An executable method for an intelligent speech and call recognition system using a machine learning-based approach

Rajarajeswari, P and Beg, OA

http://dx.doi.org/10.1142/S021951942150055X

Journal of Mechanics in Medicine and Biology

(Publisher: World Scientific)

ISSN (print): 0219-5194 | ISSN (online): 1793-6810

Accepted July 13th 2021

AN EXECUTABLE METHOD FOR AN INTELLIGENT SPEECH AND CALL RECOGNITION SYSTEM USING A MACHINE LEARNING-BASED APPROACH

Perepi Rajarajeswari*

Associate Professor, Department of Computer Science and Engineering, Kingston Engineering College, Vellore-632059,Tamilnadu, ***India.***

***(*Corresponding author- Email:** rajacse77@gmail.com)*

O. Anwar Bég

Professor of Engineering Science, Department of Mechanical and Aeronautical Engineering, SEE, Salford University, Manchester, M54WT, UK. Email: O.A.Beg@salford.ac.uk; gortoab@gmail.com

ABSTRACT: This paper describes a novel call recognizer system based on the *machine learning* approach. Current trends, intelligence, emotional recognition and other factors are important challenges in the real world. The proposed system provides robustness with high accuracy and adequate response time for human-computer interaction. Intelligence and emotion recognition from speech of human-computer interfaces are simulated via multiple classifier systems (MCS). At a higher level stage, the acoustic stream phase extracts certain acoustic features based on the pitch and energy of the signal. Here featured space is labelled with various emotional types in the training phase. Emotional categories are trained in the acoustic feature space. The semantic stream process converts speech into-text conversion in the input speech signal. Text classification algorithms are applied subsequently. The clustering and classification process is performed via a K-means algorithm.The detection of the Tone of Voice of call recognition system is achieved with the XG Boost Model for feature extraction and detection of a particular phrase in the client call phase. Speech expressions are used for understanding human emotion. The algorithms are tested and demonstrate good performance in the simulation environment.

KEYWORDS: *Machine learning, Speech recognition, Call recognition, Multilayer perceptron, XG boost model.*

1.INTRODUCTION

In daily life, the speech signal is a critical factor in enabling communication between human beings. Emotions also contribute a significant role in developing the speech recognition system. In recent years, there has been an increasing focus on “smart” or “intelligent” healthcare systems in biomedical engineering. Corporeal machine interfaces are an option for restoring communication with environmental control. Recognition of human emotions is considered as a critical challenge in the 21st century. Human emotions are considered as the main component of emotional interaction. Recently, web movies, computer tutorial applications, virtual interactive training systems can be used for

detecting the emotions of system users which leads to a more robust and comprehensive speech recognition system. (Elliot and Brzezinski 1998) ,analyzed synthetic personality, representations of emotion, social interfacing as key emotions which enrich the human–computer interface. They highlighted that increasingly new web-software consumers and increase in local computing power will accelerate the development of speech recognition and sophisticated real-time expressive graphics. (Bezooijen *et al.,* 1983) conducted fundamental work on speech recognition systems by analyzing different cultural dialects (Dutch, Taiwanese, and Japanese) to better characterize identification of vocal expressions of emotion beyond chance expectancy. They used multidimensional scaling to demonstrate that universally recognizable characteristics of vocal patterns of emotion exist which are principally associated with the activity dimension of emotional meaning. Picard [3] an electrical and computer engineer of MIT's Media Lab, pioneered a new approach in computer science termed "affective computing" which has significantly revolutionized speech recognition via embedding emotional intelligence and human emotion communication features. This has encouraged in particular modern research in autistic disorders and improving nuance recognition in human-computer interaction. (Murray and Arnott 1993) identified three types of voice parameters affected by emotion, namely voice quality, utterance timing, and utterance pitch contour. They emphasized that the pitch envelope (i.e., the level, range, shape, and timing of the pitch contour) is the key factor differentiating between the basic primary emotions, whereas voice quality is the principal factor delineating between the secondary emotions. They also observed that acoustic correlates of basic emotions are cross-cultural, whereas those of the secondary emotions are culturally specific and their findings have been subsequently deployed in the development of higher quality speech synthesizer designs in the past three decades. (Dellaert *et al.*,1996) studied a variety of statistical pattern recognition techniques to classify utterances according to emotional content, developing a novel pattern recognition technique for extracting prosodic features from speech via smoothing spline approximation of the pitch contour. Scherer *et al.*,1991 conducted tests on vocal expression patterns in naturally occurring emotions (as based on the component process theory of emotion. The speech signal is considered as the *fastest* communication process between humans (Chiriacescu 2008). Traditional speech emotion systems have been focused on *prosodic features* (which appear when sounds are put together in connected speech) or *spectral characteristics* which constitute the foundation for language processing. During recent decades, refining speaker speech systems has been a challenging task based on the changes of culture and environment of speakers due to variations in different speaker styles, dialects, modern expressions introduced in language, slang etc. It is also known that the quality of performance of speech recognition algorithms degrade in the presence of adverse environments where a speaker is under stress, emotion- this is termed the Lombard effect identified in the early 20th century. Despite

