# Therapy Bot: A Multimodal Stress/Emotion Recognition and Alleviation System

Pradeep Tiwari
Research scholar( Ph.D.)
Department of Electronics Engineering
Sardar Vallabhbhai National
Institute of Technology Surat, India
Assistant Professor
Department of Electronics and
Telecommunication Engineering
MPSTME, NMIMS University Mumbai, India

A.D. Darji
Associate Professor
Electronics Engineering
Sardar Vallabhbhai National
Institute of Technology Surat, India

## ABSTRACT

Digitalization has brought with it technological development and new opportunities for mental health care especially during the times of a pandemic where social distancing is necessary. Hence, this paper focuses on building a therapy bot application to recognize the stress/emotion of a person and provide suitable therapy. The bot is based on Multimodal Emotion Recognition (MER), which can be conceptually perceived as the superset of Speech Emotion Recognition (SER), and Textual Emotion Recognition (TER). The challenges faced in designing the therapy bot are the extraction of the discriminative features and providing the human ability of a therapist to the bot. Hence, considering these difficulties, the features are strategically selected from speech and textual modalities. The feature extracted from the speech segment is Mel-Frequency Cepstral Coefficients (MFCC), delta MFCC and acceleration MFCC while the Term Frequency-Inverse Documentary Frequency (TF-IDF) vectorization is used for the textual segment. The Support Vector Classifier (SVM) was used for calculating the confidence of the emotions from each modality. Furthermore, these confidence outputs were fused to evaluate the MER performance of the bot. The results that were calculated in real time indicated that MER performs better over SER and TER.

### General Terms

Stress Recognition, Stress Alleviation

### Keywords

Therapy Bot; Mental health; Emotion Recognition; MFCC; TF-IDF; Speech processing

## 1. INTRODUCTION

Considering the recent times and conditions, face-to-face interactions are becoming more and more strenuous and the worldwide unavailability of mental health staffs [6, 33] has urged the need of creation of a therapy bot. Emotion plays a significant role in both social interaction and clinical research. For decades, multimodal human-computer interaction researchers have worked to provide machines emotion recognition skills in order to deliver more natural, powerful, and engaging interactive experiences [30]. Human emotions may be detected by a variety of modalities, including facial expressions [15] , speech [40, 31], eye blinking [29], and posture [38], etc. The goal of MER is to create a system that can automatically recognize, comprehend, and reflect human emotions. Because MER is an interdisciplinary study topic that includes computer science, neuroscience, psychology, and cognitive science [26], thus the research in this area becomes challenging. For humans, it is easy to see the environment through a combination of sensory organs [35], but how to provide computers with similar cognitive skills is still an unanswered topic. Single-Modality deals with utilizing one feature at a time and it would not be sufficient to provide nonpareil emotion recognition results [18]. Speech-Modality deals with employing prosody-based information only while in any language, the text or sentence arrangement may also convey emotion information, thus reckoning with these points the therapy bot is co-built with two modalities, speech and textual.

The present work aimed to design and implement a therapy bot, which can identify the mental status of a person through the speech and text information provided by the user. There are primarily three challenges while developing a robust MER system for the therapy bot. First is to identify the most effective interclass distinguishable features from diverse modalities, second is to fuse the features obtained from diverse modalities and thirds is to develop a robust classifier model. The speech-based SER system aims extracting features, which show the least dependency on the speaker and the lexical content. The speech segment was used to calculate. To extract dynamic features, delta and acceleration MFCC features were obtained. The textual-based TER is first passed through removal of stop words and punctuations as a data cleaning process followed by stemming and lemmatization. The final sentence obtained is then converted to a vector using TF-IDF vectorization. Further SVM was carried out on the Indian Emotion Recognition

(IER) dataset that has been cultivated for the therapy bot, making it regional and customized. The Graphical User Interface (GUI) in the therapy bot is made using a python library named FLASK.
The remaining part of the paper is structured as follows. The section 2 is the literature review of the paper followed by the section 3 which is methodology of the proposed work. Further, the implementation; results and the discussion on the obtained results is illustrated in section 4 while the last fragment i.e. section 5 has conclusion and future scope of the paper.

## 2. LITERATURE REVIEW

The introduction of the novel Coronavirus (COVID-19) has made it difficult to move about and meet people without the stress of being infected, as it is a quick spreading virus. As the disease has progressed to pandemic level, public awareness of the COVID-19 epidemic's mental health impact has expanded substantially in recent weeks [37].

### 2.1 Need of the hour

The therapy bot application has been an attempt to provide people with emotional support during the times of distress. The Coronavirus, which affected many lives causing deaths resulting in an ugly stressful condition, economically as well as mentally. Hence, a need was found to make amends and provide emotional support during these crucial times. Moreover, recent statistical data show that the pandemic may have both immediate and long-term severe mental health consequences, particularly among healthcare personnel [37, 27]. To address this need, the therapy bot is built, which provides a medium for online therapy with the addition of text-based therapy and music therapy to alter the mood of a person. During the implementation of this paper, various works in this field were referred to relate to chatbots and feature extraction techniques.

