Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, No. 1, June : 2023

SPEECH EMOTION RECOGNITION ON LIVE CALLS

Harika Inavolu, Assistant Professor, Sreenidhi Institute of Science and Technology (SNIST), Affiliated to JNTUH, ECE Dept., Hyderabad
Gujjula Shiva Prasad Reddy,Mangu Sai Charitha,Muthyalapati Nikhil Student, Sreenidhi Institute of Science and Technology (SNIST), Affiliated to JNTUH, ECE Dept., Hyderabad

Abstract:
Speech emotion recognition is a popular study area these days, with the goal of improving human-machine connection. At the moment, the majority of effort in this area focuses on the extraction of discriminatory traits for the goal of categorising emotions. The majority of the current effort includes the speech of words for lexical analysis and emotion recognition. In our project, we use a strategy to categorise emotions into Angry, Calm, Fearful, Happy, and Sad categories. Speech Emotion Recognition, abbreviated as SER, capitalises on the fact that tone and pitch often reflect underlying emotion. The primary goal of this thesis is to recognise emotion. To create a system that will extract, characterise, and recognise information about the speaker's emotions. There are three types of features in a speech: lexical features (the vocabulary used), visual features (the expressions made by the speaker), and acoustic features (sound qualities such as pitch, tone, jitter, and so on). A Speech Emotion Recognition (SER) system is built for a dataset containing audio clips of various performers, which will analyse human emotion using speech as an input. Using the CNN Algorithm and the RAVDESS dataset, the system extracts, characterises, and recognises information about the speaker's emotions for both male and female speakers. The technology is confined to the English language and does not handle real-time input voice data. Speech emotion recognition could be used as a feedback application in hospitals to collect feedback from patients and categorise it, such as rage judgement or sadness judgement. A procedure for evaluating or predicting gender and emotions using speech is described in the proposed work.
Convolutional neural networks are used to evaluate or predict gender and emotions by displaying waveforms and spectrograms. A CNN model is developed using the input of 12162 samples in order to identify the emotions present in the speech. The suggested model's overall accuracy is calculated using just one feature, the MFCC from the speech, utilising the 4 datasets (RAVDESS, SAVEE, CREAMA-D, and TESS) in our study. The accuracy is first calculated for each emotion and gender, and then the overall accuracy is discovered.
Keywords: smart farming, Artificial intelligence, Internet of Things, sensors.

I. INTRODUCTION

A voice signal is one of the quickest and most organic ways that people can communicate with one another. The quickest and most effective way for human-machine interaction is speech signals [1]. All of the senses are employed to the best of a person's natural abilities to ensure optimum awareness of the message received. While emotional recognition in machines is incredibly challenging, it comes naturally to people. Therefore, an emotion recognition system makes use of emotion knowledge in a way that enhances communication between humans and machines [3]. The feminine or make speakers' emotions are recognised in speaking through speech. Some of the examined speech features are the linear prediction cepstrum coefficient (LPCC), fundamental frequencies, and the Mel+ frequency cepstrum coefficient (MFCC). These attributes form the foundation of speech processing. It is unclear why speech traits, in particular, are more beneficial in differentiating between various emotions, which makes emotion detection from speakers' speech exceedingly challenging. Due to the existence of various speaking rates, styles, sentences, and speakers, there is

an introduction of accosting variability that influences the aspects of speech. Even though numerous emotions can be conveyed using the same spoken word, it can be challenging to distinguish between the many parts of speech that each emotion corresponds to. Another issue arises because the way in which emotions are expressed relies on the speaker's environment and culture, which also cause disparities in speaking style. There are two types of emotions—transient and long-lasting—and the recognizer's ability to distinguish between them is unclear. Speech recognition of emotions may be speaker- or speaker-dependent. There are many classifiers available for classification, including K-nearest neighbours (KNN), Support vector machines (SVM), CNN, etc. [7].In the second half of the study, a block diagram of the speech emotion recognition system is described along with a brief introduction to speech emotion recognition. The third component includes modelling of emotions speech and various forms of speech together with some of the existing datasets that have been the subject of previous studies. The fourth section briefly describes alternative feature extraction methods for identifying speech emotions before focusing on a review of the classification portion. We have discussed KNN, SVM, CNN, recurrent neural network, etc. in this part. The application of deep learning for voice emotion recognition is briefly discussed in the sixth part.

