NadERA: A Novel Framework Achieving Reduced Distress Response Time By Leveraging Emotion Recognition From Audio

Harshil Sanghvi, Sachi Chaudhary, and Sapan H Mankad

CSE Department, Institute of Technology, Nirma University, Ahmedabad, India
{19bce238,19bce230,sapanmankad}@nirmauni.ac.in

Abstract. This paper proposes a novel framework for automatically directing the user to an appropriate helpline number in an emergency based on his/her emotional state. For emotion detection, we integrate four benchmark datasets (SAVEE, RAVDESS, TESS, and Crema-D). We further examine the impact of various features on this comprehensive dataset and see the possibility of generalization with the help of diversified data. The highest accuracy achieved by the Convolutional Neural Network (CNN) model is 93.14% using the proposed approach. The results indicate that our emotion recognition model highly depends on the choice of audio features. Finally, we use this prediction to build our single-emergency number helpline architecture which predicts the caller's emotions and directly connects them to the desired person for seeking mental help through counselors, protection with the help of police, or a general call center for any other help. This framework reduces the response time and provides a single point of connection.

Keywords: audio, MFCC, CNN, RAVDESS, SAVEE, TESS, Crema-D, emotion detection

1 Introduction

Detecting a person's emotion is important for several reasons like health applications, increasing business reach, user feedback, and making human-computer interaction much more efficient. The response time is crucial for emergency hotline services like 911 in the United States and 112 in India, where the goal is to deliver time-critical assistance to its callers. According to a study by the US Department of Homeland Security (USA), the typical shooter event at a school lasts 12.5 minutes. In comparison, the average response time for law enforcement is 18 minutes[1]. The statistics reveal that officers arrive after the crime has been perpetrated. According to 2022 figures, it takes an average of 35 minutes for a law enforcement officer to get on the scene after dialing 911 [2]. In the event of a lower-priority call, this period is extended even more. In terms of medical

[1] https://www.creditdonkey.com/average-police-response-time.html

emergencies, the average delay from the moment of a 911 call to arriving on site is 7 minutes. In remote areas, the median duration is more than 14 minutes [13].

Communication between the caller and the callee is crucial in the increased response time in such cases. During an emergency, callers are usually under a lot of stress, worry, and strain, which makes it challenging to communicate clearly and quickly with the callee. Aside from that, the fact that many nations have different helpline numbers depending on their emergency type makes this issue much more difficult since the caller is expected to remember the number suited for the scenario and then ask for assistance. Once the correct number is dialed, the help-seeker impacted by emotions and, as a consequence, is weeping, suffering shortness of breath, and so on, must ensure that all of his information is accurately logged at the emergency services desk so that he receives the fastest possible assistance. However, since there is a step of communication between the caller and the callee, human interaction always results in a longer response time and allows room for human error.

This research seeks to solve these concerns by presenting a unique framework that uses the intelligence of neural networks. We propose a framework with a single emergency number in which we detect the correct emotion from the caller's side and direct the person to the relevant department. Furthermore, we pick up on two emotions whose scenario reaction time is considerably decreased by automating the complete communication process for callers. The following are the major contributions of this work:

- A detailed literature review of the emotion recognition systems from the speech is presented.
- We then showcase the process of data pre-processing, where we take four audio datasets and concatenate them to achieve a quality dataset on speech emotion classification, which provides a high-end feature engineering model.
- A framework is proposed to provide a uni-call system using the speaker's emotions.

The rest of the paper is organized as follows. Section 2 provides a literature review of the task at hand. The framework for reducing distress response time is proposed in Section 3. Section 4 describes the data preparation and feature representation methodologies. We show experimental details in Section 5 and discuss the results in Section 6. Finally, the paper ends with concluding remarks and future directions in Section 7.

2 Related Work

Deep Learning has emerged as a versatile tool providing tremendous power and capabilities in diverse areas, from Computer Vision and natural language processing. Several deep learning models have been used in the field of medicine for leukemia detection [4, 6, 5]. Emotion recognition from speech/audio is a booming field in the current period due to advancements in technology, artificial intelligence, etc. Kanwal *et al.* [10] presented a clustering-enabled genetic algorithm to select the best-fitting features from three datasets: SAVEE, EMO-DB, and RAVDESS. The support vector machine algorithm was further used to classify the audio emotion and showed effective results.

