ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse263942020.pdf
https://doi.org/10.30534/ijatcse/2020/263942020

Emotion Detection and Therapy System using Chatbot

T. Karthick[1], A.V. Amith Sai[2], P. Kavitha[3], J. Jothicharan[4], T. Kirthiga Devi[1,*]
[1]Assistant Professor, Dept of IT, , SRMIST, Chennnai, karthict@srmist.edu.in
[2,3,4,] Dept of IT, Student, SRMIST, Chennnai
* kirthigt@srmist.edu.in

ABSTRACT

Emotions affect everything that we do, beginning with our thoughts on our actions. They represent an essential part of the life of a human being. There are many ways of make a computer understand human emotions. It can also help us improve interaction between the human computers. Everyone in this era is in some form interacting with a machine. Computers, tablets, cell phones and so on have become an important part of our lives. We have come to an era in which people without these machines cannot even live. Such computers have improved our lives. We have other drawbacks, too. Some may argue that human beings were bought closer to each other by these kinds of machines and some may not agree with this statement. That depends on the user's state of mind. These types of machines have some method by which humans can communicate with the machine as well as with one another. The main goal of this project is to increase the comfort level for encounters with human computers. People now communicate more through these devices than in person, over a day. Many people are afraid to see a therapist because they fear their privacy could be compromised because they don't have enough time, it can cost more, and so on. This initiative offers a solution to problems of this kind. This project aims at making a computer understand a human being's emotional condition and give the support he / she needs at his / her own place and at his / her own time.

Key words : Emotion Detection, Chat bot, Hamming Window, Mel Scale Filter, CNN.

1. INTRODUCTION

Emotion is defined as “A disturbance of mind or an agitation”. Some may even define emotions as an “excited mental state”. In psychological terms, emotions are defined as a state of feeling that usually result in changes that influence behavior and thinking. Neurological theories states that emotional responses are a result of activities in the brain. Charles Darwin a naturalist was the one who proposed that emotions adaptive and evolving as it helps human beings and animal to evolve and reproduce. There are several theories about human emotions and all are similar and tells the same fact that emotions are an essential part of our lives. It makes us who we are. This project deals with enabling machines understand emotions so that it can be used to help those who are in need.

The basic idea of this project is to provide help for those in need of therapy at their own place and at their own time. This project is not intended to replace psychologist. They provide a major role in helping us understand how human emotions work shown in Figure 1 and how can we help others by understand their state of emotions. We chose speech emotion recognition for this project because most of the user’s won’t prefer to disclose their faces and things like emotion detection through brain waves makes the project look more like a lab test which can spook the users for which this projected is intended to help.

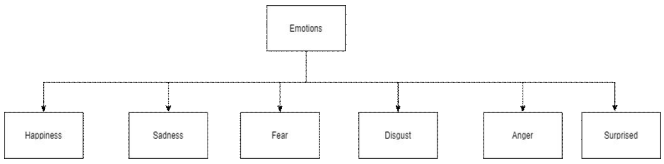


Figure 1: Basic categories of emotions

Over 75% of the affected people are affected with severe mental disorders. This becomes the reason for many suicidal problems. Around 16 to 30 years of age becomes the victim of the suicide. Bipolar disorders play a vital role in the mental disorders. To avoid this conditions and situations our model will help the patients to recover from the mental disorders and rectify the problem. Stress and Depression not only affects the affected person but also the people surrounding them. This can cause health problems and lose their lives. Not every person goes into depression but certain people are affected and couldn’t find a better solution, there are situations where they cannot discuss the matters with anyone and goes into depression. This model is not developed with the intention to replace or be an alternative to Psychiatrist. They play a very important role in the society. Certain problems require a psychiatrist doctors and no one can replace them. Due to over utilization of technology and electronic items people are obsessed and couldn’t socialize with people around them. Technology is always in a confused state whether it is boon or bane to the society.