A. EXISTING SYSTEM

The initial goal was to lean more towards a computationally productive strategy that would operate with a short training data set because the available computing sources were limited and there was only a tiny library of examples of speech that had been classified as emotionally charged. The application of transfer learning and pre-trained networks can overcome these limits, which are very common. The SER problem had to be redefined as a picture classification task in order to employ existing pre-trained networks because the majority of them are intended for image classification. Labelled voice samples were buffered into brief time blocks in order to achieve this. Each block was measured using a spectrogram array of spectral amplitude, transformed to RGB, and then fed into a convolutional neural network model that had already been trained. The Convolutional Neural Network Model can now assess many emotions thanks to extensive training. Now, every speech goes through a thorough conversion from speech to visualised image. In the trials described here, two alternative sampling frequencies—sixteen and eight kHz—and, consequently, the -low companding procedure—were used to test the SER's performance.

II. PROPOSED SYSTEM:

Machine learning (ML) is used to perform the voice emotion detection system. The operational phases are the same as for any other ML project, with additional fine-tuning systems to ensure the model performs as intended. Data collecting is the essential action, and it is very important. The data that is being used to create the model is what it will learn from, and all of the conclusions and judgements that a developed model will provide are based on supervised data. A combination of different machine learning tasks are combined in the secondary action, known as feature engineering, and applied to the collected data. These systems take different data description and data quality issues into account. The third step, where an algorithmic-based prototype is developed, is frequently investigated as the core of an ML project. This model use a machine learning (ML) algorithm to learn about the data and programme itself to respond to any new data presented to it. Estimating how well the developed model works is the final stage. Developers commonly repeat the steps of creating a model and estimating it in order to compare the performance of different methods. Measuring results enables selection of the ML algorithm most suited to the problem.

Dataset

The Torronto Emotional Speech Set (TESS), a collection of English language data, was used. The phrases were recorded to represent the following seven emotions: joyful, sad, furious, disgusted, fear,

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, No. 1, June : 2023

surprise, and neutral state. This dataset consists of 200 target words said by two women, one younger and the other older. This dataset consists of 2800 files in total. They said the words "Say the word ___." The audio quality is excellent, and the thresholds of both women are within the typical range. Kate Dupuis and M. Kathleen Pichora Fuller are the authors.

III. SYSTEM DESIGN

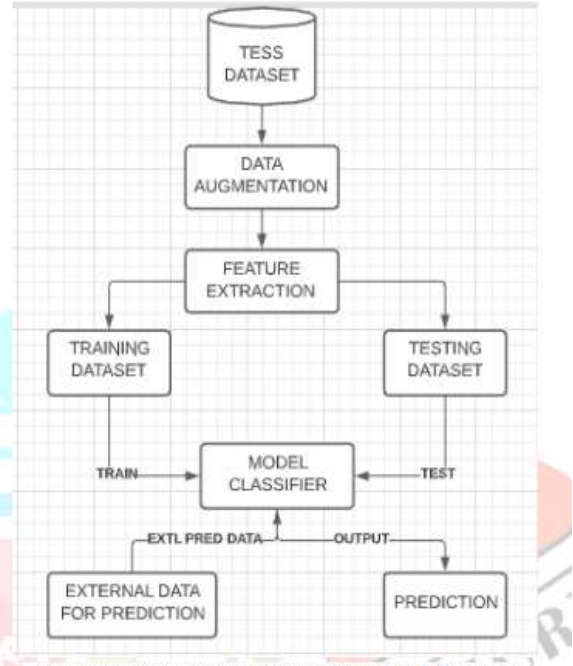


Figure 1) Architecture of Emotion Recognition from TESS Dataset Using various Machine Learning Algorithms

the file format and empty file input. The audio file will then be connected directly to python files where the output is generated in the form of emotional labels. Data visualisation displays graphic and photographic information on the provided acoustic data.Here, the initial dataset's emotional labels are separated out, and the entire set of data is then shown in spectrogram graph and wave plot diagrams Because a sign's spectrum of frequencies changes over time, a spectrumogram might serve as a visual depiction of that spectrum. Plotting waveforms of amplitude vs. time using a wave-plot, where the primary axis is amplitude and the secondary axis is time.

