Emo-Pulse: Real-time Emotion Recognition with LSTM for Pattern Analysis

1st Ishrat Jahan Easha
Dept of CSE
BRAC University
Dhaka, Bangladesh
ishrat.jahan.easha@g.bracu.ac.bd

2nd Mosa. Rabeya
Dept of CSE
BRAC University
Dhaka, Bangladesh
rabeyashammi145@gmail.com

3rd Alvi Anowar
Dept of CSE
BRAC University
Dhaka, Bangladesh
alvi.anowar@g.bracu.ac.bd

4th Md Alkawser Rahman
Dept of CSE
BRAC University
Dhaka, Bangladesh
md.alkawser.rahman@g.bracu.ac.bd

5th Sania Azhmee Bhuiyan
Dept of CSE
BRAC University
Dhaka, Bangladesh
sania.azhmee.bhuiyan@g.bracu.ac.bd

6th Farah Binta Haque
Dept of CSE
BRAC University
Dhaka, Bangladesh
farah.binta.haque@g.bracu.ac.bd

7th Annajiat Alim Rasel
Dept of CSE
BRAC University
Dhaka, Bangladesh
annajiat.alim.rasel@g.bracu.ac.bd

***Abstract*—The rise of the Internet of Things and the proliferation of voice-based multimedia applications have enabled the capture and correlation of various facets of human behavior using extensive datasets rich in trends and patterns. Within the realm of human speech lies a latent representation that encapsulates numerous expressive aspects. Extracting sentiment from human speech has become a priority through the mining of audio-based data. The ability to recognize and categorize human emotion holds paramount significance for the advancement of the next generation of artificial intelligence. LSTM, employing an RNN-based model, has been utilized to predict emotion by capturing the nuanced tone of the human voice. The model's performance in predicting categorical emotion has been rigorously evaluated, demonstrating a superior performance compared to other models by approximately 10%. Assessment of the model has been conducted using two benchmark datasets, RAVDESS and TESS, containing renditions of eight distinct emotions by voice actors. Notably, this model outperformed other state-of-the-art models, achieving an accuracy of around 80% for weighted data and approximately 85% for unweighted data.**

***Index Terms*—Machine Learning; Speech Emotion Recognition; Prediction; RNN; LSTM; Real-time Prediction**

I. Introduction

The surge in internet traffic, particularly for real-time multimedia applications like voice control on wearables, online gaming, and various services such as real-time content delivery and video conferencing, has seen a significant increase [10]. Multimedia systems encompass diverse sources, incorporating integrated audio, video streams, and texts. The expansive growth of the internet has led to the signaling of numerous human conversations, conveying a wealth of information. This surge in signaling has notably impacted the expression of human behavior and embedded emotions, with a specific emphasis on characteristics related to voice. Emotions, intricately intertwined with real-world thoughts and behavior, play a pivotal role, shaped by actual life circumstances. Human interactions, marked by a unique connection, often manifest a blend of diverse and intertwined emotions [9].

A. Research Problem

Addressing emotion identification from audio dialogue presents several challenges that need careful consideration:

- **Diverse Emotion Forms:** Current technology faces limitations in mathematically predicting various types of emotions. Moreover, there is a lack of mechanisms for forecasting mixed emotions and evaluating the accuracy of predictions [10].
- **Performance Constraints:** Emotion detection through audio feature extraction poses performance challenges compared to textual feature extraction. Additionally, there are complementary cues for identifying emotions in text and audio-based feature modalities, and the reasons for the inaccuracies in audio-based emotion detection remain undetermined [8].
- **Speaker Segmentation:** The identification of emotions from audio dialogue requires separating the voices of different speakers. This segmentation is crucial for feeding the relevant sections into the mood prediction model.

B. Research Objective

To address the challenge of extracting emotion from audio-based data, our contributions are categorized into four key areas.

- **Firstly,** we developed a Natural Language Processing-based encoding technique known as the one-hot encoding approach to incorporate high-quality audio features that differentiate eight major emotions, including happy, sad, neutral, and quiet.
- **Secondly,** we implemented a suitable attention method that optimally handled padding with feature representation inputs for the emotion detection model, specifically LSTM, in conjunction with one-hot encoding.
- **Thirdly,** to achieve satisfactory real-time speech emotion recognition accuracy, we enhanced the performance of

each component of the model using our approach, making modifications to various parameters for emotion prediction.

