SPEECH EMOTION RECOGNITION

DHARANI PRASAD S[1], BRIJ VISHAL RAJPUT[2], DHANJIT DEKA[3], ADITHYA BALAGOPALA[4],DR POOJA NAYAK S[5]

[123456] prasadsdharani@gmail.com, brijvishal.bv@gmail.com, adithyab0727@gmail.com, dhanjit1951@gmail.com, pooja-ise@dsatm.edu.in,

[12345] Student, Department of Information science and engineering, DSATM, Bangalore-88, Karnataka

[6]Faculty Department of Information science and engineering, DSATM, Bangalore-88, Karnataka

ABSTRACT

The recognition of emotions is automatic and unconscious by humans. This is a vital process for people to communicate with each other and thus, to get a better interaction from the human machine, emotions have to be considered. The importance of emotional recognition of speech increases in a number of fields. The researchers spoke about the impact of emotions in multidisciplinary applications.[1] The prediction of human emotions attracts the attention of many areas of research, which require exact predictions in uncontrolled scenarios. It has become essential to understand emotion in order to meet requirements in the actual world. Speech transmission is frequently distorted by acoustic background noise in real-world applications.[2] Although there are ways to identify emotions using machine learning methods, this study aims to identify emotions and categorize them in accordance with voice signals using deep learning, machine learning and picture classification approaches. In this research, several datasets are examined and studied for training an emotion recognition model. The study addresses a few database andcurrent methodology-related difficulties.

Keywords: *speech motion recognition; emotion recognition; automatic speech recognition; deep learning, image recognition;*

1. INTRODUCTION

The most natural form of communication is language. One of the significant information expressions and a vital component of the information in human perception is the emotional signal in the speech signal. Since robots can never fully understand the emotions of a speaker on their own, emotion identification of speech has long been a major focus of research in systems involving human-machine interaction. The quickest and most recognisable method of correspondence between individuals is the manner of speech. This fact has led academics to believe that voice is a speedy and efficient method of human and computer communication. On the other hand, this demands that the machine has sufficient knowledge to recognise human voices. Since the late 1950s, there has been a lot of study on speech mode recognition, which is nothing but the process of turning human conversation into a collection of words. Nevertheless, despite a remarkable progress in voice recognition, we are still a long way from creating a distinctive relationship between man and computer since the machine cannot understand the users emotional state. The improved interaction between humans and technology depends in large part on affect recognition.[1] Applications for voice emotion detection may be found in a variety of settings, including all centres and human-computer interaction.

Emotions can be quantified using one of three main methods: categorical, continuous, or appraisal-based. In recent years, Deep Neural Networks (DNNs) have proliferated and made revolutionary advancements in a variety of machine learning fields, including the continuous impact. The writers are recognised as first authors jointly. Recently, many fresh DNN designs have been put out in that approach, including Long-Short Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs). Speech is a complicated signal made up of several pieces of information, including the message to be sent, the speaker, the language, the location, the

emotions, etc. [2]

One of the key areas of digital signal processing is speech processing, which has applications in security, telecommunication, assistive technology, human computer interfaces, and other areas. A little amount of information from the speech signal is extracted using the feature extraction method so that it may subsequently beutilised to identify each speaker.

1.2 BACKGROUND INFORMATION

1.2.1 SPEECH EMOTION RECOGNITION

The goal of speech emotion recognition research is to guess the speaker's emotional state from speech patterns. According to many surveys, improvements in emotion recognition would simplify many processes, making the world a better place to live. SER has a unique use, which is described further below. Emotion recognition is a difficult topic because emotions vary depending on environment, culture, and individual facial reactions, and since speech corpora alone are insufficient to reliably infer emotions in a wide variety of languages.[2]

1.2.1 DEEP LEARNING

Deep Learning is a type of machine learning in which data models are created that are tied to a particular job. Numerous applications, including patternidentification, classification, decision-making, and image recognition, employ deep learning in neural networks. Other deep learning methods, such as multimodal deep learning, are widely utilized for feature extraction and easy picture identification.

1.4 APPLICATION OF EMOTIONRECOGNITION

In call centers, Emotion Recognition is used to cluster calls based on their emotional state. Conversational analysis uses emotion recognition as a accuracy measure to identify unsatisfied customers, customer happiness, and other factors. The usage of SER is an car based system that relies on the driver's mental state information to start the driver's safety and provides precaution.