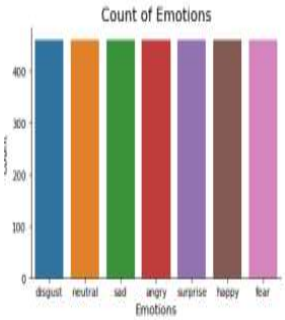


Figure 2) Bargraph mentions the no of emotions for each label vs Label

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, No. 1, June : 2023

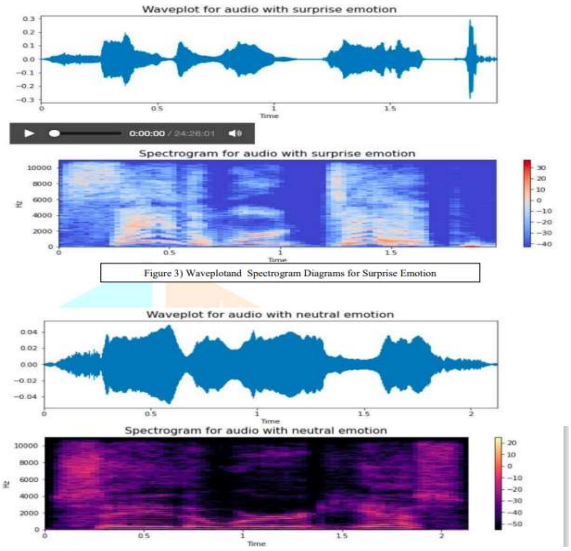


Figure 3) Waveplotand Spectrogram Diagrams for Surprise Emotion

Data Augmentation:

However, in a real-life situation, that is not the case and we can have more noisy bits in the original recorded audio file. This mostly concentrates on upsetting data where the first collected will be better tuned without noise. Therefore, we considered expanding the data by simply using the provided information and running it through two augmentation procedures, such as adding additional noise to the data and normal data values at the same time for each piece of information we gathered while maintaining the same emotional label.

Features Selection:

Features like Zero Crossing Rate, Chroma Shift, Root Mean Square Value, Mel Spectrogram, and MFCC (Mel Frequency Cepstral Coefficient) are the primary components of this project. These are a few often utilised audio features that extract the essence of a piece of audio, including emotional content, acoustic recognition, and information retrieval in the music genre. A large movement from +ve to 0 to -ve or from -ve to 0 to +ve is when the zerocrossing rate occurs. Mel Spectrograms are types of spectrograms that display sound on the Mel scale rather than in the frequency domain. The frequency of a signal must be transformed logarithmically to create the Mel Scale. Surprise Emotion Waveplotand Spectrogram Diagrams.

IV. DIFFERENT ALGORITHMS USED FOR TRAINING MODEL

Multi Layer Perceptron Classifier (MLPC):

A collection of artificial neural networks (ANNs) is known as a multilayer perceptron (MLP). MLP is used ill-definedly, rarely loosely, and rarely rigorously in systems made up of repeated layers of perceptron's (with threshold activation). Despite having a single hidden layer on average, multilayer perceptron neural networks are rarely referred to as "vanilla" neural networks. A supervised classification method for multi-band passive optical remote sensing data is presented by multilayer perceptron. The robust classifiers provided by multilayer perceptrons (MLP) may perform better compared to other classifiers, yet MLP Classifiers are usually examined for the high proportion of unstructured parameters. Long training intervals and local minima are additional MLP challenges.

Light Gradient Boosting Machine(LGBM) :

A suitable and efficient implementation of the gradient boosting algorithm is provided by the open-source Python library known as Light Gradient Boosted Machine, or LightGBM for short. By adding additional automatic feature selections, such as boosting examples with stronger gradients, LightGBM extends the gradient boosting technique. As a result, LightGBM can result in accelerated learning and improved prediction accuracy. Gradient-based one-side sampling (GOSS) is used in

LightGBM to discriminate between the perspectives used to compute the division. To maintain the effectiveness of information gain estimation, GOSS keeps those events with large gradients and only sporadiclly discards those with small gradients. When the value of information gain has a wide range, this approach can lead to a more precise gain assessment than constantly random sampling with the same target sample rate.

Gradient Boosting Classifier (GB) :

An ensemble classifier called gradient boosting is designed to work well in situations like high-dimensional data, where there are more variables than samples. It hasn't, however, been estimated for the forecasting of certain events. It is demonstrated that a tiny percentage of specimens from uncommon species are classified appropriately due to gradient boosting experiences from rare key events bias. By using subsampling in unification with an appropriate degree of shrinkage and just for a specific number of boosting repetitions and for binomial loss functions, the bias can be eliminated. It is demonstrated that when the data size is small, the number of boosting iterations where the unique events bias is excluded but cannot be evaluated well from the training data. As a result, utilising both fake and actual high-dimensional data, different improvements for the Gradient boosting's unique events bias are suggested and estimated.