- **Finally,** we developed a real-time emotion prediction system utilizing the system microphone, validating its methodology for practical applications. This system is beneficial for capturing a spectrum of emotions encountered in human or machine-operated call centers and the automated healthcare system.

The subsequent sections of this essay are organized as follows. The following part covers the foundation of the theories employed and relevant research. Section 3 provides a comprehensive description of the architecture of the LSTM model and the model used for data processing. In section 4, technical data processing methods are presented, and the data is visually presented for better comprehension. Section 5 demonstrates the utilization of the LSTM model for real-time emotion detection. The output of the model's real-time speech emotion recognition and the results are examined in section 6. Finally, the concluding section (section 7) discusses future enhancements and provides a summary of the entire research effort.

II. Background

Initially, shallow handcrafted techniques were employed for feature extraction, utilizing various statistical functions like mean, range, variance, and linear regression coefficients. These temporal characteristics paved the way for deep learning methods to unveil feature representations. Adikari et al. [7] leveraged such deep-learned characteristics for emotion recognition. Similarly, researchers focused on pitch data to identify emotions in audio, finding that Mel-scale spectrograms impaired pitch information. To address this, linearly spaced spectrogram features were introduced for improved performance [1] [6] [11]. Efforts in multimodal information integration were made to generate a comprehensive solution, combining textual, visual, and auditory modalities for rich feature representation. A CNN-based architecture in a study [5] demonstrated the impact of voice and text transcription in speech emotion identification, comparing spectrogram features with Mel-Frequency Cepstral Coefficients (MFCC). While previous studies indicated that the integration of text and audio modalities enhanced accuracy, modern multimodal emotion recognition, as shown in a publication [4], utilized a CNN single layer with textual characteristics and pre-trained word embeddings.

A. Recent Works

Contemporary emotion recognition models predominantly employ deep learning techniques and tree- or distance-based machine learning algorithms. For instance, in speech-based emotion detection, audio features distinguish between high and low to activate emotion for the initial top-level split [4]. Decision tree classifiers, particularly hierarchical ones leveraging prior knowledge, were employed in some studies [3]. Another study [2] utilized an RNN memory network with a multi-hop attention mechanism to learn an emotive summary of a situation, considering interpersonal and self-influence in the global memory of the dialogue. State-of-the-art models like DialogueRNN and its variations incorporate independent Gated Recurrent Units to collect context information, accounting for the global context of the discussion, emotion state, and speaker status.

Compared to the text-based modality, the existing work highlights a key limitation in the precision of audio modality feature representation. Additionally, the impact of audio embedding techniques on NLP capabilities remains inadequately understood. This research addresses these challenges by introducing a novel audio feature extraction technique using an RNN-based LSTM and three energy-based approaches (Root Mean Square (RMS), Zero Crossed Rate (ZCR), and Mel-Frequency Cepstral Coefficients (MFCCs)), aiming to enhance real-time emotion performance from human speech.

III. Methodology

The objective of this study is to employ artificial neural networks for the creation of a real-time emotion recognition system for human speech. As the collected data encompasses values across various scales, it undergoes a preprocessing stage involving fitting and normalization to enhance network training. This preparation phase is crucial before training a data sample using a serial-parallel architecture. Following the training operations, the serial-parallel architecture is converted into a parallelized network for predictive tasks. Manually setting and optimizing Principal Component Analysis (PCA) parameters is challenging, and this research addresses the issue by employing a sophisticated model capable of computing the significance of each past data point and providing optimized predictions. The Long Short-Term Memory (LSTM) model is implemented for this purpose.

A. LSTM Model

The model implements the LSTM cell each with three gates illustrates as the input gate, the output gate, and the forget gate. The three gates combinedly perform to learn the weights and determine the ratio of a current data sample to be remembered and past learned context should be forgotten. The cell state Ct represents the short- term and the long-term internal memory of a cell. The input gates have been implemented to select the new information to be added and stored in the current C_t. A $sigmoid$ function is implemented to reduce the input vector (i_t) values. After that, a $tanh$ function modifies each value between $[-1, 1]$ $(C - t)$. An element-by-element matrix has been multiplied by i_t and C_t which represents the information that requires to be added to the current cell. To control the output flowing to the next cell the output gate has been implemented. The output gate consists of a $sigmoid$ function and then to filter the less important information a $tanh$ function has been implemented. By this technique, the information that needs to pass through is kept. The output (o_t) is calculated by the following equation.