Thus to help the patients who are affected with mental disorders, who couldn't find an alternative solution, who couldn't seek the psychiatrist due to cost issues or being an introvert this model would have a great solution for the people and society. This model takes the users speech ,detects the emotions of the user and provides a better reply through speech. Many patients will be benefitted by this model. Several research papers were studied to get an in-depth idea about the various implementation methods that can be used to develop this particular module. The development of artificial intelligence has made us think about how to make human computer interaction as smooth as possible. We studied several research papers and found that MFCC features and Convolutional neural networks woks best. Other methods like contrastive loss using Siamese networks [1] are used to make interclass classifications which are not really necessary for our project. Contrastive loss function are based on distance loss functions and works on pairs. They are used to further classify the basic types of emotions such as fear, angry, sad, happy, etc. This project does not need interclass classifications as this projects main aim is not to classify human emotions. The emotion detection module just need to identify the six basic types of emotions which is more than enough at this stage.

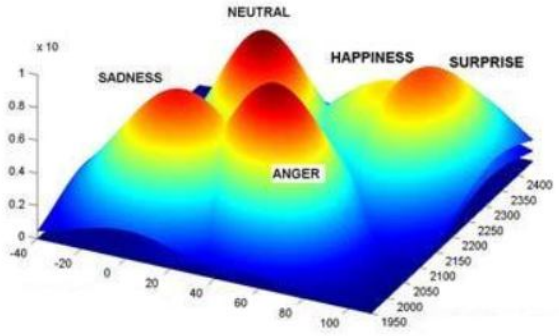


Figure 2: Graph to show the stats of the different emotions displayed by humans

Research papers like [2] concluded that collection data from a single individual rather than a group of people would improve the model's accuracy. But the downfall would be during the prediction phase, since the model is trained based on a single user's emotion rather than a group of people shown in Figure 2. This project is not intended for a single user, therefore there is a need for generalization so that our model can be used to predict the emotion of any individual. Due to technical advancements, we now have a lot of speech emotion dataset and these datasets can be combined to improve the generalization of the model thereby increasing the model's accuracy. Data augmentation techniques can be used make sure that the model does not overfit due to generalization. Emotion detection model's accuracy can also be increased by using high pass filters [3]. Mel Frequency Cepstrum Coefficient feature [5] extraction process does not remove the background noise from the data. This could affect our model's performance. High pass filters can be used to remove noise from the data. These filters only allow high frequencies to pass through and removes the low frequencies such as noise. This helped us increase our model's accuracy.

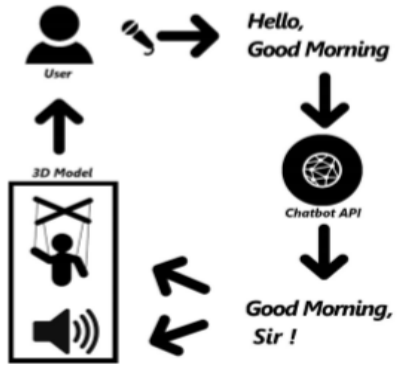
We used a sequential convolutional neural network model to make the predictions. Other methods like inception nets [4] can also be used to make the prediction. In olden days people used to stack layers on top of each other hoping that it would somehow increase the accuracy of the model. Inception net uses complex techniques to increase the model's performance. Many versions of inception nets are now available. The drawback is that inception nets are only used for images. They are not designed to handle audio data. So, for us to use inception net we have to first get the spectrogram image of the audio sample and then pass it to the inception net so that it can make the prediction. Which can become a cumbersome task.

2. WORKS RELATED TO CHATBOT MODULE

The use of Chabot's has been increasing steadily since the arrival of many messaging platforms such as WhatsApp, messenger, etc. These inventions made organizations adopt Chabot's to make communications much simpler and effective. Chabot's are conversational AI that are able to interact with humans by using natural language processing. The Chabot module in this project does not have any 3d Avatar as used in [5] but contains voice-based interaction features is shown in figure 3. 3D avatar and facial expression can be added to our model at a later stage to make the communication much more effective than it already is.

