Journal of Applied Sciences, Information and Computing

Vol. 4, Issue. 2 (2023)

© School of Mathematics and Computing, Kampala International University

ISSN: 1813-3509 https://doi.org/10.59568/JASIC-2023-4-2-02

SPEECH EMOTION RECOGNITION MODEL USING DEEP LEARNING

[1]Taiwo Kolajo[1*]
taiwo.kolajo@fulokoja.edu.ng

[2]Emeka Ogbuju
emeka.ogbuju@fulokoja.edu.ng

[3]Joy Ojochide Bello
belloojochidejoy@gmail.com

[4]Francisca Oladipo
francisca.oladipo@tau.edu.ng

[1,2,3]Department of Computer Science, Federal University Lokoja, Lokoja, Kogi State, Nigeria
[4]Department of Computer Science, Thomas Adewumi University, Oko, Kwara State, Nigeria

Abstract

Speech has long been recognized as the main form of communication between people and computers. Technology made it possible for humans and computers to interact through the development of human-computer interfaces. Although speech emotion recognition systems have advanced quickly in recent years, many difficulties have also arisen during this development, such as the inability to recognize emotions that lead to depression and mood swings, which can be used by therapists to track their patients' moods. It is necessary to create a model that detects the many emotions that contribute to depression to improve doctor-patient relationships and increase the effectiveness of spoken emotion recognition models. In this paper, over 2000 audio files were compiled. We curated a local dataset that accounts for 60% of the total dataset acquired, 40% of the dataset used was obtained from RAVDESS. To extract the proper vocal features, we employed the Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), and Root Mean Square (RMS). The Tensorflow CONV1D with Relu activation function and several sequential layers was used to build the model. The batch size was 64 and the epoch size was 50. Seven emotional states, including anger, disgust, sadness, happiness, surprise, fear, and neutrality were extracted. The accuracy of the confusion matrix, which served as the performance metrics, is 96%.

Keyword: Speech analysis, Speech emotion recognition, Machine learning, Deep learning, RAVDESS

1 Introduction

Predicting human emotions can be very difficult. Though it is easier to do these through facial expressions and gestures, it is more challenging to recognize emotions as a person gets older and more experienced, they learn to regulate their expressions. Additionally, movements and facial expressions only convey outward emotions like anger, happiness and sadness but fail to reveal emotions such as depression, disgust, boredom, and mood swings (Lalitha et al., 2015). To overcome this, there are different methods developed to recognize emotions, and speech emotion recognition (SER) is one among them.

The communicative objectives of the speaker to whom they are speaking determine the specifics of each speech act. Since speakers cannot disguise their emotions through their speech, unless they choose to pretend, it is possible to discern the speakers' emotions (Sundarprasad, 2018). Speech, however, also includes silence and noise because speakers pause at the beginning and end of sentences as well as during the transitions between syllables. This means that every utterance of communication includes both voice and non-speech. Most modern speech algorithms are capable of processing studio-recorded, neutral speech, but they fall short when it comes to emotional speech. This is because it might be difficult to represent and characterize emotions through speech. Speaking becomes more natural when emotions are present, and nonverbal communication in a conversation transmits important information like the speaker's objective. Along with the material provided through text, the way the words are spoken conveys important non-linguistic information (Nyarks and Owushi, 2022).

Depending on how it is delivered, spoken text can be interpreted in a variety of ways. In English, the term "OKAY" is used to indicate appreciation, disbelief, consent, disinterest, or a statement, for example. As a result, simply comprehending the text is insufficient to decipher the semantics of a spoken word. On the other hand, in addition to the message, speech systems must be able to handle non-linguistic information like emotions. By understanding the underlying emotions in addition to the phonetic information included in multimodal signals, humans can comprehend the anticipated message. In the case of video, non-linguistic information may be noticed through facial expressions, emotional expressions in speech, and punctuation in written text (Doma and Pirouz, 2020). The ability to control voice sounds and make a speech is one of the traits that set humans apart from other living things. It has been discovered that when humans speak, their voices change as they express their emotions. Hence, deducing human emotions through speech analysis has a reasonable chance of being useful in boosting human conversational and persuading abilities (Chunawale and Bedekar, 2020).

