iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VIII **Month of publication:** Aug 2023

DOI: https://doi.org/10.22214/ijraset.2023.55207

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

Text to Speech Conversion based on Emotion using Recurrent Neural Network

Nihar M. Ranjan
Department of Information Technology, JSPM'S Rajarshi Shahu College of Engineering, Pune

Abstract: ***Emotion based text to speech conversion system will be proved as an improved version of traditional text to speech system. Emotions help us in recognizing the message conveyed by the conveyor in more effective way. ETTS (Emotion based text to speech conversion system) will stand out from TTS (Text to speech) as it will have variation of voice according to the emotions detected in the text. ETTS will detect four basic and most used emotions of humans. These emotions are 'happy', 'sad', 'anger', 'neutral'. As ETTS will use RNN (Recurrent Neural Network) for identifying emotions, emotions will be identified with high accuracy. GRU (Gated Recurrent Unit) model of recurrent neural network helps in maintaining the high accuracy of prediction. Emotion based text to speech system will be beneficial for educational purpose like listening stories from storytelling applications for young budding children. Emotion based text to speech is going to be serviceable for visually impaired individuals. It can also be used in robots for better communication.***

Keywords: Emotion recognition, Text to speech, Recurrent Neural Network

I. INTRODUCTION

Speech synthesis is the artificial production of a human speech. A computer system used for this purpose is called as speech synthesizer, and can be implemented in software or hardware products.

A text-to-speech (TTS) system converts normal language text into speech. Other system renders symbolic linguistic representations like phonetic transcriptions into speech but emotion-based text to speech system adds emotional touch to the speech which will enhance the user experience.

Sentiment analysis and semantic analysis are processes of analysing text and determining the emotional tone they possess. It uses NLP i.e., natural language processing, text analysis, computational linguistics and Neural Network. A basic task in sentiment analysis is classifying the *polarity* of a given text at the document, sentence, or feature/aspect level whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry", "sad", and "happy".

A text-to-speech system (or "engine") composed of two parts: a front-end and a back-end.

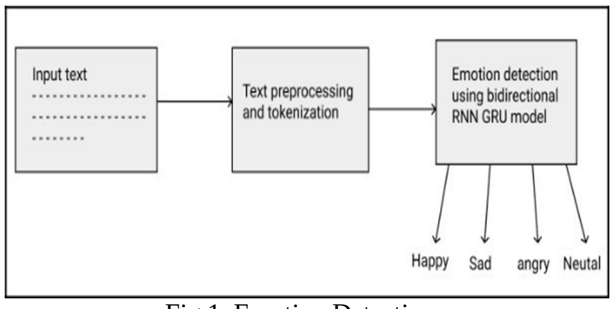


Fig 1: Emotion Detection

The front-end has a task to take text input from the user to process it into emotion-based speech.

At backend, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This phoneme or grapheme-to-phoneme conversion.

At backend, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process is often called text normalization, pre-processing, or tokenization. It then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences.

The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the back-end. The back-end often referred to as the synthesizer then converts the symbolic linguistic representation into sound. In certain systems, this part includes the computation of the target prosody (pitch contour, phoneme durations), which is then imposed on the output speech. For this process we are going to use recurrent neural network.

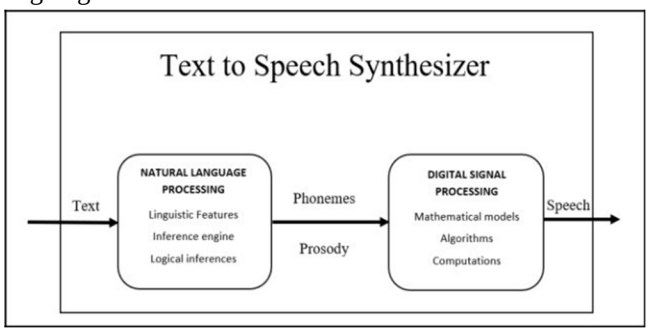


Fig 2: Existing System of Text to Speech conversion

The above figure shows a basic overview of emotion-based text to speech system or it is a backbone of ETTS system till now. It consists of two blocks namely Natural Language Processing (NLP) and Digital Signal Processing (DSP). NLP mainly consists of converting text into specific forms for processing. Here text is converted into Phenomes, Diphones to analyse and extract the emotion, meaning behind the sentence. Digital signal Processing consists of algorithms, models used for creating speech with emotions. Some of the algorithms used are syllabification algorithm, corpus-based systems etc.

