ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Volume 11 Issue 01 January 2024

Using Gated Recurrent Unit (GRU) for Speech Emotion Recognition in Children

[1] Gerald Onwujekwe, [2] Ben Haber, [3]Varoun Bajaj

[1] [2] [3] Washington University in St. Louis St. Louis, MO, USA
Corresponding Author Email: [1] geraldo@wustl.edu, [2] bhaber@wustl.edu, [3] b.varoun@wustl.edu

Abstract*— This research paper proposes a novel gated recurrent unit (GRU) neural network for speech emotion recognition (SER) in children, using the FAU-Aibo dataset of children's interactions with the Aibo robot. The GRU model shows promise in accurately predicting negative emotions in children's speech. The paper compares the GRU model with other deep learning and machine learning models, such as LSTM, SVM, and boosted trees, and shows that the GRU model achieves better accuracy, speed, and computational footprint. The paper also discusses the challenges and implications of using natural and spontaneous speech data for emotion recognition and how the GRU model can help detect negative emotions in children as a potential step toward child abuse detection.

Keywords: Children, Deep Learning, GRU, Machine Learning, Speech Emotion Recognition.

I. INTRODUCTION

Deep learning techniques have greatly increased in relevance and usage over the past several years. In recent years, the application and use of human-computer interaction have been growing and has become an important factor for an effective human emotion recognition. Human Emotion Recognition includes facial expression recognition, body language recognition, speech emotion recognition and others. Speech emotion recognition (SER) has been widely used for human emotion recognition in human-computer interaction (HCI) (Fan et al., 2021; Mustaqeem & Kwon, 2020, 2021; Wani et al., 2021). Previously, Artificial Neural Network (ANN), Support Vector Machine (SVM), and Hidden Markov Model (HHM) were used for speech emotion classification (Lin & Wei, 2005; Schuller et al., 2003). In recent years, Deep Learning methods, like LSTM Recurrent Networks and Deep Convoluted Neutral Networks (DCNN), have gained popularity and have given better performance for speech emotion classification (Abbaschian et al., 2021; Wani et al., 2021).

This paper represent a milestone in our overall goal of using deep learning for child abuse detection. We believe that detecting negative emotions from children's speech, which is the goal of this paper, is a crucial step in using deep learning for child abuse detection. As such, one of the key steps in our process is linking certain emotions to a child's actions. For instance, child abuse in certain cultures, backgrounds, or family situations causes long-lasting negative impacts on a child's physical or mental state as they grow older (Chu et al., 2013). For some victims, bullying and abuse has been linked to inhibited emotional temperaments such as such as being quite, restrained, anxious, depressed or fearful (Gladstone et al., 2006). Therefore, the widespread and potentially severe nature of these issues led us to expand upon research already conducted around this topic. Overall, our aim is to use deep learning models to recognize negative emotions in children.

The remainder of the paper is organized as follows. Section 2 discusses the extant literature on SER. Section 3 presents the methodology and explains the architecture of our network. Section 4 describes the experimental setups and section 5 discusses the results from our experiments. Section 6 concludes the paper and suggests future research directions.

II. RELATED STUDIES

In recent times, the most common approach for experiments in speech emotion recognition is through digital audio analysis using techniques such as machine learning. At extremely young ages, "Uncomfortable", "Hungry", "Pain", "Fear" and "Angry" are the basic emotions and states that can be automatically recognized by using the vocalizations of children (Nirmani 2019). There is a level of understanding that can happen between adults and toddlers without needing any further analysis. However, in experimental research, the next step taken is to analyze the audio features using different categories and levels. High level features are perceptible by humans and include features such as pitch, loudness and energy and low level features are extracted from the audio file and include features such as cepstral descriptors and spectral descriptors (Jiang & Jin, 2022). Qualitative research indicates that angry speech features has a slightly faster speaking rate and wide pitch, while sadness features a narrower pitch and slower speaking rate (Khalil et al., 2019). Furthermore, digital programs can extract thousands of these features from one audio sample, including signal energy, loudness, semitone (Eyben et al., 2010; Moffat et al., 2015; Sharma et al., 2020), and Mel-Frequency Cepstrum Coefficient or MFCC (Prabakaran and Sriuppili, 2021; Dolka et al., 2021; Patnaik, 2023).