Extra Trees Classifier (ET):

A method of ensemble learning called Extra Trees Classifier combines the results of many de-correlated decision trees (DTs) that are integrated into a "forest" to demonstrate classification effects. It performs somewhat similarly to a random forest and differs significantly from one in how it imagines the DTs inside the forest. That DT is essentially organised in Extra Trees Forest from the primary training set. As a result, a random assortment of k-features is assigned to each testing node for each tree from the feature set, from which each DT must select the key characteristics to split the statistics depending on particular numerical parameters (Gini Index). This random adoption of idiosyncrasies results in the development of various, de-correlated DTs. The normalised total reduction in the analytical standards used in the split feature judgement (Gini Index, if the Gini Index is used in the construction of the forest) is calculated for each feature during the construction of the forest in order to achieve feature preference using this particular forest arrangement. The Gini importance of the feature is the name given to this quality. Every item is regulated for adoption in descending order according to its Gini Value, and the user selects the top k features based on preference.

Random Forest Classifier (RF) :

Classifier ensembles are built on the basic premise that a group of classifiers can produce classifications that are more accurate than a single classifier. Random forest is a novel and reliable classifier that Breiman first suggested (2001). Random Forest is a machine learning system that uses different decision trees to provide decisions. a forest of randomly produced decision trees. To measure the outcome, each node in the decision tree uses a random subset of features. The final result is then created by merging the various distinct decision trees that were used in the random forest.

Decision Tree Classifier (DT):

Given that they attain constant accuracy and are very cheap to estimate, decision tree models are seen as the most advantageous in the fields of data mining, data science, and machine learning. The two phases of classification presented by the majority of decision tree classifiers are the "Tree Building Part" and "Tree Pruning Part." When creating the decision tree model, the training data set is divided repeatedly according to a sectionally optimal pattern until all or the majority of the works relevant to various obstacles maintain the same class label. Tree pruning is used to reduce the leaves and branches responsible for examination of individual or comparatively more minor data vectors in order to improve generalisation of a decision tree. The decision tree classifier (DTC) is a good

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, No. 1, June : 2023

illustration of how multistage judgement making should be done. The fundamental principle behind many multi-stage procedures is to break down a difficult judgement into a series of independent, more straightforward choices, anticipating the final outcome; this strategy would concur with the suggested aspirational solution.

Difference between Random Forest Classifier (RF) and Decision Tree Classifier (DT) :

As implied by the words "Tree" and "Forest," a Random Forest is made up of many Decision Trees. A decision tree is built using all of the relevant features from the entire dataset, whereas a random forest randomly selects the rows and features to create several decision trees before averaging the results. With more divisions when using a decision tree model on a training dataset that is provided, accuracy increases.However, one can surely overfit the data and fail to recognise when someone has over the line without performing cross-validation (on the training data set). On the other hand, the advantage of a basic decision tree is that it is easy to use, and everybody working should be aware of any variables and their values that are used to divide the data and predict a conclusion. Compared to a single decision tree, Random Forest has a longer training periodIt would be wonderful if this were acknowledged since as time goes on, the time it takes to train trees in a random forest increases along with their abundance. Decision trees are helpful because they are easier to analyse and quicker to train, despite vulnerability and dependence on a certain set of features.

V. EVALUATION & RESULT

Knowing the speaker's emotions in real time is a difficult endeavour, as we all know. A few tests were run after the models had been trained in order to assess the performance and accuracy of each model. A number testing methods are employed here, including accuracy, F1 score, MCC, recall, precision,andkappa.

Model	Accuracy	F1 Score	Kappa	MCC
Light Gradient Boosting Machine	0.9714	0.9714	0.9666	0.9667
Random Forest Classifier	0.9585	0.9585	0.9516	0.9517
Extra Trees Classifier	0.9479	0.9478	0.9392	0.9393
Gradient Boosting Classifier	0.9448	0.9449	0.9355	0.9357
Multi Layer Perceptron Classifier	0.924844	0.925	0.912259	0.912966
Decision Tree Classifier	0.8831	0.883	0.8636	0.8638

Table 1) Performance of Different Algorithms

Accuracy:

The Accuracy is generally calculated with Test data in our project, where Accuracy relates to number of correct classifications made to total number of predicted classifications made as mentioned in Figure 5.