II. RELATED WORK

In a [4], the study of speech synthesizers is done. As there are different ways to perform Speech synthesis, we can use any model according to our need. But concatenative synthesis is the method which is most widely used, as it produces naturally sounding speech. There are three sub types of concatenative synthesis, viz. Domain specific synthesis, Unit selection synthesis and Diphone synthesis. Domain specific synthesis has small database as it has selected words used for utterance. Because of this, the speech sounds natural. But it is not used for general purpose synthesis. It is limited to particular device with specified number of words. Unit selection produces less natural sound. For unit selection recorded speeches are used. They are segments of various parameters as phones, diphones, syllables, words and many more parameters. The database used for unit selection is also huge. As digital signal processing (DSP) is used for sound generation, unit selection loses the naturalness in the sound. Diphone synthesis uses minimal speech database i.e., Contain each type of diphone in each word. During processing, all these diphones superimpose with each other and results into speech. But, Diphone produced more mechanized voice and has low natural sound due to high use of digital signal processing.

In a [5], the study of Berkeley speech technologies has been done about text to speech conversion. The brief knowledge about the sound/voice generation from human vocal tract is expressed. With the understanding of the mechanism human vocal tract and sound generation by human, using linear predictive coding (LPC) technique it becomes possible to record human speech in compressed forms on chips. A single speech sample can be generated using multiplication, division and addition cycles. In 1978, Texas Instruments developed a chip which can perform these cycles and produce human like voice but, it became difficult to add synthetic parameters. Mostly this voice is used in toys and playing cars. Since there are many linguistic parameters for speech generation which are sometimes limited by computational power and sometimes by us. We are still not able to understand how the sounds are generated. For converting text into speech there are number of parameters having many stages. First the text has to normalized, then there are some words which has different pronunciation, conversion of letters to phoneme, prosody rule, voice generation, interrupt driver and lastly output hardware.

In a [6], Few students find difficulty in reading text, so with the help of text to speech technology they will get help in reading text. People are now moving towards auto reading rather than reading any text by themselves. By Text to Speech Technology (TTST) people will start reading more text and ultimately generate interest for reading. TTST will be useful for children who find difficulty in reading. Ethnography means seeing and representing. TTST uses this method so that be useful in metacognitive development of students. The self- efficacy increased to 95% using TTST. Also, students start to first listen the word and understand its meaning and then while reading for second time emphasized more on fluency.

©IJRASET: All Rights are Reserved

In a [2], Study of text is done as a syllable. Indian language text can be better segmented as syllable for speech generation. Generally, diphones are considered for breaking down the words. Syllable based synthesis does not require significant prosodic modification. Corpus based speech synthesis gives more natural speech as output with high quality. But database required for this is huge. As number of concatenations required for speech generation is less, which reduces use of DSP and gives natural sound. Hidden Semi-Markov model (HSMMs) is much better than standard HSMM. Because semi Markov model is trained on speech data recorded at normal and fast speaking rate.

In a [9], the paper describes the methods of emotion detection and different datasets used for training model. There are two types of emotion modes, viz. Discrete Emotion Models (DEMs) and Dimensional Emotion Models (DiEMs). In DEMs emotions are placed in distinct classes or categories. In DiEMs emotions are placed in a spatial space like one-dimensional or multidimensional.

III. IMPLEMENTATION DETAILS ALGORITHM

1) By using Flask API in python open the Web browser with relative HTML files and UI for user to interact.
2) Input the text using UI element.
3) If user click on TTS button then simply transfer text to relative function and play the text in form of speech by PYTTSX normal speech and go to step 8. If user clicks on ETTS button then Transfer the text to GRU model (which is defined below) for emotion detection and follow steps from 4th one.
4) Once Emotion is detected , according to emotion call the relative function to create its .wav file.
5) Once .wav file is created open it using PyDub API and do necessary modifications inn parameters like framerate(frequency) , amplitude, speech rate etc.(which details ae given below) and save the modified speech file.
6) Get back to UI and show the emotion detected in text and allow user to play relative speech file by giving him/her button option to do so.
7) Whenever user click on the given buttons perform the given action.
8) Stop

ETTS (Emotion Based Text to Speech System) is more than TTS, as it will concentrate on emotions expressed through the text. All this process will happen with the help of GRU Layer and RNN algorithm.

