International Journal of Engineering & Technology, 9 (3) (2020) 770-778

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Bag-of-words from image to speech: a multi-classifier emotions recognition system

Mai Ezz-Eldin[1,2*], Hesham F. A. Hamed[2] and Ashraf A. M. Khalaf[2]

[1]*Department of Electronics and Communication Engineering, Future High Institute of Engineering, Fayoum, Egypt*
[2]*Department of Electronics and Communication Engineering, Minia University, El-Minia, 61111, Egypt*
[*]*Corresponding author E-mail: mai.ezzeldin.89@gmail.com*

Abstract

Recently, recognizing the emotional content of speech signals has received considerable research attention. Consequently, systems have been developed to recognize the emotional content of a spoken utterance. Achieving high accuracy in speech emotion recognition remains a challenging problem due to issues related to feature extraction, type, and size. Central to this study is increasing emotion recognition accuracy by porting the bag-of-word (BoW) technique from image to speech for feature processing and clustering. The BoW technique is applied to features extracted from Mel frequency cepstral coefficients (MFCC) which enhances feature quality. The study considers deployment of different classification approaches to examine the performance of the embedded BoW approach. The deployed classifiers include support vector machine (SVM), K-nearest neighbor (KNN), naive Bays (NB), random forest (RF), and extreme gradient boosting (XGBoost). In this study, experiments used the standard RAVDESS audio dataset with eight emotions: angry, calm, happy, surprised, sad, disgusted, fearful and neutral. The maximum accuracy obtained in the angry class using SVM was 85%, while overall accuracy was 80.1 %. The empirical works have proved that using BoW achieves better results in terms of accuracy and processing time compared to other available methods.

Keywords: *bag-of-words (BoW), Mel frequency cepstral coefficients (MFCC), RAVDESS database, support vector machine (SVM), K-nearest neighbors (KNN), and extreme gradient boosting (XGBoost).*

1. Introduction

Speech is a natural modality of human machine interaction. The purpose of sophisticated speech systems should not be limited to message processing; rather they should understand the underlying intentions of the speaker by detecting expressions in speech [1]. Speech systems may reach human equivalent performance only when they can process underlying emotions effectively [2]. In the recent past, processing speech signal to recognize underlying emotions has emerged as an important research area. Effective speech emotion recognition models should be able to recognize speakers' emotion and perform the actions accordingly.

Speech emotion recognition has been used in several daily-life applications. For example, speech emotion recognition applications have been applied human–computer interaction [3], computer vision [4] , and speech recognition [5]. Speech emotion recognition systems have also been used in forensic science, to investigate and detect criminals based on their speech and emotions [6]. In addition, attempts have been made to apply speech emotion recognition to artificial intelligence machines, robotics, toys, video games and call centers [7]. Medical doctors may use the emotional contents of a patient's speech as a diagnostic tool for various disorders [6]. Recently, speech emotion recognition systems that are trained to recognize stressed speech may be used in aircraft cockpits and on-board vehicle driving systems to avoid accidents [8].

Speech emotion recognition is remains a challenging problem due to the limited understanding of speech features; consequently determining which features are effective for speech emotion is difficult. Although, many speech features have been explored in speech emotion recognition, there is not yet a general agreement on which features are the most important, and good features appear to be highly data dependent [9]. In addition, there are some limitations that degrade most emotion recognition models for almost all existing emotional speech databases [7]. The primary issue that limits recognition accuracy is the lack of benchmarking databases that can be shared among researchers. Another issue is the lack of coordination among researchers in this field; thus, the same mistakes in recording are being repeated for different emotional

Copyright ©Mai Ezz-Eldin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

speech databases.

This paper focuses on speech features and ports the bag-of-word (BoW) technique from image to speech for feature enhancement for emotion recognition. The study uses the RAVADESS audio dataset with eight emotions: angry, calm, happy, surprised, sad, disgusted, fearful, and neutral. The main advantages of using BoW are increased recognition accuracy and reduced processing time. Several classifiers, such as support vector machines (SVM), K-nearest neighbor (KNN), naive Bayes(NB), random forest (RF), and extreme gradient boosting (XGBoost) are combined with the BoW to form a complete recognition system. The proposed system is developed, evaluated, and its performance is compared to the state-of-the-art approaches like [10].

The primary contributions of this study span multiple dimensions. First, the study explores and implements the BoW technique to improve features extracted from a standard speech benchmark database using the Mel frequency cepstral coefficients (MFCC) approach. Using BoW in speech emotion recognition requires a careful design of the BoW module, specifically the input layer, processing layer, and output format. Second, five several classifiers are integrated into the recognition system, to ensure the validity of the BoW deployment. Third, the proposed recognition system is developed and evaluated using the RAVDESS dataset as a standard benchmark. The achieved accuracy shows the superiority of using BoW in speech emotion recognition.

The remainder of this paper is organized as follows: Section 2 discusses previous studies related to speech emotion recognition. The details of the proposed system, including related Background and a description of the emotion recognition methodology are provided in Section 3. Section 4 focuses on system implementation and performance evaluation with comments on the obtained results and metrics. An analysis of the obtained results is discussed in Section 5. Finally, conclusions and suggestions for future work are given in Section 6.