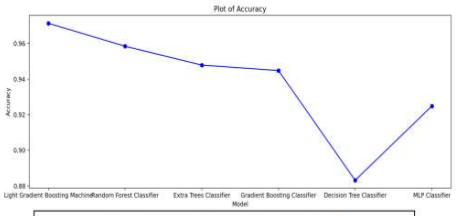
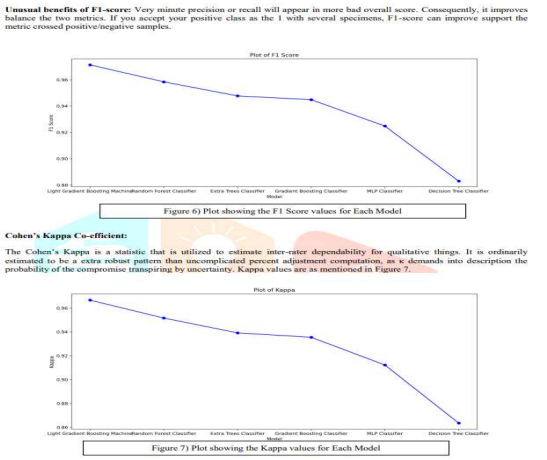


Figure 5) Plot showing the Accuracy values for Each Model

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 52, Issue 6, No. 1, June : 2023

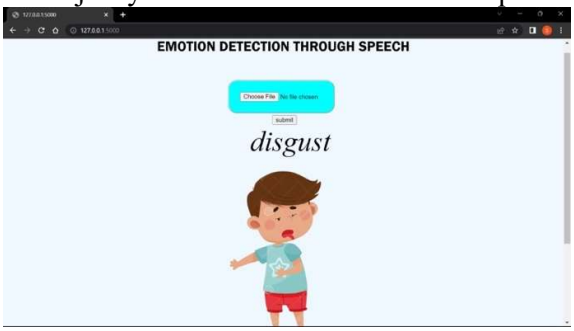
Matthews Correlation Co-efficient:



The results of the implementation can be used to draw a number of observations and inferences. The performance scores between the various techniques have improved generally. The implementation using the first strategy performed ok with a high score of 83% using the SVM algorithm, the second strategy worked well with a high score of 80% using the KNN algorithm, and the third strategy had the highest score of 90% using the SVM algorithm. The outcomes allow for the following conclusions to be drawn:

Observations:

The second technique performed less well overall than the first and third approaches. As a result, it can be inferred that utilising only the MFCC values alone cannot be a good measure to identify the emotional content of speech. This is because the selective feature technique fails to contain themajority of the information from the speech signal.



The first approach's classification report reveals that the misclassification rate is higher for the emotions of surprise and happiness. The similarities between the traits in these two categories account for this bias. This bias has been significantly reduced thanks to the dimensionality reduction phase utilised in the third strategy. The accuracy score of the suggested system is higher when

compared to the Chen et al. baseline system [4]. The successful strategy for the suggested methodology is the third experiment.

VI. REFERENCES

[1] Soegaard, M. and Friis Dam, R. (2013). The Encyclopedia of Human-Computer Interaction. 2nd ed
[2] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.
[3]L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.
[4] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.
[5] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009
[6] Shruti Bhargava, Dr. Ajay Somkuwar,noise assessment in medical imaging data ,journal of medical imaging & health informatics,Volume 6, Number 4 , 875–884 August ,2016
[7] Application of virtual reality systems to psychology and cognitive neuroscience researchCSN Koushik, SB Choubey, A ChoubeyCognitive Informatics, Computer Modelling, and Cognitive Science, 133-147
[8]Shruti Bhargava,AjaySomkuwar ,Noise Reduction Techniques Of Medical Imaging data A review, 2nd International Conference on Mechanical, Electronics and Mechatronics Engineering (ICMEME'2013) June 17-18, 2013 London (UK).
[9]Evaluation of noise exclusion of medical images using hybridization of particle swarm optimization and bivariate shrinkage methodsS Bhargava, A Somkuwar International Journal of Electrical and Computer Engineering 5 (3), 421
[10]Gesture controlled quadcopter for defense search operationsSB Choubey, A Choubey, CSN Koushik Materials Today: Proceedings 46, 5406-5411
[11]Machine Learning for Testing of VLSI CircuitA Choubey, SB Choubey VLSI and Hardware Implementations Using Modern Machine Learning Methods, 23-40
[12]Shruti Bhargava Choubey ,Abhishek Choubey,KhushbooPachori Object Detection Using Higher Quality Optimization Techniques In Video Encoders, , Lecture Notes In Networks And Systems Springer Book Series,Pp 297-304