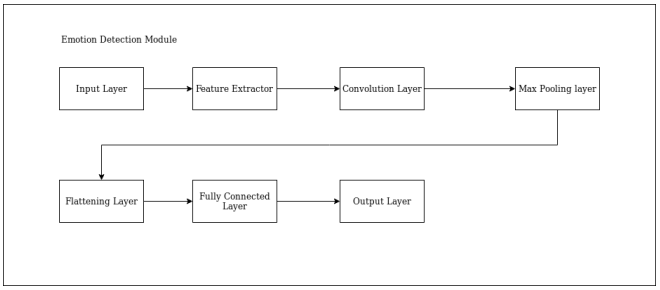
Surveys about Chabot's [8] gave us a deep insight about the different techniques that can be used to develop Chabot's. Chabot's can be classified into two types. They are; Scripted bots and AI bots. Scripted bots are bots with respond with predefined scripts from a local library. There are several drawbacks to conversational AI's such as they are not consistent and are not captivating. Chabot's which follow a particular personality [10] are more captivating and consistent than the others. This kind of conversational AI can be used to make our model more attractive and consistent than the others. Generally, Chabot's are trained using conversational datasets that are collected from various sources. In our project we have used a hybrid training method [11] in which the model is first trained using the conversational dataset and then tuned according to our need by asking the user itself what reply they would expect. This helps us train and perfect the model even more than what we already have. Our model is user centric [12] and is content drive. We allow the user to change the flow of the conversation and at the same time provide the help he or she needs based on the user's state of emotion. Since this model is content driven it enable us to make long conversations with the user on various topics[13].

Figure 3: Chatbot System Design



3. EMOTION DETECTION MODULE

We choose speech emotion detection method because this project is intended for users who are afraid of going to a therapist to get help, the flow model is shown in figure 4. These kinds of people fear to go to a therapist due to a lot of reasons such as it may cost more, lack of time, fear of being judge, etc. Speech emotion detection provides a simple way



for these kinds of users to get the help that they need.

Figure 4: Flow of Emotion Detection Module

We used Convolutional Neural Networks to identify the user's state of emotion. They user interacts with the system through his/her voice. The system then sends that audio data to the feature extractor which performs various preliminary steps to remove background noises from the data so that the important features can be extracted. The audio data contains several features such as pitch, energy, dormant frequency etc. This system extracts MFCC features from the data as it contains all the important features required to understand or identify human's state of emotion. There are several other features available such as Linear Prediction Cepstral Coefficient LPCC, Modulation Spectral Features MSF, etc. Mel Frequency Cepstral Coefficients of an audio signal are a small set of features which describes the overall shape of the envelope.

The module is shown in figure 5 is responsible for understanding or identifying a user's/human's emotion state. We used Convolutional Neural Networks to identify the user's state of emotion. They user interacts with the system through his/her voice. The system then sends that audio data to the feature extractor which performs various preliminary steps to remove background noises from the data so that the important features can be extracted. The audio data contains several features such as pitch, energy, dormant frequency etc. This system extracts MFCC features from the data as it contains all the important features required to understand or identify human's state of emotion. There are several other features available such as Linear Prediction Cepstral Coefficient LPCC, Modulation Spectral Features MSF, etc. Mel Frequency Cepstral Coefficients of an audio signal are a small set of features which describes the overall shape of the envelope. Mel Frequency Cepstral Coefficient is the most commonly used feature for speech emotion recognition as it contains all the features necessary to identify a human's state of emotion.

The main purpose of this module is to identify the user's state of emotion. The basic flow of the module is show in the figure above. The input data is passed to the feature extractor which extracts the feature from the audio data. Then the data is passed to the convolution layer which performs element wise multiplication between the feature map which is the data and the filter. Then max pooling and flattening layers are used to reduce the size of the feature map and also flatten it into a vector. At last the data is then passed to the fully connected layer which is responsible for the prediction.