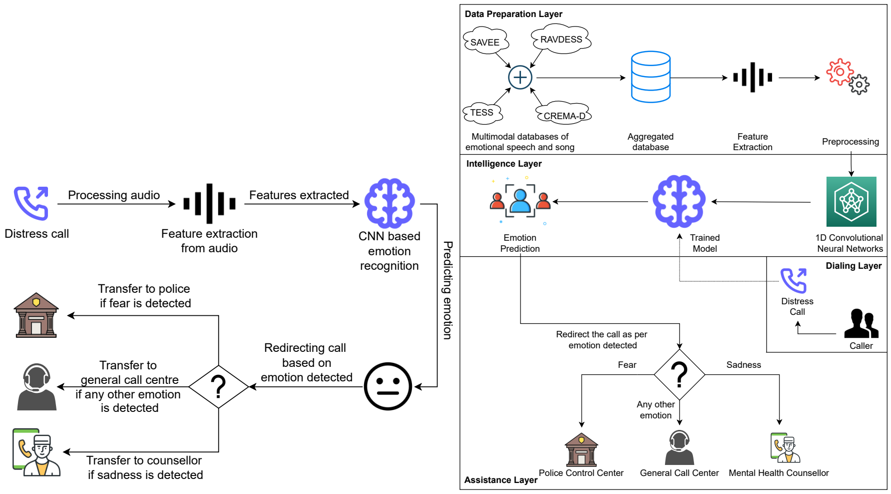
A deep feature-based layered approach outperformed conventional machine learning algorithms, as presented in [15]. Further, Mehmet [8] proposed a novel hybrid technique based on deep and acoustic features (MFCC, zero-crossing rate, and root mean square energy values) to increase the accuracy. The emotion recognition systems lack high-quality input audio data and noisy environments. To address this issue, Mingke *et al.* [17] proposed a head-fusion framework to improve the robustness of the speech recognition system and accuracy. This model was built using the RAVDESS and IEMOCAP datasets. A transfer learning and autoencoder-based 1-dimensional deep convolutional neural network-based framework for voice-based emotion recognition were presented by [14], and they obtained 96% accuracy on the TESS dataset.

This paper proposes a neural network-based emotion recognition system from audio. Based on the detected emotion, the user is redirected to the respective channel, and help is sent immediately, whether it be police requirement, depression helpline, or some other help. Four datasets are merged to bring diversity in data.

3 Proposed Framework

This section presents the system model for the proposed work to accurately detect the emotion from the caller's voice and redirect the call to the appropriate channel. Fig. 1a shows the system model for the proposed work. The person in distress calls on a single unified helpline number. The audio sample of the caller is processed, and features are extracted from the audio. These features are fed to the CNN-based classifier, which predicts the emotion. Suppose the detected emotion is *fear.* In that case, the call is automatically redirected to the police control center; else, it is redirected toward the mental health counselor if the emotion is *sadness.* Now in the scenario of any other emotion being detected, the call is redirected to a general call center. This way, the response time is drastically reduced as the caller need not remember different helpline numbers for different problems; instead, they have to call on a single number for all their problems.

As depicted in Fig. 1b, the proposed system architecture consists of four layers: the data preparation layer, dialing layer, intelligence layer, and assistance layer. The data preparation layer consists of a database formed by aggregating multimodal emotional speech databases. Then the data pre-processed in the data preparation layer is forwarded to the intelligence layer, which trains the neural network architecture on the dataset and keeps it ready for predictions on the unseen data. In the intelligence layer, the CNN model receives audio from the distressed caller and predicts the emotion in the call. The call is automatically redirected to the appropriate channel based on the predicted emotion. The following is a detailed description of the four-layered architecture.


Distress call
Processing audio
Feature extraction from audio
Features extracted
CNN based emotion recognition
Predicting emotion
Redirecting call based on emotion detected
?
Transfer to police if fear is detected
Transfer to general call centre if any other emotion is detected
Transfer to counsellor if sadness is detected
Data Preparation Layer
SAVEE
RAVDESS
TESS
CREMA-D
Multimodal databases of emotional speech and song
Aggregated database
Feature Extraction
Preprocessing
Intelligence Layer
Emotion Prediction
Trained Model
1D Convolutional Neural Networks
Dialing Layer
Distress Call
Caller
Redirect the call as per emotion detected
Fear
Sadness
Any other emotion
Police Control Center
General Call Center
Mental Health Counsellor
Assistance Layer

(a) System Model (b) Comprehensive System Architecture

Fig. 1. Proposed Framework

Data Preparation Layer First, we create a dataset in the data preparation layer by aggregating multimodal emotional speech and song databases. The four datasets that were combined are SAVEE [16], RAVDESS [11], TESS [7] and CREMA-D [3]. After the concatenation of data, several augmentation techniques, such as adding random noise, stretching, and shifting, were used to increase the number of samples. First, random noise with a rate of 0.035 was added to the original audio sample. The second technique involved stretching the audio data by a rate of 0.8. The third technique incorporated was shifting the data, wherein the first shift range was generated by taking a rate equal to 1000 and rolling the data with generated shift range. The last technique involved shifting the pitch of the given audio file by a factor of 0.7.

