University of Ibadan
Journal of Science and Logics in ICT Research

Improved Speech Emotion Recognition Using Boosting Ensemble of Class Specific Classifiers

Awoyemi, J. O., Hassan, J. B. and Oluwasola, A.M

Federal University, Oye-Ekiti, Nigeria
John.awoyemi@fuoye.edu.ng, *Joshua.hassan@fuoye.edu.ng* , *Oluwasolamercy1@gmail.com*

Abstract
Speech emotion recognition (SER) systems enable machines to understand human emotions from speech. Due to its rising popularity, it is also interesting to scholars everywhere. Researchers have explored many different techniques and methodologies to detect emotions from speech. Different machine learning approaches have been investigated since machine learning (ML) has emerged as one of the most promising methodologies. However, knowing which feature combination gives a better result has been one major challenge of speech recognition systems, as the features selected affect the classification result. Therefore, a general Random forest-based emotion detection system has been built in this study effort to categorize emotions based on the provided features. The features are selected based on the class-specific best feature for each emotion class and are identified based on the f1 measure. The selected characteristics are then combined, applying the contemporary smart ensemble approaches. Mfcc, Chroma, and Melspectogram are the features under study, and Mfcc appeared to be the best of the three case studies with an accuracy of 63% higher than other features. The three features were then Ensembled using a CatBoost algorithm which gave an accuracy of 71.62%. The result indicates that Mfcc performs better than other features listed in the study. Also, the ensemble classification result performed superior to the classifiers mentioned in the literature.

Keywords: *Speech emotion recognition, Ensemble classifier, Feature extraction, CatBoost Classifier, SVM Classifier, Boosting Ensemble.*

1. Introduction

Recently, high development in technology and information society has led to the high rapid popularity and performance of personal computers. As a result, there has been a proactive shift toward a bidirectional interface between computers and people. Therefore, a greater understanding of human emotions is necessary. It uses how the voice's tone and pitch may often be used to express underlying emotion in speech [1]. Animals like dogs and horses can understand human emotions. All call centres offer the greatest illustration of it. If you have ever noticed, contact centre agents speak in various voices. Customers influence how they pitch to and converse with them. Thus, this also occurs to regular individuals, but how does this apply to contact centres? The staff may enhance their services and increase conversion rates by identifying clients' emotions from their speech. Numerous studies have employed a variety of input streams, including audio, video, text, and others, to identify a user's emotional state [2]. According to Research on emotion recognition, a voice signal, one of the most organic forms of human communication, comprises language and paralinguistic information, including emotion constitutes a speech signal [3]. Prosodic and spectrum features are the two groups into which the speech characteristics are divided. The speech, tempo, vigour, pitch, and structure are prosodic qualities. On the other hand, spectral properties like its first derivative and the Linear Prediction Cepstrum Coefficients (LPCC) and Mel-frequency Cepstrum Coefficients (MFCC).

Numerous studies have discovered that high levels of emotional relevance are predicted by

Awoyemi, J. O., Hassan, J. B. and Oluwasola, A.M (2023). Improved Speech Emotion Recognition Using Boosting Ensemble of Class Specific Classifiers, *University of Ibadan Journal of Science and Logics in ICT Research (UIJSLICTR),* Vol. 9 No. 1, pp. 143 – 151.

©U IJSLICTR Vol. 9, No. 1, January 2023

acoustic, speech quality, and prosodic features. As the system's accuracy and performance depend on the feature extraction and the acoustic properties of the speech signal features extracted to analyze the movement, effective parallel usage of appropriate feature extraction methods is a crucial challenge in speech emotion recognition. A poor and noisy input file meant to be rejected by a rejection framework is the frame with 12 acoustical attributes on the Berlin emotional corpus EMO-DB. The general emotion identification system's accuracy is 74.70% [4].

A system with spectral and prosodic properties is also better than one with only those qualities [5]. However, the identification rate of systems that utilize energy, pitch, LPCC MFCC and MEDC features is lower than that of systems that use those features. Constructing an SVM-based emotion identification system could be considered redundant, and Research has been done on extracting features in various methods to model emotions using provided characteristics. The decision logic decodes the replies into an emotional class that delivers the most value out of all emotional classes [6].