2. Prior Research

This section focuses on the state-of-the-art for machine learning and deep learning techniques in speech emotion recognition. Three shared emotion recognition models, i.e., simple, single task hierarchical, and multi-task hierarchical models, have been proposed [11]. They used audio-visual features, i.e, energy, 41 spectral, 14 MFCC and 6 voice related LLDs for acoustic features and rasta, upper face and lower face for visual features. This study used the RAVDESS and UMSSED databases, which represent two domains (speech and song). The highest result was achieved with the RAVDESS dataset to 92% in happy class and in UMSSED dataset achieved 81% in the same class. In [12], they also used two types of datasets, the EmoDB and DES with MFCC, total energy, $F0$ (which was defined as the musical pitch of a note that is perceived as the lowest partial present) as feature inputs and achieved 77.5% and 66.8% overall accuracy for four and five class classification on the EFN dataset respectively. In addition, they achieved 67.6% accuracy on DES (five classes) and 63.5% on the EmoDB (seven classes) dataset using an ensemble of SVMs.

In [13], they proposed a bimodal emotion recognition system using a combination of facial expressions and speech signals. They used two different databases: Berline database for speech emotion recognition and FEEDTUM emotion database for facial expressions. They trained their model using different classifiers such as SVM, NB and KNN. They also used feature level fusion and match score level fusion. In [13], the classification accuracy achieved using SVM for speaker independent recognition was greater than 75% and for speaker dependent recognition, accuracy was greater than 80%.

In [14], the proposed model used RECOLA datasets, LLDs, and MFCC features and applied the Bag-of-audio-words (BoAW) technique. The highest result achieved was 75.3%.
Bombatkar et_al. [15] use a KNN classifier to classify four emotional states, i.e., anger, sadness, neutral, and happiness. The speech samples are from Berlin emotional database. The system achieved 86.02% classification accuracy for energy, entropy, MFCC, ZCC, and pitch features. Khan et_al. [16] performed emotion classification using a KNN classifier and obtained average accuracy of 91.71% with iterative forward feature selection, while the SVM classifier achieved 76.57%. Khan et_al. developed an English database comprising seven classes, i.e, Neutral, Anger, Surprise, Disgust, Fear, Happiness and Sadness. They also used Formant 0-4 (formant is defined as a peak, or local maximum, in the spectrum.), threshold entropy, sure entropy, norm entropy, median, MFCC, pitch, variance, Shannon entropy and log entropy as input features to KNN and SVM classifiers.

Another study [17] used LDC and KNN classifiers on BASE, formant $(f10)$, formant $(f15)$, and PCA features. A formant is a concentration of acoustic energy around a particular frequency in the speech wave. There are several formants, each at a different frequency. Results show that combining all the information, rather than using only acoustic information, improves emotion classification by 40.7% for males and 36.4% for females where an LDC is for acoustic information. On the other hand, in [10], RAVDESS and UMSSED datasets with LLDs features are used. They proposed four different models: simple, a single-task, MTFS/MTFL, and GMTFS/GMTFL models which achieves the highest result reached to 64.29%.

Currently, applying deep learning to emotional speech recognition is attracting significant attention. In [18] they used deep belief networks (DBN) to train the dataset and investigated speech emotion recognition based on DBNs. In that study, the recognition rate is improved further. The system's speech emotion recognition rate reached 86.5%, which was 7% higher than the original SVM method. In addition, in [19], a restricted Boltzmann machine (RBM) and DBN were applied together to audio files of a female Spanish speaker in an emotional speech database [20] as part of large project, INTERFACE, involving, English, French, Slovene, and Spanish. It consisted of repeating 184 sentences with the big six emotions together with several neutral styles. In all experiments 70% of the patterns were used for training (770 patterns), 25% for testing (275 patterns), and 5% for validation (55 patterns) using MFCC feature. The RBM and DBN provided a classification error rate of 18.37% and 24.49%, respectively for DBN-DT vs 34.69% for DT, MLP achieved 40.82%, where DBN-MLP achieved 20.41% and 25.43%, respectively for SVM vs 18.97% for DBN-SVM.
All previous studies agree that, although many speech features have been explored, which features are effective for speech emotion recognition

remains unclear. In addition, the accuracy and performance of previously proposed emotion recognizers still do not come close to reaching human levels and also consume significant processing time due to some database limitations. A previous study Theses limitations briefly mentioned in [7] indicated that there are very few benchmark databases that can be shared among researchers because most developed emotional speech databases are not publicly available. Another concern is the lack of coordination among researchers in this field; thus, the same mistakes in recording are being repeated for different emotional speech databases.

3. Proposed Recognition System

The quality of the speech samples, the extracted features, and the classification algorithm are the most important factors that influence the performance of any speech emotion recognition system. Integration of these modules provides us with an application that can recognize user emotions and give it as input to the system to respond appropriately. Classifying features involves training various emotional models to perform the classification task. We conduct several tests and experiments and analyze system accuracy in terms of speech samples, extracted features, and classification algorithms.

The basic components of the proposed system are shown in Fig. 1. The system was designed to recognize emotions using the RAVDESS audio dataset with eight emotions: angry, calm, happy, surprised, sad, disgusted, fearful, and neutral. The proposed system comprises three modules: an MFCC feature extractor, feature clustering using a BoW technique, and a classification module that uses five different classifiers types (support vector machine, K-Nearest Neighbor, Naive Bays, Random Forest, and extreme Gradient Boosting). In the classification modules, each classifier is used once to give five different systems to be tested with the proposed BoW and clustering module. Then, the output of the classifier is the recognized emotions with a BoW and cross validation technique. The methodology of our proposed models is shown in the following algorithm 1.