Deep neural networks are based on feed-forward structures comprised of underlying hidden layers between inputs and outputs, which boost the ability to link an audio sample to its respective emotion (Khalil et al., 2019). From hearing samples of infants crying, researchers used Long- and

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Volume 11 Issue 01 January 2024

Short-Term Memory neural networks, or LSTM, to classify them into five emotion classes (Jian et al., 2021) since LSTM-based recurrent neural networks are able to handle variable data (Khalil et al., 2019) better than feed-forward architectures. However, for studies attempting to create deep learning models for datasets of older children's audio samples, there could be countless more variables involved and a lower chance of obtaining significant accuracy rates without carefully considering each feature's importance within the classification phase.

Liu et al., 2018 developed a brain emotion learning model (BEL) inspired by the brain's limbic system, which is important for emotional development. The primary method to train the model according to the human brain is extracting MFCC data from each audio file. Unlike traditional frequency, MFCC accounts for how humans process what they hear much more realistically. The other most important step for the BEL model is using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for dimension reduction. PCA is unsupervised while LDA is supervised, so using a combination of both would theoretically improve discrimination between categories of audio features. With the BEL model, Liu et al. explored its effects on numerous audio datasets, including the FAU-Aibo corpus. However, they built their model to classify five different emotions from audio in this dataset: Angry, Emphatic, Neutral, Positive, and Rest. In addition, the results vastly differed when separated between speaker-dependent metrics and speaker-independent metrics. When the model was only trained to recognize one individual child speaker from the FAU-Aibo dataset, the model worked better than when all speakers were included. Another drawback with this method is the loss of information due to dimensionality reduction as important details may be discarded in the process.

III. METHODOLOGY

The FAU-Aibo dataset contains 18,216 audio files of 51 German middle-school children aged 10-13 years conversing with Sony's pet robotic dog named Aibo. The children were giving some commands to the robotic dog to follow and the dog was programmed to obey the commands 50% of the time. The reaction of the children to the dog obeying or disobeying them was recorded via a close-talk microphone and stored in small, syntactically meaningful 'chunks' as audio files. The reactions were categorized by 5 human labelers into 11 emotions. Researchers have combined some of these initial 11 emotional labels into larger, broader categories, for example, placing "touchy" and "reprimanding" labels together with "angry" to create a 4-label dataset of "Neutral," "Emphatic," "Angry," and "Motherese," and a 5-label dataset with "Neutral," "Emphatic," "Angry," and "Motherese," as well as "Rest." There is a 2-label dataset with only "Idle" and "Negative," combining all the negative emotions into one large subgroup and all the non-negative emotions into Idle. For each audio chunk, which is between 1-5 seconds in length, we utilized the openSMILE feature extraction toolkit (Eyben et al., 2010) to gather 6,600 features in the ComParE 2016 collection and 86 features in the eGeMAPS collection (Eyben et al., 2015), specifically the functionals. Dimensional reduction was required to reduce the computational power requirement and prevent overfitting. We incorporated all the extracted features into a random forest and decision tree algorithm, to calculate the relative importance and predictive power of each feature. For this step, we opted for a tree-based method instead of the popular linear discriminant analysis (LDA) because the data did not follow a normal distribution, and finding the most important variables was much simpler and accurate as LDA assumed normal distribution with linear feature distribution but tree-based methods consider non-linear feature distribution . With the decision tree method, we are able to accurately note the most important features. This process was tested separately for the ComParE 2016 and eGeMAPS collections, ensuring that our combined important features list contained data from both. For this process, we utilized the 4-label, 5-label, and 2-label datasets in separate random forests, looking for patterns to suggest which features were truly the most important in child emotion recognition. We converted all emotion labels to numbers to prepare the data for experiments. For instance, for the 2-class tests, we converted idle to 0 and negative to 1. We developed a unique gated recurrent unit (GRU) neural network that yields one of the strongest speaker-independent emotion classifications when used with the FAU-Aibo corpus. The architecture of our model is shown in figure 1.