Dataset: - We have created a dataset by combining three datasets.

They are: -

a) dailydialog
b) emotion-stimulus
c) isear
d) It is a training dataset which we are going to use to train the data. We created our own data by choosing data from given dataset in below table.
e) We split the created dataset into two datasets for training (8987 rows) and for testing(4768 rows).
f) This training dataset consist more than 8900 lines of sentences and testing dataset consist more than 4700 lines of sentences having 4 basic emotions as ‘happy’, ‘sad’, ‘angry’, ‘neutral’.

Dataset	Year	Content	Size	Emotion categories	Balanced
dailydialog	2017	dialogues	102k sentences	neutral, joy, surprise, sadness, anger, disgust, fear	No
emotion-stimulus	2015	dialogues	2.5k sentences	sadness, joy, anger, fear, surprise, disgust	No
isear	1990	emotional situations	7.5k sentences	joy, fear, anger, sadness, disgust, shame, guilt	Yes

Table 1: - Datasets

In ETTS we are going to use Recurrent Neural Network to analyse the sentiments and semantics of a sentence or paragraph. Recurrent neural network (RNN) is a class of neural networks that is helpful in modelling sequence data. Derived from feed forward networks, RNNs exhibit similar behaviour to how human brains function. As recurrent neural network keeps relation between previous word and next word it is going to be useful in analysing the actual emotion expressed in a sentence. We have divided our working of project in five parts-

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

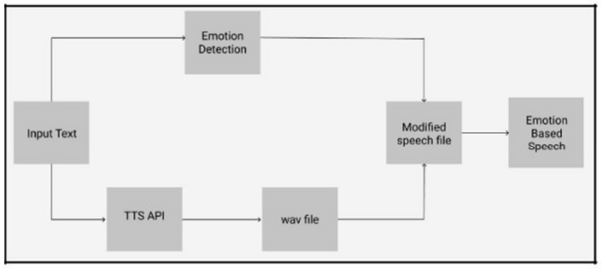


Fig 3: Proposed system for Emotion based text to speechconversion

Emotion	Speech rate	Frequency	Amplitude/Volume
Neutral	160(neutral)	Normal as it is	Normal as it is
Happy	165(slightly more)	.wav File Frequency*(2^0.5)	Original Amplitude+5
Sad	140(slow)	.wav File Frequency*(2^(0.7/2))	Original Amplitude+2
Anger	185(fast)	.wav File Frequency*(2^(0.8/2))	Original Amplitude+10

Table 2: - Emotion parameters

- *Input:* We will take input from the user in text format in the form of single sentence or paragraph. Then this data will be passed to "Text to Speech" translation. These sentences then will be pre-processed by removing stop words, converting input text into lowercase and finally tokenization of the words. After pre-processing input, this data will be passed to next part of the system i.e. emotion analyser.
- *Text to Speech Conversion:* With the help of existing text to speech APIs we will convert the input text into audio file. This audio file will have speech with monotonous voice. We will get '.wav' file using text to speech API.
- *Emotion Detection:* The word semantic itself implies meaning or understanding. Semantic layers concern with the meaning of data and not the structure of data. We will be analysing text on the basis of few parameters such as Meronomy, Polysemy, Synonyms, Homonyms.
- *Meronyms:* Sentence will be logically arranged from which system will be able to relate it to some part ofsentence.
- *Polysemy:* Using this parameter, we will be able to understandthe meaning of words and phrases.
- *Synonyms:* We will find the same meaning of the input text.
- *Homonyms:* Two words that sound same and are spelled alike but have different meanings. With the help of Relationship Extraction, we will be able to find the relation between two or more entities present in text. This will help the model in understanding the sentiments of the sentence. Also, by extracting specific keywords from the text will help inunderstanding the semantics of the sentence.
- *Modified Speech File:* Firstly, the input text will undergo text to speech process which will generate .wav file having speech of the text and by undergoing the process of emotion analysis we will have featureextracted output of the input text. Using both these outputs we will be making changes in the .wav file to add emotions by changing the voice parameters such as pitch range, volume,silence, speech rate, pitch movement and duration
- *Emotion Based Speech:* After making changes in the .wav file according to output results of emotion analysis, the file will be saved and used for speech or audio. Now, this audio will have emotions in it.