significant progress, all the emotions are still not detected clearly in any existing call recognition system. Many modern automatic speaker recognition (ASR) systems have utilized the Mel-Frequency Cepstrum Coefficients (MFCC) and Gaussian Mixture Models (GMM), developed by MIT scientist Reynolds in the early 1990s [9]. In subsequent studies by Reynolds [10], human recognition rates have been compared with reality and acted emotional speech of spectrograms. (Reynolds and Rose 1995) described the results of a detailed experimental evaluation of the Gaussian mixture speaker model on a 49 speaker, conversational telephone speech database. (Bou-Ghazale and Hansen 2000)examined the effectiveness of traditional features in recognition of speech under stress and produced novel features which enhance stressed speech recognition by coding robust features which are less dependent on the speaking conditions as an alternative to deploying conventional compensation or adaptation techniques. The stressed speaking styles were simulated as "angry" and "loud". Many researchers have as such reported on the emotional process like voice, facial expressions due to brain activities.

Machine learning is used to identify patterns in streams of inputs. Classification schemes are used to determine the category of an object which belongs to set of numerical inputs or output examples. Mathematical analysis of machine learning algorithms is a computational learning theory process and can be used to define facial recognition, object recognition etc. In daily life, speech emotion recognition has several applications which include call center conversations for analyzing the call attendants (for improving the quality of service call attendance), medical diagnosis analysis of human beings, security protocol etc (Chiriacescu 2009). As noted earlier, previously Gaussian mixture models have been implemented for observing the performance of emotions such as 93%, 85%, 89% of male and female speakers in different emotional states. Traditional methods are not able to focus accurately on the negative emotions. Currently, speech recognition can be done through many ways including facial expression, voice, body gestures, as well as body movements, in which facial emotion recognition and speech emotion recognition are key factors in designing human machine interaction. Oral communication is a rich source of information for emotion recognition when people communicate with one another. It is used in speech emotion in call center applications through mobile communication networks. A key objective in a an improved call recognition system is detecting frustration in the speaker's voice. As elaborated earlier, another problem is that the expression of emotion is strongly influenced by the speaker and their culture and environment. As there is a change in culture and environment for different speakers, their speaking style is also modified significantly (MacNeilage 2008 and Corballis 2002). Two or more types of emotions may occur simultaneously so it may be unclear for the recognizer to precisely detect which type of emotion is being experienced by human beings. To overcome these problems, intelligent speech and call recognition system are

required and these are addressed via a *machine learning methodology* in this paper. This proposed system can be used to increase the execution and performance with the help of an XG boost model. The novel principal contributions of the present study are:

1. We learn how people recognize speech and emotions to more precisely categorize features of the speech signal and also explore machine learning approaches for creating *reliable recognizers*.
2. We propose a new approach for a smart call recognition system via machine learning techniques.

3. We describe the machine learning techniques for the detection of the Tone of Voice of a call recognition system via the *XG Boost Model for Feature Extraction* and *Detection of a Particular Phrase* in the client call phase.The structure of this paper is as follows. The introduction is given in section 1. The related work in the literature is elaborated in section 2. We have described the basic machine learning algorithmic methodology in section 3. The proposed approach for the intelligent call recognition system is presented in sections 4 and 5. We have introduced Detection of the Tone of Voice in the whole File by using Machine learning with associated results in section 6. Finally Conclusions and future work are summarized with a state-of-the-art reference section at the end of the article.

2.LITERATURE SURVEY AND RELATED WORK

Computers can be used strategically to both express and interact with emotions in human perception. Neurological studies have shown over decades that “affected” computers significantly enhance the human experience. The current study embraces aspects of affective computing, which relates to emotional influences in speech recognition. It builds on previous investigations which have addressed numerous aspects of human-cumputer interaction. For example (Ayadi *et al*. 2011) studied speech emotion recognition features, databases and classification schemes. They developed a non-Gaussian approach i.e. Maximum Likelihood Parameter-based classifiers which have been shown to attain the best recognition performance on neutral emotions with a 72.40 percent efficiency.(Vogt *et al*.2008) proposed an online processing method for improving emotion recognition and extraction features, with particular focus on intra- and interspeaker variations. (Emerich and Lupuand 2009)considered facial expressions in speech recognition. Further studies are presented in (Lee and Narayanan ,2005) (on speech emotion recognition features, classification schemes). Antosik-Wójcińska *et al.* 2020) have recently studied speech recognition for depressive illness patients using smartphones and machine learning algorithms (deep neural networks). A comprehensive discussion on pattern recognition and data clustering techniques for speech recognition is given by (Jang 2020) which includes statistical variance in facial recognition. A summary of these and other investigations is categorized by

application in **Table 1**. Emotion recognition has databases featuring actions associated with emotions, natural emotions, etc. The best results can be generated by using strong emotional expressions. However such databases generally do not feature natural facial expressions. In a laboratory environment, image sequences and frontal face views can be considered. Each record has neutral state at the start and end of the testing procedure.