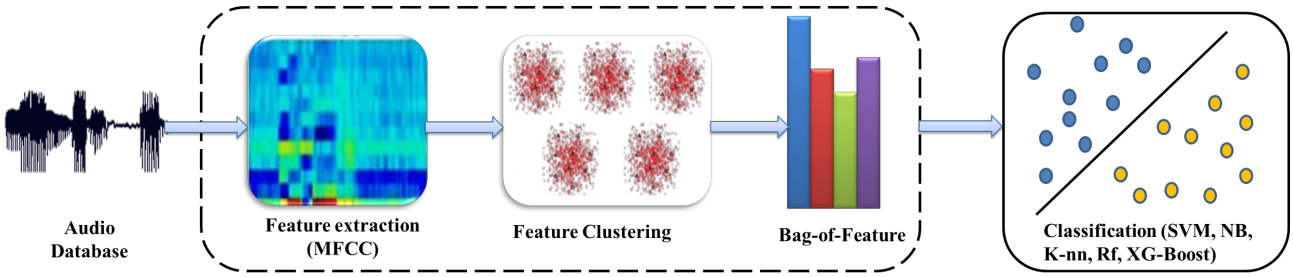


Figure 1: Proposed System Block Diagram

3.1. Feature Extraction Process

Feature extraction and the selection module that identifies the features that efficiently characterize the speech emotional contents are important issues in speech emotion recognition systems. The speech emotion extractor should not depend on the speaker's language or its lexical content. Different speech features represent different information, such as speaker gender, language, and emotion in a highly overlapped manner [9]. After extracting the emotions, one has to decide which features are to be selected; selecting the features that will be used to evaluate any of the speech systems is a crucial decision.

Classification accuracies are observed to be consistently higher for class-level spectral features compared to prosodic or utterance-level spectral features. The combination of class-level features with prosodic features improved the emotion recognition performance. Furthermore, research results have shown that spectral features computed from consonant regions contain more emotion specific information than either stressed or unstressed vowel features. It has also been found that, the average emotion recognition performance is proportional to the length of the utterance [21]. On the other hand, spectra of vowel segments obtained using Fourier and chirp transforms have been analyzed for emotion classification. The analysis demonstrated that, higher frequency regions of speech are suitable for characterizing stressed speech [22].

MFCC features are spectral features that have been used successfully for various speech processing tasks, including speech development and speaker recognition systems. Determining MFCCs consists of multi-stages, as illustrated in Fig. 2. Generally, MFCC features extracted from lower frequency speech signal (20 to 300 Hz) are proposed to model pitch variation. In this paper, we use MFCC features where the length of the analysis window is 25 ms and the step between successive windows (winstep) is 10 ms. Moreover, the default number of cepstrum to return (numcep) is 13 where the default number of filters in the filterbank (nfilt) was 26. The default Non-equispaced Fast Fourier Transform (nfft), i.e., "the FFT size" was 512. Finally, the zeroth cepstral coefficient is replaced by the log of the total frame energy.

3.2. BoW and Clustering Processes

After applying MFCC features extraction, each audio signal is converted to several frames based on the length of the signal. Using MFCC features, eeach frame is decoded into a vector of length 512. To obtain a single vector that represents each audio signal, we apply the BoW technique. By applying a clustering algorithm (e.g., K-means) on our training dataset, similar frames (represented by MFCC features vectors) are grouped into a single cluster based on Euclidean distance or similarity functions [23]. All the frames/vectors in a group should be similar and no two clusters should have similar frames. All the frames extracted from all audio signal in the training data are clustered into 250 unique audio clusters. After applying the clustering step, the result is similar to a bag-of-audio-words (audio clusters) where each audio

Algorithm 1 Proposed Model Using Different Classifiers

Require: $D_{Data} = [D_1, D_2, ...D_N]$. N is number of audio files of database.
Require: $CLASSES = [L1, L2, ...Lm]$. L_i is the label of the class and m is the number of classes.

for *i in range*$(1, N)$: **do**
    Split D_i into M frames.
    for *j in range*$(1, M)$ **do**
        decode frame-j into vector v with length 512 using MFCC feature extraction.
    end for
end for
Setting number of cluster equal to C and apply cluster algorithm on all extracted vectors v. Similar v are grouped together into one cluster based on Euclidean distance.
Build BoW using cluster centroids.
for *i in range*$(1, N)$: **do**
    applying the histogram on all vecter of D_i to get only one vector with length C.
end for
Array_fold = Split audio data into K fold.
for *y in range*$(1, K)$: **do**
    Training data = concatenate array_fold $[u]$ where U from 1 : K and u != y.
    Testing data = array_Fold $[y]$.
    Input_Training = is a matrix with size = $\Re^{len(tainingdata)*c}$.
    Output_Training = is a matrix with size $\Re^{len(tainingdata)*1}$.
    Input_Testing = is a matrix with size $\Re^{len(testingdata)*c}$
    Output_Testing = is a matrix with size $\Re^{len(testingdata)*1}$
    {# Using one of different classifiers} ▷
    Define model as one of SVM (Eq. 1), KNN (Eq. 3), NB (Eq. 2), RF, Or XGBoost (Eqs. 4, 5, 6)
    model.train(input_training, output_training)
    Output_predict = model.predict(input_testing)
    using Output_predict and Output_test, Calculate Precision, Recall and f_1 score in Testing_data for each L_i where *i in range*$(1, m)$.
end for
Calculate the average of the metrics over all the K folds.

