International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Recognition of Speech Emotion Fusion of Temporal and Spatial Features using Deep Learning

Zanje Prajakta [1], Bhosale Mohini [2], Dhas Payal [3], Thatar Pooja[4], Prof. Unde Suvarna[5]

[1,2,3,4,5]Rajiv Gandhi College of Engineering, Ahmednagar, Maharashtra

ABSTRACT

Emotional intelligence is a part of speech recognition that is becoming popular and in increasing demand. The computer aims to provide a better relationship and relationship between humans and computers. The main goal is to enable computers to understand the emotions expressed by human subjects and thus be able to provide personalized responses accordingly. While there are methods of using machine learning techniques to identify emotions, this project attempts to use deep learning and image classification to identify emotions and classify emotions based on recommendations. This document examines and explores various materials used to teach emotional intelligence.

Keywords speech recognition; deep neural network, Dataset, emotion recognition, automatic speech recognition, deep learning

I. INTRODUCTION

In the context of the rapid development of artificial intelligence (AI), human computer inter action (HCI) has been extensively studied. We live in a world where Siri and Alexa are physically close. Understanding people's emotions leads to understanding people's needs. Speech Emotion Recognition (SER) systems [1] distinguish emotions in speech and are important for human computer support, healthcare, user satisfaction products, social analysis, stress analysis, and intelligent systems. Moreover, SER system is also effective in online teaching, translation, intelligent driving and medical treatment. In some cases, humans may be replaced by computer generated characters that behave naturally and communicate convincingly by expressing human emotions. Machines need to interpret thoughts expressed in words. Only with this ability can successful communication based on the integration of human machine trust and understanding be achieved.

II. Literature Review

Recognizing and combining emotions in speech has been an important area of research for decades, and in recent years this knowledge has become even more important due to the interest of virtual assistants (such as Siri, Alexa or Google Home) and their applications in Artificial Intelligence. In the last few years, emphasis has been placed on thinking in speech. Montero et al. [1] put forward that synthesized speech cannot be marked as natural sounding in the absence of emotional features. Li and Zhao [2] used acoustic features to identify emotions in speech. They used features extracted from short- and long-term frames of utterances achieving an accuracy of 62% using Gaussian mixture models. In the study from Kandali et al. [3], the researchers used a Hidden Markov Model (HMM) and SVM to classify five different types of emotions. HMMs were used to model the sequential forward selection by identifying the best set of features. The experiments, performed on a Danish emotional speech dataset to establish gender independent predictions, recorded an accuracy rate of 88.9%. In another study by Kandali et al[4].

III. METHODOLOGY

This section describes the implementation of the feature extraction stage

3.1 Pre-processing:

3.1.1 Sampling:

Sampling is the first important step in signal processing. The signals we generally use are analog signals, that is, continuous times. So for computational purposes discrete signals are better. Sampling is required to convert these continuous time symbols into discrete time symbols.

$$f_s = \frac{1}{T}$$

fs denotes sampling frequency

3.1.2 Pre-emphasis:

The input signal usually has some low frequency, making the sample similar to neighboring samples. These components represent slow changes over time and are therefore not significant when providing signal information. Therefore, we prioritize the signal by applying a high-pass filter to highlight high energy components that represent rapidly changing signals.

This will provide important information.

3.1.3 De-silencing:

Music often has a zone of silence between the beginning and end, sometimes at higher frequencies. Squelching should be done by removing the unwanted part of the signal. Cancellation of silence is done using a threshold for the signal. We get a signal from which the voiceless part, about which we have no information, is removed and the voiced part is kept.