B. Data Modelling

This section provides an overview of the dataset and details the design of data processing for the LSTM model. The data processing architecture employed in the LSTM model involves several steps. Each audio track from the dataset is considered, and emotions are represented using numbers (e.g., 03 for happy in RAVDESS) or indicated by file names (TESS). Subsequent processes include sample rating, adjustment based on the number of audio samples per second (22.5 kHz for TESS and 48 kHz for RAVDESS), segmentation into audio segments, normalization to +5.0 dBFS, conversion into a sample array, cleanup to remove unnecessary information and silence, addition of padding for uniform length, and noise reduction.

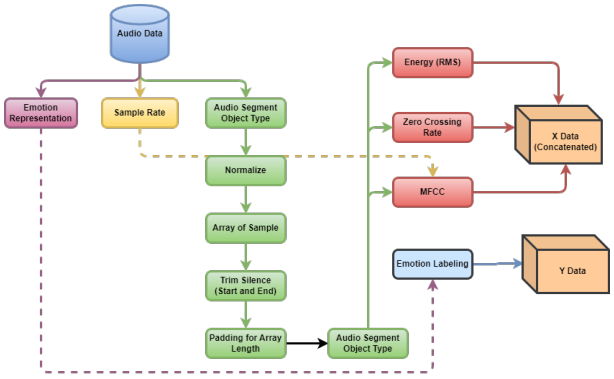

Fig. 1. Architecture of Data Processing

IV. Dataset

RAVDESS and TESS were selected as the datasets for this study. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset comprises recordings from 24 professional actors, evenly distributed with 12 men and 12 women. The actors delivered lines in a neutral North American accent, presenting two lexically similar statements. The dataset includes two levels of expression intensity: one normal and the other strong, along with an additional neutral expression. The Toronto Emotional Speech Set (TESS) dataset draws inspiration from the Northwestern University Auditory Test No. 6 (NU-6; Tillman and Carhart, 1966) and consists of 200 target words. In TESS, a spoken carrier phrase is used for recording, with two women portraying seven different emotions: neutral, happy, sad, angry, fearful, disgusted, and pleasant. The RAVDESS dataset encompasses eight emotions: neutral, calm, happy, sad, angry, afraid, disgusted, and astonished. Specifically, the TESS dataset consists of 2800 files, calculated by multiplying 2 actors by 200 phrases by 7 emotions. In contrast, the RAVDESS datasets comprise 1440 files, derived from 24 actors multiplied by 60 trials per actor.

A. Data Preprocessing

We utilized the $AudioSegment$ module from the $pydub$ library to extract the sample data for observation. Subsequently, the audio data was visualized using the $librosa$ library, and an array was incorporated as the sample. In the data visualization, the time scale of the actual audio is represented on the $x-axis$, with loudness depicted on the $y-axis$. The original audio appears faint, as indicated by the limited range on the $y-axis$, potentially introducing interference and questioning the validity of feature extraction. To address this, we normalized the sound to $+5.0$ dBFS using the $effect$ module of the $pydub$ library. The normalized track was then converted into a $numpy$ array.

Trimming, executed with the $effect$ module, is applied to reduce the starting and finishing ends of the flat line while eliminating extraneous data. For length equalization, padding has been added to the right side. The computed maximum audio duration is 243200. Despite the absence of noise in the datasets, the $noisereduce$ package provides a consistent stamp for further refinement. The resulting data is stored as the final array for analysis.

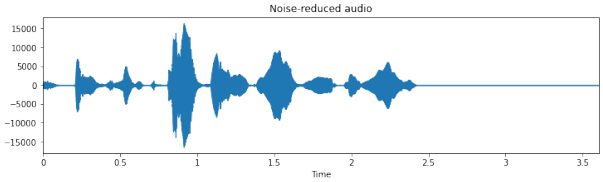

Fig. 2. Final Audio

B. Features Extraction

For the extraction of features in speech emotion identification, the $librosa$ library has been employed. The gathered features include $Root\_Mean\_Square(RMS)$, $Zero\_Crossed\_Rate(ZCR)$, and $Mel-Frequency\_Cepstral\_Coefficients(MFCCs)$. The hop_length is configured to 512, and the assumed $frame\_length$ is set to 2048. In this setup, every 2048 samples are examined, resulting in 4 sequential feature values obtained per examination. Consequently, a total of 475 sequential values, representing each feature, are returned for the entire length of the audio sequence, totaling 173056 samples, with an additional value for the last sample. The feature values are presented in the following table, with the Root Mean Square utilized for calculating the energy.