Once the augmented dataset was ready, feature extraction was carried out. All the audio files were loaded for a duration of 2.5s, with an offset of 0.6. The sampling rate was 22kHz. Once the feature set was ready, some pre-processing techniques were employed to prepare data for training. First, the target labels were encoded with integers. Secondly, data was split into train, test, and validation with a split ratio of 72:20:8. To scale down the data points to a minimal range, standard scaling was performed on the input values of the dataset. Once the values were scaled down to the range of 0 and 1, the feature set was ready for training the model.

Dialing Layer This layer comprises of helpline number powered by the automated recognition of emotion from audio using deep learning. Here the caller-in-distress calls the helpline number, and the captured audio is transferred to the intelligence layer. Once the audio passes through the intelligence layer, the intelligence layer makes an accurate prediction of the emotion. The successive tiers of the framework address the pipeline's flow.

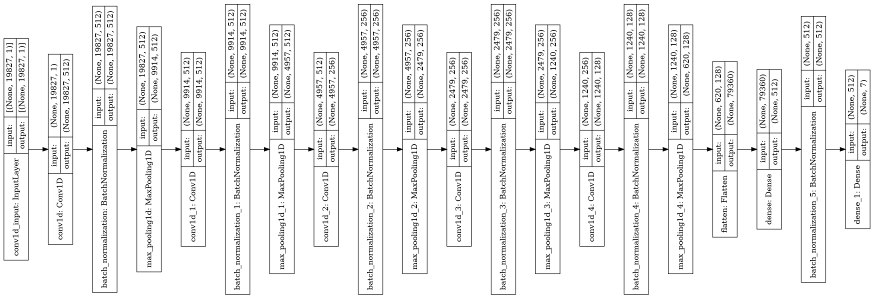

conv1d_input: InputLayer
input: [(None, 19827, 1)]
output: [(None, 19827, 1)]
conv1d: Conv1D
input: (None, 19827, 1)
output: (None, 19827, 512)
batch_normalization: BatchNormalization
input: (None, 19827, 512)
output: (None, 19827, 512)
max_pooling1d: MaxPooling1D
input: (None, 19827, 512)
output: (None, 9914, 512)
conv1d_1: Conv1D
input: (None, 9914, 512)
output: (None, 9914, 512)
batch_normalization_1: BatchNormalization
input: (None, 9914, 512)
output: (None, 9914, 512)
max_pooling1d_1: MaxPooling1D
input: (None, 9914, 512)
output: (None, 4957, 512)
conv1d_2: Conv1D
input: (None, 4957, 512)
output: (None, 4957, 256)
batch_normalization_2: BatchNormalization
input: (None, 4957, 256)
output: (None, 4957, 256)
max_pooling1d_2: MaxPooling1D
input: (None, 4957, 256)
output: (None, 2479, 256)
conv1d_3: Conv1D
input: (None, 2479, 256)
output: (None, 2479, 256)
batch_normalization_3: BatchNormalization
input: (None, 2479, 256)
output: (None, 2479, 256)
max_pooling1d_3: MaxPooling1D
input: (None, 2479, 256)
output: (None, 1240, 256)
conv1d_4: Conv1D
input: (None, 1240, 256)
output: (None, 1240, 128)
batch_normalization_4: BatchNormalization
input: (None, 1240, 128)
output: (None, 1240, 128)
max_pooling1d_4: MaxPooling1D
input: (None, 1240, 128)
output: (None, 620, 128)
flatten: Flatten
input: (None, 620, 128)
output: (None, 79360)
dense: Dense
input: (None, 79360)
output: (None, 512)
batch_normalization_5: BatchNormalization
input: (None, 512)
output: (None, 512)
dense_1: Dense
input: (None, 512)
output: (None, 7)

Fig. 2. 1D CNN Model Architecture

Intelligence Layer 1D CNN model [9] has been used for training the audio files. We conducted several experiments using different feature selection techniques, a different number of emotions, overlapping of audio frames, and averaging of audio frames. Fig. 2 depicts 1D CNN model architecture for the proposed system. The proposed system model shown in Fig. 1a uses this CNN-based emotion recognition for predicting emotions and further redirecting the calls based on emotions detected by the system.