Table 1: Literature survey summary on emotional speech recognition systems

Author	Application	Year
(Reynolds ,2008)	Probability distribution model for a biometric system	2008
(Jang ,2020)	Data clustering	2001
(Ayadi *et al.* 2011)	Classification schemes of Speech emotion recognition system	2011
(Chiriacescu *et al.* 2009)	Automatic speech emotional system	2009
(Emerich *et al.* 2009)	Facial expressional system with emotions	2009
(Ververidis *et al*. 2006)	Methods of emotional speech recognition system	2006
(Zhou *et al*. 2006)	SVM for speech emotion recognition system	2006
(Bashashati *et al.* 2007)	Signal processing algorithm	2007
(You *et al*. 2006)	Framework process of speech emotion recognition system	2006
(Lee/Narayanan ,2005)	Robust speech emotion recognition	2017

Table 1: Literature survey

In text categorization processes, a *text-analytics community* is followed. Support vector machines are used for making the learning of the text classification. Usually, each vector component value is allocated to the estimated importance of the word in a document (Schuller 2005). Previous studies have explored speech recognition systems with emotion features. The categorization was used to prioritize of voice messages. These have included identification of emotions from short utterances.These are all typical types of of Interactive Voice Response (IVR) applications. However, acoustic parameterization of speech signal can be used for the extraction of emotions from a call-center data type. In this regard machine learning provides an interesting approach which has advantages over

traditional computational algorithms. Some of the key developments in speech recognition are also visualized in **Fig. 1** for the period 1975-2015.

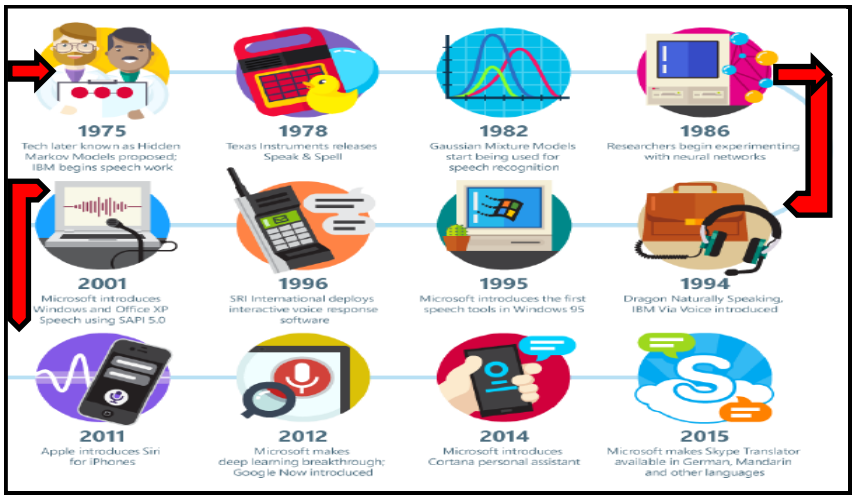

Fig. 1 Speech recognition technology progress over 4 decades

3.SYSTEM DESCRIPTION

The architecture diagram for an intelligent emotion recognition call technique is presented in **Fig. 2**.

The input voice sample is passed through two types of processing streams: (a) Acoustic stream phase (b) Semantic stream phase.

The input sample is same for both the streams for generating the two types of vectors. At a higher level stage, the *acoustic stream phase* extracts certain acoustic features based on the pitch and energy of the signal. Here featured space is labelled with various emotional types in the training phase. An extraction of features with the probability map is used for the generation of output vector v_a by space.