2. WHY DEEP LEARNING?

Recent years have seen a growing interest in the field of machine learning research known as "deep learning" [18]. Using deep learning techniques for SER offers several advantages over more conventional methods, including the ability to detect complex structures and features without manual feature extraction and tuning, the propensity to extract low- level features from the given raw data, and the ability to process unlabeled data. Input and output are separated by one or more underlying hidden layers in a deep neural network (DNN), based on a forward topology. To manage images and videos efficiently, forwarding architectures such as deep neural networks (DNNs) and convolutional neural networks (CNNs) are used. However, regressive structures, such as recurrent neural networks (RNNs).[3]

Traditional Machine Learning Flow Mechanism

Deep Learning Flow Mechanism

Deep learning techniques include multiple nonlinear parts that perform parallel computation to overcome the disadvantages of previous methods, these solutions must be organized with deeper layers of architecture. Deep learning methods like Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (RNN), Deep Belief Network (DBN), Convolutional Neural Networks (CNN) and Auto Encoder (AE) are part of basic knowledge. learning methods are used for CSR, which greatly increases the performance of the designed system. In recent years, deep learning has attracted a lot of attention as a burgeoning research area in machine learning. Some researchers have trained their separate SER models using DNN.[3]
The distinction between the regular machinelearning flow and the deep learning flowprocesses for SER is shown in the above figure. It provides a thorough comparison of Deeplearning (i.e., Deep Learning) and conventionalmethods.[3]

3. DEEP LEARNINGTECHNIQUES FOR SER:

3.1 AUTOENCODERS

One of the major aims of feature extraction, one of the key classification duties, is to find a solid data representation in the presence of noise. Autoencoders are a type of unsupervised machine learning technique that may be used for this. Because an autoencoder network normally consists of two parts: an encoder and a decoder that learns to construct a replica of the input to the output as accurately as possible, the input and output dimensions are the same. What makes such a network beautiful is the hidden layer that describes the "code" used to representthe data, or a dense representation of the real data.

3.2 MULTITASK LEARNING:

Kim et al. exploited gender and naturalness as supplementary tasks for deep neural networks in MTL in 2017. Their proposed method's high- level feature representation allows for theobservation of discrete emotional groups. They show the results of their practise on six corpora using cross- and within-corpus analysis. They used six corpora ranging from simulated to genuine datasets and two architectures, LSTM and DNN, to see how MTL influenced them. Although MTL beat STL in cross-corpora trials and gained greatly in generalisation, their modelimprovement was minimal in within-corpora testing. Adversarial Generative Learning[4]

3.2 GENERATIVE ADVERSARIALLEARNING

General, also known as Vanilla GAN [85], is composed of a generator neural network (G) and a discriminator neural network. (D). G(z) generates synthetic data from z, which is a sample from the probability distribution P. (z). The discriminator, on the other hand, utilizes the data to determine if the incoming data is manufactured or genuine. Finally, both networks reach an equilibrium in which they have a value function, with one agent attempting to maximize it and the other attempting to minimize it, as indicated in the equation below:

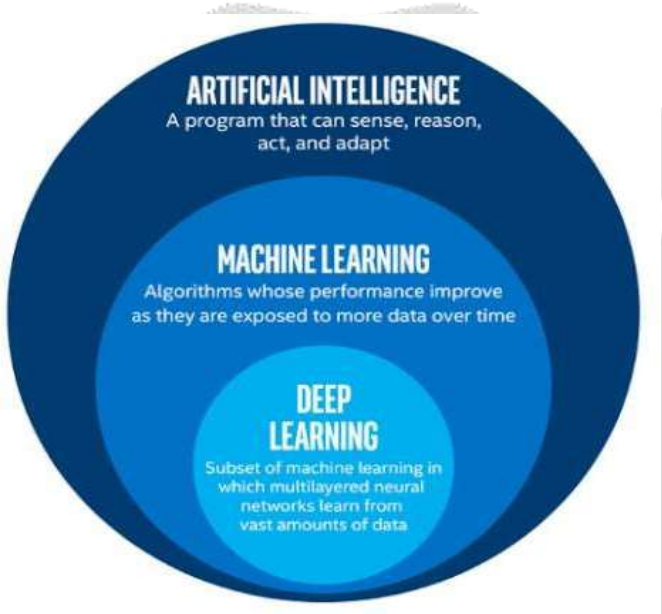
Ex P(x) [log D(x)] + Ez P(z) [log(1 D(G(z))] = E xP(x)
[log D(x)] + E zP(z) [log(1 D(G(z))][4]

3.3 TRANSFER LEARNING

The cross-domain problem with SER, when test corpora don't match train corpora, can be solved using transfer learning. To obtain two neighboring latent feature spaces for the source and target corpora and an SVM as the classifier method, Song et al. [90] practice dimension reduction and Maximum Mean discrepancy embedding optimization using transfer learning in cross-corpus speech emotion identification tasks. Five emotion categories from the EMO-DB were used as the source corpus for the experiment, while the same emotion categories from a Chinese emotion dataset were used as the test corpus. Principal component analysis (PCA) and local preserving projection have both been employed in the two proposed models to reduce their dimensions.[4] Therefore, neutral has the highest rate of recognition, whereas fear and happiness have lower rates. However, the proposed strategy outperformed the automatic recognition approach in terms of recognition rate.

4.5 ATTENTION MECHANISM

The attention mechanism for deep learning is another technique that has recently proved effective in speech emotion recognition [83,92-95]. Despite the fact that in normal deep learning algorithms for SER, all locations of a specific syllable receive equal attention, emotion is not uniformly distributed across the speech for each sample. Based on the attention weights assigned to each segment of the data that comprises an emotionally salient aspect, the classifier examines the placements of the provided samples in the attention mechanism. Mirsamadietal employed bidirectional LSTM with a weighted-polling technique to uncover more perceptive features of emotion than standard low-level descriptors (LLD) and high-level statistical aggregation functions (HSF).