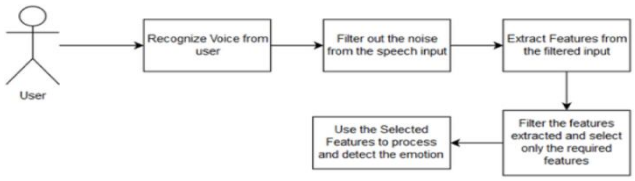


Figure 5 : Emotion Detection Module

3.1 Feature Extraction

In speech recognition, the standard method to extract features is MFCC -Mel frequency cepstral coefficients. MFCC is used commonly and by most people. The main aim of the feature extraction is to reduce the data by converting the input signal into a set of parameters while not changing the acoustic properties of the signal and provides an efficient result.

3.2 Windowing

After the first stage is completed there comes the windowing process. During the start and end of each frame there are possibilities of discontinuities, to minimize that windowing is used. Hamming window can be a preferred choice of windowing shown in Figure 6.

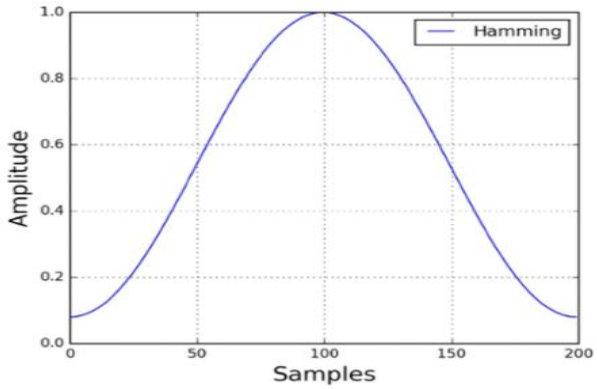


Figure 6: Hamming Window

3.3. Mel Scale Filter Bank

Mel scale filter bank is used to specify how much energy is available in particular frame and in the frequency region. The filters in the filter bank is triangular in shape and having 1 at central frequency and decrease linearly towards 0 where it goes towards the center frequency and reaches the two adjacent filters where the point is 0 is shown in Figure 7.

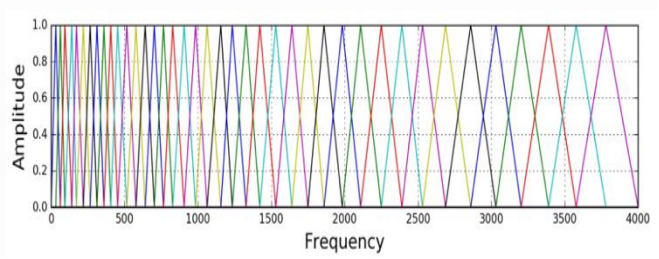


Figure 7: Mel Scale Filter Process

4. CHATBOT MODULE

We used a sequence to sequence chatbot model. When the user speaks his voice is recorded and text are generated using speech recognition software. This data is then sent to the sequence to sequence chatbot model which performs the necessary preliminary steps to convert the data into machine readable format. The feature extractor in this module is responsible for performing those preliminary steps.

The main aim of this project is to provide help to those in need as well as improve the way humans communicate with machines. The overall flow of the project is; the speech recognition software records the users voice audio and then sends the input to the Emotion Detection Module. The emotion detection module finds the users state of emotion and this information is noted. The user's text is passed to the ChatBot module which generates the response text depending upon the users state of emotion.

The chatbot module is the one responsible for providing the help that the user needs. This module first gets the users text and then based on what the user has uttered it generates the following response for it. We used a sequence to sequence chatbot model. When the user speaks his voice is recorded and text are generated using speech recognition software. This data is then sent to the sequence to sequence chatbot model which performs the necessary preliminary steps to convert the data into machine readable format. The feature extractor in this module is responsible for performing those preliminary steps.