This work uses the new smart ensemble technique to choose and evaluate the optimum acoustic features for each emotional class. The Cat-boost Algorithm of the Ensemble methods will be used as a case study.

2. Related Works

During a speech, there are three features: lexical, visual, and, therefore, acoustic [7]. The prosodic and spectra features are widely explored. Performance may be improved by comparing the recognition rate using the combination of prosodic and spectra features, which consist of energy, pitch, LPCC, MFCC, and MEDC features, respectively [5]. It is often difficult to tell which feature combination gives the best classifications. For that reason.

Gu *et. al.* [7] proposed a personal (C.S.O.) algorithm. The Cat Swarm Optimizer is used to extract only necessary features. Moreover, other feature techniques may be employed to extract characteristics, including pitch, energy, Zero crossing rate, Mel Frequency Cepstral Coefficients (MFCC), Discrete Wavelet Transform (DWT), and others (ZCR.) [8]. Different Classifiers have also been explored in the extraction of emotions, and a deep neural network was used to learn and classify the fusion features.

There are several shared traits, including speech rate, energy, pitch, format, and specific spectrum properties, such as linear prediction coefficients (LPC), linear prediction Cepstrum coefficients (LPCC), and Mel-frequency Cepstrum coefficients (MFCC), are retrieved from the data in terms of features [5].

Additionally, the sensitivity of different emotional traits varies among languages, and different combinations of defining emotional features might lead to varied rates of emotion identification. As a result, the system's rate of emotion recognition is slightly higher than the system's rate of emotion recognition when the system only uses the spectrum features of speech, such as Linear Prediction Coefficients (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and Linear Prediction Cepstrum Coefficients (LPCC) [6].

Also Ingale, and Chaudhari [9] compared the performance of five different algorithms ranging from K-nearest neighbors (KNN), Hidden Markov Model (HMM), Artificial Neural Networks(ANN), Support Vector Machine and the Gaussian Mixtures Models for extraction of prosodic features and the spectra features of speech. From this study it was deduced that the GMM is palatable for the extraction of prosodic features with an accuracy of 78.77%. With a 51.19% accuracy rate, ANN can locate the non-linear boundaries dividing emotional states. KNN provides a 64% rate of emotion recognition, whereas HMM is only appropriate for the spectral characteristics with a 76.12% accuracy. As compared to other classifier schemes, SVM offers a higher recognition rate with an accuracy of 77%. [9]. All of the real-time audio samples are affected by noise since the authors do not place enough emphasis on noise-based signals.

Vasuki [4] built an SVM-based emotion identification system that uses features to represent emotions. Based on the f1 measure, the optimal acoustical characteristics for each emotional class are determined out of the available features. The responses of the systems constructed utilizing the features are integrated.

To translate the replies into an emotional class. A rejection framework is also used to remove noise from the weak and noisy input file. The 12 acoustical characteristics on the Berlin emotional corpus EMO-DB was tested with the framework. The emotion identification system has a 70.70% accuracy rate.

After feature extraction, the classification of speech emotions is a crucial step. This essay has contrasted and evaluated the many classifiers distinguishing emotions, including melancholy, neutrality, pleasure, surprise, and rage. The Research also demonstrates the development of automatic emotion identification systems using deep neural networks. Additionally, the investigation has been carried out utilizing several ML approaches for speech emotion recognition accuracy across various languages. [10].

3.0 Methodology

3.1 Datasets (Source: kaggle.com)

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) has 7,356 files (total size: 24.8 GB) was used. Twenty-four professional actors (12 females and 12 males) perform two lexically matched sentences, all of whom have neutral North American accents. Songs feature peace, happiness, joy, sadness, wrath, fear, surprise, and disgust, whereas speech manifests these emotions. Each expression has two emotional intensity levels: normal, strong, and neutral. Available formats are audio-video (720p H.264, AAC 48kHz.mp4), audio-only (16bit, 48kHz.wav), and video-only (720p H.264, AAC 48kHz.mp4) (no sound). This portion of the RAVDESS dataset used in this investigation consists of 1440 files. The total number of trials for 24 actors is 1440. Of the 24 professional performers in The RAVDESS, 12 were male and 12 female, whose speech encompassed feelings of calmness, happiness, sadness, Anger, terror, surprise, and contempt. [11].