While the development of speech emotion recognition models has progressed quickly in recent years, several obstacles, such as a lack of appropriate data samples, have been encountered (Padi et al., 2021), reliance on feature extraction on personalized features (Li et al., 2021), the inability to eliminate noise and silence at the pauses in speech and at the intersections of words' syllables (Alluhaidan et al., 2023), as well as the incapability to recognize emotions, which can lead to mood swings and depression (Wang et al., 2021). Detecting emotions responsible for mood swings and depression can be useful to therapists to provide intelligent support services for mental health (Madanian et al. (2022). Designing a model that recognizes the many emotions that contribute to depression is necessary to increase the effectiveness of speech emotion recognition models and to improve doctor-patient relationships (Monferrer et al., 2023).

The remainder of the paper is organized as follows: Section 2 discusses the background and related work. The methodology is covered in section 3 while section 4 presents the result and discussion. The conclusion and further work are presented in section 5.

2 Background and Related Work

This section presents the background and research efforts in using machine earning and deep learning for speech emotion recognition.

2.1 Speech Emotion Recognition

In situations or applications when face-to-face human connection is not feasible or preferable, the ability to recognize human emotions through voice and speech pattern analysis can help enhance communication and persuasive abilities. Numerous research has been done on extracting emotions from aural information. The primary objective of a voice emotion detection algorithm is to identify the emotional state of a speaker from speech (Chintalapudi et al., 2023). Speech Emotion Recognition has made it possible to create automated systems that can analyse human speech with intelligence. Typically, a speech emotion recognition system focuses on collecting traits from speech signals including pitch frequency, formant features, energy-related and spectral data, then conducts a classification search to identify the underlying emotion. (Liu et al., 2021).

The successful extraction of the features has a significant impact on how well the classification of emotions performs. Numerous variables, including spectral characteristics and prosodic characteristics, were taken out by the researchers for their study. The most often employed characteristics in speech emotion identification algorithms are the prosodic and spectral aspects of speech. Examples of prosodic features include pitch, rhythm, emphasis, inflection, and pause. Spectral characteristics analyse the voice signals' frequency components. The most common classification algorithms used in speech emotion recognition include neural networks, hidden Markov models, Gaussian mixture models, and support vector machines. The majority of automatic speech emotion research focuses on choosing the best feature set and identifying the more emotive characteristics. Some of the techniques commonly used in the research community are Linear Prediction Coefficient (LPC), Spectrum and Fundamental Frequency Histogram, Instantaneous Spectra Covariance Matrix, and Mel Frequency Cepstral Coefficients (MFCC).

2.2 Related Work

This section presents the research efforts in the speech emotion recognition field. Machine learning and deep learning have gained ground in this field. The recent works done are presented subsequently.

Aouani and Ayed (2020) suggested using audio signals to recognize spoken emotions. Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Harmonic to Noise Rate (HNR), and Teager Energy Operator (TEO) were some of the audio properties that the authors retrieved. The relevant characteristics were chosen using an auto-encoder. As a classifier, SVM was utilized. The outcome revealed a 70% average accuracy rate. Madanian et al. (2022) carried out automatic voice emotion identification using machine learning for mental health digital transformation. Five machine learning classifiers were used to capture human emotions using TESS, RAVDESS, and EMO-DB databases. SVM demonstrated the best performance with 74% accuracy. In the same vein, Shaila et al. (2022) discuss different machine learning techniques to identify and classify human emotions from audio signals. Support vector machine, random forest, convolutional neural networks, multilayer perceptron and decision trees were employed. The results showed that random forest performed best with accuracy (0.85), precision (0.87), recall (0.89), and F1-Score (0.88).

Several developments in the field of speech emotion identification systems have been made possible by the application of deep learning techniques. Real-time speech emotion recognition using AlexNet was suggested by Lech et al. (2020). Berlin Emotional Speech (EMO-DB) data were used to train ALexNet to categorize speech into seven categories. The findings revealed an average accuracy of 82%. Machine learning algorithms for speech emotion identification were proposed by Rajeswari et al. (2021). The authors merged three databases—EMO-DB, TESS, and SAVEE—and used deep learning models including BLSTM, RNN/LSTM, and CNN for classification along with decision tree, SVM, and random forest to identify emotion in speech. With 94% accuracy, the integration of CNN and LSTM produced the greatest results. Singh et al. (2023) suggested a self-attention-based deep learning model for voice emotion identification. The datasets used included TESS, SAVEE, and RAVDESS. Eight emotional states could be distinguished with 90% accuracy.