Feature	Shape
Energy	(1, 339)
ZCR	(1, 339)
MFCCs	(13, 339)

TABLE I
Feature Values

V. Implementation

For data preprocessing, we utilized the scikit-learn module, Keras, and Tensorflow, with Keras serving as the front end for machine learning. The dataset employed in this research encompasses various parameters. The model was implemented

using the Keras library, featuring a dense layer with an output of 8 nodes. Two hidden LSTM layers, each with 64 nodes, were incorporated, with each node representing one of the emotional representations. The "softmax" function was employed for activation, and the "RMSProp" optimizer was used with its default settings.

A batch size of 23 was chosen as it is a factor of all the samples in the dataset and yields optimal results. The system-generated graph illustrating the shape of nodes in each LSTM layer is depicted in the figure below. The Keras library was employed to generate the output. The model is configured as a $Sequential$ with $softmax$ activation in the dense layer, acknowledging the sequential nature of audio data and utilizing LSTM for sequential data learning. The batch size was established from scratch. The training data was assessed with a maximum level of valid categorical accuracy and saved in $HD5$ format for real-time emotion recognition.

After 100 iterations without improvement, the learning rate was decreased by a factor of 0.1. The loss function was computed using $categorical_crossentropy$ and the accuracy matrix $categorical_accuracy$, with $RMSProp$ employed as the optimizer. The weights were saved after 340 iterations of the model. The model and its parameters are detailed in the table below.

Model: Sequential

Layer (type)	Output Shape	Param
LSTM	(None, 951, 64)	20480
LSTM_1	(None, 64)	33024
Dense	(None, 8)	520

TABLE II
TRAINING PARAMETERS

A. Real Time Speech Emotion Recognition

In this study, we implemented real-time speech emotion identification based on our trained model. The implementation involves creating a temporary $.wav$ file and recording audio input from the device's microphone. The recorded audio is then preprocessed, and the emotional content of the speech is returned along with a distribution. The $pyaudio$ package is utilized for audio recording.

Initially, $Keras$ and $TensorFlow$ are loaded, and the pre-trained model is compiled. Subsequent to data processing, features are retrieved from the data for each array. The number of Mel-Frequency Cepstral Coefficients (MFCC) is set at 13, and after concatenation, X is transformed into a 3D array. An emotion list is defined for legible output, and an "is silent" function is used to automatically pause audio recording and form an input after a period of silence.

This LSTM model implementation provides real-time emotion prediction as speech emotion recognition from the audio output. The device's sound card is used for audio recording. The session begins with $pyaudio$ joining the input channel, and the audio signal is recorded using $pyaudio$ and $wave$ if there is no silence. The recorder stops after a set time, begins processing, and starts the subsequent recording for preprocessing. An array of 8 emotion probabilities is returned when the $model.predict$ is invoked, represented as, for example, [array([p_neutral, p_calm, p_happy, p_sad, p_angry, p_feaful, p_disgust, p_suprised], dtype=float32)]. Following this, predictions are visualized briefly and saved in a list. The expected result is displayed using the $matplotlib$ package. If there is silence for more than two seconds, the session automatically ends, and the summary shows the total session time while terminating the input connection. The sampling rate is 24414, the chunk represents the hop length, and the sample size matches the model. The 32-bit audio recording format accepts input only from a mono channel.

VI. RESULTS

To assess the model's performance during the training phase, the display of loss and categorical accuracy values has been utilized. Additionally, a confusion matrix has been generated to depict the percentage of correct predictions for each emotion in both the validation and test sets. The accuracy rates for each emotion in the test and validation sets are presented in the following figure, illustrating the training and validation loss.

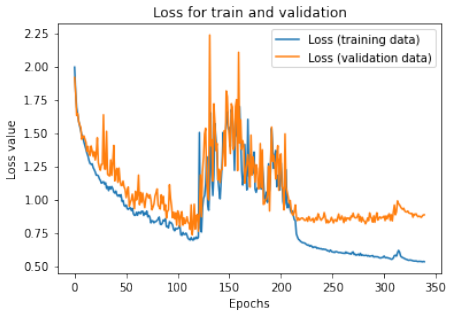

Fig. 3. Training and Validation Loss

From the figure, we can see the training loss is less than the validation set after 200 epochs. In the next figure, the accuracy of the model is shown.