### 2.2 Chatbots in Literature

In a paper by Takeshi Kamita et al., [17], a Structured Association Technique (SAT) was utilized to develop a counselling technique into digital material and a self-guided emotional healthcare system using a Virtual Reality (VR) head-mounted display (HMD), which resulted in a positive stress reduction evaluation. The throwback of this system was that extra and strenuous elements were used and the installation was cumbersome. Only limited locations were available and the practicability of such a system was very restricted. In a paper by Simon D'Alfonso et al [3], a case study was formulated which was based on an ongoing Horyzons site. A web application and interface were outlined using Natural language analysis and chatbot technologies. However, privacy was the major drawback of this system. Eileen Bendig et al [5] have discussed Clinical Psychotherapy and Counseling using Chatbots to Improve Mental Wellness and found that while current bots showed promise in terms of practicability, practicality, and acceptability, they aren't yet ready to be transferred to psychotherapy environments. In a paper by Gilly Dosovitsky et al., [8], it was seen that a chatbot was used to provide guidance to the people facing depressive episodes. Alaa Ali Abd-Alrazaq et al., [1] demonstrated the efficacy and safety of utilizing chatbots to enhance mental health, observing that chatbots were beneficial in alleviating depression, anxiety, tension, and acrophobia. However, the results were still conflicting regarding severity of anxiety and positive and negative effects of the chat box. In the paper by Kien Hoa Ly et al., [22], an automatic conversational chatbot was constructed for improving mental health. The results were efficient but due to minor differences in demographic characteristics between the two sets, more improvement was expected. A digital psychotherapy chatbot capable of depression analysis was created in a study by Bhuvan Sharma et al. [28]. This study, which was designed to reach 300 million individuals throughout the world, yielded substantial results while also proposing several treatments for depression. However, this study and the chatbot was only restricted to the state of depression. The therapy bot devised in this paper deals with various emotions such as angry, sad, happy, and neutral tone. The bot first recognizes the mood of the individual and accordingly suggests remedies such as a music playlist to provide music therapy.

### 2.3 Feature Extraction from Different modalities

The two modalities considered in the making of this bot are speech-based modality and textual based modality. The extensively used speech emotional features can be classified into voice quality, prosody and spectral features [24, 23]. Zhang et al., [41] deployed the technique of producing audio-visual segment features using CNN based features. Noroozi et al., [25] designed an audio-visual MER system with features like MFCC, Energy and facial landmarks from key frames extracted from video clips. In the paper by Avots et al., [4], the eNTERFACE and SAVEE database was processed employing the audio-visual emotion recognition technique. A cross-corpus evaluation was made using MFCC for the audio part and faces were extracted using the Viola-Jones face detection algorithm. In the work of Wu et al., [36] proposed an emotion recognition framework combining bi-stage fuzzy fusion and CNN. Haq et al., [14] did the audio-visual processing where energy, pitch, duration and MFCC features were used as audio features. In the work proposed by Gera et al. [10], the properties of varying the audio features like energy, pitch, MFCC and its derivatives and methods for a database were discussed. In the paper by Wang et al., [34], a MER system is employed using the pitch, intensity features along with the first 13 MFCC features from the speech samples. In the work by Venkataraman et al [32] for the RAVDESS database, Log-Mel Spectrogram, MFCC, pitch and energy were considered in the audio modality. Based on SAVEE database Kim et al., [19] proposed an Informed Segmentation and Labelling Approach (ISLA) where speech signals were used to change the dynamics of the upper and lower facial area. Audio-Visual Emotion recognition was carried out using ISLA technique. Zhalehpour et al., [39] based on eNTERFACE database employed an automatic peak frame selection from audio-visual channels. For textual modality in the paper by Ruijun Liu et al. [21] surveyed sentiment analysis from text using concepts of Natural Language Processing (NLP) to identify the emotion from the text. Peters et al. [20] proposed a method called Embeddings from Language Models (EML) to obtain word vectors from text and created the bidirectional LSTM model for classification. Akhtar et al., [2] proposed a mixed deep learning approach for text sentiment analysis. Day and Lin [7] analyzed the Google Play consumer reviews of Chinese text from the LSTM deep learning model and concluded that it LSTM shows better performance over statistical models like Naive Bayes classifier and SVM. From the different approaches discussed above, CNN is the highly popular algorithm for emotion recognition used along with MFCC of speech and NLP approaches from text. However, due to high complexity, CNN models usually require high convolutional layers in order to increase accuracy [12]. The intricacy of the network and the training time, which can expand exponentially with the addition of each layer, are the primary drawbacks of bigger network depth. Considering the accuracy and computation time, to get real time response from the chat bot, MFCC, delta MFCC and acceleration