The naming convention for files:

Each of the 1440 files has a unique filename. A seven-part number identity serves as the filename. (e.g., 03-01-06-01-02-01-12.wav). These filenames specify the features of the stimulus: file names with identifiers. Modality (01 is full-AV, 02 is just video, and 03 is just audio. Voice channel (01 for speaking, 02 for music). Emotion (01) is neutral, (02) is peaceful, (03) is pleased, (04) is sad, (05) is furious, (06) is afraid, (07) is disgusted, and (08) is astonished. Emotional intensity (01) (intense), (02) (normal).

NOTICE:
The 'neutral' mood lacks substantial intensity. Youngsters are conversing by the door in the statement (01), while dogs sit by the door (02). Repeat (01 is the first iteration, 02 is the second repetition). Actor (born 2001–24). Male actors have odd numbers; female performers have even numbers. The dataset is audio-based and in the wav file format.

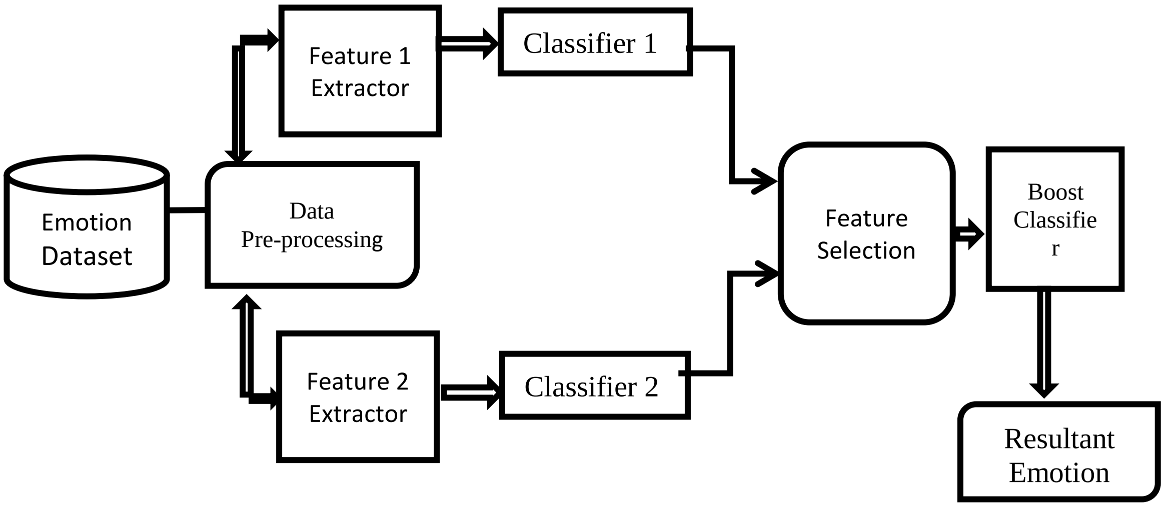


Figure 1 System architecture

3.2. Dataset pre-processing

Before collecting the signal's characteristics from the voice samples, pre-processing is used [12]. Speech samples include undesirable information during the speech's recording, such as noise and certain environmental variations. By using the filters at this point, these may be eliminated. In this study, the steps carried out in the pre-processing to provide syntactic data for audio are noise injection, time shifting, pitch and speed adjustments.

3.3 Feature Extraction

Mel-frequency cepstral coefficients (MFCC)
A representation of the short-term power spectrum of a sound is called MFCC, and it is based on a nonlinear mel scale of frequency and a linear cosine transform of a log power spectrum. The frame size is 25 msec at a tempo of 10 msec. The frames are windowed with a Hamming function, using a pre-emphasis with k = 0.97. 26 Mel-bands derived from the FFT power spectrum are used. The (12+1) MFCC are calculated. The Mel-spectrum has a frequency range of 0 to 8 kHz [13]. Figure 2 shows the extraction processes of the MFCC features.