Model has four layers that help in processing the text andidentifying emotions.

➢ *Embedded Layer*

In input layer of ETTS we have used a pretrained model named Word2Vec. Word2vec is a technique for natural language processing. The word2vec algorithm uses a neural network model to learn word associations from a large corpus of text. Once trained, such a model can detect synonymous words or suggest additional words for a partial sentence. As the name implies, word2vec represents each distinct word with a particular list of numbers called a vector. The vectors are chosencarefully such that a simple mathematical function (the cosine similarity between the vectors) indicates the level of semantic similarity between the words represented by those vectors.

➢ *GRU Layer*

Bidirectional GRU's are a type of bidirectional recurrent neural networks with only the input and forget gates. It allows for the use of information from both previous time steps and later time steps to make predictions about the current state.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VIII Aug 2023- Available at www.ijraset.com

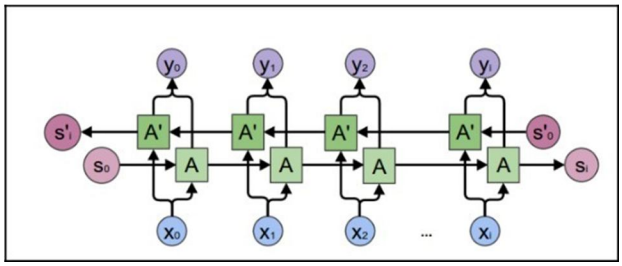


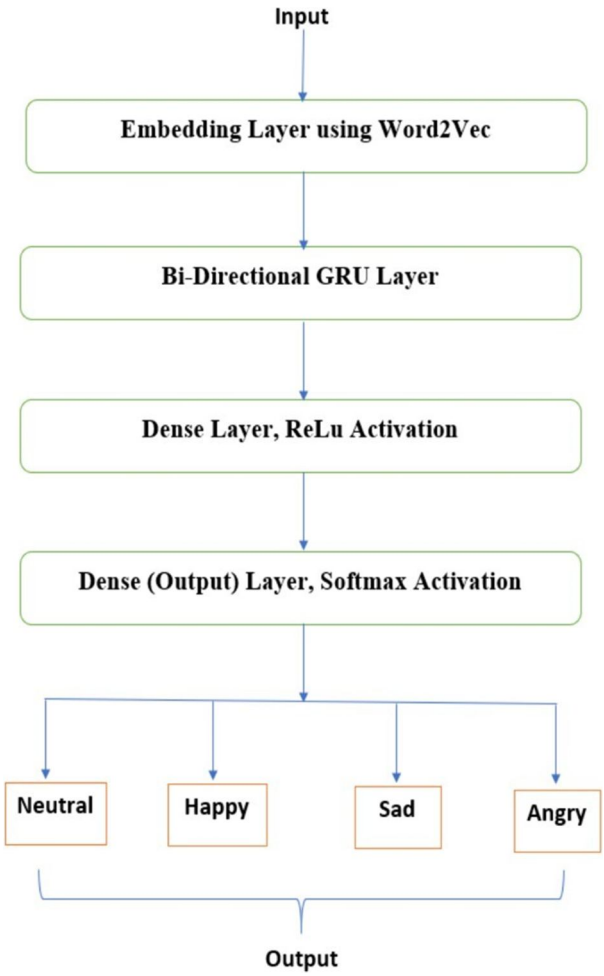
Fig 4. Gated Recurrent Unit

➢ *Dense layer with RELU activation*

Dense layer is the regular deeply connected neural network layer. It is most common and frequently used layer. Dense layer does the below operation on the input and return the output. output = activation (dot(input, kernel) + bias).

The rectified linear activation function or ReLU for short is a piecewise linear function that will output the input directly if it is positive, otherwise, it will give output zero.

The rectified linear activation function overcomes the vanishing gradient problem, allowing models to learn fasterand perform better



➢ *Dense layer with SoftMax activation*

SoftMax is a mathematical function that converts a vector of numbers into a vector of probabilities, where the probabilities ofeach value are proportional to the relative scale of each value in the vector.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Dataset is divided as training and testing dataset.

