Emotion Recognition From Speech and Text using Long Short-Term Memory

Sonagiri China Venkateswarlu
Dept. of Electronics and Communication Engineering, Institute of Aeronautical Engineering, India
cvenkateswarlus@gmail.com

Siva Ramakrishna Jeevakala
Dept. of Electronics and Communication Engineering, Institute of Aeronautical Engineering, India
jsrkrishna3@gmail.com (corresponding author)

Naluguru Udaya Kumar
Dept. of Electronics and Communication Engineering, Marri Laxman Reddy Institute of Technology and Management, India
joyudaya@gmail.com

Pidugu Munaswamy
Dept. of Electronics and Communication Engineering, Institute of Aeronautical Engineering, India
sidduvamsi@gmail.com

Dhanalaxmi Pendyala
Dept. of Electronics and Communication Engineering, Institute of Aeronautical Engineering, India
nanisp197@gmail.com

Received: 2 May 2023 | Revised: 22 May 2023 | Accepted: 23 May 2023

 DOI: https://doi.org/10.48084/etasr.6004

ABSTRACT

Everyday interactions depend on more than just rational discourse; they also depend on emotional reactions. Having this information is crucial to making any kind of practical or even rational decision, as it can help to better understand one another by sharing our responses and providing recommendations on how they may feel. Several studies have recently begun to focus on emotion detection and labeling, proposing different methods for organizing feelings and detecting emotions in speech. Determining how emotions are conveyed through speech has been given major emphasis in social interactions during the last decade. However, the real efficiency of identification needs to be improved because of the severe lack of data on the primary temporal link of the speech waveform. Currently, a new approach to speech recognition is recommended, which couples structured audio information with long-term neural networks to fully take advantage of the shift in emotional content across phases. In addition to time series characteristics, structural speech features taken from the waveforms are now in charge of maintaining the underlying connection between layers of the actual speech. There are several Long-Short-Term Memory (LSTM) based algorithms for identifying emotional focus over numerous blocks. The proposed method (i) reduced overhead by optimizing the standard forgetting gate, reducing the amount of required processing time, (ii) applied an attention mechanism to both the time and feature dimension in the LSTM's final output to get task-related information, rather than using the output from the prior iteration of the standard technique, and (iii) employed a powerful strategy to locate the spatial characteristics in the final output of the LSTM to gain information, as opposed to using the findings from the prior phase of the regular method. The proposed method achieved an overall classification accuracy of 96.81%.

Keywords-emotion recognition; speech recognition; MFCC; LSTM; deep learning

I. INTRODUCTION

Applications that rely on human-machine connections have made emotional reactions a priority since they are such an integral part of human interactions. Reactions can be analyzed and interpreted scientifically in a variety of ways, including facial characteristics, bodily indicators, and language. To

achieve more natural and open interactions between humans and computers, it is necessary to regularly recognize and appropriately retain emotions represented by audio signals [1, 2]. In the last two decades, several studies have developed and refined several Machine-Learning (ML) approaches to the problem of emotion interpretation, such as Speech Emotion Recognition (SER). There is a wide variety of uses for speech recognition technology. The effectiveness of audio interfaces and advisory services is measured by frustration identification. Online businesses try to tailor their offerings to the unique needs of their customers on an emotional level [3]. Monitoring the stress levels of flight crews has been shown to reduce the number of aircraft accidents. Many studies used facial-expression recognition tools in their products to improve the user experience of people interacting with computers and increase user engagement [4]. Web-based interfaces have improved the precision of emotion detection using real-time facial identification and emotion prediction, intending to attract as many customers as possible by making changes based on their preferences [5].

The primary emphasis is on evaluating input data, audio, and video evidence to determine the subject's state of mind and provide advice. There are two main factors to consider while designing an SER system: (i) locating and extracting relevant features from an effective emotional text database, and (ii) constructing a trustworthy LSTM model using ML techniques. In actual use, the difficulty with which an SER program extracts emotional features is a major issue. Several studies have defined basic speech characteristics that convey speech content, such as power, tone, amplitude intensity, time domain power spectrum values, Mel-Frequency Cepstrum Coefficients (MFCC), and amplification features. For this reason, the vast majority of specialists favor using mixed features, which are made up of various characteristics that together convey more information. In addition, using composite features increases the likelihood of mistakes, as it makes training more difficult for most deep learning algorithms due to the large size and repetition of voice signals. Therefore, eliminating high-level speech duplication requires careful feature selection. Feature extraction or feature selection may help improve the precision and efficiency with which an ML model is trained, decrease unnecessary effort, focus on a specific area, and minimize internal requirements. Expressions of emotion in spoken language are ultimately classified. Emotion recognition is achieved by applying energy spectrum features to real-world voice data. Emotional nuances are reflected in the vocal signal in a multitude of ways. One of the trickiest parts of emotion analysis is deciding which features to use.