The power spectrum of each frame is then calculated using a periodogram; it is made to resemble the human cochlea, an ear organ that vibrates in various locations according to the frequency of external noises. Start by obtaining the frame's Discrete Fourier Transform to do this.

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{-j2\pi kn/N}$$

:
where:

- The frame time signal is called si(n) (iframes)
- In a Hamming Window, The sample size is N.
- h(n) stands for the Hamming Window.
- K is the DFT's length.

To determine the periodogram estimate of the power spectrum.

$$P_i(k) = \frac{1}{N}|S_i(k)|^2$$

Chroma: An octave-invariant number termed "chroma" and a "pitch height" that denotes the octave the pitch is used to deconstruct the chroma characteristic of a pitch class, which refers to the "colour" of a musical pitch. A chroma vector is a feature vector that normally has 12 elements and represents the energy of each pitch class, C,C#, D, D#, E,..., B, in the signal. A feature vector with perceptual motivation is the chroma vector. The cyclic helix model of musical pitch perception employs the idea of chroma. Hence, a normal chromatic scale's twelve pitch classes are represented by the Chroma vector [14]. Figure 3 shows the chromagram obtained from a sample of the audio recordings.

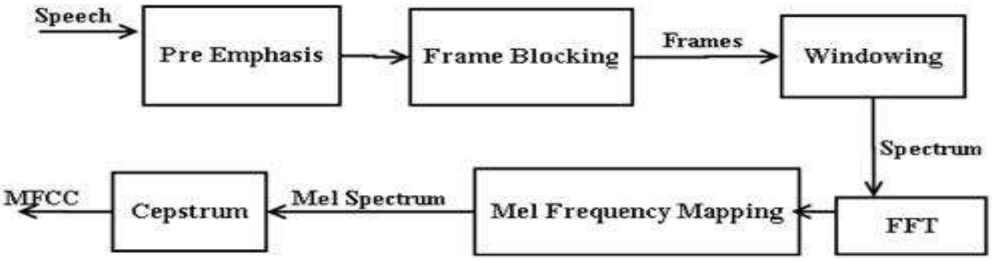


Figure 2: MFCC extraction from a speech signal

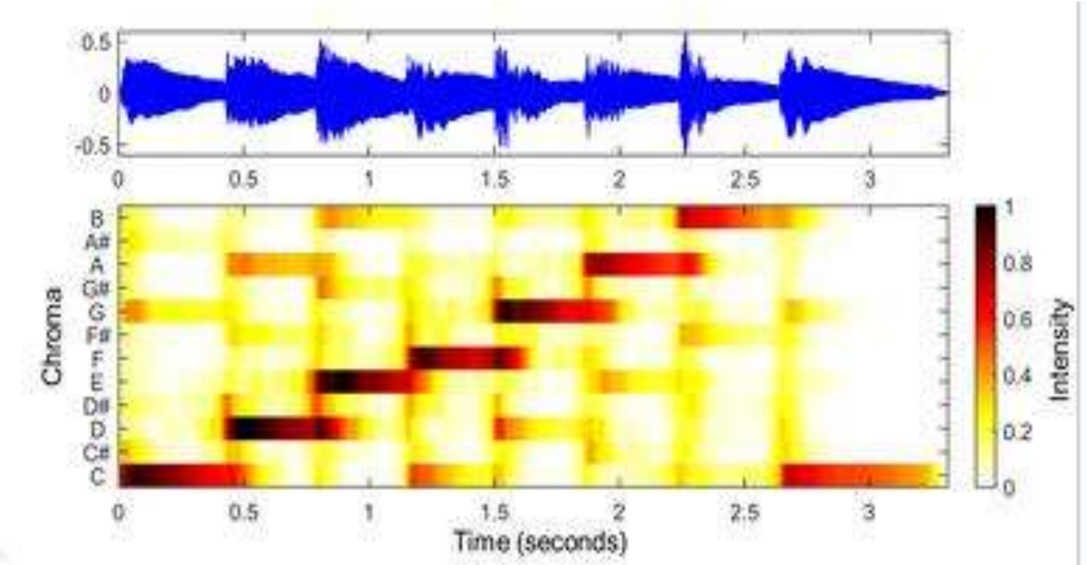


Figure 3: Chromagram obtained from the audio recording.

