Vocal Analysis and Sentiment Discernment using AI

Praveen Singh*, Preeti Nagrath

Bharati Vidyapeeth's College of Engineering, India

Emails: praveensingh3129@gmail.com; preeti.nagrath@bharatividyapeeth.edup

***Correspondence: praveensingh3129@gmail.com**

Abstract

One of the major factors for personal development and growth is understanding human emotions, and therefore it plays an important role in imitating human intelligence. Vocal and Sentiment analysis is the major focus points for advancement in Artificial Intelligence (AI). Sentiment analysis provides major help to data analysts of big enterprises to measure public opinion, conduct market research, understand customers' experiences, and view brand and product reputation. Emotion recognition provides an opportunity to grasp the general people's sentiments about social events, marketing strategies, political views, and product liking. In this paper, we have used various AI models on a variety of audio datasets to recognize and analyze the sentiments of the speaker. Our dataset includes some audio songs sung by some singers and some audio clips of a few actors. We trained CNN and LSTM models to analyze our dataset and predict their accuracy. The ever-growing need for sentiment analysis coincides greatly with the extension of social media such as forum discussions, and social networks like Facebook, Twitter, Instagram, and many other similar platforms.

Keywords: Vocal Analysis; Sentiment Discernment; Artificial Intelligence; Personal development

1. Introduction

The process of identifying and diagnosing human emotion is known as emotion recognition. The accuracy of people varies at recognizing the emotion of others. Helping people with emotion recognition and the use of technology is a relatively growing research area. The necessity of an interface between computers and human is a growing demand due to the ever-thickening use and demand of computers. Casually, Emotion recognition [1] [2]with the help of speech can also be described as the detection of emotions by feature extraction of voice transferred by humans. As computers improve in predicting and analyzing the emotional state of the human speaker, interaction between humans and computers can get more personalized and interactive, helping them in differentiating the contextual meaning of the same word.

Various classifiers are used for emotion detection among which the most highly used models are the Support Vector Model (SVM), k-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), etc [3][4][5]. Various uses of emotion recognition involve: - psychiatric, mental conditioning, intelligent kid toys, lie detection, and many more[6].

Doi : https://doi.org/10.54216/FPA.070204

Received: January 19, 2022 Accepted: April 15, 2022

Emotion in the speech of a speaker can be considered as a communication channel featuring various parts involving the expression of the emotion of the speaker or the way it is delivered by the speaker which conveys the emotion felt by the speaker[7]. Diverse methods to analyze vocal behavior like emotion, moods, and stress, concentrating on the verbal and non-verbal aspects of speech is known as Speech Emotion Analysis[8].

Now the most common assumption is that there exists a voice parameter that shows the affective state a person is experiencing. This statement appears appropriate given that the most effective states include physiological reactions, which help in modifying various aspects of the process of voice production. Speech emotion analysis is considered complicated as the vocal expression is evolutionarily old, continuously varying, and evolving[9]. We tried to study and understand the recognition of emotion through human speech and voice analysis using various models like CNN and LSTM.

2. Literature Survey

Mucahit Buyukyilmaz [1] 2016 proposed a gender recognition solution using an MLP deep learning model which achieved an accuracy of 96.74%, this was performed on 3,168 recorded samples of male and female voices which were produced using acoustic analysis. Speech signals have seen a wide amount of research related to emotions, Ramdinmawii [10] used signal processing methods, zero-frequency filtering, and short-time energy to analyze four emotions from speech signals from the IIT kGP Telugu and german dataset. Nicholson [11] also developed an emotion recognition one-class-in-one neural network which achieved 50% accuracy for eight emotions, while Al-Talabani [12] used SVM and LDC on a Kurdish and berlin dataset which showed SVM was more accurate. Huang [5] proposed that DNN could be applied for emotion recognition, their result showed how DBN of DNN in speech emotion recognition has a huge advantage. Leh Luoh [13] developed a system that could detect 5 emotions but without training it in many steps on a short utterance word with an average accuracy of 55%. Soltani [14] developed speech emotion detection using neural networks on Berlin data with word utterances achieving an accuracy of 77%.