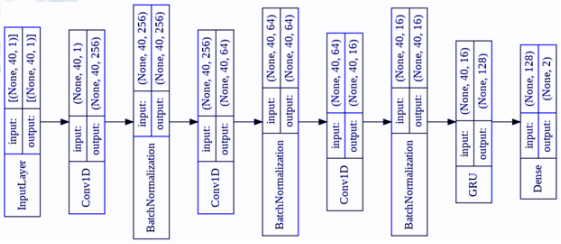

Fig. 1. GRU Model Diagram

```
Model: "model_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_2 (InputLayer)         [(None, 40, 1)]           0

conv1d_3 (Conv1D)            (None, 40, 256)           768

batch_normalization_3 (Batc  (None, 40, 256)           1024
hNormalization)

conv1d_4 (Conv1D)            (None, 40, 64)            32832

batch_normalization_4 (Batc  (None, 40, 64)            256
hNormalization)

conv1d_5 (Conv1D)            (None, 40, 16)            2064

batch_normalization_5 (Batc  (None, 40, 16)            64
hNormalization)

gru_1 (GRU)                  (None, 128)               56064

dense_1 (Dense)              (None, 2)                 258

=================================================================
Total params: 93,330
Trainable params: 92,658
Non-trainable params: 672
_________________________________________________________________
```

Fig. 2. GRU Model Summary

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Volume 11 Issue 01 January 2024

Compared to other networks, our GRU trained with the data much faster and yielded more accurate results. One of the strengths of our GRU model is that it uses fewer training parameters which use less memory, and provides great benefits for smaller datasets and low power devices. The internal processes within our GRU are more efficient as well, with a reset gate to determine how much past information is needed to neglect. GRU models in general are better suited to take data in chunks, performing well when dealing with audio files short in length.

IV. EXPERIMENTS

Of the 18,216 files total, 5,283 are labelled negative (1) and 12,933 are idle (0). We ran the early iterations of our models with many more idles than negatives. The class imbalance affected the performance of the model. We used the data-level class imbalance adjustment method in Leevy et al., 2018 to adjust for the imbalance, randomly sampling from the idles such that the entire new dataset had 5,823 files for each class .

Similarly, when testing with the 4-class and 5-class labels, we resampled the data such that there was an equal amount of each label present. For each audio chunk, the model outputs a number corresponding to the emotion class it best belongs in. For example, for the 2-class models, the outputs fall between 0 and 1, then mapped to either “Idle” or “Negative.” By default, values greater than or equal to 0.5 are predicted negatives, and less than 0.5 are predicted idles; the overall accuracy in our classification report gives the percentage of correct predictions. For our model, changing the threshold to a value lower than 0.5 vastly improved the recall, or the rate of predicting “Negative” correctly. For the purposes of our research, we find it more important to correctly predict “Negative” even if the rate of predicting “Idle” becomes slightly less precise. This procedure vastly decreases the Type-II error and ensures that wrongly classified “Negative” chunks are as few as possible.

Many of the experiments described in the literature review, including one from Zhao et al. (2019), used the CNN + LSTM deep learning models for speech recognition. We built several models that researchers use often to compare with the performance of our model. We tested CNN + LSTM models. We also tested our model against a Support Vector Machine (SVM), Boosted Decision Trees, and Logistic Regression models. The proposed GRU model outperformed these other methods and we discuss the results in the next section. For all our experiments, we used a system with an Nvidia Tesla K80 64GB GPU, 2vCPU@2.2GHz, 12.6GB RAM, and 33 GB disk space.

V. RESULTS

We present the results of our experiments in this section. First, we will present our results running our GRU model on the 2-class, 4-class, and 5-class on the Aibo dataset, followed by LSTM, SVC, and boosted trees on the 2-class dataset.

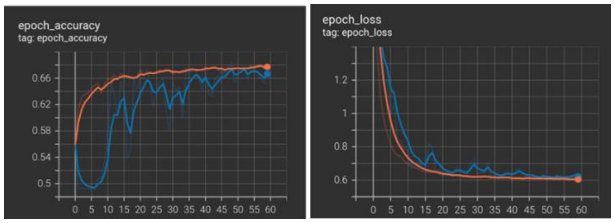

Fig. 3. Accuracy and Loss Graphs for 2-class GRU model

Fig. 3 shows the accuracy and loss plot for the 2-class GRU model. The orange plot is the train data and the blue plot is the test data. The GRU loss function is stable with less fluctuation for the training data than the testing data. The threshold for negative is X_test > 0.3, increasing the recall for predicting negative emotion. The model achieved a smoothed accuracy rate of 66.65% and a a raw accuracy rate of 67.58% on the test set. The smoothed loss value is at 0.6216 and the raw loss value is 0.6113 on the test set.