Assistance Layer In this layer, the trained neural network architecture predicts the emotion from the caller's audio and takes action, as mentioned earlier in the framework. In this manner, the proposed framework eliminates the need for remembering multiple helpline numbers and automates the entire redirection by leveraging deep learning.

4 Methodology

4.1 Dataset Selection

We have used four datasets. RAVDESS dataset is a multi-actor dataset of different emotion audios. The dataset comprises 24 actors, each having emotions of sadness, happiness, fear, calm, neutral, surprise, and disgust. This dataset is rich in emotions with different ranges without gender bias. SAVEE dataset consists of four male actors showing seven different emotions, as in the RAVDESS dataset. The TESS dataset consists of seven emotions, the same as RAVDESS, expressed by two women. The SAVEE dataset is a multi-speaker database comprising audio-video data and text descriptions. The dataset expresses emotions such as happiness, sadness, anger, surprise, fear, disgust, and neutrality. The CREMA-D dataset comprises audio files from 91 actors having varied ages, gender, race, and backgrounds. The emotions it focuses on are sadness, happiness,

Table 1. Features used during this work

Feature	Dimensions	Feature	Dimensions
Zero Crossing Rate	108	Bark Frequency Cepstral Coefficient (BFCC)	1781
Short Term Energy	1	Gamma Tone Frequency Cepstrum Coefficient (GFCC)	1300
Entropy of Energy	1	Magnitude-based Spectral Root Cepstral Coefficients (MSRCC)	1300
RMS	108	Normalized Gammachirp Cepstral Coefficients (NGCC)	1300
Spectral Centroid	108	Linear Frequency Cepstral Coefficients (LFCC)	1300
Spectral Flux	1	Power Normalized Cepstral Coefficient (PNCC)	7150
Spectral Rolloff	108	Phase-based Spectral Root Cepstral Coefficients (PSRCC)	1300
Chroma STFT	1296	Linear Predictive Coefficients (LPC)	1365
Mel Frequency Cepstral Coefficients (MFCC)	1300		

anger, fear, disgust, and neutrality in four different speech levels. These datasets provide a varied class availability from different genders and varied data of different emotions. Hence, it helps in achieving generalization.

Various datasets used in this research have varied shortcomings. An emotion's expression depends on the speaker's accent, language, background, etc. The model trained to recognize emotion in the English language will not equally work to classify the emotion in any other language like Chinese or Indian. RAVDESS dataset suffers from selection bias as it was formed by speakers belonging to different regions, like Canada, who have a rigid American accent. Also, these datasets are created by trained speakers, and therefore emotions from natural language speakers may vary from the dataset. Because of this reason, the architecture should be validated before deployment in the real world. Another limitation is that the datasets contain only a few statements spoken with different emotions. This limits the scope of the dataset.

4.2 Data Preparation

The audio files of each dataset contain different emotions, which are appended together. Each audio recording comprises approximately 4 to 5 seconds of audio. Further, data is set to be noise-free by different techniques. Each audio file is then augmented by adding noise where the rate is set to 0.035, stretched by setting the rate as 0.8, shifted by setting the rate as 1000 for both low and high pitches, and then the pitch is set with the rate as 0.7. No class balancing was performed.

Due to space constraints, we have not included a detailed description of these features. The list of various features used in our proposed work is shown in Table

1 along with the dimension of the features taken. Interested readers may follow Librosa [1] and Spafe [12] for more details.

5 Experiments

Librosa and Spafe libraries were used to extract different features. All features were extracted with a frame length of 2048 ms and a hop length equal to 512 ms.