A Modular neural network on speech utterances developed by Bhatti [15] achieved the best overall classification of 83.3%. A convLSTN_RNN model proposed by Kurpukdee [16] is an advanced method in its class that extracts phoneme-based features from raw input signals to recognize four emotions of the IEMOCAP dataset. Attention LSTM by Yu [17] showed that LSTM provides an accuracy of 73% on the IEMOCAP dataset. Berlin emotional dataset with seven categorical emotions is widely used for better acoustic feature extraction on it and is used by Lech [18] who demonstrated a unique method of audio classification by generating time-based images of audio files and using them to train a CNN model with frequency scaling. This unique method displays accuracy of 82%. On the other hand, this dataset is used by Kerkeni [19] who used MLR and SVm to obtain an accuracy of 83%. Automatic effect sensing [20] uses both audio and visual information on the RECOLA database. Leila Kerkeni [21] developed an RNN classifier and compared its performance with the MLR and SVM by them before [22]on the berlin database with a better recognition rate of 90.5%. Classifiers like HMM and GMM [16 show good results on speaker normalization for emotion detection which is a unique technique that improves the performance of speaker-independent emotion classification. Like the berlin database, the IEMOCAP dataset released in 2008 by USC is also widely used for research purposes, Gaurav Sahu David R [23]used models like Random forests, SVM, LSTM, and logistic regression to prove a lighter ML-based model trained over few handcrafted features can achieve performance comparable to the current deep learning-based state of the art method for emotion recognition.

3. Methodology

3.1. Mel Frequency Cepstrum Coefficients (MFCC)

Feature extraction is the beginning step of any automated speech recognition system which is to identify the elements of the audio signal that are responsible for identifying the linguistic content and ignoring all the other elements which contain information like emotion, background noise, etc[24]. The major factor of speech is that human sound gets automatically filtered by the shape of their vocal tract which includes their tongue, teeth, and other features. The sound coming out depend on this shape. If the shape can be determined precisely, this should

Doi : https://doi.org/10.54216/FPA.070204

Received: January 19, 2022 Accepted: April 15, 2022

give us an accurate depiction of the phoneme being produced[25]. The job of MFCCs is to accurately represent the envelope, which is the shape of the vocal tract which manifests itself in an envelope in a short time power spectrum.

3.2. Convolutional Neural Network (CNN)

CNN is a popular technique used in deep learning and AI. The most basic component of a CNN is convolutional layers. Its basic purpose is to take an input, transform it in some way and then output the transform input to the next layer. Different weights are assigned to the neurons from which each layer is made. They can detect patterns and images. A 3D network is produced by these layers, as they are aligned on top of each other[26].

In CNN, the neurons are not connected to all the neurons of the adjacent layer, but rather to a fixed number of neurons only. We have the same value for weights for all the neurons. Therefore, across the whole layer same pattern is filtered. This layer is called the convolution layer. It is followed by two sets of layers called the activation layer and pooling layer. These assist in building a simple pattern based on what the convolutional layer discovers. The activation layer is used for proper training and the latter for dimensionality reduction.[27] These layers are beneficial in finding complex patterns, but they do not have the capability to understand these patterns. A well-linked layer is used to understand the pattern completely. It allows us to classify and identify data samples.

3.3. Long Short-Term Memory (LSTM)

LSTM is a genre of Recurrent Neural Network (RNN). In RNN, the output obtained from the last step is used as input in the current step[28]. The problem of long-term dependencies is tackled by the LSTM in which the RNN is not able to predict the word stored in the long-term memory but can provide more precise predictions from the current information. As the gap increases, RNN performance diminishes. LSTM can retain the information for a longer period[29]. The information is used for processing, predicting, and classifying on the criteria of time series data. In simple words, an LSTM workflow is like RNN. It processes and passes the data simultaneously providing information as it propagates forward. LSTM's cell operation is where the difference lies. This operation gives LSTM the ability to keep or forget the information.