Fig 5 Training dataset

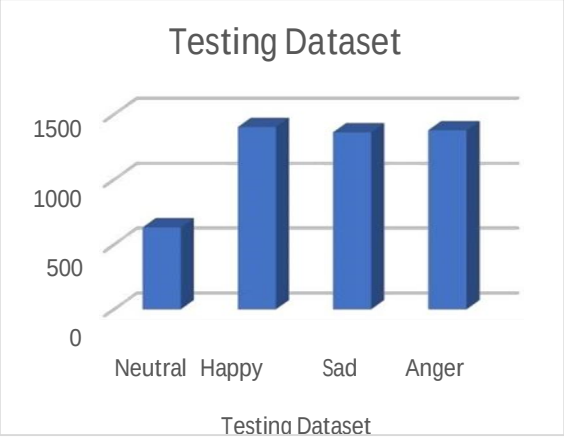
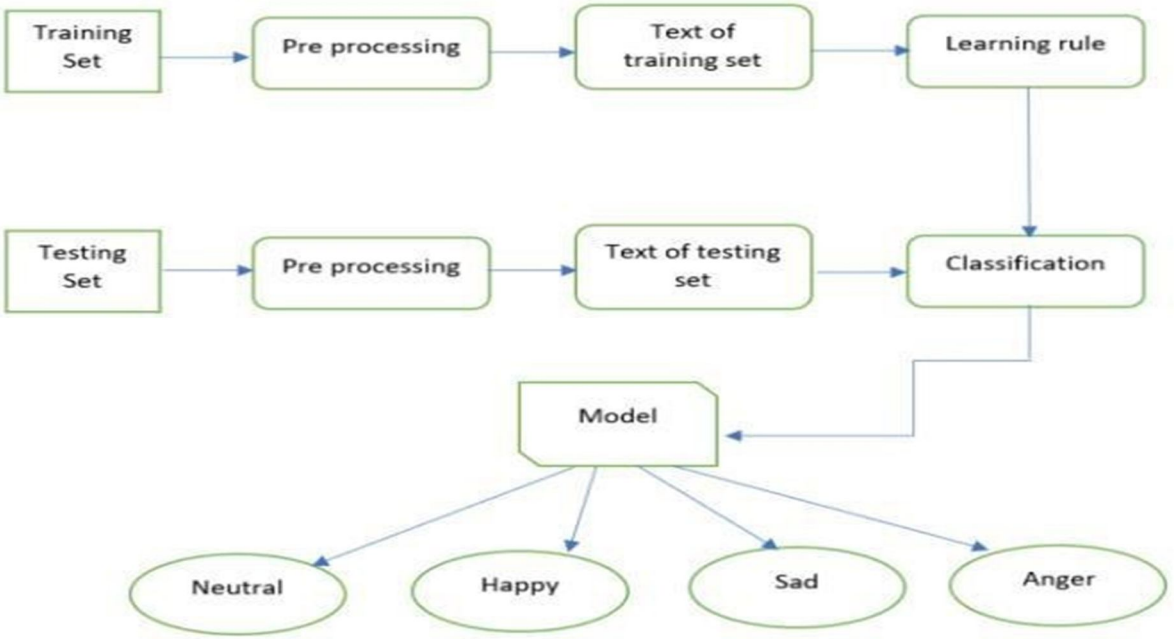


Fig 6. Testing dataset

Model flow diagram :



©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Lets take example, Sentence: Whether the cries of anguishfor 1991 alone are justified remains to be seen
Pre-processing of text :
{'whether': 1, 'the': 2, 'cries': 3, 'of': 4, 'anguish': 5, 'for': 6, '1991':
7, 'alone': 8, 'are': 9, 'justified': 10, 'remains': 11, 'to': 12, 'be':
13,
'seen': 14}
found 14 unique tokens
[[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,
14]] 14
[[00000000000000000000000
00000000000000000000000
…… 00000000000000000000000
00000000000000000000000
00000012345678910111213 14]]
shape of data= (1, 500)
after this process the above vector/matrix is sentforward to model for classification of emotions.

IV. RESULTS AND DISCUSSIONS

The trained model is able to classify text into four emotions (happy, sad, angry, neutral). Trained model hasan accuracy of 86.77% and model is able to identify the emotions. graphical representation is shown in the belowdiagrams.