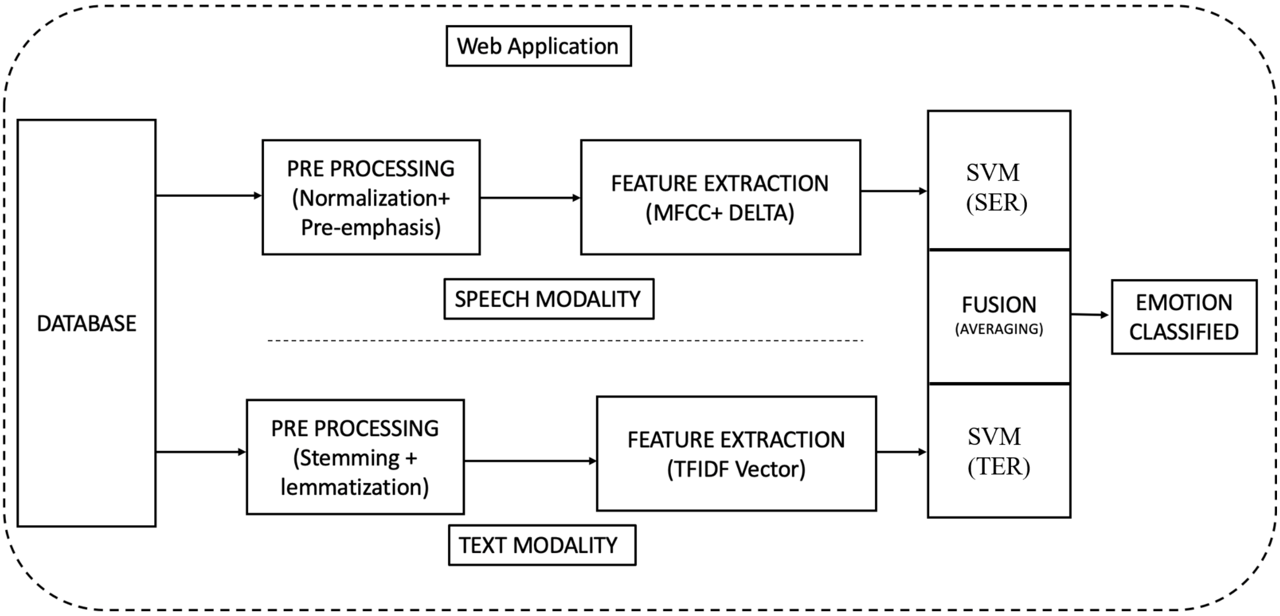


Fig. 1: Block diagram of Therapy Bot application

MFCC was proposed to be used for speech modality and TF-IDF for text modality.
This section emphasizes on making the process as limpid as possible and hence, intricate attributes of the proposed MER system setup is carefully explained. As illustrated in the Fig. 1, it is broken into steps such as database creation, SER/TER Feature Extraction, classification, fusion and emotion recognition.

## 3. METHODOLOGY

The aim of the bot is to find Multimodal Emotion Recognition (MER), which can be conceptually realized as the superset of Speech Emotion Recognition (SER), and Textual Emotion Recognition (TER). Initially the local database is created. This local database is then pre-processed after converting into text. The database results into two madalities first original recored speech and second to the text. The feature extracted from the pre-processed speech segment is Mel-Frequency Cepstral Coefficients (MFCC), delta MFCC and acceleration MFCC while the Term Frequency-Inverse Documentary Frequency (TF-IDF) vectorization is used for the pre-processed textual segment. The Support Vector Classifier (SVM) was used for calculating the confidence of the emotions from each modality which is finally used to identify the emotion.

### 3.1 Database

The local database is created and named as the Indian Emotion Database (IED) used in this paper is segregated into two modalities and four unlike emotions viz, sad, happy, angry and neutral. The two different modalities are for speech and textual modality. The Speech part of IED is formulated by 40 different people contributing their audio recordings in four different manners depicting the basic emotions of happiness, sadness, anger and neutral.

### 3.2 Speech Modality for SER

The Speech Emotion Recognition (SER) segment is divided into two major blocks. The first block deals with the pre-processing with consists of the normalization process and the pre-emphasis. The second block then deals with the feature extraction, which consists of MFCC, Delta, and Acceleration features. In SER channel, initially the base features were extracted from speech frame, and then the mean of all the frames were taken. A speech file consists of 200 frames would give $200 \times 40$ dimension MFCC feature vector. Now the mean of all frame would give 40 MFCC feature from each wav file. The MFCC features are then combined with delta and acceleration MFCC. The SER is designed in such a way that both the static as well as dynamic features are used for the correct assessment of the emotions.

*3.2.1 Pre-processing* . Pre-processing comprises of normalization and pre-emphasis. In SER, many cues like background noise can be considered which have different ranges hence; to eliminate this normalization is performed. The implementation of normalization makes it easier to spot the changes in the vocal spectrum. Pre-emphasis has been used to eliminate the noise and to balance the frequency spectrum.

*3.2.2 Feature Extraction from Speech.* MFCC is a speech feature extraction algorithm and attained by the combination of Mel filter bank and the power spectrum obtained from speech sample [12]. The block diagram for MFCC feature extraction is as shown in Fig. 2.