4. Workflow

Extracting human emotions very much variable on the kind of dataset and the language[30]. For our work, we have used a combination of two RAVDESS datasets, a speech audio dataset, and a song audio dataset thus having a lot of variations in frequency and actor. The prime step before moving on to the important part is structuring the dataset. We collected the two datasets and combined them into a single array of data and also created corresponding label array data. Here, the label data consisted of 5 characteristics, particularly sad, calm, happy, angry, and fearful. We reduced our labels to 5 from 7 to reduce complexity in our learning rate and therefore give better accuracy. After data collection, the next step is to determine how to extract features. Identification of unique emotions from audio signals is the tricky part as audio signals can't be understood by the models. These signals need to be converted into a format that could be used in the models and also exhibit unique characteristics which enable our learning model to differentiate between them. Thus, Mel-Frequency Cepstral Coefficients Characteristics are created for each audio signal and thereby converting our input audio signal array to an MFCC characteristic matrix which now could be used in our models.

Doi : https://doi.org/10.54216/FPA.070204

Received: January 19, 2022 Accepted: April 15, 2022

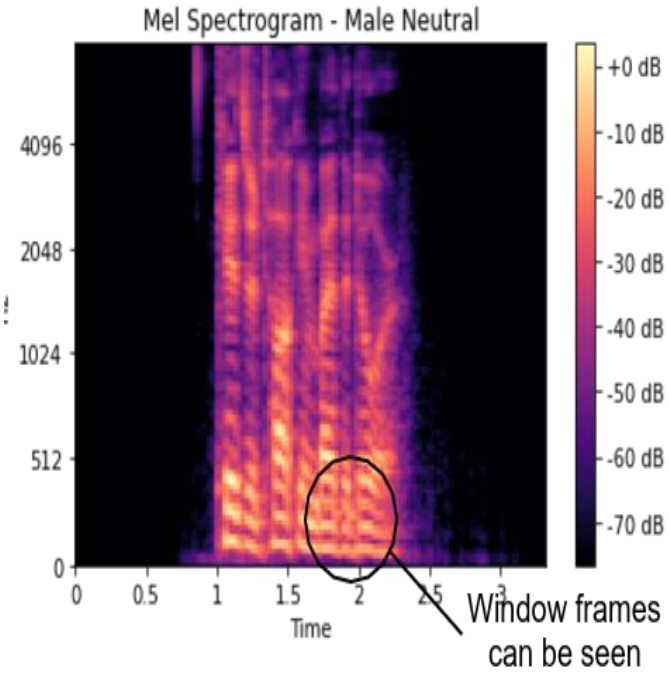

Figure 1: MFCC spectrogram

The Mel-frequency cepstral coefficient (MFCC) is used here because it has 39 features, and the feature count is small enough to force us to learn the information of the audio.