Table I: Precision-Recall-F1score For 2-Class Gru Model

	precision	recall	f1-score	support
IDLE	0.60	0.90	0.72	1478
Negative	0.53	0.98	0.69	1434
micro avg	0.56	0.94	0.70	2912
macro avg	0.56	0.94	0.70	2912
weighted avg	0.57	0.94	0.70	2912
samples avg	0.60	0.94	0.71	2912

Fig. 4 and Table I shows the result for the confusion matrix and the precision-recall-fscore. The precision for the "IDLE" class is 0.6, which means that 60% of the instances predicted as "IDLE" were actually true positives, while for the "Negative" class it is 0.53, indicating that 53% of the instances predicted as "Negative" were actually true negatives.

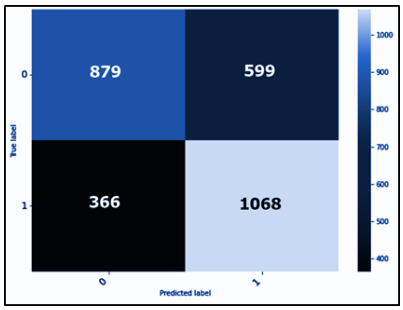

Fig. 4. Confusion matrix for the 2-class GRU model

The recall for the "IDLE" class is 0.9, indicating that 90% of the actual "IDLE" instances were correctly identified, while for the "Negative" class it is 0.98, indicating that 98% of the actual "Negative" instances were correctly identified. In our model, the F1-score for the "IDLE" class is 0.72, while for the "Negative" class it is 0.69. For the confusion matrix, 0 =“Idle” and 1 = “Negative.” 879 instances of the Idle emotions were correctly recognized while 1068 of the

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Volume 11 Issue 01 January 2024

negative emotions where classified correctly. The GRU model maintains the lowest loss score when compared with the LSTM, Boosted Trees and Support Vector Classifier (SVC). The rest of the results are shown in table II.

Table II: 2-Class Model Comparison

Model Type	Test Set Size	Smoothed Acc. (%)	Raw Acc. (%)	Smoothed Loss	Raw Loss
GRU	2912	66.65	67.58	0.6216	0.6113
LSTM	2912	62.52	64.32	0.6719	0.6490
Boosted Trees	2912	-	68.13	-	0.6960
SVC	2912	-	69.71	-	0.8035

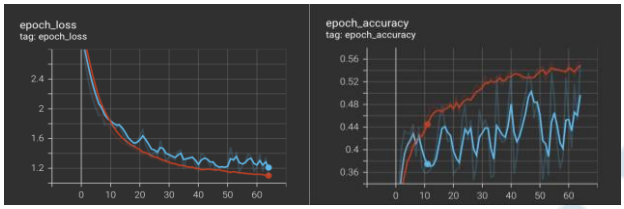

Fig. 5. 4-class GRU model loss and accuracy graphs: Red = Training and Blue = Validation.

On the four emotional class experiments, our GRU model results in a testing loss of 1.32. There is higher fluctuation among the test sets with an overall model accuracy of 55.05%.

Table III: Precision-Recall-F1score For 4-Class Gru Model

	precision	recall	f1-score	support
Neutral	0.41	0.33	0.37	140
Emphatic	0.53	0.53	0.53	152
Angry	0.54	0.61	0.58	144
Motherese	0.68	0.72	0.70	150
macro avg	0.54	0.55	0.54	586
weighted avg	0.54	0.55	0.55	586