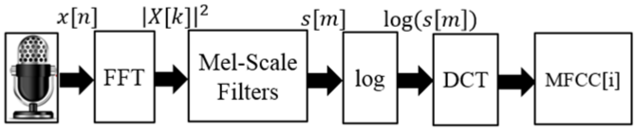


Fig. 2: MFCC Feature Extraction

Mel-scale ($S_k$) is obtained using equation (1) where 'f' is the frequency in Hz. .

$$S_k = Mel(f) = 2595 \times (log_{10}(1 + \frac{f}{700})) \quad (1)$$

To extract perception based features, 128 triangular Mel filter banks are obtained using equation (1). The square of the absolute value of the discrete Fourier transforms of the discrete-time voice input signal $x[n]$ is called the power spectrum. For the input signal $x[n]$ at discrete time instances n, short-time Fourier transform produces $X[k]$ at discrete frequency instances k for a frame of length N. Now, the Power spectrum $X[k]^2$ is applied on triangular filters (M=128) of the filter bank denoted by $H_m[k]$ to estimate Mel Scaled power spectrum $S[m]$ using equation (2.

$$S[m] = \sum_{k=0}^{N-1} X[k]^2 H_m[k], \quad 0 \leq m \leq M \quad (2)$$

MFCC of a speech sample is the logarithm of the Mel- power spectrum S[m] converted back to the cepstrum domain by using discrete cosine transform. The MFCC formula is given in equation (3).

$$MFCC[i] = \sum_{m=1}^{M} log(S[m]) \cos[i(m - \frac{1}{2})\frac{\pi}{M}] \quad i = 1, 2, ...L \quad (3)$$

The value of 'L' signifies the number of MFCC coefficients from each frame whereas 'M' indicates the speech frame length. In the present work, AER performance was analysed for MFCC with L=40 because emotion-related information can also be found in the high frequencies. Since MFCC features were calculated for consecutive speech frames thus they gave static information of that particular frame [13]. However, computing the first and second derivatives of base features, on the other hand, might provide additional information about the signal's temporal dynamics [9]. The first and second-order derivatives of MFCC coefficients called delta MFCC and acceleration MFCC respectively were used to get dynamic MFCC feature vectors.

### 3.3 TEXT Modality for TER

The motive behind the addition of Textual Emotion Recognition (TER) is to make the therapy bot more reliable and to provide a sense of privacy to the user. Text in a language also indicate positive or negative emotions. The process of TER is divided into two blocks viz, pre-processing and feature extraction.

*3.3.1 Pre-processing.* Pre-processing in TER is broken down into data cleaning process, stemming and lemmatization. Initially the text sentence is passed through removal of stop words and punctuations as a data cleaning process and then stemming. The objective behind using stemming was that for a quick emotion recognition scanning and understanding all the words is not necessary. Hence, addition of stemming reduces the inflection in words to their root forms. Further, Lemmatization reduces the modified words suitably and confirms that the root word belongs to the same language. After using both stemming and lemmatization the possibility of redundant feature extraction is reduced.

*3.3.2 Feature Extraction from Text.* The TF-IDF is calculated by the multiplication of two metrics: Term Frequency (the frequency of a word's phrase) and Inverse Document Frequency (the frequency of a word's inverse document) [16]. The amount of times a word is seen in a document is indicated by 'Term '. The frequency function is used to determine how many times a word is present in a text. The frequency can be modified further based on the document's length. The word's inverse document frequency i.e., IDF over a group of documents refers to how frequent or infrequent a word appears in the entire document set. The more frequent a term is, the closer it gets to zero. Before using the logarithm, multiply the total number of documents by the number of documents that include a word. Thus, if the word is extensively used and appears in a big number of documents, this score will be near to zero. It will be near to 1 if not. The greater the score indicates that the term inside that particular document is the more important. Mathematically, the TF-IDF score for the word denoted by 't' in a document denoted by 'd' from document set indicated by 'D' is computed using equation (4).

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D) \quad (4)$$

In equation (4), $TF(t, d) = log(1 + frequency(t, D))$ and $IDF(t, D) = log(\frac{N}{count(d \in D : t \in d)})$. This calculated TF-IDF is further fed to SVM which correspondingly to produces the confidence of each emotion class. For obtaining TF-IDF, random sentences were collected online which was used to create a database of 17436 words. After removing stop words and repeating words, 6840 unique words remained in the database. To obtain the TF-IDF vector of a sentence, first the sentence is to be pre-processed and then the unique words in the sentence is obtained. These unique words are used to calculate the TF-IIDF vector for the sentence.

*3.3.3 Support Vector Machin.* A pattern classifier called SVM library [16] has been used to classify the emotion class of the utterance. SVM has utilized linear kernel for generating feature representative models based on training vectors. The model developed is used for the recognition of emotions from the test speech. SVM in this case is used for finding confidence values of each emotion from both the modalities.