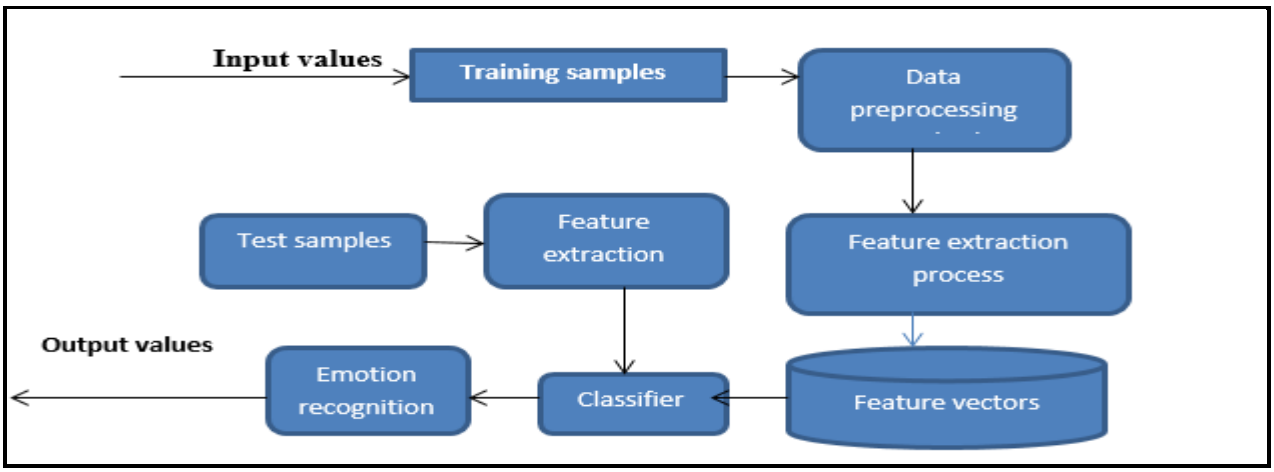

Fig 2. System architecture for smart emotion recognition speech system

The vector size is defined with different types of emotions. Emotional categories are to be trained in the acoustic feature space. The semantic stream process can be performed to convert a speech into-text conversion in the input speech signal. Later in the text classification, algorithms are applied to classify the utterance *Ui* to the specified emotional categories. The output value of this classification is also probabilistic and during the training phase this feature space is labelled with the different types of emotional categories. Utterance containing content of particular emotion which represents the hood probability vector. Output value is the *combination of two streams weights*. U_o is calculated for *N* types of emotional categories. Here U_o is the output vector.

$$U_0^i = Wa * Ua^i + Wa * Us^i \quad (1)$$

Where $i=1 \ldots N$ and the following notation applies.

$U_0 = \{U_0^1, U_0^2, \ldots U_0^N\}$

W_a represents the *confidence* of acoustic stream for input utterance *Ui*

W_s represent the confidence of the semantic stream for input utterance *Ui*

The values of W_a and W_s are calculated from the distribution of the scores generated in W_a and W_s respectively. Machine learning techniques are applied for the detection of the Tone of Voice in the of call recognition system. The XG Boost Model can be used for *feature extraction detection* of a *particular phrase* in the client call phase. The neural network classifier can be used for Detection of a Particular Phrase in the client call phase.

4.SYSTEM OUTLINE

The system to be analyzed consists of the following stages, each of which are addressed in due course:• Data recording and data collection process

• Pre-processing method

• Feature extraction method

• Classification method

• Recognition of call

a) ***Data recording and Data collection process*** **:** in this study the focus is on the development of a speech call recognizer system based on the following states in **Table 2**:

Elation-Despair state
Sadness-Happy state
Boredom state
Shame-pride state
Hot anger relation state
Cold anger, sadness state

Table 2: Emotion states examined

Non–professional actors are recorded for different telephone messages. We have created 15 neural network recognizers by taking 15 feature inputs with 20 node architectures. The accuracy rate of recognizers has 78% with 15 features input nodes with 20 node architectures.The emotion recognition (ER) system includes databases, decision support systems, media with voice messages, e-mails,cyberspace etc. The system consists of three processes:

A)Monitoring the wave files

B)Voicemail center

C)Prioritization of messages.

For every 10 seconds the voice messages are monitored by wave file monitors.

Summary file: Emotions are described in the form of the message, length and silence of the message. The emotion description file has 3 seconds for chunked messages which describes the emotional data. The priorities can be used to process the summary files of messages and sorts them. The call recognizer system refers to identify the speaker for translating speech and "trains" the specific person's voice. It can be used for the speaker identification with authentication .The following approach can be used for a call recognizer system based on the machine learning approach.

b)*Feature Extraction Techniques*

The following fundamental steps are executed in the feature extraction process(Schuller etal 2005):

Steps 1: *Pre-emphasis Phase*

In this step the signal is flowing through a filter with first order. It increases the signal energy with maximum frequency. In the time domain, with input i[n] and $0.9 \leq k \leq 1.0$, we have:

$$j[n]=i[n]-k[n-1] \quad (2)$$

Here k is the pre-emphasis parameter.

Steps 2: ***Framing Phase***

This entails the segmentation of speech samples for the conversion of analog to digital conversion with small frame and minimum time. Here signals are divided in *N* number of frames. Adjacent frames are separated by *M* (*M*<*N*).

Step 3*: Windowing Phase*:A tapered window is applied to each frame. Windowing process can be used to make waveform more smoother. The window equation is :

$$W(n),\ 0 \leq n \leq N-1 \quad (3)$$

Here:

N = number of samples in each frame,*Y*[n] = output signal parameters,*X*(n) = input signal parameters

W (n) = Hamming window (an extension of the Hann window which improves the quality or harmonics of the sound and is directly compatible with the discrete fast Fourier transform).