In conclusion, the reviews demonstrate that much effort has been made to guarantee that the emotion identification system can discern emotions from a speaker's voice. To make sure that the speech emotion recognition system can function smoothly and ideally, there is still a lot of work to be done. When creating enhancements to the speech emotion detection system, certain important aspects should be considered, such as detecting emotions from native language corpora and recognizing sadness and mood swings from speech.

3 Methodology

We developed a model that can recognize the emotional state of a speaker through speech signals and match them to corresponding emotions. To overcome the problems of the existing system, the proposed system creates a consistent speech emotion recognition system. The methodology steps are described subsequently.

3.1 Data Collection

Over 2000 audio files were compiled from different sources. The local dataset comprises 60% of the total dataset acquired, 40% of the dataset used was obtained from a multimodal database of emotive speech and music called RAVDESS. Ten professional Nigerian actors from films and television shows make up the locally created dataset. Part of the dataset was generated by recording the emotions of people while they act emotionally.

3.2 Data Selection

This aspect is concerned with the selection of data from available data sources. The dataset we make use of is in audio format and was curated, it contains over two thousand audio files with each statement expressed in a variety of eight (8) different emotions that a depressed person feels: disgust, anger, calm, fear, happy, surprise, sad and neutral.

3.3 Preprocessing

This part has to do with the preparation of data in the form that will be fit for advanced evaluation and processing. The dataset's naming scheme followed a pattern, with odd actors designating the male sex, and even actors designating the female sex. The emotion that was identified in the audio clip is the target variable. Each recording lasted around three (3) seconds. Noise and pauses between words and the juncture between syllables of the words were removed from audio recordings. Voice Activity Detection and Filtering signal processing technique was used to reduce the noise. To minimize background noise (also known as additive noise), the Voice Activity Detection and Filtering method, a common spectral subtraction technique was used to subtract the noise spectrum’s estimate from the noisy speech spectrum. In order to provide a clear representation of the underlying signal, Wiener filtering was employed to remove the noise from the damaged signal.

3.4 Feature Selection

Time-domain features and frequency-domain characteristics are two of the numerous categories of audio features. The Short-term energy of signals, Maximum Amplitude, Zero Crossing Rate, Entropy of Energy, and Minimum Energy are examples of time-domain properties. These characteristics offer a more straightforward method of audio signal analysis and are relatively simple to extract. Spectrograms, MFCC, Spectral Entropy, Spectral Centroid, Chroma Coefficients, and Spectral Roll-off are examples of Frequency-Domain Features. These features reveal a more complex pattern in the audio signal, which may aid in determining the signal's underlying mood. Due to its benefits, we adopted the frequency-domain feature MFCC in this research. MFCC represents the short-term power spectrum of a sound by altering the audio signal through a series of phases to mimic the human cochlea. The Mel scale is preferred over linear scales because it closely resembles how sound is perceived by humans.

