A Brief Survey on Emotion Based Text to Speech Conversion System

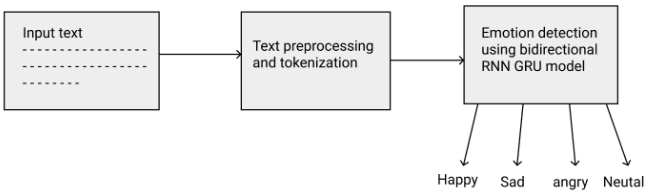
Supriya Dhanaraj Dhumale, Manjiri Vitthal Khopade, Bhushan Dhimate, Avadhoot Yogesh Dhere

Abstract: ***Text to speech conversion is one of the applications of machine learning. It is widely used in search engines, standalone applications, web applications, chatbots and android applications. But still there is need to upgrade text to speech system so that we can get more interactive and user-friendly application. Traditional text to speech application has monotonous voice as output which does not has emotions in it and seems to be more mechanized. So, there is need to improvise the existing system by embedding the flavour of emotions in it. Existing text to speech cannot be used in story telling applications also it does not provide effective communication. Most of the Text to Speech systems are developed using algorithms such as Support Vector Machine (SVM), Naïve Bayes etc. Emotion Based Text to Speech System will help to improvise the existing Text to Speech system. With the help of machine learning and deep learning algorithm such as Recurrent Neural Network can be used for performing sentiment analysis and semantic analysis on the input text. We are going to use neural network which is more effective and help to maintain a relation between previous word and next word. Emotion based text to speech system will be able to identify four emotions 'happy', 'sad', 'angry' and 'neutral'. 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.***

Keywords: ***Emotion recognition, Text to Speech, GRU.***

I. INTRODUCTION

Speech synthesis is the artificial production of a human speech. A computer system is 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 render symbolic linguistic representations like phonetic transcriptions into speech but formatted paper, volume no/ issue no will be in the right top corner of the paper. In the case of failure, the papers will be declined from the database of journal and publishing house. It emotion-based text to speech system add emotional touch to the speech which will enhance the user experience. Sentiment analysis and semantic analysis are processes of analyzing 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.


Input text

Text preprocessing and tokenization
Emotion detection using bidirectional RNN GRU model
Happy
Sad
angry
Neutal

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 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 . 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.

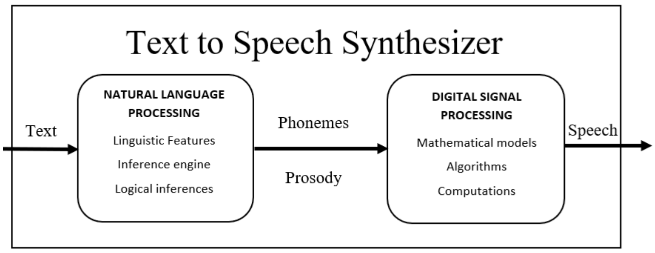


Fig 2: Existing System of Text to Speech conversion

Manuscript received on September 09, 2021.
Revised Manuscript received on September 12, 2021.
Manuscript published on September 30, 2021.
* Correspondence Author

Bhushan Hemant Dhimate, Department of Computer Science, Savitribai Phule Pune University, Pune (Maharashtra), India.

Manjiri Vitthal Khopade, Department of Computer Science, Savitribai Phule Pune University, Pune (Maharashtra), India.

Avadhoot Yogesh Dhere, Department of Computer Science, Savitribai Phule Pune University, Pune (Maharashtra), India.

Supriya Dhanaraj Dhumale, Department of Computer Science, Savitribai Phule Pune University, Pune (Maharashtra), India.

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. LITERATURE REVIEW

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 whichever text is been read by the system the reader will be able to see and understand it properly. This method proved to 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. 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 unidimensional or multidimensional.

III. PROPOSED SYSTEM

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

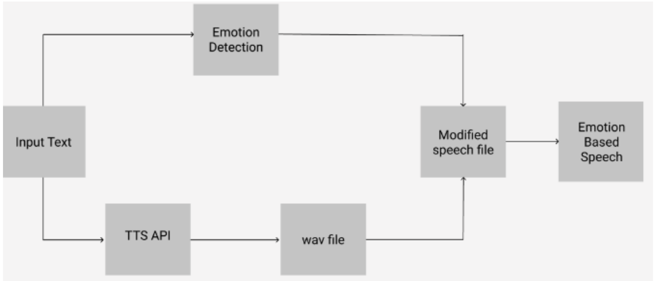


Fig 3: Proposed system for Emotion based text to speech conversion

A) 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 analyzer.