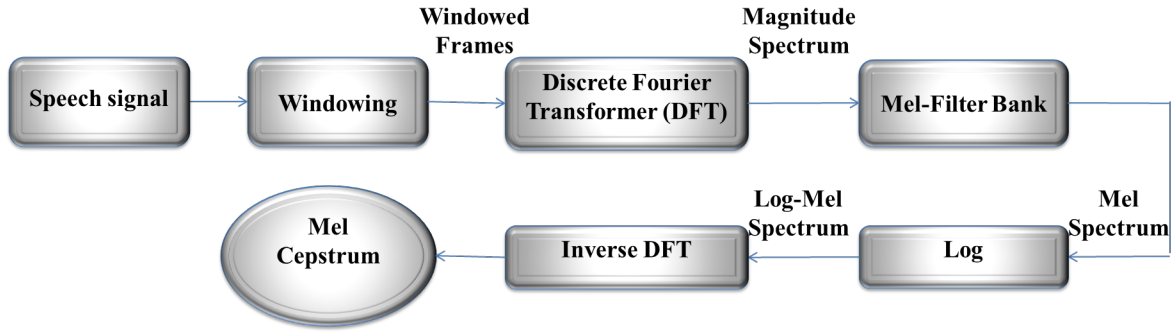


Figure 2: MFCC Block Diagram

cluster represents similar frames. Now, each audio signal can be represented by a histogram of the unique audio clusters based on its audio frames. In the proposed model, the audio words and their frequency of occurrence in the audio signal will be considered a feature vector for the audio signal.

3.3. Classification Process

Generally, most current speech emotion recognition research focuses on the classification process because it represents the interface between the classification techniques and the problem domain. Because each classifier has advantages and disadvantages, it is difficult to determine which classifier is performs better than the other. Although traditional classifiers have been used in almost all existing speech emotion recognition systems and have achieved good accuracy, the search for better classifiers remains an important task for researchers in diverse fields. There are two important stages in speech emotion recognition systems. The first stage, i.e., the front-end processing unit, extract appropriate features from the available (speech) data. The second stage involves classification to determine the underlying emotion of the speech utterance. The database used in classification process is divided into training and testing sets. The training dataset is used to obtain better boundary conditions, which are used to determine each target class. Once the boundary conditions are determined, the next task is to predict the target class.

3.3.1. Support Vector Machine Classifiers

A SVM is a popular binary classification technique that is used in emotional speech recognition to identify patterns and analyze the data for classification and regression analysis. SVM also use a kernel function to transform the original set of features to higher dimensional feature

space, which is necessary to obtain optimum classification in this new feature space. The goal of a SVM is to find the optimal separating hyperplane that maximizes the margin of the training data which contains eight classes as mentioned previously. We used the BoW output vector as the input to the SVM. Here, the output is a vector that represents a probability of each class from the eight classes in the RAVDESS dataset. The margin can be obtained by the following equation:

$$Margin = ||X_{+1} - X_{-1}|| = \lambda ||W|| = \frac{2||W||}{||W||^2} = \frac{2}{||W||} \tag{1}$$

3.3.2. Naive Bayes Classifier

In statistics and computer science literature and NB classifier is a simple "probabilistic classifiers". It known by various names, including simple Bayes and independence Bayes. The reason for these various names is that NB is based on applying Bayes' theorem with strong (naive) independence assumptions between the features, which is a way of going from $P(X|Y)$, known from the training dataset, to find $P(Y|X)$. The conditional probability can be decomposed as follows:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{2}$$

where $P(y|x)$ is the posterior probability of class $(y, target)$ given predictor $(x, attributes)$, $P(y)$ is the prior probability of class, $P(x|y)$ is the likelihood, i.e., the probability of predictor given class and $P(x)$ is the prior probability of the predictor.
NB classifiers are highly scalable and require a number of linear parameters relative to the number of variables $(features/predictors)$ in a learning problem. We use an NB classifier due to its statistical probability characteristics, which enables emotion classification using the BoW output as inputs. The output of this classifier is represented as vectors of feature values.

3.3.3. k-Nearest Neighbor Classifier

KNN is a non-parametric method that is frequently used due to its ease of interpretation and low calculation time. KNN can be used for both classification and regression predictive problems. In both cases, the input consists of the k closest training examples in the future space. Therefore, we can conclude that the KNN algorithm assumes that similar things exist in close proximity. The distance between the item and the first nearest neighbor can be calculated as follows.

$$d(p,q)^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2 \tag{3}$$

The output of KNN depends on whether it is used for a regression or classification process. In this method, the BoW output vector is used as input. The output is a vector that represents the distance between the tested label and the first neighbor from the training classes of dataset.