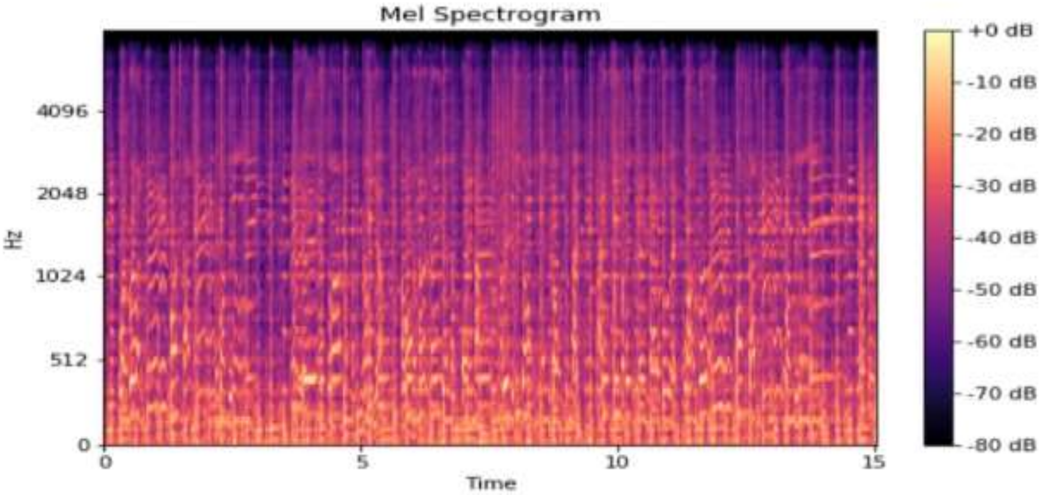


Figure 4: Melspectogram obtained from the audio recording.

Mel Spectrogram

A mel spectrogram is produced by translating the frequencies of a spectrogram to the mel scale. I understand. Who dared to guess? Remarkably, only a few lines of code are required to create a mel spectrogram. A sample of a mel spectogram gotten from a speech sample is shown in Fig 4.

3.4 Feature Selection

We have used the Random Forest Algorithm to identify the best classifying feature for an emotional class. First, by noting which feature best describes an emotion class. For the Ensemble, the only features picked are those with the greatest recognition rate and suggest at least one emotional class. The optimum classifier for each emotional category is then identified as a consequence. Finally, the responses from these classifiers are combined to get the ensemble result. The ensemble classifier's aggregate response, then, reflects representations of each emotional class's dominant features.

3.5 Ensemble techniques

By using ensemble techniques, more learning algorithms are compared to what would be feasible with only one of the separate learning algorithms utilized for prediction. In contrast to a statistical ensemble in statistical mechanics, which is usually limitless, a machine learning ensemble only consists of a specific restricted collection of various models. Which enables a far more flexible structure to exist among those alternatives [15]. Cat-boost Algorithm is a boosting ensemble technique that learns the learners' weaknesses and iteratively improves observations.

Training: The training set's speech input utterances are folded twice and given to the system for training. Each learner is generated using a distinct feature. Every emotional lesson is tested and imitated by all features. Each Classifier's performance on the development set is tracked and measured. The top feature in each emotional class is picked using one metric. The best test takers' replies are merged to provide a combined response. As a result, the normalized average measure of the best learners is used to determine and record the weight of each learner participating in the ensemble.

Ensemble: The straightforward fusion only averages the classifiers' responses. The boosting Algorithm combines the weight factor with the classifiers' responses. Every classifier's starting weight is fixed based on their single measure. Updates are made to the weight vector (Weights of all classifiers) using the training set of data.