Several conversation recognition methods have been proposed, and a variety of Deep Neural Networks have been introduced to facilitate automated discourse recognition with low power consumption [6]. The deep spectrum characteristic performs a comprehensive analysis of a novel acoustic classification produced by running data through a neural network audio classification and building a feature map from the activation of the final fully connected layer [7]. The features metrics for the identification of level 2 and level 5 speech-based emotions were compared with the traditional acoustic representation. It can be beneficial for people with autism who can use portable devices to understand their own feelings and emotions and possibly adjust their social behavior accordingly [8]. In [9], an MFCC was used to analyze spectral characteristics in audio data to classify the 7 emotions using the Logistic Model Tree (LMT) algorithm, showing an accuracy of 70%. In [10-12], ML classifiers were used to improve the classification accuracy of fNIRS signals, decoding cognitive states, and classification of power system stability. These approaches focus on some features and neglect others, while their accuracy cannot exceed 70%, which can influence performance in recognizing emotion in speech. This study used the LSTM parameter to extract features from a speech and text dataset.

II. METHODOLOGY

This kind of technology is called voice emotion recognition, as talking is the way people communicate with computers and convey their feelings [13]. Emotional content in spoken language may be extracted by combining several approaches to signal analysis. There are many models for analyzing speech signals to predict and determine the underlying mood. This study used a recurrent neural network model with LSTM, for training and analysis of audio files with sequential data. This study aimed to design a system that can achieve good accuracy in detecting embedded emotions in speech.

A. Data Collection

Data collection is fundamental to the development of data-related activities. Overfitting is a key issue that has to be mitigated when using a large dataset and deep learning methods. There is a plethora of speech recognition datasets accessible to download from a variety of websites online, and most studies use freely accessible resources. This study used a text-based dataset [14] that contained text-based emotion details for various emotions.

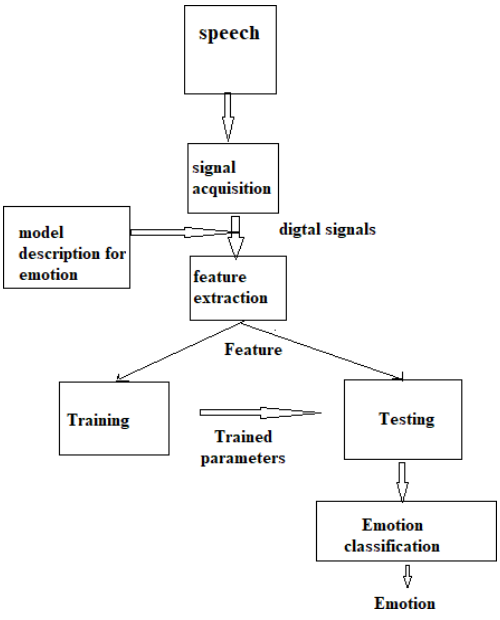

Fig. 1. The LSTM model for speech recognition.

Participants were given the option to record audio that expressed a variety of feelings, including happiness, calmness, sadness, anger, surprise, fear, and disgust. Using the Python library Keras, LSTM models were constructed sequentially. Making this model requires only a few simple procedures, as shown in Figure 1. The layers of a neural network are the basic building blocks, and the sequential class provides the structure for this model. At first, a new sequential class instance was created. Next, a stack of layers was constructed and set to communicate with each other. The memory cells in the recurrent neural layers were denoted by the symbol LSTM(). The dense layer was a fully linked layer used to generate results before the LSTM stack of layers. After a network has been constructed, it must be compiled. One of the benefits of compilation is the reduction in time it takes to complete.

Figure 2 shows a block diagram of a continuous speech recognition system based on the pattern recognition paradigm. The speech signal is analyzed as a resulting sequence of feature vectors grouped in speech unit (phonemes or triphones) patterns. Each obtained pattern is compared with reference patterns, pre-trained, and stored with class identities. These pre-trained patterns, obtained in a learning process, are the acoustical models for speech units. The outcome of the speech recognition stage is the recognized word sequence. The series of primitive layers is transformed into a highly optimized collection of matrix transformation values. Due to the way Keras is configured, the syntax of this transformation must typically be one that can be executed by a CPU. Additionally, the optimizer and error functions must be defined before the model is compiled.