Fig 7: - Training Accuracy

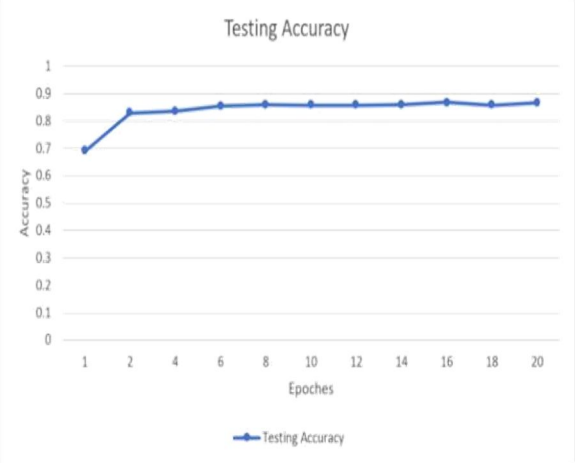


Fig 8: - Testing Accuracy

following graphs give information about the loss duringtraining and testing data.

Fig 9: Losses during Training

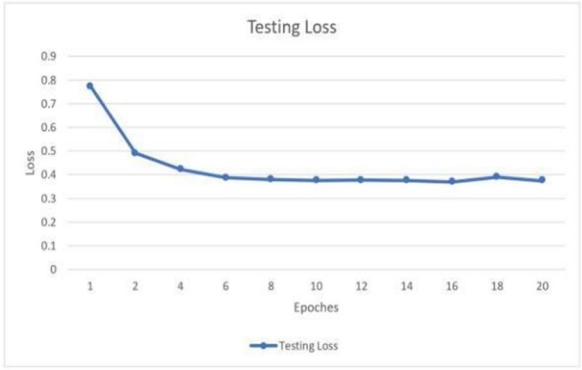


Fig 10: Losses during Testing

1) *Accuracy:* It is calculated using the ratio of total no. true positive and true negative to the ratio of the total observations. The accuracy of the trained model is 86.77 %.
2) *F1-Score:* It is the weighted average of recall and precision.The f1-score of the model is 86. 77.

a) *Input Text:* Maybe that's why of all the things to frighten girls and boys its ghosts who terrify me the most. Not because they can cause damage, or are frightening in themselves, its their existence at all that upsets me. To be trapped, in such a small snippet of existence, a snapshot of a moment, built on trauma or pain. Forever chained to the moment of transition, unable to let go. I get shivers just thinking about it. In a way it is a kind of immortality. Note, that I did not say a way to live forever. The whole world moves on but you as a ghost are stuck, stranded, untouched and untouchable. Unable to influence the world forever,rejected and left on the outside. Who wouldn't choose nothing, compared to that?

b) *Output:* Sad Emotion is detected and Speech file with sademotion based modifications is created and saved.