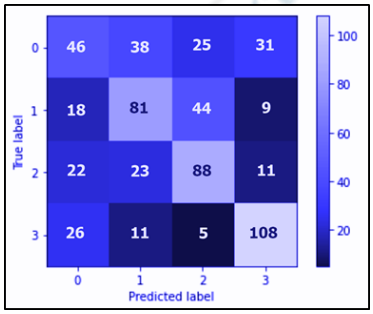

Fig. 6. Confusion matrix for the 4-class GRU model

Figure 6 and Table III shows the confusion matrix and the precision-recall-f1score results for the 4-class GRU model. Overall, our GRU model did well in predicting 'angry' emotions and was able to distinguish it from motherese emotion and neutral emotion. However, it did struggle with separating angry emotion and emphatic emotion and classified several instances of emphatic as angry. The Emphatic and Angry classes also have reasonable precision, recall, and f1-score values, while the Neutral class has relatively low values for these metrics. The macro avg and weighted avg values provide an overall view of the model's performance across all classes, with the weighted avg taking into account the imbalance of class sizes in the dataset. The GRU model show the greatest accuracy in identifying Angry and Motherese according to F1-score, while also being notably precise for Angry speech. The model also predicted the "Emphatic," emotions with a decent level of accuracy.

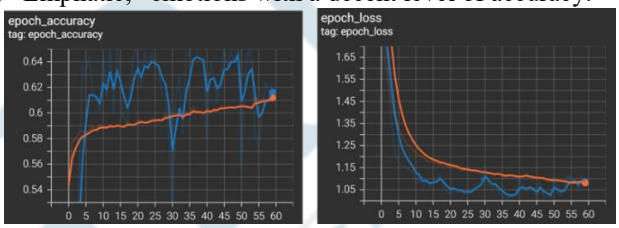

Fig. 7. Accuracy and Loss plots of the 5-class GRU model

Figure 7 shows the accuracy and loss plots for the GRU model using the five emotional class dataset. The orange plot is for the training and the blue plot is for the test performance. The first 5-class GRU model shows a loss value reaching 1.083. The smoothed accuracy fluctuates but exhibits a general increase with more epochs, reaching 61.62% while the raw accuracy stands at 62.44%.

The 5-class model performed similarly to the 4-class, and certain emotion classes were predicted much more frequently than others. The reports show a 64 percent recall score for "Emphatic" and 67 percent recall for "Motherese," while also severely lowering recalls of other classes.

Table IV: Precision-Recall-F1score For 5-Class Gru Model

	precision	recall	f1-score	support
Neutral	0.29	0.23	0.26	225
Emphatic	0.38	0.64	0.48	213
Angry	0.58	0.33	0.42	230
Rest	0.34	0.13	0.18	214
Motherese	0.42	0.67	0.51	230
macro avg	0.40	0.40	0.37	1112
weighted avg	0.40	0.40	0.37	1112

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Volume 11 Issue 01 January 2024

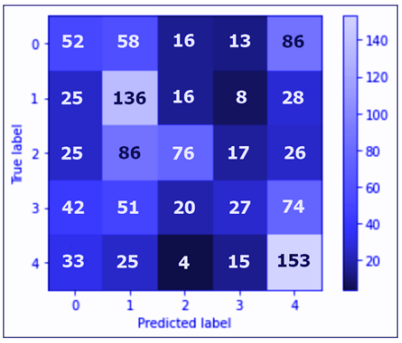

Fig. 8. Confusion matrix for the 5-class GRU model

However, it does show more promise in specifically being able to predict "Angry" emotions correctly, with a 58 percent precision score. The model does have difficulty in predicting "Rest," perhaps because it lacks the strength of "Angry" or "Emphatic," which are more easily identified both with machine learning and with human knowledge. Many "Rest" emotional samples were incorrectly predicted as "Motherese" for example, so it is also an emotion that shares many qualities with others and thus, is difficult to predict on its own. The 5-class model also utilized a smaller sample than the 2-class to ensure that each category was equally represented. While the 4-class and 5-class datasets may better test the classification power of our GRU model on the Aibo dataset, restricting the data to two classes greatly improves overall accuracy while simplifying the classification process.

Table V benchmarks the results of our 5-class models with other research works that used the FAU Aibo dataset for emotion recognition.

