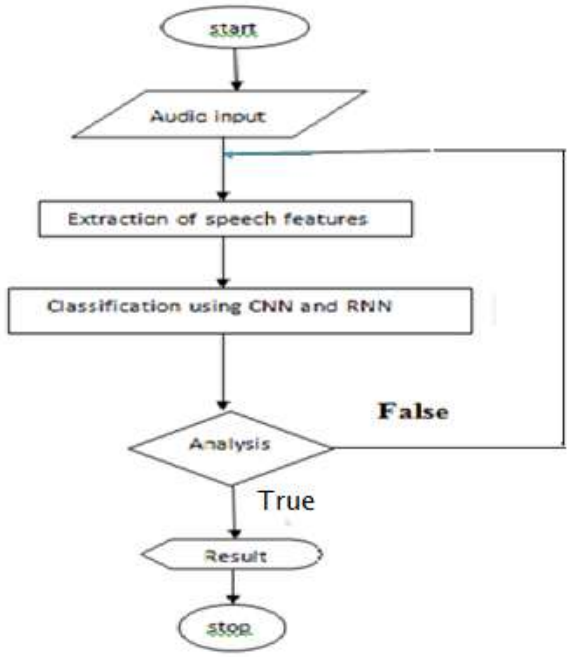

Fig 2: Data Flow Diagram

2. MODULES

2.1 List of modules

- Speech Signal
- Speech Dataset
- Speech Emotion

2.2MODULES DESCRIPTION

Speech Signal

Speech processing typically involves a basic representation of a speech signal in a digital domain which requires limiting the band width of the signal, sampling it at a certain corresponding rate and storing each sample with an adequate resolution. But our focus in the field of speech processing is in communication.

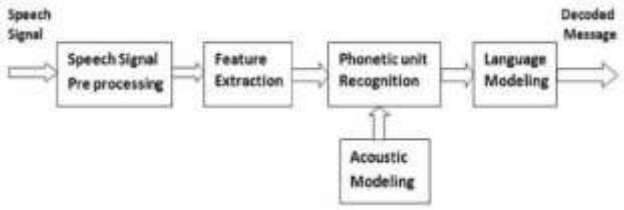

Fig 3: Speech signal processing diagram

Speech Dataset

A speech corpus (or spoken corpus) is a database of speech audio files and text transcriptions. In speech technology, speech corpora are used, among other things, to create acoustic models (which can then be used with a

speech recognition engine). Corpora is the plural of corpus (i.e. it is many such databases).

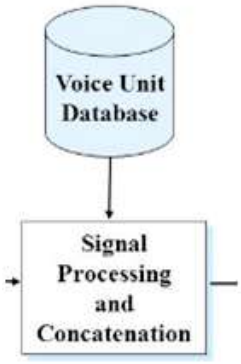

Fig 4: Speech dataset diagram

Speech Emotion

Speech emotion Recognition (SER) can be defined as the extraction of the emotional state of the speaker from his or her speech signal. There are few universal emotions-including Neutral, Anger, Surprise, Fear, Happiness, Sadness which any intelligent system with finite computational resources can be trained to identify or synthesize as required.

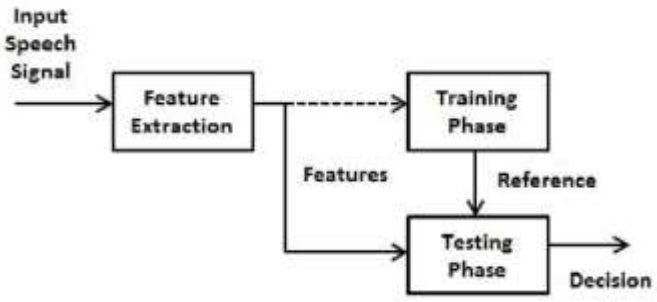

Fig 5: Speech emotion processing

3. ALGORITHM:

Recurrent Neural Network (RNN)

- Convert abstracts from list of strings into list of lists of integers (sequences)
- Create feature and labels from sequences
- Build LSTM model with Embedding, LSTM, and Dense layers
- Load in pre-trained embeddings
- Train model to predict next work in sequence
- Make predictions by passing in starting sequence

CONVOLUTIONAL NEURAL NETWORK (CNN)

- Understand the type of data.
- Preprocess audio data.
- Construct the neural network model.
- Adding a dropout layer.
- Re-evaluate your new model.
- Predictions on test data.

4. CONCLUSION:

The field of machine learning is sufficiently new to still be rapidly expanding, often from innovation in new formalizations of machine learning problems driven by practical applications. However, recognizing emotions from speech is still a challenging problem. In this paper, we proposed the CNNs, RNNs and time distributed CNNs based network without using any traditional hand-crafted features to classify emotional speech. For SER, we combined a deep hierarchical CNNs feature extraction architecture with LSTM network layers. Moreover, we investigated the recognition result by comparing with the basic CNNs and LSTM based emotion recognition results. We verified that CNNs-based time distributed networks show better results. This comparison of results provides a baseline for future research, and we expect that it can give a better result when using more concatenated CNNs. In future, we are planning to study the audio/video based multimodal emotion recognition task.

REFERENCES

1. Priyanka M; A. Milton,"cross corpus speech emotion recognition", 2019.
2. ChangjiangJiang; Rong Mao," Speech Emotion Recognition based on Multiple Feature Fusion", 2019.
3. Boris Knyazev ; Roman Shvetsov ; Natalia Efremova ; Artem Kuharenko," Leveraging Large Face Recognition Data for Emotion Classification", 2018.
4. Michal Chmulik ; Roman Jarina ; Michal Kuba ; Eva Lieskovska," Continuous Music Emotion Recognition Using Selected Audio Features",2019.
5. Sevedehsamaneh Shojaeilangari;Wei-Yun Yau;Karthik Nandhakumar,"Robust Representation and Recognition of Facial Emotions Using Extreme Sparse Learning", 2018.
6. Tian Kexin; Huang Yongming," Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition.", 2019.

7. Kołakowska. ,"A review of emotion recognition methods based on keystroke dynamics and mouse movements". Xi Ouyang ; Srikanth Nagisetty ; Ester Gue Hua Goh ; Shengmei Shen ; Wan Ding ; Huaiping Ming," Audio-Visual Emotion Recognition with Capsule-like Feature Representation and Model-Based Reinforcement Learning",2018.

8. K. Takahashi, and R. Nakatsu.,"Emotion recognition from speech: a review".

9. P. Schaich and J. Williams,"Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System".

10. Yue Xie ; Ruiyu Liang ; Zhenlin Liang ; Chengwei Huang ; Cairong Zou ; Bjrn Schuller ," Speech Emotion Classification Using Attention-Based LSTM",2019.

11. Rohan Rajak ; Rajib Mall," Emotion recognition from audio, dimensional and discrete categorization using CNNs",2019