*3.3.4 Decision Level Fusion.* After obtaining confidence values from text and speech modality they are supposed to be combined together to provide the final emotion recognition result. The calculated MER accuracy is the average of confidence values obtained from both the modalities as given in equation (5).

$$MER_{acc} = \frac{(SER_{confidence} + TER_{confidence})}{2} \quad (5)$$

## 4. IMPLEMENTATION, RESULTS AND DISCUSSION

The implementation process began with the data acquisition. The IER database was curated by segregating it into two segments the Speech segment and the Textual segment. For the Speech segment, a collection of 4 emotions which are happy, sad, angry and neutral experienced by the subjects is used. It is a custom database that is created using the recordings of the acquaintances. A code was programmed to record the voices of different people including various genders. 40 people agreed from the vast majority that was approached. The different existing databases were surveyed and 25 sentences were picked that depicted 4 emotions. i.e., 4000 speech recordings were obtained. These sentences were labelled as per their emotion. The MER includes two steps: Training and Testing.

### 4.1 Training

In total, 4000 recordings were used for training the model. For speech modality, these speech samples or recordings underwent normalization and pre-emphasis as a pre-processing step. Then MFCC feature vector was evaluated from the pre-processed speech data. Further to obtain the dynamic attributes of the speech the delta and acceleration features of MFCC were also included, 40 points MFCC are extracted and then passed through delta and acceleration which correspondingly yielded 40 points respectively. Thus, a feature vector of dimension 120 was formed from each speech sample. Hence, for 4000 speech samples the training data of 4000x120 was applied to the SVM classifier model. For text modality, a Google-API was used to convert the speech recordings into text thus we obtained 4000 sentences as text samples. These text sentences were first passed through removal of stop words and punctuations as a data cleaning process and after stemming and lemmatization the final sentence obtained is then converted to a text training vector of dimension 4000 x 1 using TF-IDF vectorization. A SVM classifier model would be trained using TF-IDF features.

### 4.2 Testing

A web application is constructed to access this Therapy Bot because the web is the most widely utilized networking tool that meets the needs of all sorts of users to solve any type of problem. While creating web portal, the appearance of the site increases the importance of the development. A web application's attractiveness may easily attract a larger number of visitors, resulting in a web portal's success. Thus, the web application is built using flask which is a python library which can fulfil technological needs of a decent web portal [11]. Real time approach was carried out for testing. Initially, the user has to make an account, sign in and speak to the bot about how they feel. The application is speech independent, hence the user is free to speak any sentence. The login page of GUI of the therapy bot is shown in Fig. 3. For the user logging first time, sign up is required. After user has signed up, they can login successfully. Now next page representing MER result will appear as illustrated in Fig. 4. There are three options as shown with color buttons: 'Record' in blue for recording the speech utterance, 'Pause' in yellow for taking break while recording and 'Stop' for stooping the ongoing recording. The sampling rate of the recording is kept at 48 KHz. This recorded speech sample would be a test sample for recognizing the emotion. After recording the test speech file, it can be played using audio play sign present on the GUI to verify if the recording was successfully done. If the recording is appropriate, it is uploaded using upload button. Now in backend, for SER the speech file would be pre-processed and the feature extraction is done. This speech feature is then applied to trained SER SVM model to give confidence value of each emotion. For TER, recorded speech is converted into text using google API. This text file is pre-processed and then text feature is extracted from the converted text sequence. This text feature vector is then applied to trained TER SVM model to give confidence value of each emotion.