The relations for the windowing signal computations are given below:

$$Y(n) = X(n) \times W(n) \quad (4)$$

$$W(n) = 0.54 - 0.46\ Cos\ (2\Pi n/L);\ 0 \leq n \leq L-1 \quad (5)$$

$$= 0,\ otherwise.$$

Step 4: ***Fast Fourier Transform***

This process can be used to convert N samples of each frame from time domain to frequency domain. Here X[K] represents the magnitude of frequency for N discrete frequency bands. XG Boost Model for Feature extraction: XG Boost model requires a fixed amount of features for each file. Some capabilities are shown in **Fig. 3.**

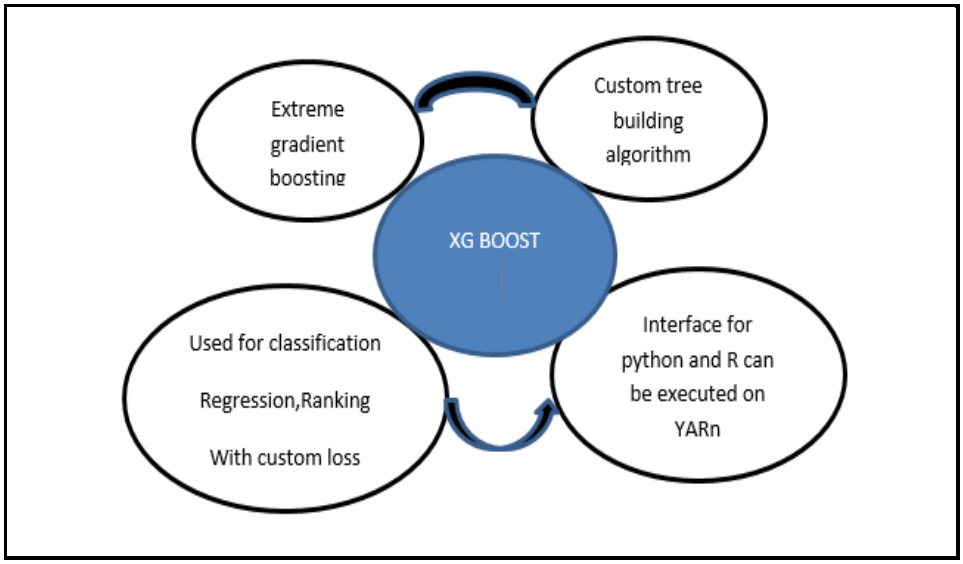

Figure 3: XG Boost features

We have created several signals and statistics (parameters) for achieving the desired computational needs. In Table 3,The following list of signals are implemented.

Step1. Calculate the mean value of the signals.
Step 2. Calculate the mean value of the first 10 seconds of the signal and Compute the Mean value of the last 3 seconds of the signal.
Step 3. Compute the mean value of the local maximum value of the signal.
Step 4. Compute the mean value of the local maximum value of the first 10 seconds of the signal and calculate the mean value of local maximums of the last 3 seconds of the signal.
Step 5. All steps are calculated for each signal. The total amount of features is 36 excluding recording length. All in all, 37 numerical features are available for each recording. Prediction accuracy of the algorithm is 0.869.

Table 3: Signals implemented in XG Boost modelling

Machine learning techniques are applied for the detection of the Tone of Voice of the call recognition system. The XG Boost Model can be used for Feature extraction Detection of a Particular Phrase in

the client call phase. The neural network classifier can be used for Detection of a Particular Phrase in the client call phase.

c)*Classification and Recognition method*:

In this method, six types of emotions are considered and form datasets from all types of speakers (humans). For example, the pitch range is in between 90Hz to 150Hz &160Hz to 380Hz. Here 18 data sets are considered. Normally the hierarchical approach can be used for the improvement of speaker-independent emotion to classify the utterances from the "Berlin emotional database" [27]. It is applied to improve *speaker-independent emotion*.Table 4 gives Classification and recognition method.

1.	Measure the traffic from the extraction of data. It is per hour basis.
2.	Execute the normalization process for monthly traffic of each base station to its maximum value.
3.	Generate a Key- Value pair to perform map-reduce operations.
4	Clustering and classification process can be performed by using K-means algorithm
5	The system regulates order of data pairs which is observed by the LB.
6	First of all initialize the load balancing (LB) during the training phase and assign an empty training set to LB. Store the data pairs by using **Straining pairs** and to train the LB. The controller shows the data pairs of LB is at k =0 of the event pair. Observation of LB at the k^{th} pair, k ∈ [0, ∞], if, the system waits and the controller shows the next data pair. The pair is added to Straining. Parameters are found at the training stage and return to the normal stage. LB value is at the index value k =0. The expected number of trials are conducted before finding the useful one by using the following formula in step 7.
7	The value of $E(k) = (1-p)/P$ in a geometric distribution. Define the system algorithm by using *K-means clustering:* $f(d) = \min d(Y, Z)$ $F(d)$ is a function employed to calculate and store the minimum Euclidean distance and is given as: $$F(d) = \sqrt{\sum (x_r - y_t)^2}$$ The minimum value of the Euclidean distance decides the *emotional test sample state*.