Table V: 5-class model benchmark

Paper Title	Model Type	Avg. Accuracy(%)
Zhao et al., 2019	BLSTM-CTC	43.0
Ibrahim et al., 2022	Bidirectional ESN	46.0
Thirumuru et al., 2022	SVM	59.9
Deb & Dandapat, 2019	Extreme Learning Machines	53.4
Attabi & Dumouchel, 2013	GMM + Euclidean	47.44
Proposed Model	**GRU**	**62.44**

VI. CONCLUSION

In our experiments, we tested the capability of multiple types of deep learning and machine learning models in speech emotion detection. Our GRU model with approximately 90,000+ trainable parameters yielded better results and runs 29.29% faster than the LSTM model, making it a more time-saving and memory-efficient model. The GRU model for the 2-class speaker-independent FAU-Aibo dataset classifies between idle and negative emotions with an accuracy of 67.58% and recall at nearly 1.0 for the negative emotional class. The 4-class and the 5-class models also performed significantly well with the 5-class model achieving an accuracy of 65% on the Aibo dataset. We must note that the FAU-Aibo dataset has a distinct characteristics that are challenging for machine prediction, compared to others widely used in the field. Primarily, all the reactions and conversations between the children and the Aibo robot were organic and naturally generated, unlike other datasets which utilize professional actors imitating these emotions, making it ideal for studying natural emotional responses and recognition but challenging at the same time because the emotional responses are sometimes not very distinct. In addition, Children's emotional expressions can be subtle and nuanced, making it challenging to accurately classify them.

While our unique model removes certain specificity for pure emotion classification power, we built it to specifically maximize the accuracy to detect negative emotions. The ultimate goal of our research is to locate as many negatives as possible which is a milestone step in taking necessary action to help children who are feeling stressed or uncomfortable. Negative emotional feeling among children should be detected quickly and the cause should be traced and resolved. The unique GRU model, at its core, helps in this detection phase.

Future research would benefit from expanding the dataset used for speech emotion recognition. This could involve collecting a larger and more diverse dataset that includes a wider range of emotions, exploring more fine-grained emotion recognition and distinguishing between specific emotions like anger, sadness, fear, etc. This would provide more detailed insights into children's emotional states and potentially improve the detection of child abuse signs. Future research could also incorporate other modalities such as facial expressions, and text recognition to improve the accuracy and robustness of child speech emotion recognition.

REFERENCES

[1] Eyben, Florian, Martin Wöllmer, and Björn Schuller. "Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor." In Proceedings of the International Conference on Multimedia - MM '10, 1459. Firenze, Italy: ACM Press, 2010. https://doi.org/10.1145/1873951.1874246.

[2] Jian, Tianye, Yizhun Peng, Wanlong Peng, and Zhou Yang. "Research on LSTM+Attention Model of Infant Cry Classification." Journal of Robotics, Networking and Artificial Life 8, no. 3 (October 9, 2021): 218–23. https://doi.org/10.2991/jrnal.k.210922.013.

[3] Khalil, Ruhul Amin, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. "Speech Emotion Recognition Using Deep Learning Techniques: A Review." IEEE Access 7 (2019): 117327–45. https://doi.org/10.1109/ACCESS.2019.2936124.

[4] Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep Learning Techniques for Speech Emotion Recognition,

from Databases to Models. Sensors, 21(4), Article 4. https://doi.org/10.3390/s21041249

[5] Attabi, Y., & Dumouchel, P. (2013). Anchor Models for Emotion Recognition from Speech. IEEE Transactions on Affective Computing, 4(3), 280–290. https://doi.org/10.1109/T-AFFC.2013.17

[6] Chu, D. A., Williams, L. M., Harris, A. W. F., Bryant, R. A., & Gatt, J. M. (2013). Early life trauma predicts self-reported levels of depressive and anxiety symptoms in nonclinical community adults: Relative contributions of early life stressor types and adult trauma exposure. Journal of Psychiatric Research, 47(1), 23–32. https://doi.org/10.1016/j.jpsychires.2012.08.006

[7] Deb, S., & Dandapat, S. (2019). Emotion Classification Using Segmentation of Vowel-Like and Non-Vowel-Like Regions. IEEE Transactions on Affective Computing, 10(3), 360–373. https://doi.org/10.1109/TAFFC.2017.2730187

[8] Dolka, H., M, A. X. V., & Juliet, S. (2021). Speech Emotion Recognition Using ANN on MFCC Features. 2021 3rd International Conference on Signal Processing and Communication (ICPSC), 431–435. https://doi.org/10.1109/ICSPC51351.2021.9451810

[9] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., & Narayanan, S. S. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190–202.