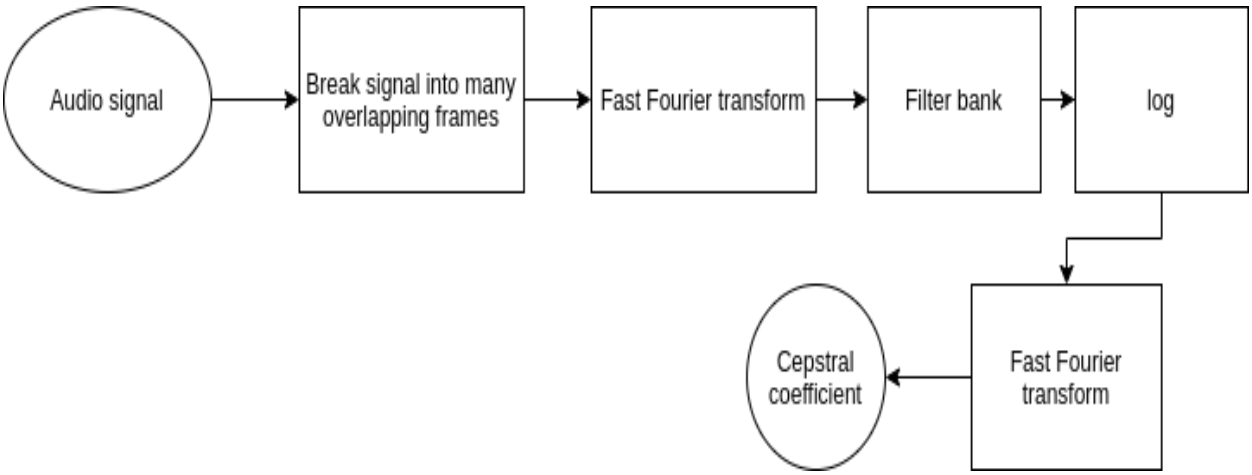

Figure 2: MFCC working

To extract the MFCC characteristics from the signal, we used the librosa library which has prebuilt functions to implement it.

Doi : https://doi.org/10.54216/FPA.070204

Received: January 19, 2022 Accepted: April 15, 2022

Since our overall aim is the classification of an audio signal into different emotions, thus we can use classification algorithms mainly in the deep learning field since we need to select the features from the MFCC characteristics. CNN and LSTM have been used for achieving our results. LSTM RNNs are selected as it is suitable for time series data. Also, simple RNNs suffer from vanishing gradient problems which increase with the length of the training sequences, and thus LSTM was preferred.

CNN is a common model for classification purposes and shows tremendous results and hence was preferred. We structured the CNN as shown in figure 1.3 for the MFCC input matrix of our audio signals.

