Acoustic Intelligence In Conversational Solutions Emotion Detection From Speech

.A. Chitra[1], Ashok Raj V[2] , Mr. Vishnubalaji R K[3]

{*ctr.mca@psgtech.ac.in , ashokrajv@outlook.com, vishnubalaji3698@gmail.com}

Professor and Head,Computer Applications, PSG College of Technology, Coimbatore, India[1],PGStudent,Computer Applications, ,PSG College of Technology, Coimbatore,India[2],PG Student,Computer Applications, PSG College of Technology, Coimbatore, India[3]

Abstract:Speech is one of the natural ways of expressing ourselves as humans. It is defined as the ability to convey thoughts, ideas, or other information by means of articulating sound into meaningful words. In computer technology, speech processing paved the way for an infinite number of innovations such as Siri,Alexa, etc., which are artificial intelligence systems that are embedded in mobile devices that recognize and respond to human commands. Speech emotion can be perceived with the message of utterance and independent language of utterance. Speech emotion recognition systems are the collection of methodologies that process and classify speech signals to detect human emotions. Speech emotion recognition systems use speech signals labeled with emotions as a dataset. In this proposed system, supervised deep learning techniques such as Long Short Term Memory (LSTM) and Bidirectional LSTM models are used as classifiers to classify emotions. The emotions that are considered for classification are angry, neutral, sad, and excited. Speech emotion recognition has its applications in robotics, call centers, etc

Keywords: Emotion recognition; IEMOCAP; Recurrent Neural Networks; Long short term memory

1 Introduction

Multimedia pattern recognition is an emerging technology that can extract and analyze large amounts of multimedia information from video and audio sources. In recent years, there has been a drastic growth in the application of machine learning technology using deep learning to solve various recognition problems. Speech Emotion Recognition (SER) is an especially significant task in understanding the characteristics of speech in media. However, recognizing emotions from the speech is a very challenging problem because people express emotions in different ways, so with the help of time-domain, spectral domain, mfcc, and chroma feature, emotions are recognized. These features can be used to identify hepitc, rhythm, notes, and melody of an audio signal.

1.1 Purpose of theProject

The purpose of this project is to design and develop an efficient system that classifies human emotions from audio signals. The system uses an audio dataset that is labeled with emotions from

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2314555

the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database. These audio signals are preprocessed to get relevant data such as energy, pitch, tone, and rhythm of the audio which are used to classify the emotions. The audio features such as time domain and spectral domain features are extracted from the audio signal. The emotions that are considered for classification are angry, excited, sad, neutral.

1.2 Scope of theProject

The scope of this project is that it uses audio signals to classify the emotions of the speaker. This project can be effectively used in call centers, customer relationship management to detect the emotions of the customer/ caller in a conversation which in turn helps the callee to respond accordingly. This helps in improving and maintaining customer relationships. The deliverable of this project is a dashboard that takes audio signal as input and produces predicted emotion of that audio signal as output.

1.3 Limitations

The limitation of this project is that it uses only audio features to classify emotions. When combining the words that are spoken along with audio features might produce better results. Therefore, including the text transcriptions of the data might result in a better classification of emotions.

2 Literaturesurvey

Suraj Tripathi et al. (2019) [1] proposed a speech emotion recognition method based on speech features and speech transcriptions (text). Speech features such as Spectrogram and Mel-frequency Cepstral Coefficients (MFCC) from audio signals. The text helps capture the semantic meaning of the audio signals. The best result of recognition rate was 69.62%, achieved by combining the MFCC-Text Convolutional Neural Network (CNN) model proved to be the most accurate in recognizing emotions in IEMOCAP data.

Mehmet Berkehan et al. (2019) [2] proposed methods and techniques to extract features from audio signals such as energy, tone, etc., to classify emotions from the different databases available for speech emotion recognition. Some of the features include time-based and spectral-based features. The best result of recognition rate was 90.05 %, achieved by combining the Mel-frequency Cepstral Coefficients (MFCC) and Modulation Spectral (MS) features for the RNN model in the Spanish emotional database.