The main objective of this model is to provide the necessary help that the user need in their own time and at their own place. Sequence to Sequence model has become the most used model for Machine translations and Dialog systems. It consists of an encoder and a decoder. The main aim of the encoder is to get only the necessary information from the data and discard all unwanted data. There are a number of hidden layers and each layer influences the next one and the final layer comprises the summary of information. The decoder depends upon these generated sequences called context.The overall flow of the project is shown in fiure 8, the speech recognition software records the users voice audio and then sends the input to the Emotion Detection Module. The emotion detection module finds the users state of emotion and this information is noted. The users text is passed to the ChatBot module which generates the response text depending upon the users state of emotion.

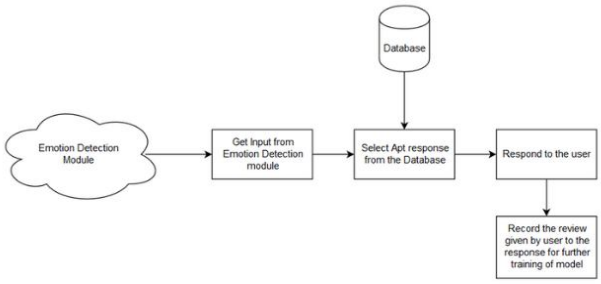


Figure 8: Flow of Chatbot Module

5. EMOTION DETECTION MODULE

This module is responsible for understanding or identifying a user's/ human's emotion state. We used Convolutional Neural Networks to identify the user's state of emotion. The user interacts with the system through his/her voice. The system then sends that audio data to the feature extractor which performs various preliminary steps to remove background noises from the data so that the important features can be extracted. The audio data contains several features such as pitch, energy, dormant frequency etc. This system extracts MFCC features from the data as it contains all the important features required to understand or identify human's state of emotion. There are several other features available such as Linear Prediction Cepstral Coefficient LPCC, Modulation Spectral Features MSF, etc. Mel Frequency Cepstral Coefficients of an audio signal are a small set of features which describes the overall shape of the envelope.

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 32, 32, 32)	896
conv2d_2 (Conv2D)	(None, 30, 30, 32)	9248
max_pooling2d_1 (MaxPooling2	(None, 15, 15, 32)	0
dropout_1 (Dropout)	(None, 15, 15, 32)	0
conv2d_3 (Conv2D)	(None, 15, 15, 64)	18496
conv2d_4 (Conv2D)	(None, 13, 13, 64)	36928
max_pooling2d_2 (MaxPooling2	(None, 6, 6, 64)	0
dropout_2 (Dropout)	(None, 6, 6, 64)	0
conv2d_5 (Conv2D)	(None, 6, 6, 64)	36928
conv2d_6 (Conv2D)	(None, 4, 4, 64)	36928
max_pooling2d_3 (MaxPooling2	(None, 2, 2, 64)	0
dropout_3 (Dropout)	(None, 2, 2, 64)	0
flatten_1 (Flatten)	(None, 256)	0
dense_1 (Dense)	(None, 512)	131584
dropout_4 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 10)	5130

Total params: 276,138
Trainable params: 276,138
Non-trainable params: 0

Figure 9: Chatbot Model Summary

Mel Frequency Cepstral Coefficient is the most commonly used feature for speech emotion recognition as it contains all the features necessary to identify a human's state of emotion.

The feature extractor first removes the background noise from the audio sample. Then the MFCC Mel Frequency Cepstrum Coefficient features are extracted from the audio data by first splitting it into frames of size 20-40ms which is more than enough to get the appropriate spectral estimate. During this period the audio signal is considered to be stationary. To make things similar to how humans process this information we then try to find out the power spectrum of each frame. The human ear also processes the information in the same way. The human ear has an organ called cochlea. Cochlea vibrates at different spots based on the audio signal. The mel filter bank then finds out how much energy exists in different frequency regions. There is still a lot of unwanted information in the audio signal shown in figure 10.