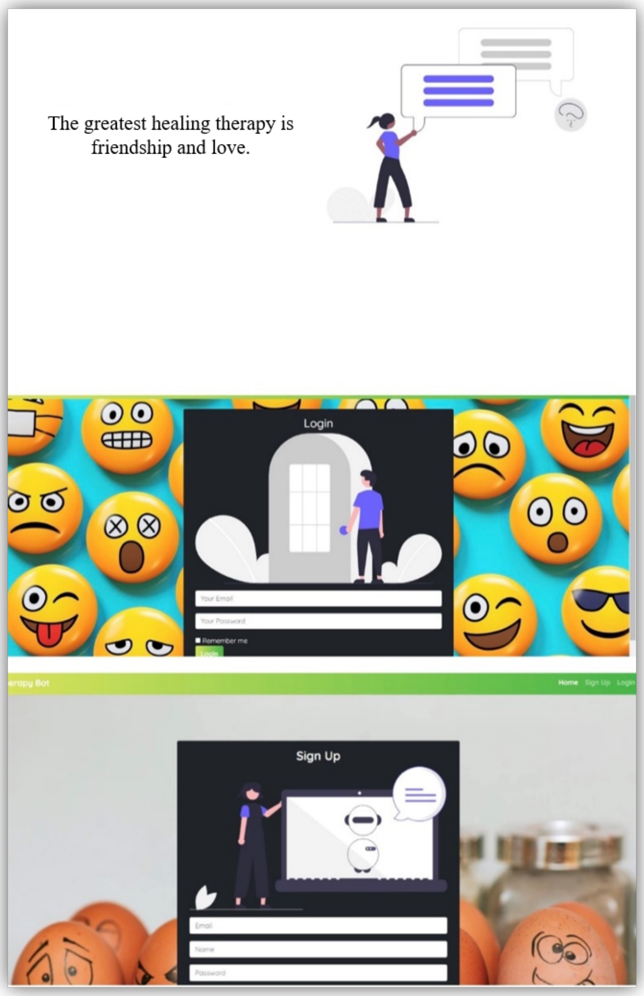


Fig. 3: Login page of GUI of the Therapy bot

Further these confidence values obtained from SER and TER is fused at decision level using averaging technique. Finally, the detected emotion would appear on the GUI. As visible in Fig. 4, 'Emotion Detected- Happy' appeared for a speech recorded with happy emotion. To identify the performance of the model, 60 test samples were recorded, Accuracy of 63.5% is found for MER while the confidence score of 0.69 for SER and 0.58 for TER was observed. Thus, the bot provides an accuracy of 63.5% for MER. The restriction on the accuracy cropped up due to the restriction in the hardware i.e., training T0 understand the performance of the BOT, the implementation summarized in the Table 1.
In addition to this, based on the emotion detected a MEME which is a modern source of entertainment and a Spotify playlist is suggested to elevate the mood of an individual which is also visible in Fig. 4 . According to the emotion recognized the motivational quote, meme and playlist is provided. Also, mantra meditation and other yoga techniques is suggested based on the emotion recognized.