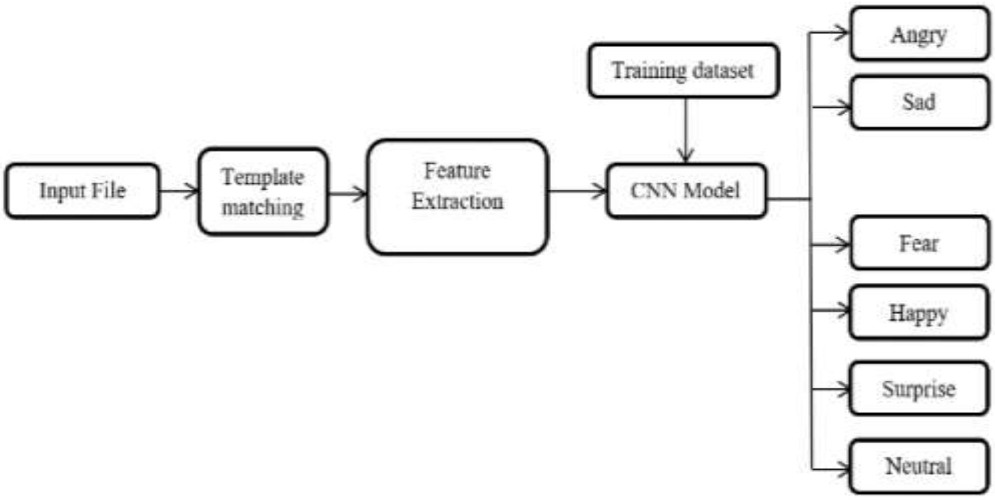
3.1.4 Framing:

For analytical purposes, observable stable signals are preferred. If we look at the speech signal short time, we can obtain a smooth signal. We divide the input signal into small pieces at certain times. In general, it has been determined that duration of 2030 ms is reached for speech. This ensures two things; The first one ensures that the problem is stable in a short time, and the second one ensures that the signal does not change too much in a short time.

3.1.5 Windowing:

Most digital signals are so large and infinite that they all need to be analyzed at once. The signal value of each field must be available for better calculation and performance. A window must be opened to convert a large digital signal into a small signal to be processed and analyzed and arrive at the final signal. Rectangular, Hamming, Blackman etc

IV. WORKING PROCEDURE



CNN Algorithm:

//Anaconda With Jupyter Notebook Tool In Python Language.

Step 1: The Sample Audio Is Provided As Input.

Step 2: The Spectrogram And Waveform Is Plotted From The Audio File.

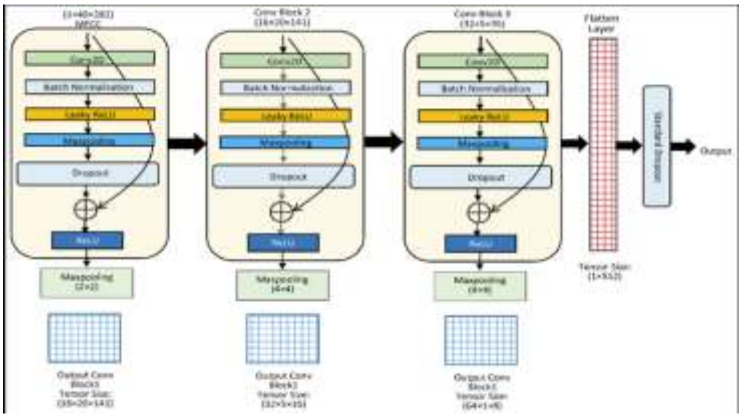
Step 3: Using The LIBROSA, A Python Library We Extract The MFCC (Mel Frequency Cepstral Coefficient) Usually About 10–20.

//Processing Software

Step 4: Remixing The Data, Dividing It In Train And Test And There After Constructing A CNN Model And Its Following Layers To Train The Dataset.

Step 5: Predicting The Human Voice Emotion From That Trained Data (Sample No. - Predicted Value - Actual Value)

V. System Implementation



Methodology Used In this work, we make a comprehensive comparison of different approaches to speech-based emotion recognition. Analyzes were conducted using recordings from the Ryerson Audiovisual Affective Speech and Song Database (RAVDESS). After the raw data is preprocessed, features such as logarithmic mel spectrogram, mel cepstral coefficients (MFCC), noise and power are taken into account. The importance of these features for classification is compared using methods such as Long Short Term Memory (LSTM), Convolutional Neural Network (CNN), Hidden Markov Model (HMM) and Deep Neural Network (DNN).