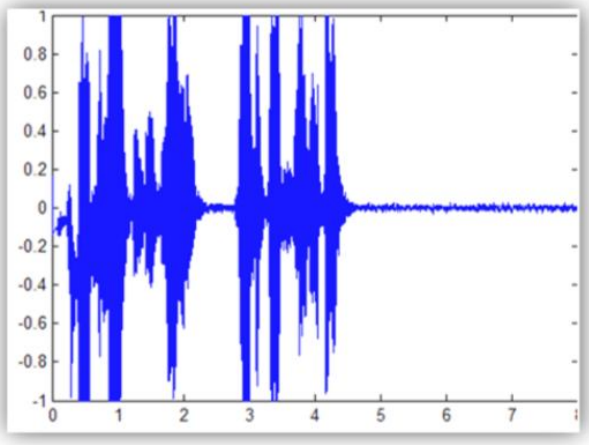


Figure 10 |: Waveform of recorded sample

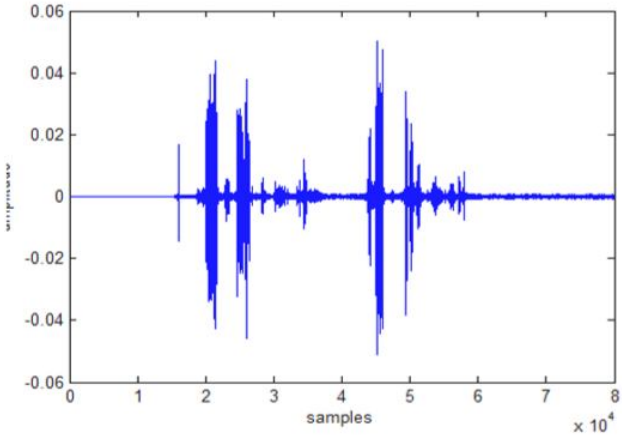


Figure 11 : After applying high pass filter , waveform of recorded sample

The MEL filter filters and takes only the needed features from the audio signal. Once we have the energies, we then take logarithm of them which is also similar to how human beings process these kinds of information. Then we compute the Discrete Fourier Transform on the signal to get the required features. The data is collected and stores in a csv file. The data is then passed to the CNN model. The Convolutional Neural Network based model is the one responsible for identifying the features that are required to identify or understand the user's state of emotion. This model comprises of 6 convolution layers followed by activation function rectified linear unit and max pooling layers shown in figure 11.

The convolution layer performs element wise multiplication between the feature map and the filters. Then an activation function called Rectified Linear Unit is used to remove the negative values from the feature map as they are not important. Replacing them with zeros would not affect our model. The max pooling layers are used to reduce the size of the feature map without ignoring any important information. The layers are stacked on top of each other and data augmentation methods are used to make sure that the model does not overfit. The main purpose of this module is to identify the user's state of emotion. The basic flow of the module is show in the figure above. The input data is passed to the feature extractor which extracts the feature from the audio data. Then the data is passed to the convoltion layer which performs element wise multiplication between the feature map which is the data and the filter. Then max pooling and flattening layers are used to reduce the size of the feature map and also flatten it into a vector. At last the data is then passed to the fully connected layer which is responsible for the prediction which is shown in figure 12.