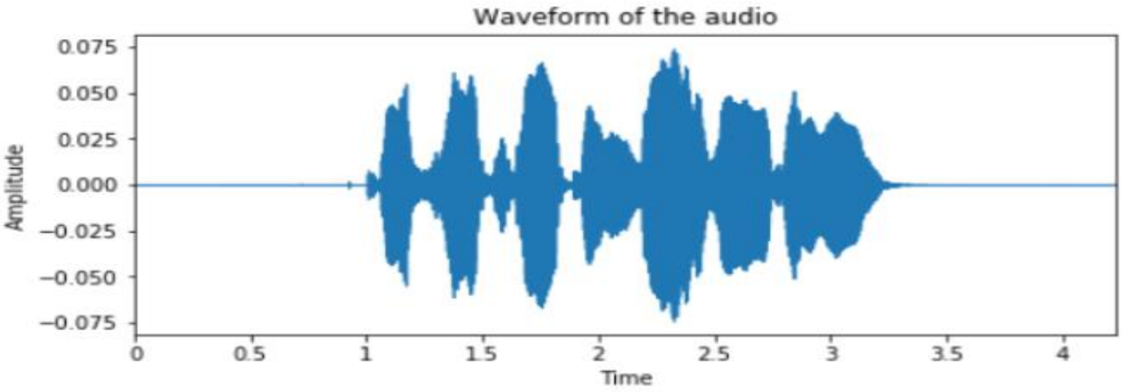
3.5 Model Implementation

Two sets were created from the available dataset. The training set was 70% and the remaining 30% was the test set. During the development of this research work, Google colab notebook was used to combine executable code. The training data was fit into the machine, and the test set was used against the training set to obtain accuracy. In this work, a variety of models were trained to utilize the MFCC, MFCC + Delta, and MFCC + Delta-Delta coefficients configurations. Using MFCC traits and Log-Mel Spectrogram data represented as images, an audio file's mood was classified using CNN. CNNs allow us to explicitly assume that the inputs are pictures, which enables us to systematically embed certain characteristics into the architecture. To bring invariance, max-pooling is introduced to CNNs. As a result, only the pertinent high-dimensional features are learnt for a variety of tasks including segmentation, classification, etc. It uses shared kernels (weights) to take advantage of the 2D correlated structure of image data and automatically learns important features from high-dimensional input pictures. CNN may also be used to recognize audio. Linearity is established on the log-frequency axis during Fast Fourier processing using the Log Mel-filter bank Transformation (FFT) representation of raw speech signals, allowing convolution operation along the frequency axis. With this, the model can quickly and efficiently understand the underlying patterns. This characteristic, along with CNN's greater representation capability, enables the model to learn the underlying patterns quickly and efficiently, leading to state-of-the-art performance in speech-based emotion recognition systems. The 250 ms frame duration in 3D audio was chosen since it was determined to be the shortest effective time for evoking emotions. To find patterns in the sound waveform, multi-layer CONV1D was also trained on raw audio.

4 Results and Discussion

4.1 Data Preprocessing Results

The preprocessing stage involved removing background noise, pauses between words, and syllable breaks from audio recordings. The noise was reduced using the Voice



Activity Detection and Filtering signal processing approach. To provide a clear representation of the underlying signal, Wiener filtering was employed to remove the noise from the damaged signal. Pure audio pre- and post-data cleaning waveforms are depicted in Figures 1 and 2, respectively.

Figure 1. The waveform of pure audio pre-data cleaning

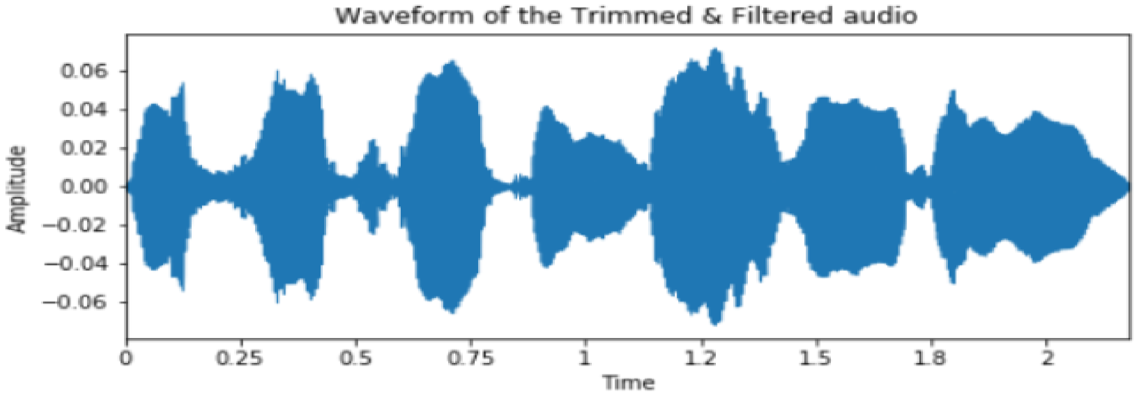


Figure 2. The wave of pure audio post-data cleaning

4.2 Model Performance

Tensorflow CONV1D with Relu activation function was used to build the model. Multiple sequential layers were used, with an epoch size of 50 and a batch size of 64. The Root Mean Squared Propagation (rms-prop) optimizer, which is an extension of gradient descent and accuracy based on precision and recall was also utilized during compilation. While Figure 3 displays the training and testing accuracy, Figure 4 depicts the training and testing loss.