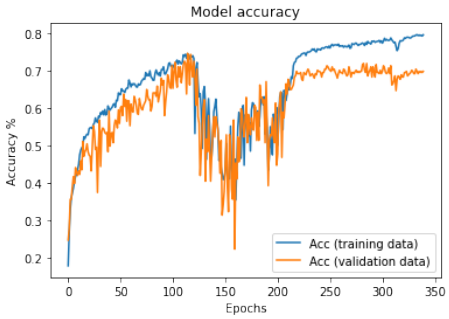

Fig. 4. Accuracy

A. Validation Set Evaluation

The validation set has been assessed using a confusion matrix, where the emotions are predicted based on the validation

set. The confusion matrix is presented below. Upon inspection of the matrix, it is evident that in the majority of cases, our predictions are accurate. Notably, our model performs well when predicting happiness, followed by high accuracy in predicting sadness and anger.

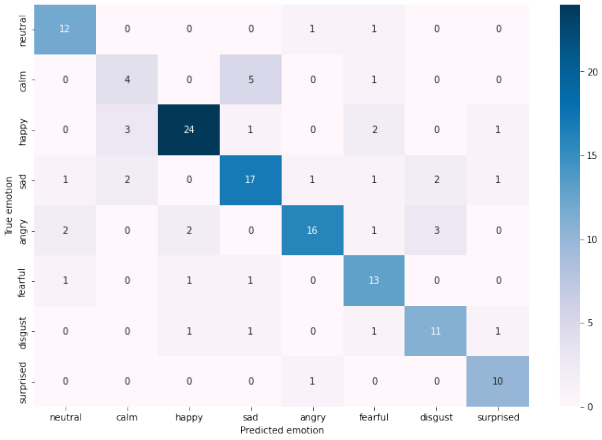

Fig. 5. Confusion Matrix for Validation Set

B. Test Set Evaluation

The confusion matrix has been employed to assess the emotions in the test set and serves as a foundation for those emotions. The following image illustrates the confusion matrix. The matrix indicates that our model accurately predicted the happy emotion, followed by sadness, anger, fear, and neutral emotions, in that order.

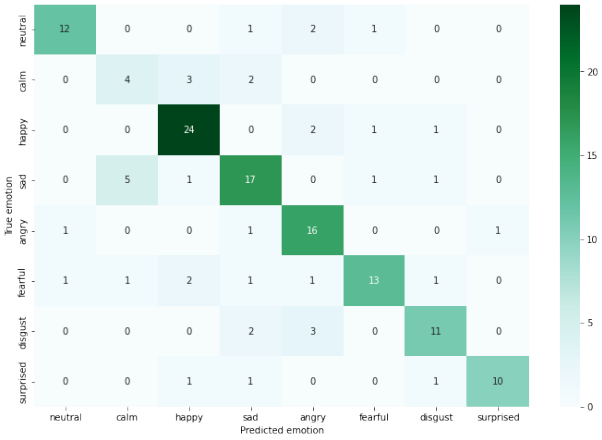

Fig. 6. Confusion Matrix for Test Set

C. Real Time Performance Evaluation

We gathered input for real-time evaluation using the device's microphone and tested the model with three distinct emotional tones, achieving the desired results. The upcoming graph illustrates the outcome when intentionally setting the input to a deep, somber voice. We utilized standard speech for typical scenarios and included a single shout to assess the model's performance in various situations. The graph below presents the model's output and corresponding predictions.

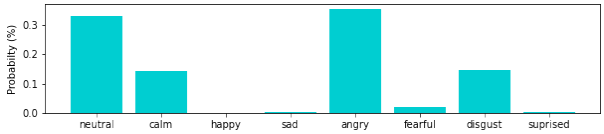

Fig. 7. Output

Finally, after the session ended, we were able to experience every emotion that had been predicted for the entire time, with calm being the emotion that had been maximized. The session summary is depicted in the following graphic.