CatBoost Classifier

CatBoost is a decision tree method that makes use of gradient boosting. It is the replacement for the MatrixNet Algorithm, which was created within the company, Yandex researchers are frequently employed for task grading, forecasting, and suggestion making. It may be used everywhere and for a variety of problems.

Random Forest

An ensemble learning method for classification, regression, and other issues, random forests, often called random decision forests, entails training several decision trees. The class selected by the majority of trees is the output of the random forest for classification problems. The mean or average forecast of the individual trees is returned for regression tasks. Random decision forests address the problem of decision trees overfitting their training set. Random forests often perform better than choice trees, although they are less accurate than gradient-enhanced trees. As opposed to that, On the other hand, data attributes may affect how well they work [16].

Random Forest generates many decision trees, merged to get a more precise forecast. The premise of the Random Forest model is that a collection of uncorrelated models (single decision trees) performs much better than a single one. Each tree casts a classification or "vote" in a Random Forest classification. The forest chooses the classification with the most "votes". Finally, regression uses a Random Forest, which chooses the mean of all tree outputs. The main finding is that there is little to no correlation between the smaller Random Forest model's decision trees and its component models. While some decision trees may be inaccurate individually, most of the group will be accurate, leading to a successful overall result. Figure 5 shows the diagram of a random forest decision tree.

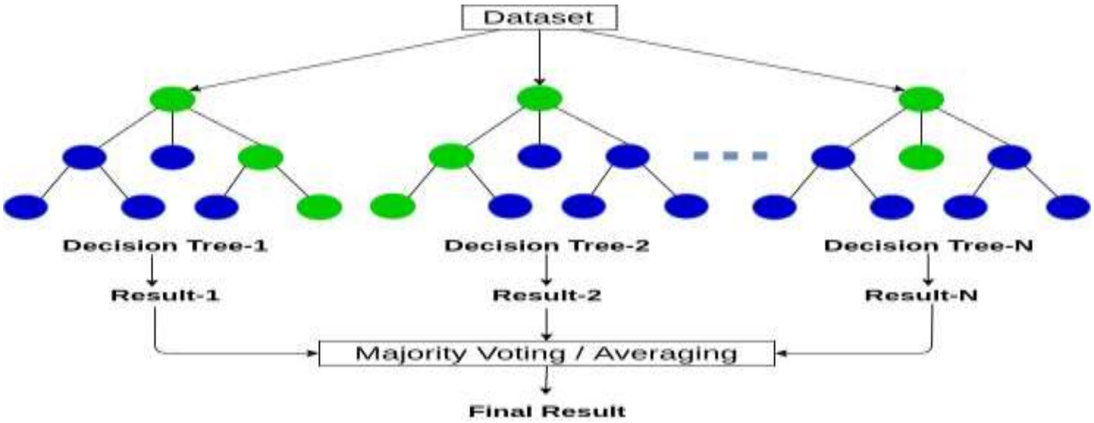


Figure 5: Random Forest decision tree.

4. Results and Discussion

4.1 Results

At this point, the performance of each feature extracted has been assessed and the classification algorithms have been used both at the data level and classification level. For this reason, researchers have utilized a variety of measurements, including accuracy, precision/recall, fallout, error, and many more. The following list includes a handful of these actions [17]:

Precision: is characterized as a pertinent section of the speech.

Recall: is the percentage of pertinent speech emotion recalled, according to its definition.

True Positive: implies that the appropriate class has been assigned to the spoken emotion.

False Negative: implies that the class for speech emotion is incorrect.

True Negative: signifies that speech emotion is misclassified and does not belong in that class [18].

Accuracy: Accuracy is the degree to which measurement findings come close to the actual value; it is the degree to which measurements come close to a certain value. It is the result of adding true Positives and True negatives. The mathematical representation is seen below [19].