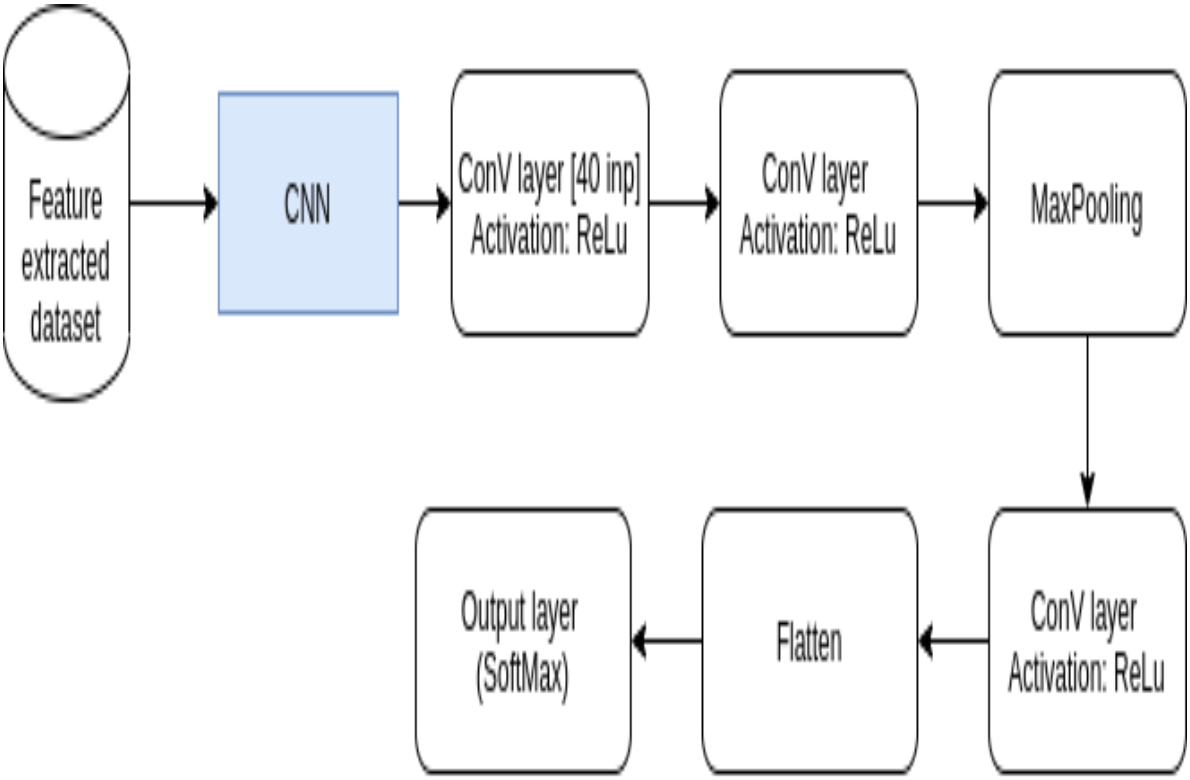

Figure 3: CNN model

Thus, after combining all the steps discussed above then our model structure comes out to be as shown in figure 1.4. With feature extraction as the most complex task in the case of audio signals and then creating a feature selector and classifier model using Deep learning techniques, to generate the most optimized audio emotion classification model.

Doi : https://doi.org/10.54216/FPA.070204

Received: January 19, 2022 Accepted: April 15, 2022

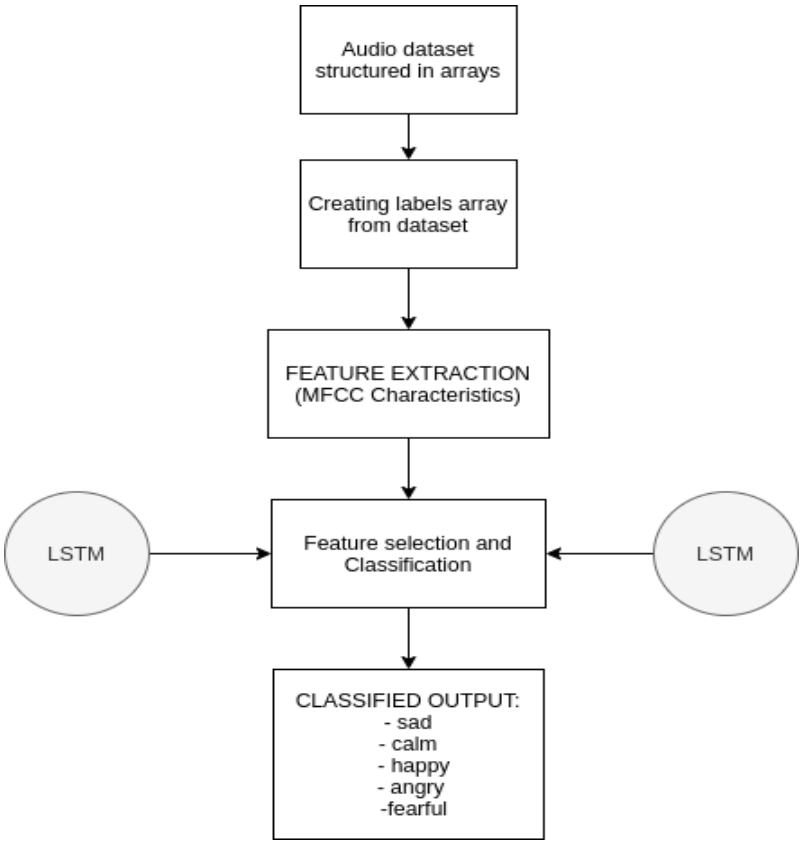

Figure 4: Overall workflow

5. Experimental Results and Analysis

By implementing and comparing various deep learning and machine learning algorithms for vocal analysis and sentiment recognition on the Ravdess dataset which includes audio of speech and audio of songs, we have concluded that in deep learning models, the best results are obtained using CNN, followed by LSTM.

The training accuracy of 80.35% and validation accuracy of 82.72% is achieved with CNN using 100 epochs with a training loss of 0.62% and validation loss of 0.75%. LSTM on the other hand has a training accuracy of 80.72% and validation accuracy of 82.13% is achieved using 70 epochs with a training loss of 0.56% and validation loss of 0.51%.