3.3.4. Random Forest Classifier

RF uses bootstrap aggregation (bagging) and random subspace (feature bagging). The goal of bootstrap is to reduce variance. For classification trees, the variance is reduced by a majority vote taken for each predicted class. Random subspace attempts to reduce the correlation between each tree. Reducing correlation between each tree is achieved by selecting features from a random subset of all features when building the tree rather than considering all features. The trees also tend to over fit and not be as robust, meaning that they are sensitive to data changes [24]. However, a solution to this problem is using tree ensemble methods [24]. Tree ensemble builds on the idea of building a "strong" model by combining an ensemble of "weak" learners. Two common tree ensemble algorithms are RF and Boosted Trees.
The RF classification algorithm comprises multiple decisions trees that use bagging and features to build each individual tree. Theses bagging and feature are used randomness. The RF algorithm input is the BoW output vector, and the output is an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree

3.3.5. Extreme Gradient Boosting Classifier

XGBoost is a variant of the gradient tree boosting proposed by Friedman [25]. Gradient tree boosting is a tree ensemble boosting method that combines a set of weak classifiers to create a strong classifier. The strong learner is trained iteratively starting with a base learner [26]. Both gradient boosting and XGBoost follow the same principal. The key differences between them lie in implementation details. XGBoost achieves better performance by controlling the complexity of the trees using different regularization techniques [26]. Boosting consists of three simple steps:
**An initial model $F0$ is defined to predict the target variable y. This model will be associated with a residual $(y–F0)$
**A new model $h1$ is fit to the residuals from the previous step
**Now, $F0$ and $h1$ are combined to give $F1$, the boosted version of $F0$. The mean squared error from $F1$ will be lower than that from $F0$:
To improve the performance of $F1$, we could model the residuals of $F1$ and create a new model $F2$.

$$f_m(1) < -f_0(x) + h_1(x) \tag{4}$$

This can be done for 'm' iterations, until residuals have been minimized as much as possible:

$$f_2(x) < -f_1(x) + h_2(x) \tag{5}$$

Here, the additive learners do not disturb the functions created in the previous steps. Instead, they impart information of their own to reduce errors.

$$f_m(x) < -f_{m-1}(x) + h_m(x) \tag{6}$$

In this experiment, gradient boosting is used to predict the residuals or errors of prior models and then added together to make the final prediction. In this method, the BoW output vector is used as input, and the output is a vector that uses a gradient descent algorithm to minimize the loss when adding new models.

4. Simulation Experiments and Results

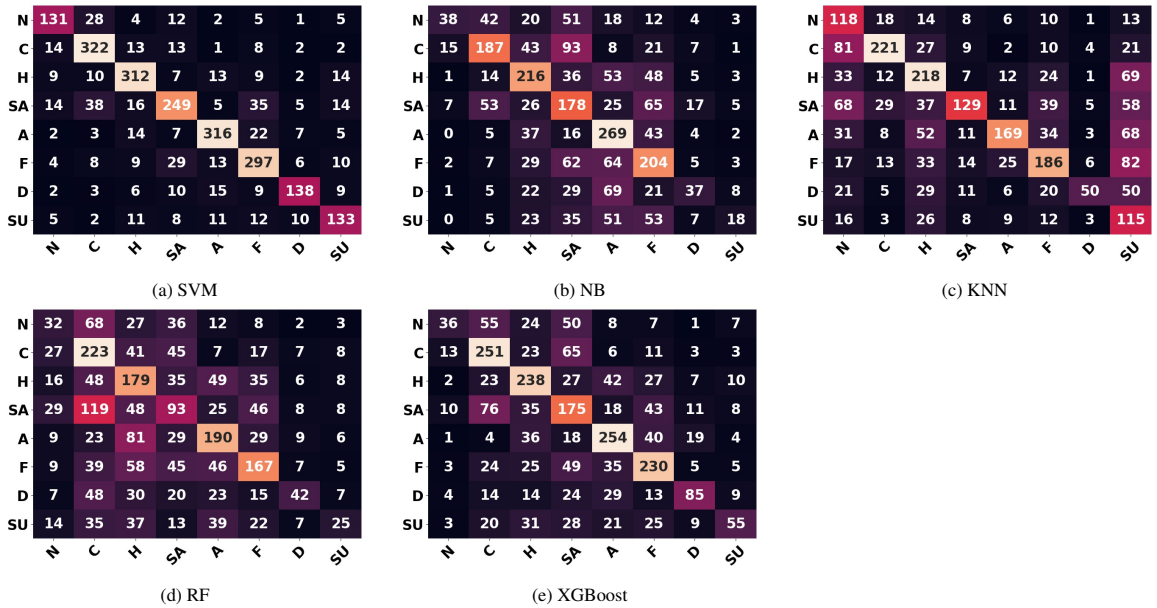
In this section, we describe experiments that use RAVDESS dataset [27], a publicly available dataset comprising simulated English expressions for training and testing. We used two parts of this dataset i.e., speech and song. The speech recordings comprise eight classes divided by class (neutra 96, calm 186, happy 192, sad 192, angry 192, fearful 192, disgust 192, and surprise 192). The speech was recorded by 24 professional actors (12 male, 12 female). The song recordings, which were recorded by 23 professional actors and divided by class, comprise consist of the first six emotions (neutral 90, calm 180, happy 180, sad 180, angry 180, and fearful 180) recorded by 23 professional actors. The same sentences were spoken and sung at normal and strong emotional intensity, each with two repetitions. The statistics for class, speaker, number of files, and language for speech and song are listed in Table 1.