Accuracy = (True positives + True Negatives) (True positives + False Negatives + False Positives + True Negatives) (1)

Recall = True Positives / (True Positives + False Negatives) (2)

Precision = True Positives / (True Positives +False Positives) (3)

F1 = (2×precision× recall) / (precision + Recall) (4)

Table 1 displays the f1 value of the random forest classifier learned on the train set after training with the three features at various emotional classes. Mfcc has the highest fi score for each emotion class, and it classifies calm best with an f1 score of 75%. On the other hand, the Chroma feature is very poor on the classification of each of the emotion classes, with 35% being the highest of the emotional class, while mel-spectrum performed averagely on each of the emotion classes; it classifies Calm with an f1 score of 66%. from this results it can be deduced that Mfcc classifies all the emotional classes better compared to the remaining two features.

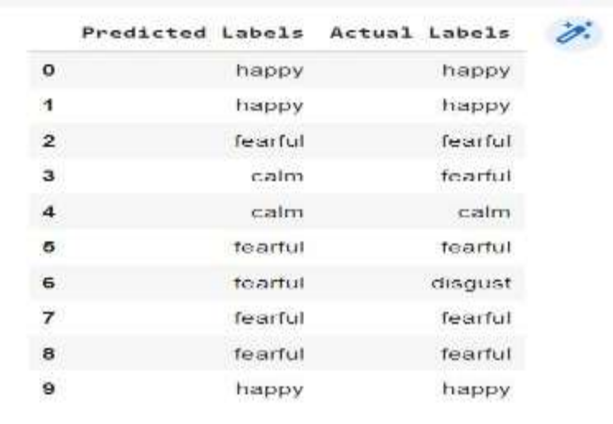
Table 1: performance of classifiers constructed using individual feature

Feature	Anger	calm	disgust	fearful	happy	neutral	sad	surprised
Mfcc	**0.69**	**0.75**	**0.58**	**0.58**	**0.56**	**0.51**	**0.58**	**0.63**
Chroma	0.22	0.35	0.27	**0.35**	0.28	0.19	0.32	0.21
Mel-spectrum	0.63	0.66	0.45	0.53	0.45	0.50	0.40	0.35

Table 2: test set output built on ensemble classifier (Cat-boost classifier)

feature	Anger	calm	disgust	fearful	happy	neutral	sad	surprised
Precision (%)	0.73	0.74	0.71	0.76	0.69	0.71	0.72	0.65
Recall(%)	0..79	0.91	0.77	0.69	0.69	0.47	0.70	0.60
F1-score(%)	**0.76**	**0.82**	**0.74**	**0.72**	**0.69**	**0.57**	**0.71**	**0.63**
support	48	56	48	51	48	32	44	50

Table 2 illustrates the results of the Cat-boost classifier constructed on individual features and an ensemble of the best features (Mfcc, Chroma, Mel-spectrum). The initial data was separated into the train (80%) and test (20%) sets after being segmented into features (X) and labels (Y) [18]. As a result, the algorithms were evaluated on the 20% test set after being trained on the 80% train set. Also, test Accuracy scores, F1 scores, Recall and Precision were compared. From the results of the Ensemble, it is denoted that the Ensemble of these three features improved the classifier's performance for each of the emotional classes. Taking the f1 score of Anger from the classification of individual features as a case study. The results obtained are 0.69, 0.22 and 0.63, respectively. At the same time, the new smart booting Ensemble of the features gave an f1 score of 76% and an accuracy of 71.62%, which shows that the new smart booting Ensemble performs better compared to individual features. Figure 6 shows the predicted label vs the actual label.



	Predicted Labels	Actual Labels
0	happy	happy
1	happy	happy
2	fearful	fearful
3	calm	fearful
4	calm	calm
5	fearful	fearful
6	fearful	disgust
7	fearful	fearful
8	fearful	fearful
9	happy	happy

Figure 6: Predicted label vs actual label

5. Conclusion

We have used an ensemble method to enhance speech emotion recognition. We evaluated the features(Mfcc, Chroma, Mel spectrogram) and deduced that Mfcc performed better than the Chroma and Mel spectrogram feature. Also, the Ensemble of the three features gives better accuracy for each emotional class as much as the ensemble method worked well for the Emotion recognition system both at the selection and classification levels. However, there are still misclassifications due to resembling emotions. For example, it can be deduced in Figure 6 that there are misclassifications of emotions of the test samples, resulting from resembling emotions. Using the first ten rows as a case study, the system classified fearful as calm and disgust as fearful. In future Research, we plan to work on resembling emotions and explore more feature extraction techniques to get the best feature combinations for the speech emotion recognition system.