Doi : https://doi.org/10.54216/FPA.070204

Received: January 19, 2022 Accepted: April 15, 2022

Table 1: Analysed data after training the model

Model Name	**TRAINING %**		**VALIDATION %**	
	Accuracy	**Loss**	**Accuracy**	**Loss**
CNN	80.35	0.56	82.72	0.51
LSTM	80.72	0.62	80.13	0.75

CNN model provided us with an accuracy of 82.72% on the validation dataset with 100 epochs. Even though LSTM provided us with higher accuracy on the training dataset at 80.72%, its result wasn't better compared to CNN using the validation dataset. The accuracy obtained in the CNN model is adequate compared to other models and without augmentation.

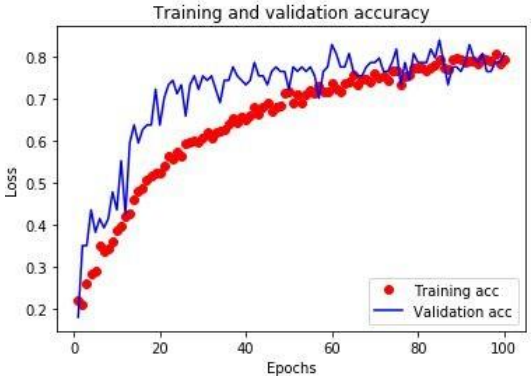

Figure 5: CNN training and validation acc

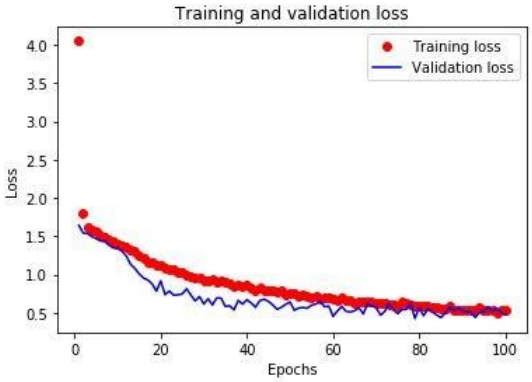

Figure 6: CNN training and validation loss

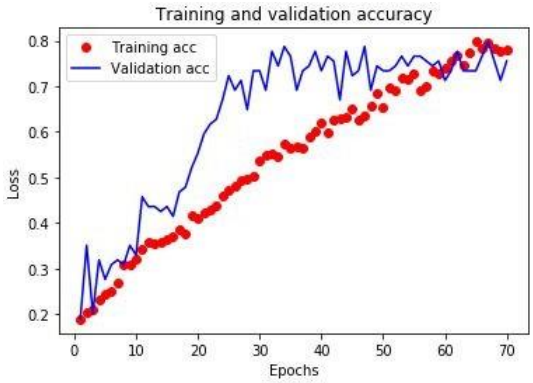

Figure 7: LSTM training and validation acc

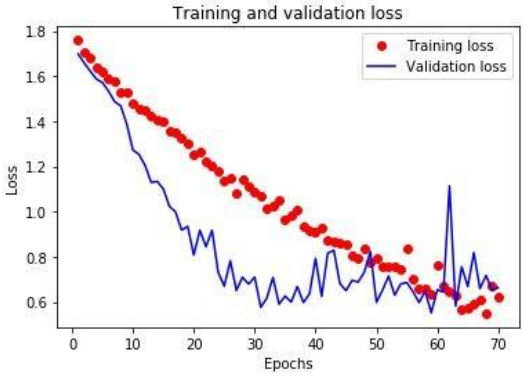

Figure 8: LSTM training and validation loss

Doi : https://doi.org/10.54216/FPA.070204

Received: January 19, 2022 Accepted: April 15, 2022

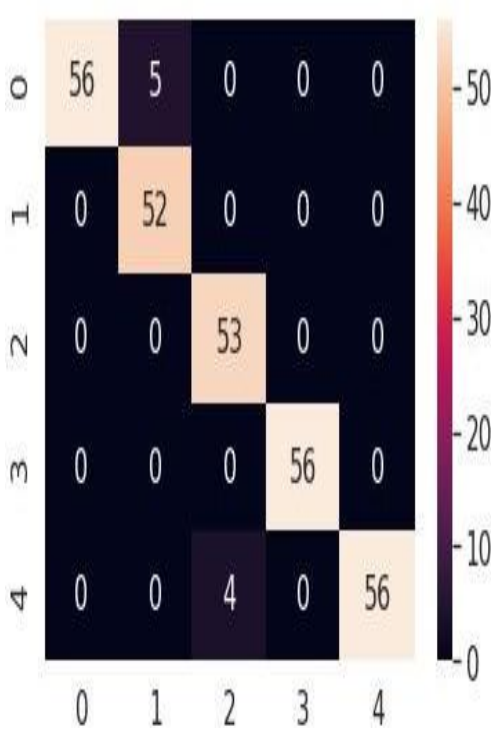

Figure 9: CNN Confusion matrix

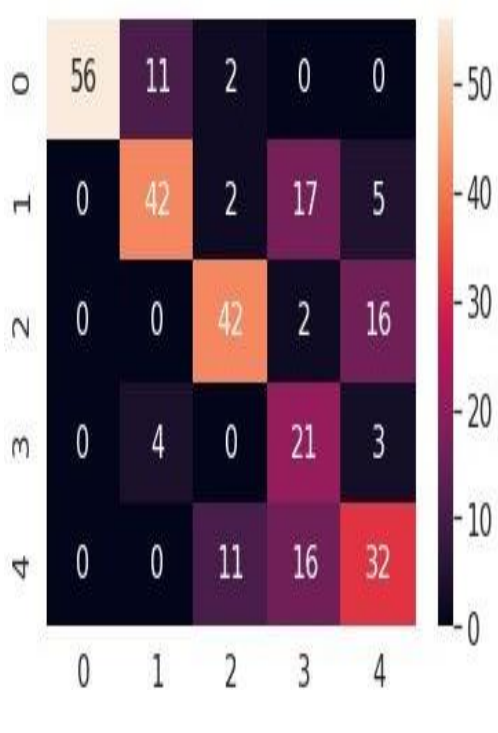

Figure 10: LSTM Confusion matrix

References

[1] Buyukyilmaz, M., & Cibikdiken, A. O. (2016). Voice Gender Recognition Using Deep Learning.