[10] Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The munich versatile and fast open-source audio feature extractor. Proceedings of the 18th ACM International Conference on Multimedia, 1459–1462.

[11] Fan, W., Xu, X., Xing, X., Chen, W., & Huang, D. (2021). LSSED: A Large-Scale Dataset and Benchmark for Speech Emotion Recognition. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 641–645. https://doi.org/10.1109/ICASSP39728.2021.9414542

[12] Gladstone, G. L., Parker, G. B., & Malhi, G. S. (2006). Do Bullied Children Become Anxious and Depressed Adults?: A Cross-Sectional Investigation of the Correlates of Bullying and Anxious Depression. The Journal of Nervous and Mental Disease, 194(3), 201. https://doi.org/10.1097/01.nmd.0000202491.99719.c3

[13] Jiang, Y., & Jin, X. (2022). Using k-Means Clustering to Classify Protest Songs Based on Conceptual and Descriptive Audio Features. In M. Rauterberg (Ed.), Culture and Computing (pp. 291–304). Springer International Publishing. https://doi.org/10.1007/978-3-031-05434-1_19

[14] Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. Journal of Big Data, 5(1), 42. https://doi.org/10.1186/s40537-018-0151-6

[15] Lin, Y.-L., & Wei, G. (2005). Speech emotion recognition based on HMM and SVM. 2005 International Conference on Machine Learning and Cybernetics, 8, 4898-4901 Vol. 8. https://doi.org/10.1109/ICMLC.2005.1527805

[16] Liu, Z.-T., Xie, Q., Wu, M., Cao, W.-H., Mei, Y., & Mao, J.-W. (2018). Speech emotion recognition based on an improved brain emotion learning model. Neurocomputing, 309, 145–156. https://doi.org/10.1016/j.neucom.2018.05.005

[17] Moffat, D., Ronan, D., & Reiss, J. D. (2015). An Evaluation of Audio Feature Extraction Toolboxes.

[18] Mustaqeem, & Kwon, S. (2020). A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. Sensors, 20(1), Article 1. https://doi.org/10.3390/s20010183

[19] Mustaqeem, & Kwon, S. (2021). Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network. International Journal of Intelligent Systems, 36(9), 5116–5135. https://doi.org/10.1002/int.22505

[20] Patnaik, S. (2023). Speech emotion recognition by using complex MFCC and deep sequential model. Multimedia Tools and Applications, 82(8), 11897–11922. https://doi.org/10.1007/s11042-022-13725-y

[21] Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model-based speech emotion recognition. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., 2, II–1. https://doi.org/10.1109/ICASSP.2003.1202279

[22] Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. Applied Acoustics, 158, 107020. https://doi.org/10.1016/j.apacoust.2019.107020

[23] Thirumuru, R., Gurugubelli, K., & Vuppala, A. K. (2022). Novel feature representation using single frequency filtering and nonlinear energy operator for speech emotion recognition. Digital Signal Processing, 120, 103293. https://doi.org/10.1016/j.dsp.2021.103293

[24] Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., & Ambikairajah, E. (2021). A Comprehensive Review of Speech Emotion Recognition Systems. IEEE Access, 9, 47795–47814. https://doi.org/10.1109/ACCESS.2021.3068045

[25] Zhao, Z., Bao, Z., Zhang, Z., Cummins, N., Wang, H., & Schuller, B. W. (2019). Attention-Enhanced Connectionist Temporal Classification for Discrete Speech Emotion Recognition. Interspeech 2019, 206–210. https://doi.org/10.21437/Interspeech.2019-1649

[26] Nirmani, Shashiwadana. Child Emotion and State Recognition by Voice, 2019. https://doi.org/10.13140/RG.2.2.29139.04648.

[27] Prabakaran, D., and S. Sriuppili. "Speech Processing: MFCC Based Feature Extraction Techniques- An Investigation." Journal of Physics: Conference Series 1717, no. 1 (January 2021): 012009. https://doi.org/10.1088/1742-6596/1717/1/012009.