Table 1: RAVDESS database details (Speech and Song)

Pattern	Speech	Song
Classes	Neutral, Happy, Calm, Angry, Sad, Fearful, Disgust, Surprised	Neutral, Happy, Calm, Angry, Sad, Fearful
Speakers	24 professional actors	23 professional actors
No. files	1440	1012
Language	English	English

In our experiment, the RAVDESS corpus was divided into training (80%) and test (20%) sets. The RAVDESS dataset has a relatively small number of data points, therefore, we used 5-fold cross validation to measure the average performance of the models under different data splits to avoid overfitting in a specific fold. In each round, a single fold serves as test data, and the training data are composed of the other folds. We use MFCC features and the Python programming language to implement all simulation experiments. In this paper, we used MFCC features where the length of the analysis window was 25 ms and winstep " the step between successive windows " equal to 10 ms. numcep " the number of cepstrum to return" is used as default 13 where nfilt " the number of filters in the filterbank" was as default 26. nfft " the FFT size" also as default was 512. Finally, the zeroth cepstral coefficient is replaced with the log of the total frame energy. winfunc – the analysis window to apply to each frame. All experiments were performed using an HP laptop with an Intel(R) Core(TM) i5-3210M CPU @2.50GHZ process and 4.00 GB RAM running the Windows 10 Pro 64-bit operating system.

Confusion matrices of the classification results for multiple classes using SVM, NB, KNN, RF and XGBoost are shown in Fig. 3 for all data from speech and song. The darker the color, the lower the accuracy/confusion. As shown in Fig. 3, the SVM achieves the highest results in all classes. RF, NB, and KNN returned the lowest results for neutral, happy, sad, and fearful classes, calm, disgust, and surprised classes, and the angry class, respectively.



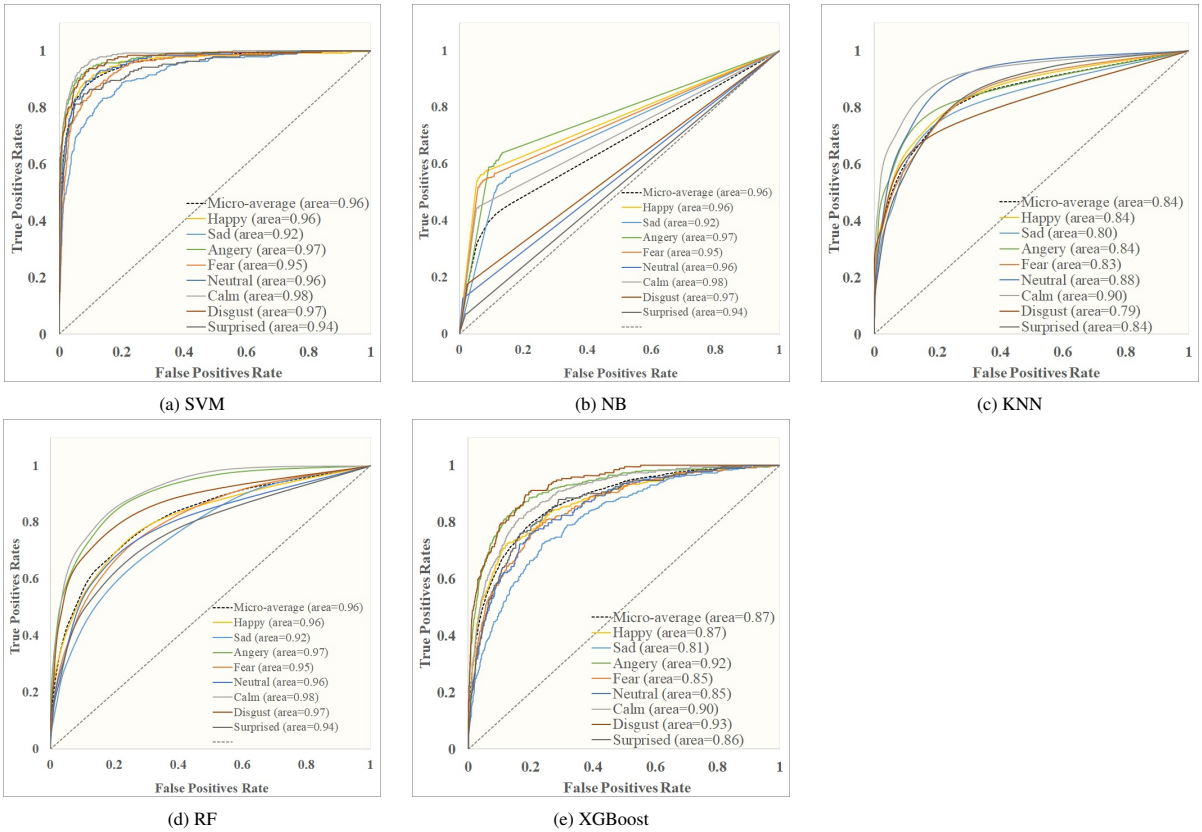
(a) SVM (b) NB (c) KNN

(d) RF (e) XGBoost

Figure 3: Confusion matrix of eight classes for traditional models; (a) SVM, (b)NB, (c) KNN, (d) RF, and (e) XGBoost. **Abbreviation:-** N: Neutral, C: Calm, H: Happy, SA: Sad, A: Angry, F: Fearful, D: Disgust and SU: Surprised. (with taking the same order in row from left to right and in column from top to down).