[2] Byun, S. W., & Lee, S. P. (2021). A Study on a Speech Emotion Recognition System with Effective Acoustic Features Using Deep Learning Algorithms. *Applied Sciences 2021(Vol. 11,No. 4,p. 1890*).

[3] Bhatti, M. W., Wang, Y., & Guan, L. (2004). A neural network approach for human emotion recognition in speech. *Proceedings - IEEE International Symposium on Circuits and Systems, 2*.

[4] Langari, S., Marvi, H., & Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction. *Informatics in Medicine Unlocked*(Vol. *20*,p. 100424)

[5] Huang, C., Gong, W., Fu, W., & Feng, D. (2014). A research of speech emotion recognition based on deep belief network and SVM. *Mathematical Problems in Engineering, 2014*.

Doi : https://doi.org/10.54216/FPA.070204

Received: January 19, 2022 Accepted: April 15, 2022

[6] Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., & Wróbel, M. R. (2014). Emotion Recognition and Its Applications. *Advances in Intelligent Systems and Computing*, *300*, 51–62.

[7] Mauchand, M., & Pell, M. D. (2020). Emotivity in the Voice: Prosodic, Lexical, and Cultural Appraisal of Complaining Speech. *Frontiers in Psychology*, *11*

[8] Tawari, A., & Trivedi, M. M. (2010). Speech emotion analysis: Exploring the role of context. *IEEE Transactions on Multimedia*, *12*(6), 502–509.