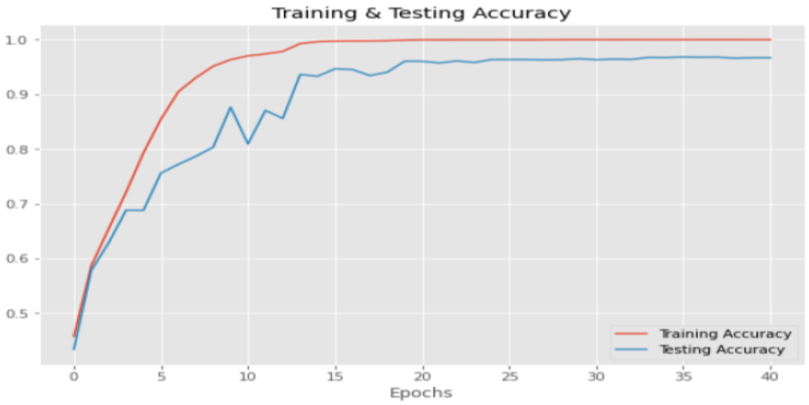


Figure 3. Model Training and Testing Accuracy

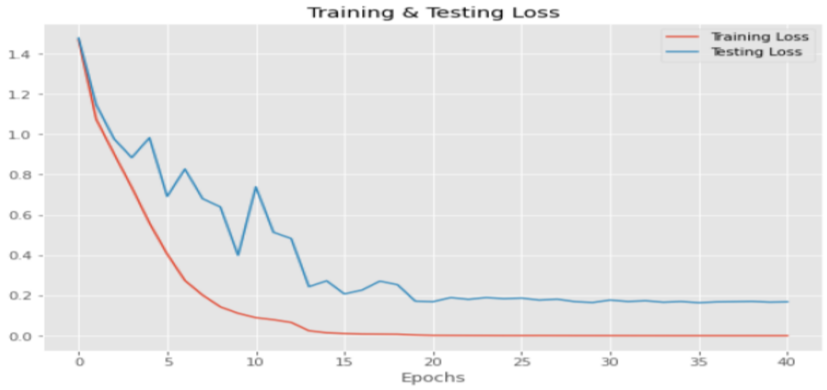
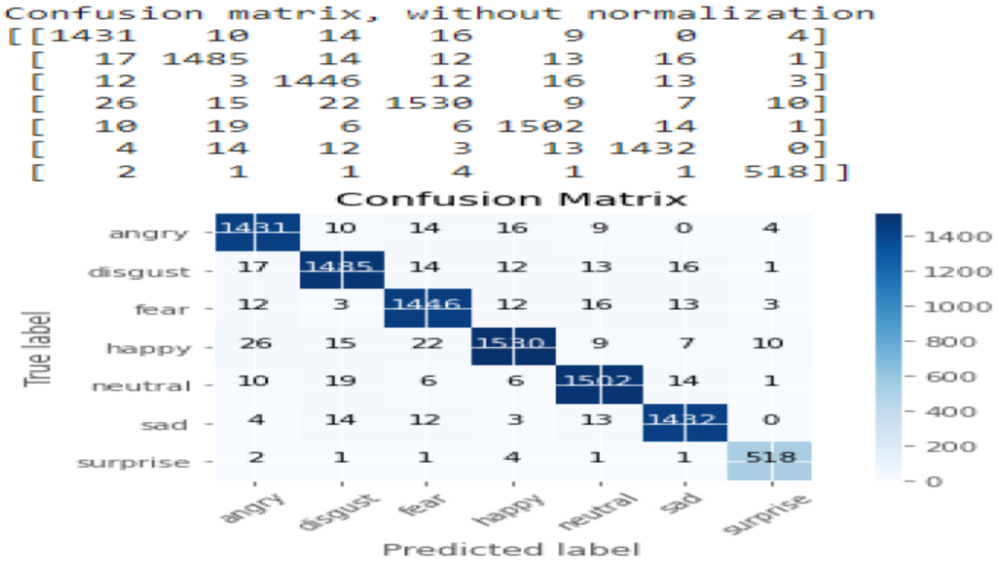


Figure 4. Model Training and Testing Loss

Using the confusion matrix as a performance indicator on the test data, the model's accuracy was verified. On the test set, the model was able to reach an accuracy of 96%. Figure 5 presents the result of the test data accuracy without normalization.



```
Confusion matrix, without normalization
[[1431   10   14   16    9    0    4]
 [  17 1485   14   12   13   16    1]
 [  12    3 1446   12   16   13    3]
 [  26   15   22 1530    9    7   10]
 [  10   19    6    6 1502   14    1]
 [   4   14   12    3   13 1432    0]
 [   2    1    1    4    1    1  518]]
```