Table 4: Classification and recognition method

5.PROPOSED ALGORITHM FOR NEW CALL RECOGNIZER SYSTEM:

In this section the proposed call recognizer system has been designed. To achieve this goal, first, one selects all the necessary files which are then arranged in a distributed manner. The proposed approach is illustrated in the following steps in **Table 5**:

Step 1. Input values:
Y= (Y_1, Y_2 .., Y_j. . .Y_m) Z = (Z_1, Z_2 .., Z_j. . . Z_m).
Xj= new mean value
Y is the object vector *Z* is the cluster vector, *N*=objects, *m*= dimension vector
Step 2. Data pre-processing
a. Set files on Distributed File System.
b. Choose the fields for doing the analysis.
c. To make the extraction process, transformation processes, load operation.
d. Measure the traffic from the extraction of data. It is per hour basis.
e.Normalization process for monthly traffic of each base station to its maximum value.
f.To generate a Key - Value pair to perform map-reduce operations.
Step 3. Clustering process with K-means algorithm
The system regulates the order of data pairs which are observed by the LB.
First of all initialize the LB during the training phase and an empty training set is assigned to LB. Store the data pairs by using Straining pairs and to train the LB. The controller shows the data pairs of LB is at $k = 0$ of the event pair. Observation of LB at the k^{the} pair, $k \in [0, \infty]$; if, the system waits the Controller shows the next data pair. The pair is added to Straining during the contrary phase. Parameters are found at training stage and return to normal stage. LB value is at index $k = 0$. Expected number of trials are conducted before finding the useful one by using the following formula.
The value of E (k) = (1-p) /P in a geometric distribution.
Define the system algorithm by using K-means clustering
$f(d) = \min d(Y, Z)$
F (d) = function to calculate and store minimum Euclidean distance
K = clusters, d (Y, Z) = function to calculate Euclidean distance between object and cluster vector. Cluster variation is performed in a graphical way.
For each type of emotional pair: All input set are clustered by using K-Means clustering ($k = 2$)
Minimization of Distance: L1 norm, Squared Euclidean, Correlation and Cosine are used to define the k means function.
Initial Cluster Centroids: A UDC centroid which reduces the distance for the input features of emotion pairs.
Maximum Number of Iterations: 1 to 100 iterations for random and UDC centroids.
The error value is calculated by using the recognition accuracy.

Table 5: Algorithm execution in proposed call recognizer system

5.1 *Experimental analysis:*

In this section**,** experimental details have been illustrated in table 6. The proposed approach has been implemented in a symbolic simulation environment i.e. **MATLAB 9.0**. Datasets originated from the UCI machine learning repository. In addition a supplementary confusion matrix for male and female speakers with emotions is also given in **Table 6a**.

All speakers						
Experiment analysis	**Features**	**Distance measure in norm**	**Centroid**	**Iterations**	**Recognition accuracy**	**Variance**
Elation-Despair state	MFCC	L1	UDC and Random	100	76.1%	1.75%
Sadness-Happy-state	MFCC	L1	UDC and Random	1	74.5%	14.54%
Boredom state	Pitch	L1	UDC and Random	100	72.5%	2.45%
Shame-pride state	MFCC	L1	UDC and Random	1	71.1%	3.45%
Hot anger relation state	MFCC	L1	UDC and Random	1	72.3%	10.45%
Cold anger sadness state	MFCC	L1	UDC and Random	1	76.4%	3.54%
Male speakers						
Experiment analysis	**Features**	**Distance measure**	**Centroid**	**Iterations**	**Recognition accuracy**	**Variance**
Elation-Despair-state	MFCC features	L1	UDC and Random	1	87.98	8.96%
Sadness-Happy-state	MFCC features	L1	UDC and Random	1	84.56	6.75%
Boredom state	Pitch features	L1	UDC and Random	1	88.98	10.45%
Shame-pride state	MFCC features	L1	UDC and Random	1	82.34	14.54%
Hotanger relation state	MFCC features	L1	UDC and Random	1	82.54	15.45%
Cold anger sadness state	MFCC features	L1	UDC and Random	1	81.25	16.54%
Female speakers						
Experiment	**Features**	**Distance measure**	**Centroid**	**Iterations**	**Recognition accuracy**	**Variance**
Elation-Despair-state	MFCC features	L1	UDC and Random	1	80.42%	9.65%
Sadness-Happy-state	MFCC features	L1	UDC and Random	1	76.45%	10.69%
Boredom state	Pitch features	L1	UDC and Random	1	70.56%	15.45%
Pride-Shame-state	MFCC features	L1	UDC and Random	1	86.71%	4.54%
Hot anger relation state	MFCC features	L1	UDC and Random	1	79.62%	5.67%
Cold anger sadness state	MFCC features	L1	UDC and Random	1	75.45%	16.57%