The receiver operating characteristic (ROC) curves, shown in Fig. 4, are created by plotting the false positive rate (FPR) against the true positive rate (TPR) at various threshold settings using SVM, NB, KNN, RF, and XGBoost classifiers. The TPR, also known as sensitivity,

measures the proportion of correctly identified emotions. Similarity, the FPR, also known as specificity, measures the proportion incorrectly identified emotions. Performance for each emotion can be measured by the area under the ROC curve, that indicates how each emotion is distinctively classified compared to other emotions.



(a) SVM (b) NB (c) KNN

(d) RF (e) XGBoost

Figure 4: ROC Results figure: shows the Receiver operating characteristic to multi-class using SVM, NB, Knn, RF and XGBoost

5. Discussions

As illustrated in the confusion matrices, the classification results for emotion recognition using the SVM classifier achieves the highest result compared to all other classifiers. The best result with the SVM classifier is obtained for the angry class (85%). The SVM achieved 80.04% overall precision results. On the other hand, RF classifier returned the lowest overall precision results (39%). The accuracy (%) achieved by state-of-the-art's models and the proposed model (%) is listed in Table 2. In this table, we list only studies that used the RAVDESS dataset with eight classes (only speech, only song or both of speech and song). From Table 2, it can be seen that, with only speech files, the maximum accuracy (79.5%; 5 minutes training time) achieved by a state-of-the-art model occurred with a CNN classifier [28]. The proposed SVM with MFCC and BoW achieved 79.36% accuracy with 5 minutes training time. On the other hand, with both RAVDESS speech and song files, using GResNets achieved 64.48 % accuracy [29], while the proposed SVM + BoW model achieved to 80.1 % overall accuracy.

In addition, from the results presented in Table 2, it is evident that song return the highest result because songs files have hard and clear tones; therefore, it is easy to train the model and predict the emotions in each class. The proposed SVM model achieved the highest overall accuracy (88.48%).

6. Conclusions and Future Work

In the field of human-computer interaction, automatic speech emotion recognition has emerged as an important research area in the recent past. Emotion recognition in speech is a challenging problem because it is unclear which features are effective for recognition. In this paper, we have proposed a method aims to recognize emotions using the RAVDESS audio dataset with speech files (eight emotions: angry, calm, happy, surprised, sad, disgusted, fearful and neutral) and song files (six emotions: angry, calm, happy, sad, fearful and neutral). In the proposed speech emotion recognition model, MFCC, SVM, KNN, NB, RF, and XGBoost classifiers are used together with BoW and clustering modules. As illustrated in Table 3, the best result is achieved in all classes using the SVM classifier, and the highest result (85%) occurs in the angry class. With the neutral class, the SVM classifier achieved 73 %. The recognition rate for the calm class was 78%. For the happy, sad, fear, disgust, and surprise classes, the rates were 81%, 75 %, 82% and 69%, respectively. Overall precision with the SVM classifier was 80.1 %. The other proposed models achieved overall precision ranging from 58% to 39%. The lowest overall precision (39%) occurred with the RF classifier.

Few studies that have considered applying BoW on deep learning to speech emotion recognition. We believe that this research direction should be explored further.

Table 2: Comparison of state-of-the-art and proposed models (%) (using RAVDESS database, eight classes).

Authors	Feature	classifier	ACC.% (Speech)	ACC.% (Song)	ACC.% (All)
[28]	Spectrogram	CNN	79.5	–	–
[29]	Spectrogram	GResNets	–	–	64.48
[?]	MFCCs, spectral centroids and MFCC derivatives	SVM	75.69	–	–
[30]	eGeMAPS, supervector, log-spectrogram, F0, MFCC and log-Energy	BLSTM	69.4	–	–
[31]	CWT & prosodic	SVM	60.1	-	–
Proposed model	MFCC + BoW	SVM	79.36	88.48	80.1
		NB	42.96	58.08	46.83
		KNN	44.63	66.16	49.30
		RF	39.19	48	38.80
		XGBoost	50.56	48.58	54.12

Table 3: Comparison between 8-class emotion classification Precision of different models (%).

Classifier	Neutral	Calm	Happy	Sad	Angry	Fear	Disgust	Surprised	ALL
SVM	73	78	81	75	85	75	82	69	80.1
NB	61	59	52	35	48	44	43	43	48
KNN	31	71	50	65	71	55	70	24	58
RF	22	37	36	29	49	49	48	36	39
XGBoost	51	54	56	40	62	58	62	54	54

References

[1] M. Schröder, Emotional speech synthesis: A review, in: EUROSPEECH Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, Vol. 7, September 3-7, 2001, pp. 561–564.

[2] D. O'Shaughnessy, Speech communication: Human and machine addison wesley 20 (November 30, 1999) 548 pages.

[3] M. Joshi, S. Srivastava, Human computer interaction using speech recognition technology, in: National Conference on Mathematical Analysis and Computation (NCMAC) 2015, At JAIPUR, 2015.

[4] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, arXiv preprint arXiv:1207.0580 (3 Jul 2012).

[5] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada, IEEE, 26 - 31 Mar 2013, pp. 6645–6649.

[6] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk, IEEE transactions on Biomedical Engineering 47 (7 Jul 2000) 829–837. doi:10.1109/10.846676.

[7] M. El Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition 44 (3) (March 2011) 572–587. doi:10.1016/j.patcog.2010.09.020.