REFERENCES

[1] Speech Emotion Recognition with Librosa.(2015, October 5). retrieved from https://data-flair.training/blogs/python-mini-project-speech-emotion-recognitionition Project.

[2] Liu, M., Li, S., Shan, S., Wang, R., & Chen, X. (2014)Deeply Learning Deformable Facial Action Parts Model for Dynamic Expression Analysis. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; pp. 143–157.

[3] Amarjeet, S,(2020). Speech Emotion Recognition Using Enhanced Cat Swarm,p.6.

[4] Vasuki,P. (2019). Speech Emotion Recognition Using Adaptive Ensemble of Class Specific Classifiers, Department of Information Technology, SSN College of Engineering, Chennai 603110, India

[5] Yixing, p., Peipei, S., & Liping, S.(2020). Department of Computer Technology Shanghai Jiaotong University, Shanghai, China panyixiong@sjtu.edu.cn,shen@sjtu.edu.cn, lpshsen@sjtu.edu.cn

[6] Zixing, Z., Weninger, F., Wollmer, F., & Schuller, B. (2011). Unsupervised learning in cross-corpus acoustic emotion recognition. Proceeding of the IEEE Workshop on Automatic Speech Recognition and Understanding. Waikoloa, HI, pp: 523-528.

[7] Gu, Y., Chen, S., & Marsic, I. (2018). "Deep multimodal learning for emotion recognition in spoken language," *IEEE International Conference on Acoustics, Speech, and Signal Processing* Proceedings, Calgary, Canada, April 2018.

[8] Anusha, K., Hima, B., Valiveti1, A., & Kumar, B. (2021). Feature extraction algorithms to improve the speech emotion recognition rate © Springer Science+Business Media, LLC, part of Springer Nature 2020

[9] Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. International Journal of Soft Computing and Engineering (IJSCE), 2(1), 235–238

[10] Dr Yogesh, K., & Dr Manish, M (2019). Machine Learning Based Speech Emotions Recognition System, International Journal of Scientific & Technology Research · July 2019

[11] Banse, R., & Scherer, K. R. Acoustic profiles in vocal emotion expression. J. Pers. Soc. Psychol. 1996, 70, 614–636.

[12] Kurpukdee, N., Kasuriya, S., Chunwijitra, V., Wutiwiwatchai, C., & Lamsrichan, P. (2017). A study of support vector machines for emotional speech recognition. In *2017 8th international conference of information and communication technology for embedded systems* (*IC-ICTES*) (pp. 1–6). IEEE.

[13] Ibrahim, N. J., Idris, M. Y., Yakub, M., Yusof, Z. M., Rahman, N.A., & Dien, M.I.(2019). Robust feature extraction based on spectral and prosodic features for classical Arabic accents recognition. Malaysian J.Comput. Sci. 2019, 46–72.

[14] Kumari, S., Perinban, D., Balaji, M3., Gopinath, D4., & Hariharan, S.J5. (2021). Speech Emotion Recognition Using Machine Learning Assistant Professor, Department of Information Technology, Panimalar Engineering College, Anna University,

[15] CatBoost — A new game of Machine Learning | by Affine | Medium. https://affine.medium.com/catboost-a-new-game-of-machine-learning-72a7dcea0ac4

[16] Madeh. S., Chakrabarti, C., & Spanias, A.(2020).A multimodal approach to emotion identification utilizing undirected topic models," in *Proceedings of the 2020 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 754–757, Melbourne, Australia, June 2014.

[17] Selvaraj, M., Bhuvana, R., & Padmaja, S. (2016). Human speech emotion recognition. International Journal of Engineering and Technology, 8, 311–323.

[18] Power, C.K., & Henn. J. M., .(2021). Department of Information Technology, Panimalar Engineering College, Anna University, Chennai.

[19] Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. International Journal of Soft Computing and Engineering (IJSCE), 2(1), 235–238.