Training Model and Testing Model:

A training data is fetched to the system which consists the expression label and Weight training is also provided for that network. An audio is taken as an input. Thereafter, intensity normalization is applied over the audio. A normalized audio is used to train the Convolutional Network, this is done to ensure that the impact of presentation sequence of the examples doesn‘t affect the training performance. The collections of weights come out as an outcome to this training process and it acquires the best results with this learning data. While testing, the dataset fetches the system with pitch and energy, and based on final network weights trained it gives the determined emotion. The output is represented in a numerical value each corresponds to either of five expressions.

Speech Database:

In this research, different speech data were used to analyze speech perception. It is widely used in all datasets of Berlin and AIBO. Burkhardt et al. It w as written in German by the actors. The recording facility is the Technical Acoustics Department of TU Berlin. 5 German male artists and 5 German fe male artists contributed by reading one of the selected articles. The different emotions recorded are anger, fear, neutrality, hatred, happiness and sadnes s. Another piece of emotional data, called Aibo, was collected in the real world when 51 children interacted and played with Sony's Aibo robot, which was controlled by a human, and answered the phone to delete the children's conversations. AIBO records five moods: positive, neutral, angry, relaxed a nd anxious.

VI. CONCLUSION

SER technology, one of the key technologies in humancomputer interaction, has attracted great attention from researchers at home and abroad in recent years due to its ability to analyze behavior and thus improve the human-computer relationship. In this paper, we propose a deep

learning algorithm with SER fusion features. For data processing, we used four four-process

RAVDESS datasets containing 5760 audio samples using white Gaussian noise (AWGN). For the network model, we converted both convolutional neu ral networks (CNN) to extract spatial features and encoder network to extract physical features to classify emotions into eight groups.

VII. REFERENCES

1. Montero, J.M.; Gutierrez-Arriola, J.M.; Palazuelos, S.E.; Enriquez, E.; Aguilera, S.; Pardo, J.M. Emotional speech synthesis: From speech database to TTS. ICSLP 1998, 98, 923–926. [Google Scholar]

2. Li, Y.; Zhao, Y. Recognizing emotions in speech using short-term and long-term features. In Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 98), Sydney, Australia, 30 November–4 December 1998. [Google Scholar]

3. Lin, Y.-L.; Wei, G. Speech emotion recognition based on HMM and SVM. In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; pp. 4898–4901. [Google Scholar]

4. Kandali, A.; Routray, A.; Basu, T. Emotion recognition from Assamese speeches using MFCC features and GMM classifier. In Proceedings of the TENCON 2008—2008 IEEE Region 10 Conference, Hyderabad, India, 19–21 November 2008; pp. 1–5. [Google Scholar]

5. Speech Emotion Recognition Methods: A Literature Review Babak Basharirad, And Mohammadreza Moradhasel

6. M. M. H. El Ayadi, M. S. Kamel, and F. Karray, ―Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models,‖ in 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, 2007, vol. 4, pp. IV–957–IV–960.

7. Harár, P.; Burget, R.; Kishore Dutta, M. Speech Emotion Recognition with studies. In Proceedings of the 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2–3 February 2017; pp. 137–140.

8. Kun Han, Dong Yu, and Ivan Tashev, ―Speech emotion recognition using deep neural network and extreme learning machine.,‖ in Interspeech, 2014, pp. 223–227.

9. S. Wu, T. H. Falk, and W. Y. Chan, ―Automatic recognition of speech emotion using long-term spectro-temporal features,‖ in 16th International Conference on Digital Signal Processing, (Santorini- Hellas), pp. 1–6, IEEE, 5-7 July 2009. DOI: 10.1109/ICDSP.2009.5201047

10. 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. .