Figure 5. Result of Confusion Matrix without Normalization on Test Data

We created a list of labels called cm_plot_labels and plotted a confusion matrix graph using the y_predicted value of the test set and the y_actual value of the test set. Using the confusion matrix graph generated as shown in Figure 5, the model had a more wrong prediction for angry, disgust and fear as happy. We then normalized the prediction by making Sad, Happy, Angry, Fear, Disgust, and Surprise emotions to be displayed as true (1) and made Neutral emotions to be displayed as false (0). So, the value (1) means that the model recognizes an emotion, and the value (0) means that it recognizes Neutral emotion. This is presented in Figure 6.

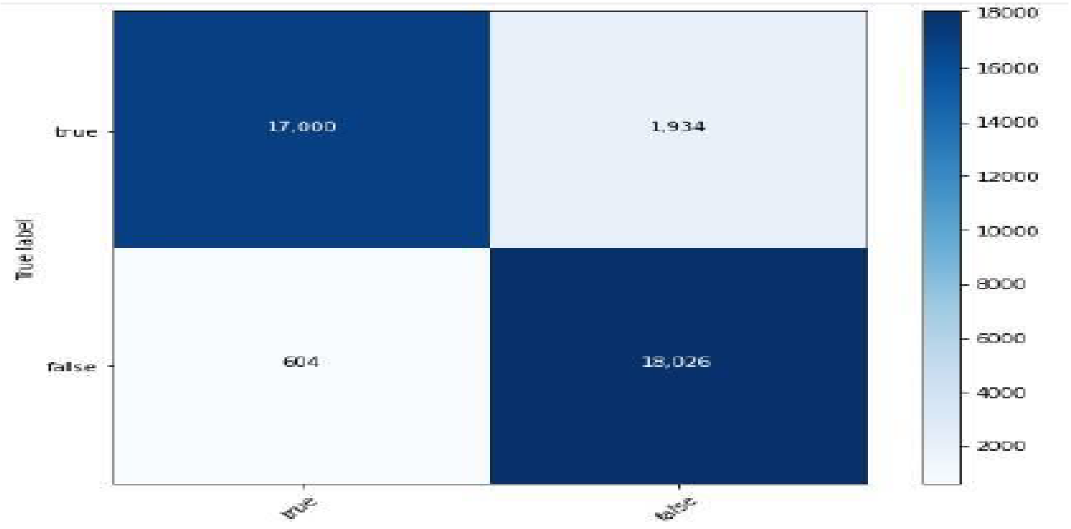


Figure 6. Result of Confusion Matrix with Normalization on Test Data

5 Conclusion and Further Work

In this research, we conducted an in-depth analysis of speech emotion recognition systems. A model that recognizes emotions and produces a corresponding emotion in the appropriate text format based on the voice input provided by the user based was developed. Features extraction was carried out following pre-processing of the raw audio recordings. A convolutional neural network was used as the classifier to identify these emotions. Tensorflow CONV1D with Relu-activation function and several sequential layers was used to build the model. The batch size was 64 and the epoch size was 50. The outcome demonstrates that accuracy of 96% utilizing the confusion matrix as the performance measure. Additionally, we found that the choice of audio features has a considerably greater influence on the outcomes than model complexity does on emotion recognition. The output of this research can be used in many aspects for the betterment of the people. For instance, therapists and psychiatrists can use this model to detect the mood of their patients by simply listening to their voices. This will go a long way to reducing cases of depression. The contributions of this research include 1) curating a local dataset for speech emotion recognition, and 2) building a model for speech emotion recognition based on a greater percentage of locally curated audio data.

During this research, we found that speech emotion recognition is a very wide field with lots of untapped potential waiting to be harnessed. In future work, we recommend that recognizing speech emotion should not be based on English speakers alone but should be extended to low-resource languages. For instance, carrying out speech emotion recognition in local languages will go a long way to help people who can only speak their local dialect communicate with a psychologist effectively. In addition, research efforts can be directed to the recognition of compound emotions such as happily disgusted, happily surprised, sadly angry, sadly fearful, sadly disgusted, sadly surprised, and angrily fearful.