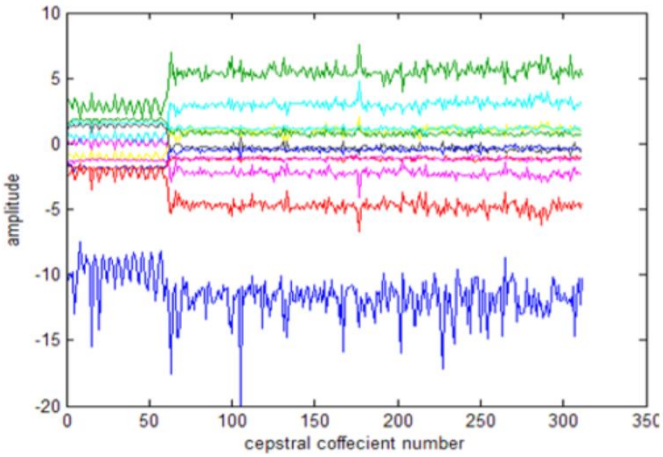


Figure 12: MFCC Feature Extraction

6.CONCLUSION

The main objective is to provide help to the user's in need at their own place and at their own time. We are also trying to improve human computer interaction. This project is not developed with the intention to replace therapist. They play an important role and in future we may need the help of a therapist to generate the response texts. Now a days people communicate through machines more. People are afraid to go to a therapist as they find the cost to be high, or they feel that they may be judged or due to lack of time. This project is designed to help those kinds of people so that they can get the

necessary help that they may need. The efficiency is good in noisy environment as well. As feature extraction plays an important role in the model the acoustic properties of the signal are undisturbed and provides an efficient result.

REFERENCES

1. Z. Lian, Y. Li, J. Tao, and J. Huang. 2018. Speech emotion recognition via Contrastive Loss under Siamese Networks. In The Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data Workshop (ASMMCMMAC'18), October 26, 2018, Seoul, Republic of Korea. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3267935.3267946
2. M Anandan, M Manikandan, T Karthick, Advanced Indoor and Outdoor Navigation System for Blind People Using Raspberry-Pi, Journal of Internet Technology, 2020, vol 21, pp:183-195
3. Davletcharova, Assel & Sugathan, Sherin & Abraham, Bibia & James, Alex. (2015). Detection and Analysis of Emotion From Speech Signals. Procedia Computer Science. 10.1016/j.procs.2015.08.032.
4. Mishra, Pawan & rawat, Arti. (2015). Emotion Recognition through Speech Using Neural Network. International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE).
5. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019.
6. M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2017, pp. 2257-2260.
7. Yenigalla, Promod & Kumar, Abhay & Tripathi, Suraj & Singh, Chirag & Kar, Sibsambhu & Vepa, Jithendra. (2018). Speech Emotion Recognition Using Spectrogram and Phoneme Embedding. 3688-3692. 10.21437/Interspeech.2018-1811.
8. Petrushin, Valery. (2000). Emotion recognition in speech signal: Experimental study, development, and application. ICSLP. 222-225.
9. A., Sameera & John, Dr. (2015). Survey on Chatbot Design Techniques in Speech Conversation Systems. International Journal of Advanced Computer Science and Applications. 6. 10.14569/IJACSA.2015.060712.
10. P. A. Angga, W. E. Fachri, A. Elevanita, Suryadi and R. D. Agushinta, "Design of chatbot with 3D avatar, voice interface, and facial expression," 2015 International Conference on Science in Information Technology (ICSITech), Yogyakarta, 2015, pp. 326-330.
11. Zhang, Saizheng & Dinan, Emily & Urbanek, Jack & Szlam, Arthur & Kiela, Douwe & Weston, Jason. (2018). Personalizing Dialogue Agents: I have a dog, do you have pets too?. 2204-2213. 10.18653/v1/P18-1205.
12. Liu, Bing & Tür, Gokhan & Hakkani-Tur, Dilek & Shah, Pararth & Heck, Larry. (2018). Dialogue Learning with Human Teaching and Feedback in End-to-End Trainable Task-Oriented Dialogue Systems. 2060-2069. 10.18653/v1/N18-1187.
13. Fang, Hao & Cheng, Hao & Sap, Maarten & Clark, Elizabeth & Holtzman, Ari & Choi, Yejin & Smith, Noah & Ostendorf, Mari. (2018). Sounding Board: A User-Centric and Content-Driven Social Chatbot. 96-100. 10.18653/v1/N18-5020.