B) 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.

C) 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 analyzing 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 of sentence.

Polysemy:

Using this parameter, we will be able to understand the 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 in understanding the semantics of the sentence.

D) 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 feature extracted 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.

E) 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.

IV. CONCLUSIONS

Emotion Based Text to Speech conversion system, using neural network is useful and efficient way in extracting emotions from the text. Emotion based text to speech system will be the modified and improvised system of normal text to speech system. 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 will be able to convert text into emotion-based speech which can be used in story telling applications, Audio books and it will be more helpful for visually impaired individual for analyzing emotions from any text

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, application number: 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. Michael H. O'Malley, Berkeley Speech Technologies, Text-to-Speech Conversion Technology, Published in IEEE journal.
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. Michelann Parr, The future of text-to-speech technology: How long before it's just one more thing we do when teaching reading? , Published in International Conference on Education and Educational Psychology (ICEEPSY 2012) by SciVerse ScienceDirect, Procedia – Social and Behavioral Sciences 69 (2012) 1420-1429.
6. KashfiaSailunaz, Manmeet Dhaliwal, Jon Rokne, Reda Alhajj, Emotion detection from text and speech: a survey, Social Network Analysis and Mining, Published in Springer.
7. Ravi Kalyan Bhakat, N. P. Narendra, and KrothapalliSreenivasa Rao, Corpus Based Emotional Speech Synthesis in Hindi, Published in Springer, LNCS 8251, pp, 390-395.
8. Francisca Adoma Acheampong, Chen Wenyu, Henry Nunoo-Mensah, Text-based emotion detection: Advances, challenges, and opportunities publishedby John Wiley &Sons, Ltd.
9. Caining Yu, Qingxi Tian, Fang Cheng, and Shiqing Zhang, Speech Emotion Recognition Using Support Vector Machines, Published in Springer, CCIS 152, pp. 215-220.
10. Nihar Ranjan, Midhun Chakkaravarthy, " Evolutionary and Incremental Text Document Classifier using Deep Learning", International Journal of Grid and Distributed Computing, ISSN 2005-4262, Vol 14 No. 1, pp. 587-595
11. Nihar Ranjan, Kunal Phaltane, " A Survey on Techniques in NLP", International Journal of Computer Application, ISSN 0975- 8887, Volume 134 No 8, January 2016, pp 6-9
12. Nihar Ranjan, Rajesh S. Prasad, "Automatic text classification using BPLion-neural network and semantic word processing, Imaging Science Journal, ISSN 1368-2199, September 2017, pp. 1-15
13. Nihar Ranjan, Vishnu Panickar, " Text Document Classification using Convolution Neural Network", Journal of Emerging Technologies and Innovative Research, ISSN 2349-5162, Volume 7, Issue 6, June 2020, pp. 329-332.
14. Nihar Ranjan, Midhun Chakkaravarthy " A Brief Survey of Machine Learning Algorithm for Text Document Classification on Incremental Database" TEST: Engineering and Management, ISSN 0193- 4120, Volume 83, May -June 2020, pp. 25246- 25251
15. Nihar Ranjan, Rajesh Prasad, " LFNN: Lion Fuzzy Neural Network based evolutionary model for text classification using context and sense based features", Applied soft computing, ISSN 1568 – 4946, July 2018, pp. 994-2018

AUTHOR PROFILE

Bhushan Hemant Dhimate is currently in his final year of bachelor's degree in computer science. He is pursuing this degree from JSPM Narhe Technical Campus, Savitribai Phule Pune University, India. His area of interest is in Machine Learning and Artificial Intelligence.

Manjiri Vitthal Khopade is currently in her final year of bachelor's degree in computer science. She is pursuing this degree from JSPM Narhe Technical Campus, Savitribai Phule Pune University, India. Her area of interest is in Machine Learning and Artificial Intelligence.

Avadhoot Yogesh Dhere is currently in his final year of bachelor's degree in computer science. He is pursuing this degree from JSPM Narhe Technical Campus, Savitribai Phule Pune University, India. His area of interest is in Machine Learning and Artificial Intelligence.

Supriya Dhanaraj Dhumale is currently in her final year of bachelor's degree in computer science. She is pursuing this degree from JSPM Narhe Technical Campus, Savitribai Phule Pune University, India. Her area of interest is in Machine Learning and Artificial Intelligence.