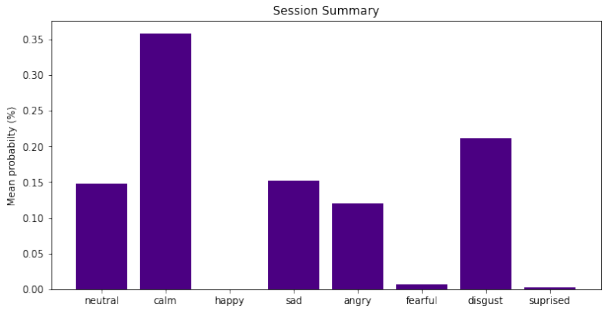

Fig. 8. Summary

VII. Conclusion and Future Work

The successful resolution of the key challenge of emotion identification from audio data holds significant real-world benefits. This paper provides multiple theoretical contributions for future research. The study emphasizes the effectiveness of one-hot encoding-based feature extraction with padded parameters, well-suited for an RNN-based LSTM model in the complex task of emotion detection from speech. While the approach involves three feature engineering methods along with one-hot encoding-based embedding for audio emotion detection, our model evaluation demonstrates superior performance compared to other state-of-the-art models discussed in the literature review. The model achieves a promising 81% weighted accuracy and approximately 80% unweighted accuracy after adjusting the learning rate at factor 0.1 following 100 epochs.

Given the diverse features, the combination of three modalities could potentially enhance accuracy. However, it might hinder sectors such as healthcare and customer service contact centers that solely rely on audio as a modality. As our model is exclusively trained on audio-based modalities, its optimal performance is expected in that context. Comparatively, the performance of the limiting factor used for real-time automatic speaker diarization falls slightly short when compared to other state-of-the-art models. Future work could potentially construct a fully automated emotion recognition pipeline from human speech by optimizing model complexity and parameter settings.

This research suggests a feature pipeline for real-time emotion identification from human speech in manual diary conversations, offering a solution in the absence of reliable real-time approaches. Integrating this suggested system with a real-time speaker diarization methodology could open avenues for significant commercial applications. A method for creating speech embeddings from a large corpus of speech data may enhance accuracy by reflecting rich information, especially compared to the high-dimension and high-sparsity nature of one-hot encodings.

This work serves as a valuable innovation, opening doors for various tech-driven industries. Recognizing both simple and complex human emotions by machines could improve services in sectors such as virtual call centers, elder care, and AI-driven robots. Companies focused on customer satisfaction could leverage this innovation to gain advantages, studying how their representatives handle clients exhibiting a range of emotions during conversations. While humans can easily interpret complex emotions through dialogue, machines face challenges in predicting emotions solely from audio without facial expressions. Enabling machines to comprehend human emotions through audio dialogue represents a significant advancement in human-computer interaction through enhanced exploitation of massive audio data.

REFERENCES

[1] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2227–2231. doi:10.1109/ICASSP.2017.7952552.

[2] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," Speech Communication, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[3] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 2594–2604. doi: 10 . 18653 / v1 / D18 - 1280. [Online]. Available: https://aclanthology.org/D18-1280.

[4] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep learning based emotion recognition system using speech features and transcriptions,"arXiv preprint arXiv:1906.05681, 2019.

[5] A. Satt, S. Rozenberg, and R. Hoory, "Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms," in Proc. Interspeech 2017,2017, pp. 1089–1093. doi:10.21437/Interspeech.2017-200.27

[6] I. Madhavi, S. Chamishka, R. Nawaratne, V. Nanayakkara, D. Alahakoon, and D. De Silva, "A deep learning approach for work related stress detection from audio streams in cyber physical environments," in 2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), vol. 1, 2020, pp. 929–936. doi: 10.1109/ETFA46521.2020.9212098.

[7] A. Adikari, G. Gamage, D. De Silva, N. Mills, S.-M. J. Wong, and D. Ala-hakoon, "A self structuring artificial intelligence framework for deep emotions modeling and analysis on the social web," Future Generation Computer Systems, vol. 116, pp. 302–315, 2021.

[8] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 112–118. doi: 10.1109/SLT.2018.8639583.

[9] C. E. Izard, Human emotions. Springer Science Business Media, 2013.

[10] S. A. Alvi, B. Afzal, G. A. Shah, L. Atzori, and W. Mahmood, "Internet of multimedia things: Vision and challenges," Ad Hoc Networks, vol. 33, pp. 87–111, 2015. doi: 10.1016/j.adhoc.2015.04.006.

[11] Hochreiter, Sepp Schmidhuber, Jürgen. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.