VARIOUS DATABASES THAT HELP TORENDER AND TEST SER:

5.1 Crowd-sourced Emotional MultimodalActors Dataset (Crema-D)

A collection of audio-visual data that is particularly well suited for the investigation of multimodal emotion expression and perception. The data collection includes phrases uttered in a variety of fundamental emotional states with various facial and vocal emotional expressions. Multiple raters evaluated 7,442 footages of 91 performers from differentethnic origins in the audio, visual, and audio-visual modes.[5]

5.2 Ryerson Audio-Visual Database of EmotionalSpeech and Song (Ravdess)

A multimodal library of songs and speeches with strong emotions is called the RAVDESS. 24 professional actors with a gender-balanced representation in the database read lexically matched lines with a neutral North American accent.[6]

5.3 Surrey Audio-Visual Expressed Emotion (Savee)

The Surrey Audio-Visual Expressed Emotion (SAVEE) database fills a gap in the development of an automatic emotion recognition system. The collection includes 480British English utterances recorded from four male actors depicting seven different emotions. The phrases were selected from the normal TIMIT corpus and

phonetically balanced for each mood. In a visual media lab, high-quality audio-visual equipment was employed to record the data, which was subsequently analysed andlabelled.

5.4 Toronto emotional speech set (Tess)

These stimuli were inspired by the Northwestern University Auditory Test No. 6. (NU-6; Tillman & Carhart, 1966). Two actresses (26 and 64 years old) recited a set of 200 target words in the carrier phrase "Say the word_," and recordings of the set were made to depict each of the seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). Thereare a total of 2800 stimuli.

6. FUTURE SCOPE

This field has a lot of shortcomings and has a lot of room for growth. The intricacy of speech signal preprocessing makes it difficult to identify emotions from speech signals. This field of study is capable. Only seven fundamental emotions have been successfully recognized thus far. For the purpose of identifying more than seven emotions, research should be conducted. The work will be expanded in the future to improve the accuracy of emotion recognition systems. An accurate implementation of the speaking pace can be investigated to see if it can address some of the model's shortcomings.

7. APPLICATIONS

The speech emotion recognition system has a wide range of uses. A few of them are mobile- based voice recognition systems, intelligent toys, mental diagnosis, and dialogue with robots. Call center emotion recognition, which allows for the identification of consumer emotions and improves service quality. Intelligent tutoring software, in-car boards, prosody in dialogue systems, voice mail sorting, computer games, lie detectors, learning environments, educational software, and the ability to identify the emotional state of call center conversations in order to provide feedback to an operator or supervisor for monitoring purposes are all applications of speech emotion recognition. There is little interaction between teachers and students in the online learning environment. Teachers can better comprehend their students' responses to matching learning content by identifying the emotional states of the students through vocal signals in the online learning environment.[8]

8. CONCLUSION

The application of deep learning, machine learning, and picture classification techniques for emotion recognition has been the subject of numerous studies and surveys. In transdisciplinary applications, an essential component is understanding emotions. Emotion identification of speech has long been a key area of research in systems involving human-machine interaction because robots will never fully understand the emotions of a speaker on their own.

In this project, autoencoders, multitasking, general, transfer learning, and multiple databases including Crema-D, Savee, and Tess have all been investigated. The intricacy of speech signal preprocessing makes it difficult to identify emotions from speech signals. Research should be conducted to improve the accuracy of emotion recognition systems. Other augmentation.

9. REFERENCES

[1]- Speech Emotion Recognition Using Machine Learning 1Amitha Khan K H, 2Ankitha Chinnu Mathew, 3Ansu Raju, 4Navya Lekshmi M, 5Raveena R Maranagttu, 6Rani Saritha R

[2] - Speech Emotion Recognition
Ramya. T1, Erica Davey2, Kavya. M3, Uzma Taj4, Mrs. Supriya5

[3] - Speech Emotion Recognition Using Deep Learning Techniques: A Review RUHUL AMIN KHALIL1 , EDWARD JONES 2 , MOHAMMAD INAYATULLAH BABAR 1 , TARIQULLAH JAN1 , MOHAMMAD HASEEB ZAFAR 3 ,

AND THAMER ALHUSSAIN4

[4] - Review Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models Babak Joze Abbaschian, Daniel Sierra-Sosa and Adel Elmaghraby

[5] - https://github.com/CheyneyComputerScience/CRMA-D

[6] - https://zenodo.org/record/1188976#.Y60hrXZBxD8

[7] - AUTOMATIC AUDIO BASED EMOTION RECOGNITION SYSTEM: SCOPE AND CHALLENGESChaitanya Singlaa* , Sukhdev Singhb ,Monika Pathakc

[8] – A Literature Review on Emotion Recognition usingVarious Methods By Reeshad Khan & Omar Sharif