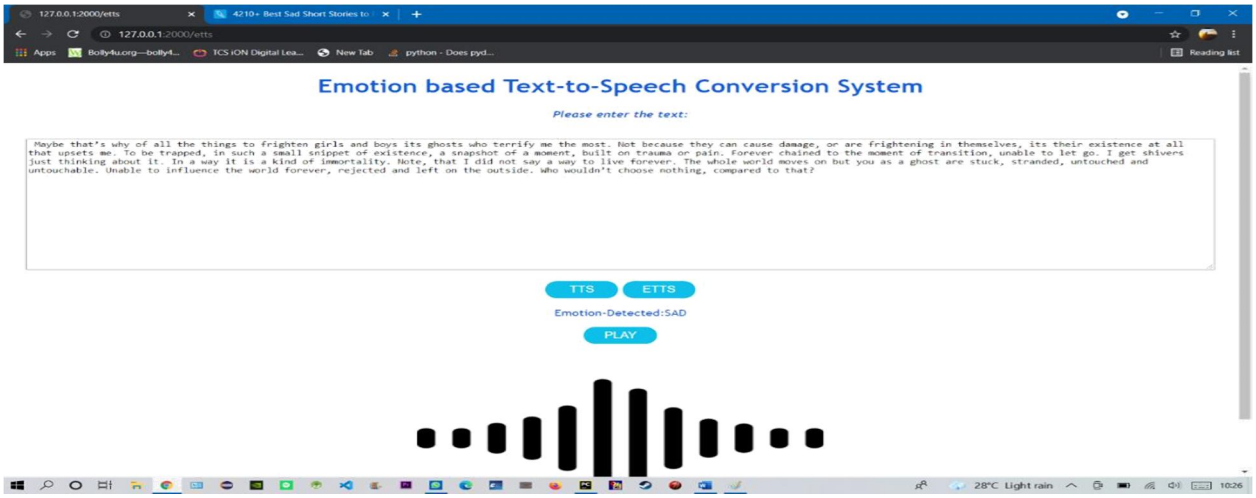


Fig 11: Example input to system and its output

V. CONCLUSIONS

Our model is able to classify text with an accuracy of 86.77% on combined Dataset. The model is able to classify text on the basis of four classes 'happy', 'sad', 'angry; and 'neutral'. Using word embedding along with Recurrent Neural Network helped in maintaining the correlations between the words.

With the help of neural network, we find the correlation between the words in the sentence and understand the meaning and emotions present in them. Neural network helped in classifying data for sentiment and semantic analysis.

By implementing our proposed method, we willbe able to convert text into emotion-based speech which can be used in story telling applications, Audio books andit will be more helpful for visually impaired individual foranalysing emotions from any text.

VI. FUTURE SCOPE

1) Standalone application can be developed
2) More deep emotions can be added such asjealously, shyness etc.
3) Background Music for certain emotions can beadded.

Applications:

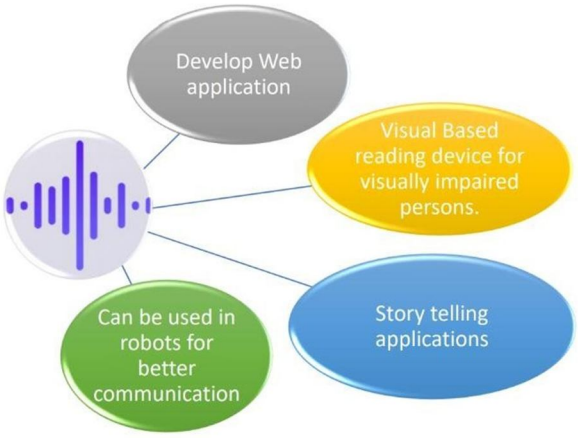
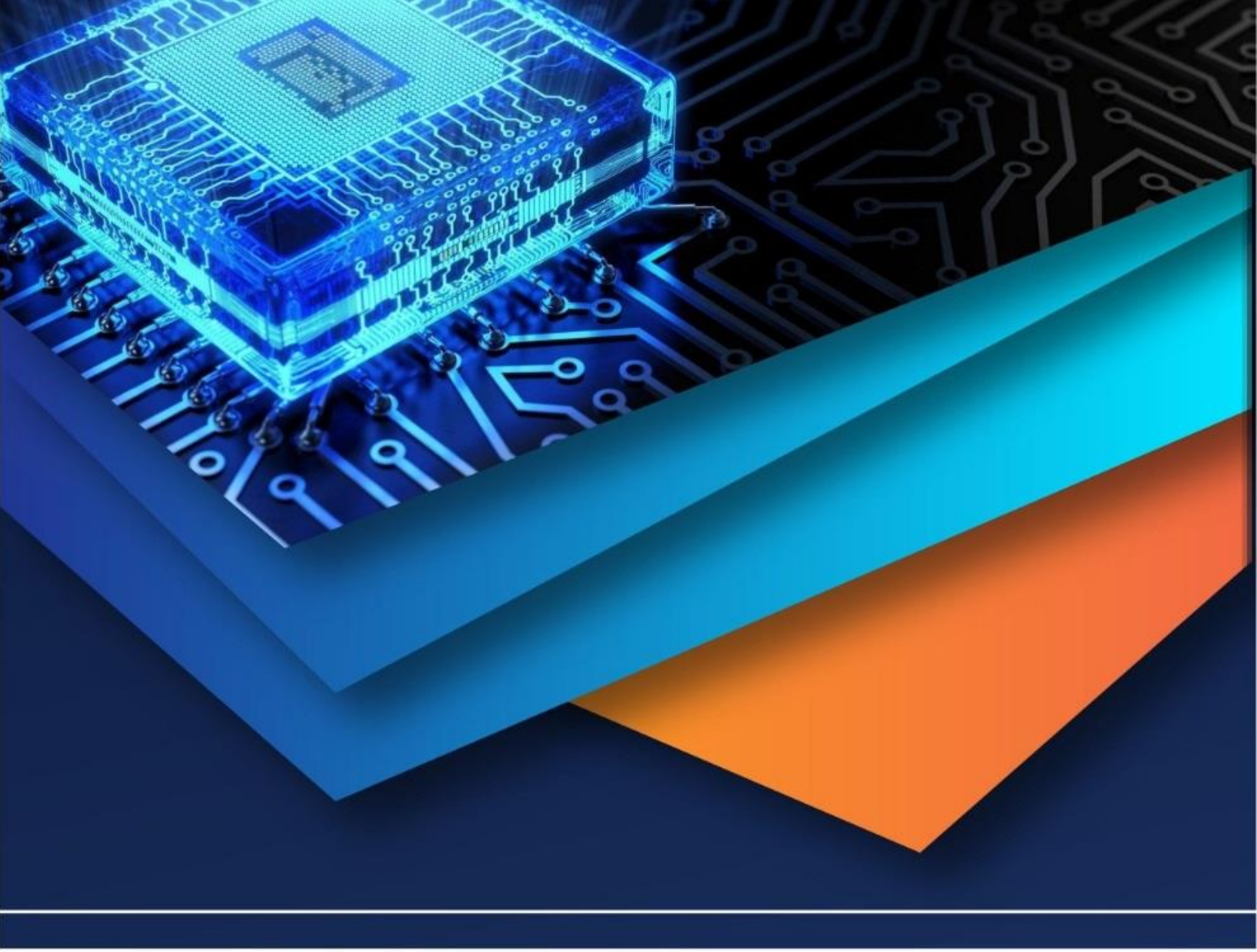