Wootaek Lim et al. (2016) [3] proposed a hybrid method with CNN and LSTM neural networks that uses a spectrogram of the audio signal as input which is allowed to flow into the Convolution Neural Network followed by LSTM models for predicting the emotions. The models resulted in 94.26% accuracy.

3 Dataset Description

IEMOCAP (Interactive Emotional Dyadic Motion Capture), collected at the University of Southern California (USC), is one of the standard datasets for emotion recognition. It consists of twelve hours of audio and video recordings performed by 10 professional actors (five women and five men) and organized in 5 sessions of dialogues between two actors of different genders, either playing a script or improvising. Each sample of the audio set is an utterance assigned with an emotion label. Labeling was made by six students of USC, three at a time for each utterance. The annotators were allowed to assign multiple labels if necessary. The final true label for each utterance was chosen by

majority vote if the emotion category with the highest vote was unique. Since the annotators reached consensus more often when labeling improvised utterances (83.1%) than scripted ones (66.9%), we concentrate only on the improvised part of the dataset. For the sake of comparison with the prior state-of-the-art approaches, we predict four of the most represented emotions: neutral, sadness, anger, and happiness, which leave us 2280 utterances in total. The IEMOCAP database is annotated by multiple annotators into categorical labels, such as anger, happiness, sadness, neutrality, as well as dimensional labels such as valence, activation, and dominance. The training dataset consists of 3838 audio files and the test dataset consists of 1098 audio files. The emotion labels fall into four categories, which are: angry, sad, excited, andneutral.

4 Proposed method

The dataset used in classifying the emotions, data preprocessing techniques, and implementation of LSTM models and Bi-directional LSTM Models. Feature Extraction The time-domain features and spectral-domain features of the audio signal are extracted from the IEMOCAP dataset. Every audio signal is split into small windows of size 200 milliseconds. For every window, the audio features are extracted. Figure 5.3 shows the windowing of the audio signal with a 50% overlap where green represents one window and red represents the other window. In order to make every input of the same size, the maximum window size is set to 100 where an audio signal having less than the maximum preferred window size will be padded with zero at the end.
The feature-length of each audio file is 33 (3 - Time domain features + 5 - spectral domain features, 13 MFCCs, 12 Chroma features). The total shape of input is three-dimensional (Number of audio files, frame per audio, feature-length). Figure 1 shows the windowing of the audio signal.