Table 6a: Experimental analysis for male and female speakers with emotions

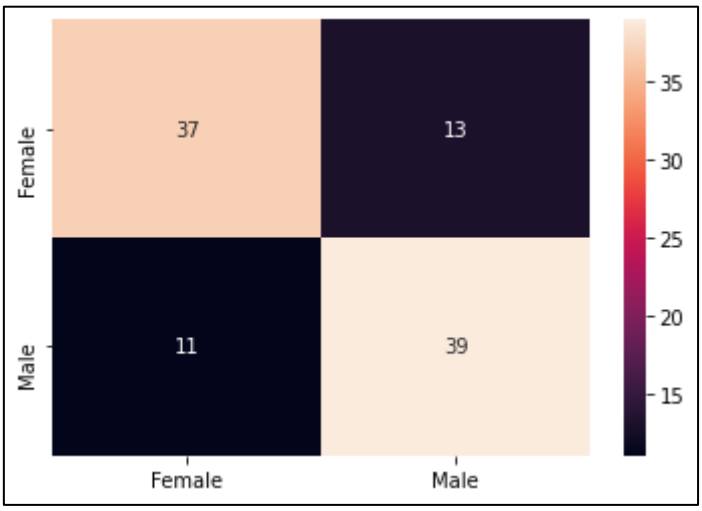

Table 6b) Confusion matrix for male and female speakers with emotions.

5.2 *Results and discussion:*

Herein the new plan base station behavior and behavior of the different base stations is checked. Clusters 1 to 2 are shown in **Fig. 4** and clusters 3 and 4 in **Fig. 5**. Each consists of different base stations and the associated centroid pattern of each cluster is given.

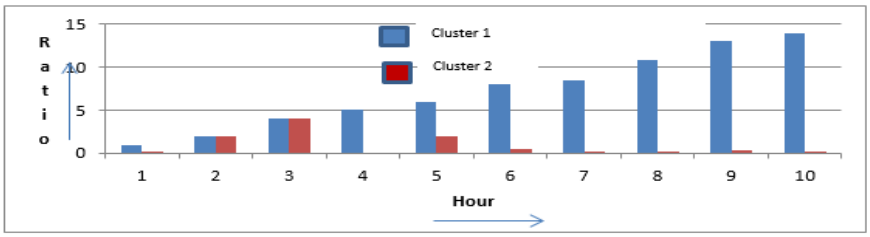

Figure 4. Centroid patterns of clusters 1 and 2

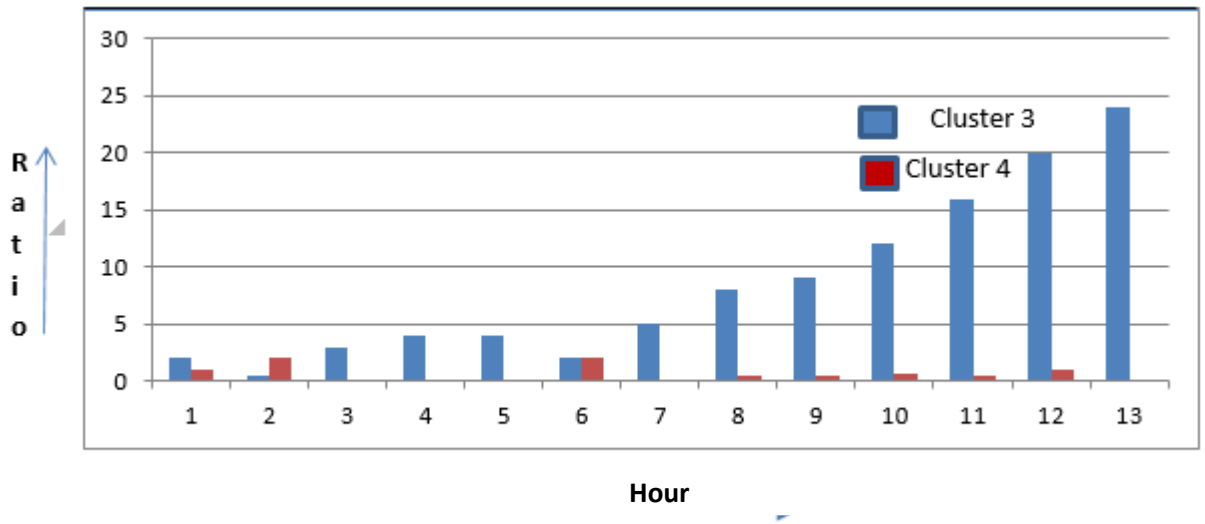

Figure 5. Centroid patterns of clusters 3 and 4.