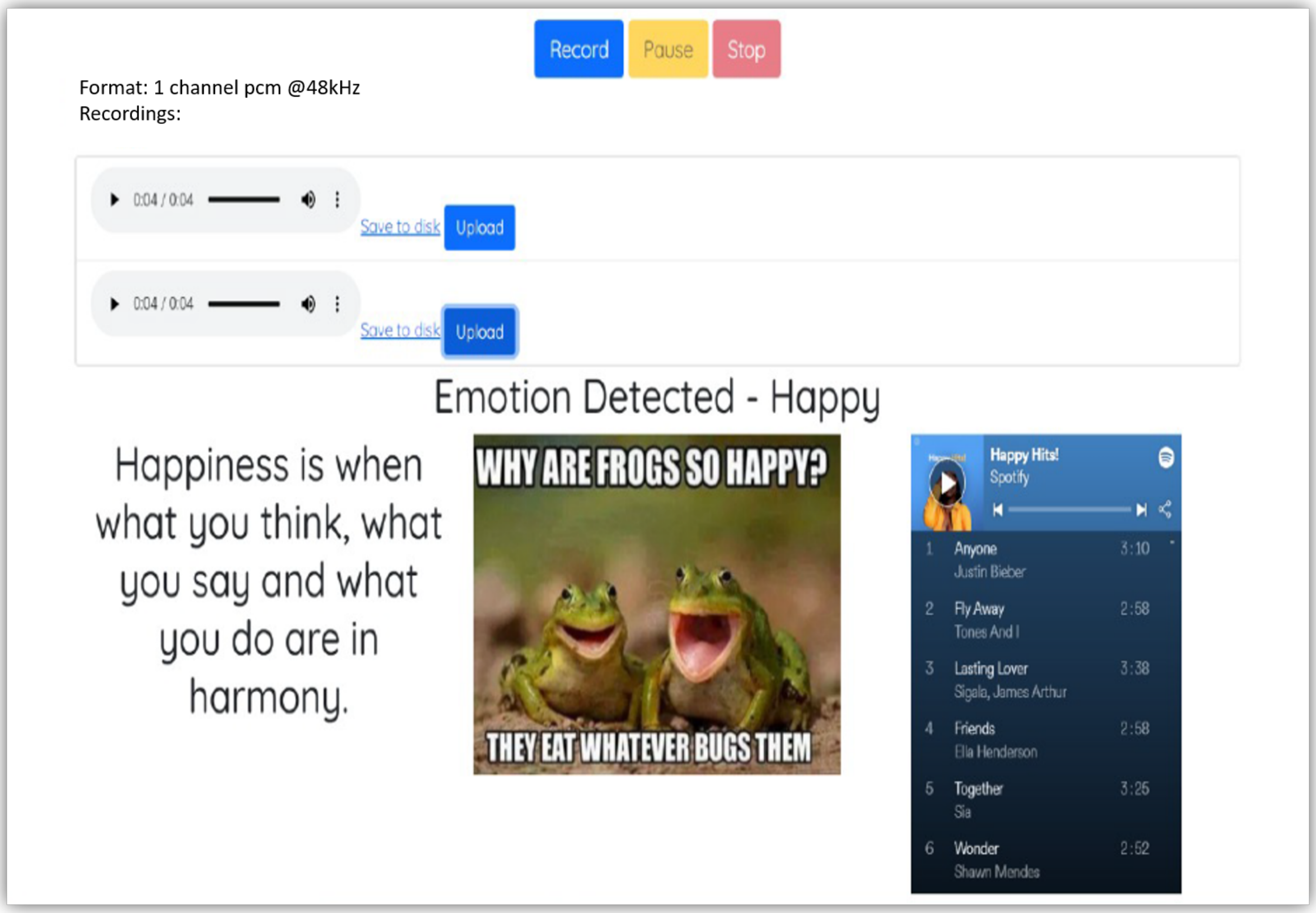


Fig. 4: Emotion detected and suggestions corresponding to the emotions

Table 1. : Performance Evaluation of Bot for Emotion Recognition

| Modality | Training Samples | Test Samples | Feature Dimension | Confidence Score | Final Accuracy (%) |
|---|---|---|---|---|---|
| Speech | 4000 | 60 | 120 | 0.69 | 63.5 |
| Text | 4000 | 60 | 1 | 0.58 | |

## 5. CONCLUSION

The mental issues have become the common disease of the age especially in this pandemic time. Due to the social distancing regulations in person therapy has become difficult. The main aim was to create a real time bot and the objective was met by using MER which considered the speech and textual aspects. Therapy bot is capable of finding mental status of a user through 4 emotions and suggesting appropriate therapy. Based on the obtained emotions appropriate therapy was suggested. The way a person displays emotions varies from user-to-user. The therapy bot designed here is a general model. In future it can be customized according to each user to make it more accurate and to provide better results. The therapy bot covers only 4 emotions, the emotions such as disgusts, grief etc. could be beneficial for people's usage. The bot exists as a web application; in future an android/iOS application could be developed to cater the crowd.

## 6. REFERENCES

[1] Alaa Ali Abd-Alrazaq, Asma Rababeh, Mohannad Alajlani, Bridgette M Bewick, and Mowafa Househ. Effectiveness and safety of using chatbots to improve mental health: systematic review and meta-analysis. *Journal of medical Internet research*, 22(7):e16021, 2020.

[2] Md Shad Akhtar, Ayush Kumar, Asif Ekbal, and Pushpak Bhattacharyya. A hybrid deep learning architecture for sentiment analysis. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 482–493, 2016.

[3] Mario Alvarez-Jimenez, Sarah Bendall, Peter Koval, Simon Rice, Daniela Cagliarini, Lee Valentine, Simon D'Alfonso, Christopher Miles, Penni Russon, David L Penn, et al. Horyzons trial: protocol for a randomised controlled trial of a moderated online social therapy to maintain treatment effects from first-episode psychosis services. *BMJ open*, 9(2):e024104, 2019.

[4] Egils Avots, Tomasz Sapiński, Maie Bachmann, and Dorota Kamińska. Audiovisual emotion recognition in wild. *Machine Vision and Applications*, 30(5):975–985, 2019.

[5] Eileen Bendig, Benjamin Erb, Lea Schulze-Thuesing, and Harald Baumeister. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health–a scoping review. *Verhaltenstherapie*, pages 1–13, 2019.

[6] Marco Colizzi, Antonio Lasalvia, and Mirella Ruggeri. Prevention and early intervention in youth mental health: is it time for a multidisciplinary and trans-diagnostic model for care? *International journal of mental health systems*, 14(1):1–14, 2020.

[7] Min-Yuh Day and Yue-Da Lin. Deep learning for sentiment analysis on google play consumer review. In *2017 IEEE international conference on information reuse and integration (IRI)*, pages 382–388. IEEE, 2017.

[8] Gilly Dosovitsky, Blanca S Pineda, Nicholas C Jacobson, Cyrus Chang, and Eduardo L Bunge. Artificial intelligence chatbot for depression: Descriptive study of usage. *JMIR Formative Research*, 4(11):e17065, 2020.

[9] H Fayek. Speech processing for machine learning: filter banks, mel-frequency cepstral coefficients (mfccs) and what's in-between, 21 april 2016, 2018.

[10] Abhishek Gera and Arnab Bhattacharya. Emotion recognition from audio and visual data using f-score based fusion. In *Proceedings of the 1st IKDD Conference on Data Sciences*, pages 1–10, 2014.

[11] Miguel Grinberg. *Flask web development: developing web applications with python*. " O'Reilly Media, Inc.", 2018.

[12] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[13] Song Guanjun, Zhang Shudong, and Wei Feigao. Research on audio and video bimodal emotion recognition fusion framework. *Computer Engineering and Applications*, pages 1–9, 2019.

[14] Sanaul Haq and Philip JB Jackson. Multimodal emotion recognition. In *Machine audition: principles, algorithms and systems*, pages 398–423. IGI Global, 2011.

[15] Haiping Huang, Zhenchao Hu, Wenming Wang, and Min Wu. Multimodal emotion recognition based on ensemble convolutional neural network. *IEEE Access*, 8:3265–3271, 2019.

[16] Thorsten Joachims. Making large-scale svm learning practical. Technical report, Technical report, 1998.