6. References

[1]Alluhaidan, A. S., Saidani, O, Jahangir, R., Nauman, M. A., & Neffati, O. S. (2023).Speech emotion recognition through hybrid features and convolutional neural network. *Applied Sciences, 13*(8), 4750. https://doi.org/10.3390/app13084750.

[2]Aouani, H., & Ayed, Y. B. (2020). Speech emotion recognition with deep learning. *Procedia Computer Science, 176*, 251-260.

[3]Chintalapudi, K. S., Patan, I. A. K., Sontineni, H. V., Muvvala, V. S. K., Gangashetty, S. V., & Dubey, A. K. (2023). Speech emotion recognition using deep learning. *2023 International Conference on Computer Communication and Informatics (ICCI)* (pp. 1-5). Coimbatore, India.

[4]Chunawale, A., & Bedekar, M. V. (2020). Human emotion recognition using physiological signals: a survey. *2nd International Conference on Communication & Information Processing (ICCIP)* (pp. 1-9). SSRN.

[5]Doma, V., & Pirouz, M. (2020). A comparative analysis of machine learning methods for emotion recognition using EEG and peripheral physiological signals. *Journal of Big Data, 7*, 18. https://doi.org/10.1186/s40537-020-00289-7.

[6]Lalitha, S., Geyasrutia, D., Narayanana, R., & Shravani, M. (2015). Emotion Detection using MFCC and Cepstrum Features. *4th International Conference on Eco-friendly Computing and Communication Systems* (pp. 29-35).

Amrita Vishwa Vidyapeetham, Bangalore, Karnataka, India.

[7]Lech, M., Stolar, M., Best, C., & Bolia, R. (2020). Real-time speech emotion recognition using a pre-trained image classification network: effect of bandwidth reduction and companding. *Frontiers in Computer Science, 2*, 14.

[8]Li, D., Zhou, Y., Wang, Z., Gao, D. (2021). Exploiting the potentials of features for speech emotion recognition. *Information Sciences, 548*, 328-343.

[9]Liu, Z., Rehman, A., Wu, M., Cao, W., & Hao, M. (2021). Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence. *Information Sciences, 563*, 309-325. https://doi.org/10.1016/jins.2021.02.016.

[10]Madanian, S., Parry, D., Adeleye, O., Poellabauer, C., Mizra, F., Matthew, S., & Schneider, S. (2022). Automatic speech emotion recognition using machine learning: digital transformation of mental health. *PACIS 2022 Proceedings, 45*, 1630.

[11]Monferrer, M., Garcia, A. S., Ricarte, J. J., Montes, M. J., Fernandez-Caballero, A., & Fernandez-Sotos, P. (2023). Facial emotion recognition in patients with depression compared to healthy controls when using human avatars. *Scientific Reports, 13*, 6007. https://doi.org/10.1038/s41598-023-31277-5

[12]Nyarks, A., & Owushi, J. N. (2022). Assessment of the factors that cause pronunciation difficulties to learners of English as a foreign language. *World Atlas International Journal of Education & Management, 5*(1), 52-60.

[13]Padi, S., Sadjadi, S. O., Sriram, R. D., & Manocha, D. (2021). Improved speech emotion recognition using transfer learning and spectrogram augmentation. *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI'21)*. ACM, New York, NY, USA.

[14]Rajeswari, S. S., Gopakumar, G., Nair, M. (2021). Speech emotion recognition using machine learning techniques.

[15] Sharma, H., Saraswat, M., Yadav, A., Kim, J. H., Bansal, J. C. (eds) *Congress on Intelligent Systems (CIS2020). Advances in Intelligent Systems and Computing, 1335*. Springer, Singapore. https://doi.org/10.1007/978-981-33-6984-9_15

[16]Singh, J., Saheer, L. B., & Faust, O. (2023). Speech emotion recognition using attention model. *International Journal of Environmental Research and Public Health, 20*, 5140. https://doi.org/10.3390/ijerph20065140.

[17]Sundarprasad, N. (2018). *Speech emotion detection using machine learning techniques*. Master's Projects. San Jose State University. https://doi.org10.31979/etd.a5c2-v7e2.

[18]Wang, H., Liu, Y., Zhen, X., & Tu, X. (2021). Expression speech recognition with a three-dimensional convolutional network. *Frontiers in Human Neuroscience, 15*, 713823.