[9] Pérez-Espinosa, H., Zatarain-Cabada, R., & Barrón-Estrada, M. L. (2022). Emotion recognition: from speech and facial expressions. *Biosignal Processing and Classification Using Computational Learning and Intelligence*(pp. 307–326).

[10] Ramdinmawii, E., Mohanta, A., & Mittal, V. K. (2017). Emotion recognition from speech signal. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*-2017 (pp.1562–1567).

[11] Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural Computing and Applications*, *9*(4), 290–296.

[12] Al-Talabani, A., Sellahewa, H., & Jassim, S. A. (2015). Emotion recognition from speech: tools and challenges. *Mobile Multimedia/Image Processing, Security, and Applications 2015*, *9497*, 94970N.

[13] Luoh, L., Su, Y. Z., & Hsu, C. F. (2010). Speech signal processing based emotion recognition. *2010 International Conference on System Science and Engineering, ICSSE 2010*, 487–490

[14] Soltani, K., & Ainon, R. N. (2007). Speech emotion detection based on neural networks. *2007 9th International Symposium on Signal Processing and Its Applications, ISSPA 2007, Proceedings*.

[15] Nam, Y., & Lee, C. (2021). Cascaded Convolutional Neural Network Architecture for Speech Emotion Recognition in Noisy Conditions. *Sensors 2021(* Vol. 21, No. 13, p. 4399)

[16] Kurpukdee, N., Koriyama, T., Kobayashi, T., Kasuriya, S., Wutiwiwatchai, C., & Lamsrichan, P. (2018). Speech emotion recognition using convolutional long short-term memory neural network and support vector machines. *Proceedings - 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (*pp. 1744–1749)

[17] Yu, Y., & Kim, Y. J. (2020). Attention-LSTM-attention model for speech emotion recognition and analysis of IEMOCAP database. *Electronics (Switzerland)*, *9*(5).

[18] Lech, M., Stolar, M., Best, C., & Bolia, R. (2020a). Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Frontiers in Computer Science*, *2*.

[19] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M., & Cleder, C. (2020b). Automatic Speech Emotion Recognition Using Machine Learning. *Social Media and Machine Learning*.

[20] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017). End-to-End Multimodal Emotion Recognition using Deep Neural Networks. *IEEE Journal on Selected Topics in Signal Processing*, *11*(8), 1301–1309.

Doi : https://doi.org/10.54216/FPA.070204

Received: January 19, 2022 Accepted: April 15, 2022

[21] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., & Mahjoub, M. A. (2018). Speech emotion recognition: Methods and cases study. *ICAART 2018 - Proceedings of the 10th International Conference on Agents and Artificial Intelligence*, *2*, 175–182.

[22] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M., & Cleder, C. (2020a). Automatic Speech Emotion Recognition Using Machine Learning. *Social Media and Machine Learning*.

[23] Home | Cheriton School of Computer Science | University of Waterloo. (n.d.).

[24] MFCC Technique for Speech Recognition - Analytics Vidhya. (n.d.).

[25] Kadiri, S. R., & Alku, P. (2019). Mel-frequency cepstral coefficients of voice source waveforms for classification of phonation types in speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (pp.2508–2512).

[26] Lech, M., Stolar, M., Best, C., & Bolia, R. (2020b). Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Frontiers in Computer Science* (Vol. 2, p. 14).

[27] Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., … Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data 2021 8:1*, *8*(1), 1–74.

[28] LSTM | Introduction to LSTM | Long Short Term Memor. (n.d.).

[29] Narváez, F. R. (2021,Decemeber). *Smart technologies, systems and applications : Second International Conference, SmartTech-IC 2021*.

[30] Guo, J. (2022). Deep learning approach to text analysis for human emotion detection from big data. *Journal of Intelligent Systems*, *31*(1), 113–126.

Doi : https://doi.org/10.54216/FPA.070204

Received: January 19, 2022 Accepted: April 15, 2022