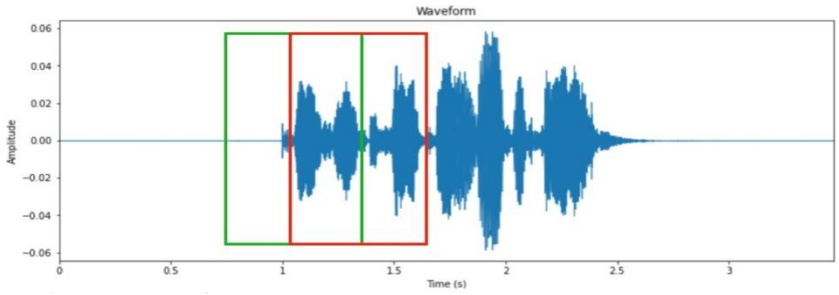

Figure 1 Windowing of audio signal

4.1 LSTMModel

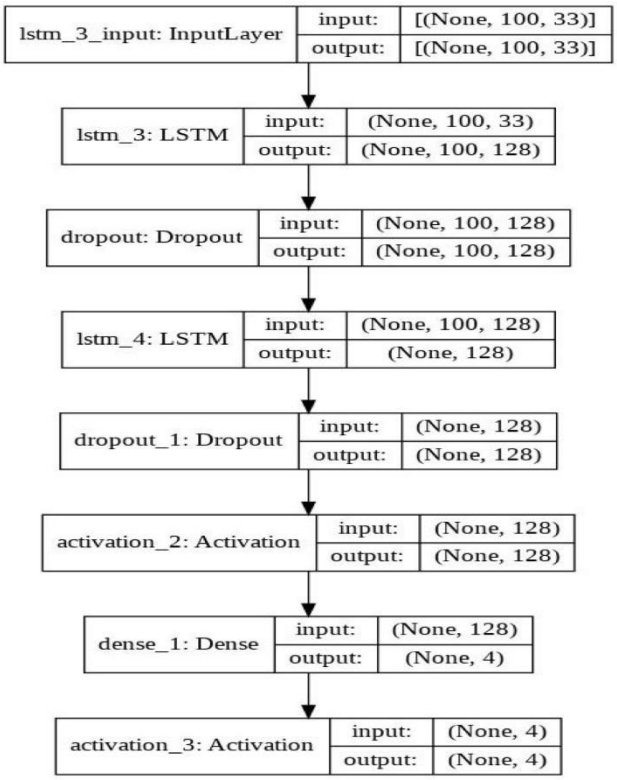

Figure 2 LSTM Network Architecture

The implementation of Long Short-Term Memory models is the classifiers of emotion that use the speech features as input and classify the emotions as angry or sad or excited or neutral. A dataset is preprocessed and stored as arrays namely X_train, y_train, X_test, y_test which represents the training dataset and testing dataset. The emotions haveto be encoded to number as it is easy for the machine to process. The LSTM model as shown in figure 2 is built with two LSTM layers stackedwith128unitseach.Themodeltakestheinputofshape [3838, 100, 33] which represents the total size of the input to train as 3838. Every layer is added with a dropout layer to reduce the over fitting of the model. The activation function is added for transforming the summed weighted input from the nodes into the activation of the node. 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 output zero. The output layer consists of 4 units which represent the number of emotions that can be classified. Adding an activation layer with soft max assigns decimal probabilities to each class in a multi-class problem. Adam optimizer is used to change the weights and learning rate of the neural network to reduce losses. Figure 4.2.1 shows the LSTM network architecture. After building the LSTM model successfully, the model is trained with the dataset of size 3838 along with the validation dataset of size 1098. The trained LSTM model is tested by a test dataset.