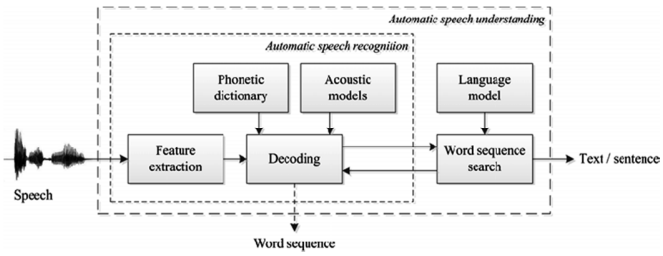

Fig. 2. The LSTM model for speech-to-text recognition.

The datasets, consisting of a matrix of input patterns *X* and an array of output patterns *y*, must be specified before training the network. In addition, the constructed network was trained with a backpropagation method and optimized using a loss function and optimization strategy supplied by the model's construction. This approach requires training over a certain number of epochs, as shown in Figure 3. Once training was complete, the network was tested using non-overlapping training data. The accuracy of the prediction was used as a statistic because it helps in forecasting the performance of a constructed model. Data predictions were made using the predict() command after the model's efficacy was assessed.

B. Model Training

The model was trained using the fit() function with the following parameters: train *X*, target *X*, validation data, and a number of epochs. The test set included in the dataset was partitioned into *X*_test and *y*_test for validation purposes. The model iterated the data a certain number of times, as defined by the epochs parameter. Up to a point, the more the epochs, the better the model will become. From that point on, the model will no longer progress with each epoch. The model was trained for 31 epochs.

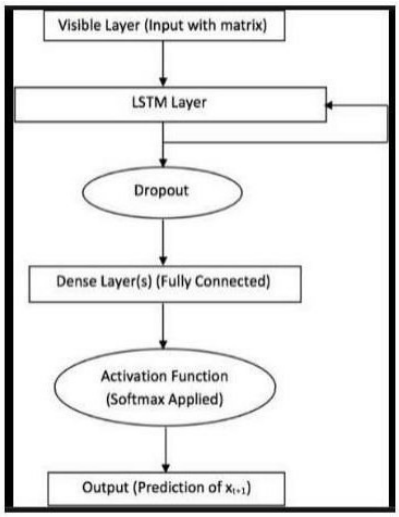

Fig. 3. Model training for LSTM.

Keras defines neural networks as a series of layers. The sequential class serves as a framework for these levels of layers. The first step is to build a sequential instance of a class, followed by building a stack of layers and arranging them in the sequence in which they should be interlinked. The LSTM recurrent neural layers are made up of memory cells which are known as LSTM () cells. The dense layer is a fully connected layer that frequently precedes the LSTM stack of layers and is used to produce a result. \

III. RESULTS

The results of the proposed LSTM ML model. were compared with those of [12], which used SVM and auto-encoder and acquired a 74.07% accuracy rate. The network was evaluated once it was trained with a separate set of training data. After evaluating the performance of the model, it was used to make predictions. This action was carried out using the predict() function, and the output format was the same as specified by the output layer.

The results of the model show its effectiveness in achieving good results compared to other emotion recognition studies, as it achieved an accuracy of 96.82%. A web app was created for the real-time application of the proposed speech emotion recognition model. A user can enter the text in the prompt displayed to predict and display the output emotion. The user can select a speech input file from the required location and press transcribe. Then, the speech will be converted to text and the emotion is displayed as output. The user can also select to use his microphone to record the voice input, convert it into text, and display the emotion as output. Table I shows the possible texts and emotions considered in this study.

There are three types of inputs; the first is by entering the text message and then displaying the predicted emotion for the given text. The second is by uploading a voice file and then displaying the predicted emotion output, and the third is by allowing the mic to record the voice, analyzing it as a text, and then displaying the predicted emotion output. Every text, speech, and voice has an emotion. Table II shows the overall accuracy of the emotional classification model.