Table 2. Configuration for various experiments performed

Feature	**Exp-1**	**Exp-2**	**Exp-3**	**Exp-4**	**Feature**	**Exp-1**	**Exp-2**	**Exp-3**	**Exp-4**
ZCR	✓	✓	✓	✓	LFCC	✓	✓	✓	✓
Mean Energy	✓	✓	✓	✓	MFCC	✓	✓	✓	✓
Entropy of Energy	✓	✓	✓	✓	MSRCC	✓	✓	✓	✓
RMSE	✓	✓	✓	✓	NGCC	✓	✓	✓	✓
SPC	✓	✓	✓	✓	LPCC	✓	✓	✓	✓
SPC Flux	✓	✓	✓	✓	PNCC	✓	✓	✓	✓
SPC Roll-off	✓	✓	✓	✓	PSRCC	✓	✓	✓	✓
Chroma STFT	✓	✓	✓	✓	Number of Emotions	7	7	6	7
BFCC	✓	✓	✓	✓	Overlapping	✓			
GFCC	✓	✓	✓	✓	Averaging	✓	✓		
iMFCC	✓	✓			Total features size	1861	1861	19827	19827

5.1 Train, Test and Validation Criteria

Using a stratified approach, we performed the train and test data split based on the 80:20 ratio. The training data was then split into a 90:10 ratio where the first part was taken as the training dataset, and the remaining portion was used as the validation dataset. Then, this validation data consisting of audio emotions from all four datasets were used to predict the accuracy [9] and performance of the model. All the data pre-processing and cleaning steps were performed on the merged dataset before splitting into train and test datasets.

5.2 Methodology

After merging and distributing audio samples from all the datasets, these audio samples were trimmed or padded to 2.5 seconds to keep a uniform duration. This was followed by feature extraction. We obtained features from each audio sample's four (augmented) variations. Once these features were extracted, we appended them to the list and thus created a new dataset containing features for each sample and having size four times the original dataset. Once this dataset was prepared, we partitioned the dataset into the train, test, and validation sets. Then we performed standard scaling on the training data to shift the distribution to have zero mean and unit standard deviation. Once the dataset was ready, we converted target labels into one-hot encoded vectors. This data was fed into a one-dimensional CNN model, and results were obtained.

Different experiments are performed using the above features and emotions, as shown in Table 2. The number of emotions used varies in the experiments. Also, overlapping and averaging of audio frames have been done in some experiments. The respective values of coefficients from different frames are averaged and considered in the averaging configuration. Thus, in the averaging scenario, from each frame, a set of features were extracted and then averaged over an audio sample, whereas in the other case, no averaging was done. Therefore, the number of coefficients in the averaging configuration is much less than in the other case. Similarly, multiple frames were formed based on hop length during segmentation in the overlapping configuration. There was no overlapping for a hop length equal to the window size.

In a nutshell, we performed two types of settings for these experiments: (i) with all seven emotions present in the database, Experiments 1, 2, and 4 correspond to this approach; (ii) with only six emotions, by dropping the corresponding samples of *surprise* class, to avoid the class imbalanced problem. Experiment 3 showcases this approach. The rationale for these two approaches was to look into the proposed system's ability to generalize to unknown samples.

Firstly, experiment 1 used all the features and seven emotions. Overlapping and averaging audio samples were also performed, and an accuracy of 87.74% was obtained after training the dataset on the CNN model. Experiment 2 was carried out using all the features and emotions. However, the difference is that the overlapping of audio files was not carried out, only averaging was performed, and an accuracy of 86.83% was obtained. In experiment 3, all the features were used except IMFCC features, and six emotions were considered, dropping the *surprise* class. The surprise class is dropped as it causes a class imbalance problem, and overlapping along with averaging technique on audio samples is also not performed. The accuracy in this experiment was 94.92%. Experiment 4 was carried out on all seven emotion types with all features except IMFCCs; overlapping and averaging were not performed, and the accuracy obtained was 93.14%. It can be observed from the F1-score that the proposed model with Experiment 4 performs adequately well even if one class is imbalanced. Thus, it has the capability to behave well in the presence of unseen samples. Table 3 illustrate the comparison of different performance metrics, such as Loss, Accuracy, and F1 score for all experiments.

6 Results and Discussion

The one-dimensional CNN model gave an accuracy of 93.14% on the testing dataset comprising the merged dataset from four audio emotion datasets: SAVEE, RAVDESS, Crema-D, and TESS. The high performance indicates that the proposed model consisting of four differently sourced datasets is highly robust and can efficiently recognize the audio emotions of any race, gender, age, and emotion.

From the results, we can see that applying frame overlapping while extracting features improves the performance by almost 1% (from 86.83% to 87.74% accuracy). However, from experiment 4 results, it is observed that the performance drastically improves from 87.74% to 93.14% when overlapping and averaging are not done. We achieved an accuracy of 93.14% for experiment 4 using all seven emotions. Experiment 4 gives us the best performance based on these metrics using all seven emotions compared to Experiment 3, which gives 1.78% higher accuracy than Experiment 4 but only using six emotions, where the surprise class has been dropped for better class balancing. As experiment 4 involves the presence of imbalanced class *surprise*, its performance should not be judged only based on accuracy measure; thus, we also calculated the F1 score for all experiments. Surprisingly, the performance of Experiment 4 based on the F1 score is also at par with Experiment 3. So the experiment chosen for further system model is Experiment-4.