4.2 Bi-directional LSTMModel

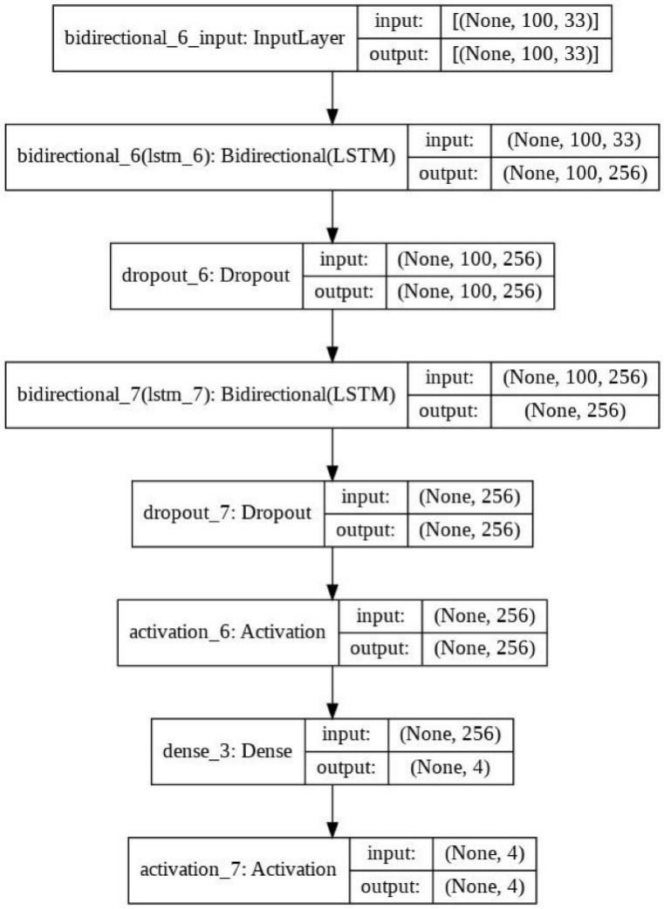

Figure 4.3.1 BiLSTM Network Architecture

The BiLSTM model is built with two layers stacked with 128 units each. The model takes the input of shape [3838, 100, 33] which represents the total size of the input to train as 3838. Every layer is added with a dropout layer to reduce the overfitting of the model. The output layer consists of 4 units which represent the number of emotions that can be classified. Similar to the architecture of the LSTM model, the BiLSTM model has been built with adam optimizer, the loss function being sparse categorical cross-entropy and the activation function of the last hidden layer is ReLu and the activation function for the output layer is softmax. Figure 4.3.1 shows the LSTM network architecture After building the BiLSTM model successfully, the model is trained with the dataset of size 3838 along with the validation dataset of size 1098. The trained BiLSTM model is tested by a test dataset. We discussed the dataset used for the SER and the preprocessing techniques involved in extracting the audio features which are used to determine the pitch, energy, tone, and rhythm that play an important role in classifying the emotions. The steps involved in the implementation of the LSTM and BiLSTM are used as emotion classifiers that were trained and validated with preprocessed data from the IEMOCAP data set.

5 Experimental results

The predictions and result analysis of the LSTM and BiLSTM model that is built to classify the emotions using the IEMOCAP dataset.

5.1 LSTMResults

Emotions	Anger	Excited	Neutral	Sad
Anger	**83**	36	49	3
Excited	48	**82**	158	11
Neutral	17	22	**313**	32
Sad	6	13	111	**115**

Figure 3 Confusion matrix of LSTM Model

The accuracy of the LSTM model is calculated as TP + TN / Total number of samples in the test dataset which resulted in 54.01% model accuracy. Figure 3 shows the Confusion matrix of the LSTM Model.

5.2 BiLSTMResults

The accuracy of the BiLSTM model resulted in 69.62% accuracy. Figure 4 shows the Confusion matrix of the LSTM Model. The long Short-Term Memory model and the Bidirectional Long Short-Term Memory model are implemented and take the preprocessed dataset as input and produce the emotion as the output. The models were trained with a dataset that produced an accuracy of about 54.01% and 69.62%. From this result analysis, BiLSTM gives more Accuracy in predicting the emotion than the LSTM.

Emotions	Anger	Excited	Neutral	Sad
Anger	**75**	48	39	8
Excited	41	**121**	117	20
Neutral	33	46	**230**	75
Sad	3	24	67	**151**

Figure 4 Confusion matrix of LSTM Model

CONCLUSION

These datasets have been preprocessed to extract speech features that determine the energy, pitch, tone, and rhythm of the audio signal. The preprocessed input of the model involves the time and spectral domain features of the audio signal. Deep learning concepts, Long Short-Term Memory model, and Bidirectional Long Short-Term Memory model are implemented and take the preprocessed dataset as input and produce the emotion as the output. The models were trained with a dataset that produced an accuracy of about 54.01% and 69.62%. Speech emotion recognition can be used in call centers to determine the emotions of the caller and callee during the conversation from which callee can respond to the caller accordingly.

REFERENCES

[1] Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag Singh, PromodYenigalla, "Deep Learning-based Emotion Recognition System Using Speech Features and Transcriptions", Proceedings of the 2019 International Conference on Computational Linguistics and Intelligent TextProcessing(CICLing).

[2] Mehmet BerkehanAkçay, Kaya OğuZ, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers", December 2019, Speech Communication Volume 116, Publication of European Association for Signal Processing (EURASIP) and International Speech Communication Association(ISCA).

[3] WootaekLim, Daeyoung Jang and TaejinLee, Audio and Acoustics Research Section, ETRI, Daejeon, Korea, "Speech Emotion Recognition using Convolutional and Recurrent Neural Network", Proceedings of the 2016 ConferenceofAsia-PacificSignalandInformationProcessingAssociationAnnualSummit (APSIPA).

[4] Yenigalla, P.; Kumar, A.; Tripathi, S.; Singh, C.; Kar, S.; Vepa, J. Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In Proceedings of the INTERSPEECH, Hyderabad, India, 2–6 September 2018.

[5] Bandela, S.R.; Kumar, T.K. Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC. In Proceedings of the 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, India, 3–5 July 2017; pp.1–5.

[6] Aldeneh, Z.; Provost, E.M. Using regional saliency for speech emotion recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp.2741–2745.