[8] B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, Vol. 1, IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Que., Canada, 17-21 May 2004, pp. I–577. doi:10.1109/ICASSP.2004.1326051.

[9] L. Devillers, L. Vidrascu, L. Lamel, Challenges in real-life emotion annotation and machine learning based detection, Neural Networks 18 (4) (May 2005) 407–422. doi:org/10.1016/j.neunet.2005.03.007.

[10] B. Zhang, E. M. Provost, G. Essi, Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning approach, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, IEEE, 20-25 March 2016, pp. 5805–5809. doi:10.1109/ICASSP.2016.7472790.

[11] B. Zhang, G. Essl, E. M. Provost, Recognizing emotion from singing and speaking using shared models, in: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, IEEE, 21-24 Sept. 2015, pp. 139–145. doi:10.1109/ACII.2015.7344563.

[12] T. Danisman, A. Alpkocak, Emotion classification of audio signals using ensemble of support vector machines, in: International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, Berlin, Heidelberg, Springer, 2008, pp. 205–216. doi:10.1007/978-3-540-69369-7_23.

[13] S. Emerich, E. Lupu, A. Apatean, Emotions recognition by speech and facial expressions analysis, in: 2009 17th European Signal Processing Conference, Glasgow, Scotland, IEEE, 24-28 August 2009, pp. 1617–1621.

[14] M. Schmitt, F. Ringeval, B. W. Schuller, At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech., in: Interspeech, San Francisco, USA, 8–12 September 2016, pp. 495–499. doi:10.21437/Interspeech.2016-1124.

[15] A. Bombatkar, G. Bhoyar, K. Morjani, S. Gautam, V. Gupta, Emotion recognition using speech processing using k-nearest neighbor algorithm, International Journal of Engineering Research and Applications (IJERA), Cranfield, Bedfordshire, United Kingdom (12-13 April 2014) 2248–9622.

[16] M. Khan, T. Goskula, M. Nasiruddin, R. Quazi, Comparison between k-nn and svm method for speech emotion recognition, International Journal on Computer Science and Engineering 3 (2) (Feb 2011) 607–611.

[17] C. M. Lee, S. S. Narayanan, Toward detecting emotions in spoken dialogs, IEEE transactions on speech and audio processing 13 (2) (22 February 2005) 293–303. doi:10.1109/TSA.2004.838534.

[18] B. Chen, Q. Yin, P. Guo, A study of deep belief network based chinese speech emotion recognition, in: 2014 Tenth International Conference on Computational Intelligence and Security, Kunming, China, IEEE, 15-16 Nov 2014, pp. 180–184. doi:10.1109/CIS.2014.148.

[19] M. E. Sánchez-Gutiérrez, E. M. Albornoz, F. Martinez-Licona, H. L. Rufiner, J. Goddard, Deep learning for emotional speech recognition, in: Mexican conference on pattern recognition, Springer, Cham, 2014, pp. 311–320. doi:10.1007/978-3-319-07491-7_32.

[20] J. M. Montero, J. M. Gutierrez-Arriola, S. Palazuelos, E. Enriquez, S. Aguilera, J. M. Pardo, Emotional speech synthesis: From speech database to tts, in: Fifth International Conference on Spoken Language Processing, Sydney, Australia, 30th November - 4th December 1998.

[21] D. Bitouk, R. Verma, A. Nenkova, Class-level spectral features for emotion recognition, Speech communication 52 (7-8) (July–August 2010) 613–625. doi:10.1016/j.specom.2010.02.010.

[22] M. Sigmund, Spectral analysis of speech under stress, IJCSNS International Journal of Computer Science and Network Security 7 (2007) 170–172.
[23] L. Liberti, C. Lavor, N. Maculan, A. Mucherino, Euclidean distance geometry and applications, SIAM Review 56 (1) (2014) 3–69. doi:10.1137/120875909.
[24] G.James, An introduction to statistical learning: with applications in R, Springer Verlag New York, 2013. doi:10.1007/978-1-4614-7138-7.
[25] J. H. Friedman, Greedy function approximation: a gradient boosting machine, Annals of statistics 29 (2001) 1189–1232. doi:10.1214/aos/1013203451.
[26] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, California, USA, ACM, 13 - 17 August 2016, pp. 785–794. doi:10.1145/2939672.2939785.
[27] S. Livingstone, F. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, Vol. 13, 2018, p. e0196391. doi:10.1371/journal.pone.0196391.
[28] M. ., S. Kwon, A cnn-assisted enhanced audio signal processing for speech emotion recognition, Sensors 20 (2019) 183. doi:10.3390/s20010183.
[29] Y. Zeng, H. Mao, D. Peng, Z. Yi, Spectrogram based multi-task audio classification, Multimedia Tools and Applications 78 (3) (2019) 3705–3722. doi:10.1007/s11042-017-5539-3.
[30] M. A. Jalal, E. Loweimi, R. K. Moore, T. Hain, Learning temporal clusters using capsule routing for speech emotion recognition, in: Proc. Interspeech 2019, Graz, Austria, 15–19 September, 2019, pp. 1701–1705. doi:10.21437/Interspeech.2019-3068.
[31] P. Shegokar, P. Sircar, Continuous wavelet transform based speech emotion recognition, 2016. doi:10.1109/ICSPCS.2016.7843306.