[17] Takeshi Kamita, Tatsuya Ito, Atsuko Matsumoto, Tsunetsugu Munakata, and Tomoo Inoue. A chatbot system for mental healthcare based on sat counseling method. *Mobile Information Systems*, 2019, 2019.

[18] Jonghwa Kim and Elisabeth André. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, 30(12):2067–2083, 2008.

[19] Yelin Kim and Emily Mower Provost. Isla: Temporal segmentation and labeling for audio-visual emotion recognition. *IEEE Transactions on affective computing*, 10(2):196–208, 2017.

[20] Sandeep P Kishore, Evan Blank, David J Heller, Amisha Patel, Alexander Peters, Matthew Price, Mahesh Vidula, Valentin Fuster, Oyere Onuma, Mark D Huffman, et al. Modernizing the world health organization list of essential medicines for preventing and controlling cardiovascular diseases. *Journal of the American College of Cardiology*, 71(5):564–574, 2018.

[21] Ruijun Liu, Yuqian Shi, Changjiang Ji, and Ming Jia. A survey of sentiment analysis based on transfer learning. *IEEE Access*, 7:85401–85412, 2019.

[22] Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. A fully automated conversational agent for promoting mental wellbeing: a pilot rct using mixed methods. *Internet interventions*, 10:39–46, 2017.

[23] Muharram Mansoorizadeh and Nasrollah Moghaddam Charkari. Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*, 49(2):277–297, 2010.

[24] Youssef Mroueh, Etienne Marcheret, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE, 2015.

[25] Fatemeh Noroozi, Dorota Kaminska, Tomasz Sapinski, and Gholamreza Anbarjafari. Supervised vocal-based emotion recognition using multiclass support vector machine, random forests, and adaboost. *Journal of the Audio Engineering Society*, 65(7/8):562–572, 2017.

[26] Panagiotis C Petrantonakis and Leontios J Hadjileontiadis. A novel emotion elicitation index using frontal brain asymmetry for enhanced eeg-based emotion recognition. *IEEE Transactions on information technology in biomedicine*, 15(5):737–746, 2011.

[27] Jianyu Que, Jiahui Deng Le Shi, Jiajia Liu, Li Zhang, Suying Wu, Yimiao Gong, Weizhen Huang, Kai Yuan, Wei Yan, Yankun Sun, et al. Psychological impact of the covid-19 pandemic on healthcare workers: a cross-sectional study in china. *General psychiatry*, 33(3), 2020.

[28] Bhuvan Sharma, Harshita Puri, and Deepika Rawat. Digital psychiatry-curbing depression using therapy chatbot and depression analysis. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 627–631. IEEE, 2018.

[29] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing*, 3(2):211–223, 2011.

[30] Tengfei Song, Wenming Zheng, Cheng Lu, Yuan Zong, Xilei Zhang, and Zhen Cui. Mped: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access*, 7:12177–12191, 2019.

[31] Ivona Tautkute, Tomasz Trzcinski, and Adam Bielski. I know how you feel: Emotion recognition with facial landmarks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1878–1880, 2018.

[32] Kannan Venkataramanan and Haresh Rengaraj Rajamohan. Emotion recognition from speech. *arXiv preprint arXiv:1912.10458*, 2019.

[33] Milton L Wainberg, Pamela Scorza, James M Shultz, Liat Helpman, Jennifer J Mootz, Karen A Johnson, Yuval Neria, Jean-Marie E Bradford, Maria A Oquendo, and Melissa R Arbuckle. Challenges and opportunities in global mental health: a research-to-practice perspective. *Current psychiatry reports*, 19(5):28, 2017.

[34] Yongjin Wang, Ling Guan, and Anastasios N Venetsanopoulos. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia*, 14(3):597–607, 2012.

[35] Guanghua Wu, Guangyuan Liu, and Min Hao. The analysis of emotion recognition from gsr based on pso. In *2010 International symposium on intelligence information processing and trusted computing*, pages 360–363. IEEE, 2010.

[36] Zunjing Wu and Zhigang Cao. Improved mfcc-based feature for robust speaker identification. *Tsinghua Science & Technology*, 10(2):158–161, 2005.

[37] Jiaqi Xiong, Orly Lipsitz, Flora Nasri, Leanna MW Lui, Hartej Gill, Lee Phan, David Chen-Li, Michelle Iacobucci, Roger Ho, Amna Majeed, et al. Impact of covid-19 pandemic on mental health in the general population: A systematic review. *Journal of affective disorders*, 2020.

[38] Jingjie Yan, Guanming Lu, Xiaodong Bai, Haibo Li, Ning Sun, and Ruiyu Liang. A novel supervised bimodal emotion recognition approach based on facial expression and body gesture. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 101(11):2003–2006, 2018.

[39] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. Baum-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 8(3):300–313, 2016.

[40] Shiqing Zhang, Shiliang Zhang, Tiejun Huang, and Wen Gao. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590, 2017.

[41] Zixing Zhang, Eduardo Coutinho, Jun Deng, and Björn Schuller. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):115–126, 2014.