TABLE I. TEXT AND ITS CONSIDERED EMOTIONS

Text	Emotion
I didn't feel humiliated	Sadness
I am feeling grouchy	Anger
I am grabbing a minute to post I feel greedy wrong	Anger
I am ever feeling nostalgic about the fireplace	Love

TABLE II. LSTM OVERALL ACCURACY

Testing	Speech	Text	Voice
Accuracy	96.81%	96.81%	96.81%

IV. CONCLUSION AND FUTURE SCOPE

The primary goal of this study was to employ recurrent neural networks with LSTM to determine a person's emotional state. This study used a dataset of text files that depict a wide range of emotions, such as happiness, sadness, fear, contempt, surprise, and apathy. As most ML models take numeric values as input, the data was transformed into arrays to use them for feature extraction. The Libros package was used to extract the file and MFCC features were used in this model. The collected values were then fed into an LSTM model, which used these characteristics to provide an overall anticipated emotion. The model achieved an overall accuracy of 96.81% for emotion recognition using speech, text, and voice data. This model was used to enhance a real-time speaker identification system using a digital signal processor. The volume of the system may be adjusted according to the environment. This system can be used for aid to disabled persons. Applications and websites can use this approach to gauge user sentiment and decide how to best tailor their offerings to their audience. Additionally, this approach can be used in voice-based virtual assistants, chatboxes, and call centers for handling customer complaints.

REFERENCES

[1] Mustaqeem and S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition," *Sensors*, vol. 20, no. 1, Jan. 2020, Art. no. 183, https://doi.org/10.3390/s20010183.

[2] A. M. Badshah *et al.*, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 5571–5589, Mar. 2019, https://doi.org/10.1007/s11042-017-5292-7.

[3] 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," *IEEE Access*, vol. 7, pp. 117327–117345, 2019, https://doi.org/10.1109/ACCESS.2019.2936124.

[4] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition." arXiv, Apr. 10, 2015, https://doi.org/10.48550/arXiv.1409.1556.

[5] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. de Albuquerque, "Cloud-Assisted Multiview Video Summarization Using CNN and Bidirectional LSTM," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 77–86, Jan. 2020, https://doi.org/10.1109/TII.2019.2929228.

[6] B. Liu, H. Qin, Y. Gong, W. Ge, M. Xia, and L. Shi, "EERA-ASR: An Energy-Efficient Reconfigurable Architecture for Automatic Speech Recognition With Hybrid DNN and Approximate Computing," *IEEE Access*, vol. 6, pp. 52227–52237, 2018, https://doi.org/10.1109/ACCESS.2018.2870273.

[7] J. Huang, B. Chen, B. Yao, and W. He, "ECG Arrhythmia Classification Using STFT-Based Spectrogram and Convolutional Neural Network," *IEEE Access*, vol. 7, pp. 92871–92880, 2019, https://doi.org/10.1109/ACCESS.2019.2928017.

[8] E. Sucksmith, C. Allison, S. Baron-Cohen, B. Chakrabarti, and R. A. Hoekstra, "Empathy and emotion recognition in people with autism, first-degree relatives, and controls," *Neuropsychologia*, vol. 51, no. 1, pp. 98–105, Jan. 2013, https://doi.org/10.1016/j.neuropsychologia.2012.11.013.

[9] A. A. A. Zamil, S. Hasan, S. MD. Jannatul Baki, J. MD. Adam, and I. Zaman, "Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames," in *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)*, Dhaka, Bangladesh, Jan. 2019, pp. 281–285, https://doi.org/10.1109/ICREST.2019.8644168.

[10] M. M. H. Milu, M. A. Rahman, M. A. Rashid, A. Kuwana, and H. Kobayashi, "Improvement of Classification Accuracy of Four-Class Voluntary-Imagery fNIRS Signals using Convolutional Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 13, no. 2, pp. 10425–10431, Apr. 2023, https://doi.org/10.48084/etasr.5703.

[11] S. R. Jeevakala and H. Ramasangu, "Classification of Cognitive States using Task-Specific Connectivity Features," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10675–10679, Jun. 2023, https://doi.org/10.48084/etasr.5836.

[12] N. A. Nguyen, T. N. Le, and H. M. V. Nguyen, "Multi-Goal Feature Selection Function in Binary Particle Swarm Optimization for Power System Stability Classification," *Engineering, Technology & Applied Science Research*, vol. 13, no. 2, pp. 10535–10540, Apr. 2023, https://doi.org/10.48084/etasr.5799.

[13] S. R. Bandela and T. K. Kumar, "Emotion Recognition of Stressed Speech Using Teager Energy and Linear Prediction Features," in *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)*, Mumbai, India, Jul. 2018, pp. 422–425, https://doi.org/10.1109/ICALT.2018.00107.

[14] "Emotion Detection from Text." https://www.kaggle.com/datasets/pashupatigupta/emotion-detection-from-text.