Fig 12: Applications of the Model

REFERENCES

[1] Caroline G. Henton, Santa Cruz, Calif, Method and apparatus for automatic generation of vocal emotion in a synthetic text to speech system, patent number: 5860064, applicationnumber: 805893.
[2] Tejashree M. Shinde, V. U. Deshmukh, P. K. Kadbe, Text to Speech Conversion Using FLITE Algorithm published in Internation Journal of Science and Research (IJSR)ISSN(Online): 2319-7064.
[3] Professor Tadashi Kitamura, Associate Professor Keiichi Tokuda, Simultaneous Modelling of Phonetic and Prosodic Parameters, and Characteristics Conversion for HMM- BASED Text-to-speech Systems published in ResearchGate.
[4] ItunuoluwaIsewon, JeliliOyelade, Olufunke Oladippupo,Design and Implementation of Text to Speech Conversion for Visually Impaired People, Published in International Journal of Applied Information Systems (IJAIS)-ISSN: 2249-0868.
[5] Michael H. O'Malley, Berkeley Speech Technologies, Text- to-Speech Conversion Technology, Published in IEEE journal.
[6] Tejashree Mane, Dhanashree Ghone, Nihar Ranjan, et al., "Text Document Classification by using WordNet Ontology and Neural Network" International Journal of Computer Applications (0975 – 8887) Volume 182 – No. 33, December 2018
[7] R.V. Darekar, Nihar Ranjan, et al, "A hybrid meta-heuristic ensemble based classification technique speech emotion recognition" Advances in Engineering Software , Elsevier, Volume 180, June 2023
[8] Deepak Mane, Ranjeet Bidve, Nihar Ranjan., "Traffic Density Classification for Multiclass Vehicles Using Customized Convolutional Neural Network for Smart City" Lecture Notes in Network and Systems, Springer Nature , September 2022, pp 1015-1030
[9] Nihar Ranjan, Zubair Ghouse "A Multi-function Robot for Military Application" Imperial Journal of Interdisciplinary Research (IJIR) Vol-3, Issue-3, ISSN: 2454-1362, pp 1785-1788, 2017
[10] Nihar Ranjan, Midhun C., " Evolutionary and Incremental Text Document Classifier using Deep Learning" International Journal of Grid and Distributed Computing Vol. 14, No. 1, pp. 587-595, 2021
[11] Shikha Singh, Priti Sarote, et al., "Detection of Parkinson's Disease using Machine Learning Algorithm", International Journal of Computer Applications (0975 – 8887) Volume 184 – No.6, April 2022
[12] Bendre S, Bhosale M, Gunjal Y, Ranjan N. IOT Based Irrigation Automation And Nutrients Recommendation System. IJIACS. 2017 Nov;6(11):43-6.

.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)