Table 3. Results obtained from various experiments performed on the merged dataset

Metric	**Exp-1**	**Exp-2**	**Exp-3**	**Exp-4**
Loss	0.6846	0.7563	0.2774	0.3973
Accuracy	87.74%	86.83%	94.92%	93.14%
F1 Score	87.81%	86.98%	94.97%	93.17%

Although the performance metrics favor Experiment 3, we recommend the system in Experiment 4, which has a higher potential to withstand and show robust performance in realistic scenarios.

7 Conclusion and Future work

This paper presents a novel framework for the automatic and seamless redirecting the distress calls to the corresponding help center through the emotional state of the caller's voice. A one-dimensional CNN model was implemented to experiment on several audio representations derived from the audio's time domain and frequency domain characteristics. We carried out different experimental scenarios to examine the impact of frame overlapping and frame averaging while extracting features for detecting emotions from audio. Results obtained from the experiments on these scenarios convey that the system exhibits better performance without averaging and overlapping.

In the future, we plan to extend our work by employing oversampling techniques and mel-spectrogram-based methods. In addition, we wish to add a feature in our framework where the caller's location coordinates will automatically be notified to the appropriate agency and use blockchain-based smart contracts to store these coordinates to enhance the safety of the proposed framework. Parallelly, we will conduct more research and analysis to map different emotions to different distress response agencies, e.g., hospitals and fire brigades, to increase the use cases of our novel framework.

References

1. McFee et al., B.: librosa/librosa: 0.9.2 (Jun 2022)
2. Briggs, R.W., Bender, W., Marin, M.: Philadelphia police response times have gotten 4 minutes longer, about 20% worse (Feb 2022)
3. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing **5**(4), 377–390 (2014)
4. Das, P.K., A, D.V., Meher, S., Panda, R., Abraham, A.: A systematic review on recent advancements in deep and machine learning based detection and classification of acute lymphoblastic leukemia. IEEE Access **10**, 81741–81763 (2022)
5. Das, P.K., Meher, S.: An efficient deep convolutional neural network based detection and classification of acute lymphoblastic leukemia. Expert Systems with Applications **183**, 115311 (2021)
6. Das, P.K., Meher, S.: Transfer learning-based automatic detection of acute lymphocytic leukemia. In: 2021 National Conference on Communications (NCC). pp. 1–6 (2021)
7. Dupuis, K., Pichora-Fuller, M.K.: Toronto emotional speech set (TESS). University of Toronto, Psychology Department (2010)
8. Er, M.B.: A novel approach for classification of speech emotions based on deep and acoustic features. IEEE Access **8**, 221640–221653 (2020)
9. Gajjar, P., Shah, P., Sanghvi, H.: E-mixup and siamese networks for musical key estimation. In: International Conference on Ubiquitous Computing and Intelligent Information Systems. pp. 343–350. Springer (2022)
10. Kanwal, S., Asghar, S.: Speech emotion recognition using clustering based ga-optimized feature set. IEEE Access **9**, 125830–125842 (2021)
11. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one **13**(5), e0196391 (2018)
12. Malek, A., Borzì, S., Nielsen, C.H.: Superkogito/spafe: v0.2.0 (Jul 2022)
13. Mell, H.K., Mumma, S.N., Hiestand, B., Carr, B.G., Holland, T., Stopyra, J.: Emergency medical services response times in rural, suburban, and urban areas. JAMA surgery **152**(10), 983–984 (2017)
14. Patel, N., Patel, S., Mankad, S.H.: Impact of autoencoder based compact representation on emotion detection from audio. Journal of Ambient Intelligence and Humanized Computing **13**(2), 867–885 (2022)
15. Suganya, S., Charles, E.Y.A.: Speech emotion recognition using deep learning on audio recordings. In: 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer). vol. 250, pp. 1–6 (2019)
16. Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G.: Combining frame and turn-level information for robust recognition of emotions within speech. pp. 2249–2252 (01 2007)
17. Xu, M., Zhang, F., Zhang, W.: Head fusion: Improving the accuracy and robustness of speech emotion recognition on the iemocap and ravdess dataset. IEEE Access **9**, 74539–74549 (2021)