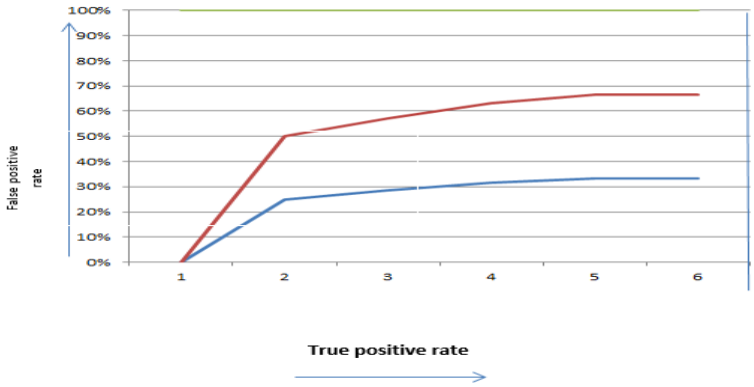

Fig 6: Detection of the Tone of Voice of call recognition system

Fig. 6 demonstrates the sensing of the tone of voice of call recognition system. It can be found with the help of the true positive rate and false positive rate.

5.3 *Detection of a Particular Phrase in the client calls phase:*

It is necessary to detect the words in *audio files* for different phrases. This is applied to observe the call center agent files. It acknowledges the client calls within the first 10 minutes of a call. We have 200 phrases with an average duration of 1.5 seconds to call center agents for bringing in concepts. A great deal of time is taken to conduct checks and *marking files* are recorded. Datasets are utilized via augmentation. We have transformed each file randomly 6 times, adding noise, changing frequency and changing volume. The resulting dataset contains 1500 samples. The prediction accuracy of this algorithm is 0.8. Calls were conducted between June-August 2020, as documented in **Table 7.**

Values	June	July	August	Total
Calls Handled	16,491	16,418	17,573	50,482
Average Handle Time	753	757	757	756
%Transfer	8.1%	7.0%	7.1%	7.4%
% Offer	92.5%	92.6%	92.7%	92.6%
% Accept	75.9%	75.3%	75.2%	75.4%
% Applied	70.1%	71.9%	71.5%	71.2%
Breakage	21.0%	19.6%	19.8%	20.1%
Applied per call	49.2%	50.1%	49.8%	49.7%
Call back within 2 days	19.6%	19.9%	19.8%	19.8%

Table 7: Call center snapshot chronology

The data set information from this source is available at [27].

5.3-1 ***Payout matrix:*** This matrix is applied per call and call regeneration. The values are given below in **Table 8**.

≥ % to Target	Pay out
120%	$450
110%	$350
100%	$250
90%	$150
80%	$75

Below 80%	$50

Table 8: Applied per call

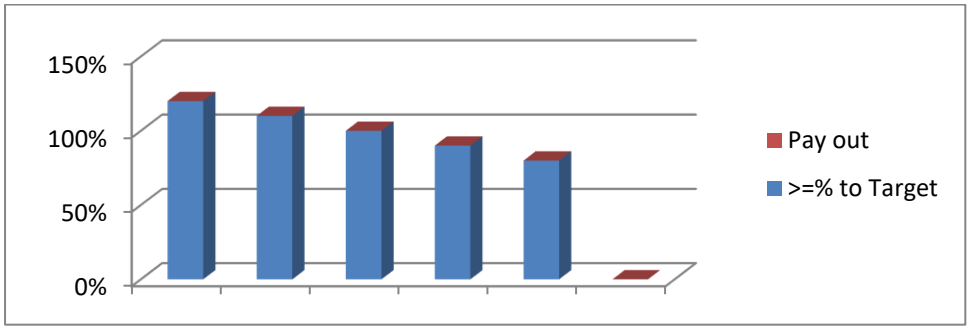

Fig 7. Applied per call

Figure 7 shows the Applied per call.

≥ % to Target	Pay out
120%	$225
110%	$175
100%	$125
90%	$50
Below 90%	$0

Table 9: Call regeneration

Table 9 shows the call regeneration process with respective to target and Payout**.**

5.4 *Experimental analysis:*

We receive 205 files for testing with our deep neural network (DNN) classifier with the sample of 205 files. 177 files are neutral and 28 files are suspicious. DNN has a single one among the various files which belongs to the scheme. It correctly identifies the 172 neutral files as neutral.

- 9 neutral files were identified as suspect

- Correctly 15 suspicious files were identified and 17 suspicious files were neutral.

A 2 X 2 *confusion matrix* is shown in **Table 10** to relate the performance of the proposed work.

	Neutral file	**Suspicious file**
Neutral File	73% True negative	13% True positive
Suspicious file	4% False negative	10% False positive

Table 10. Performance of proposed work

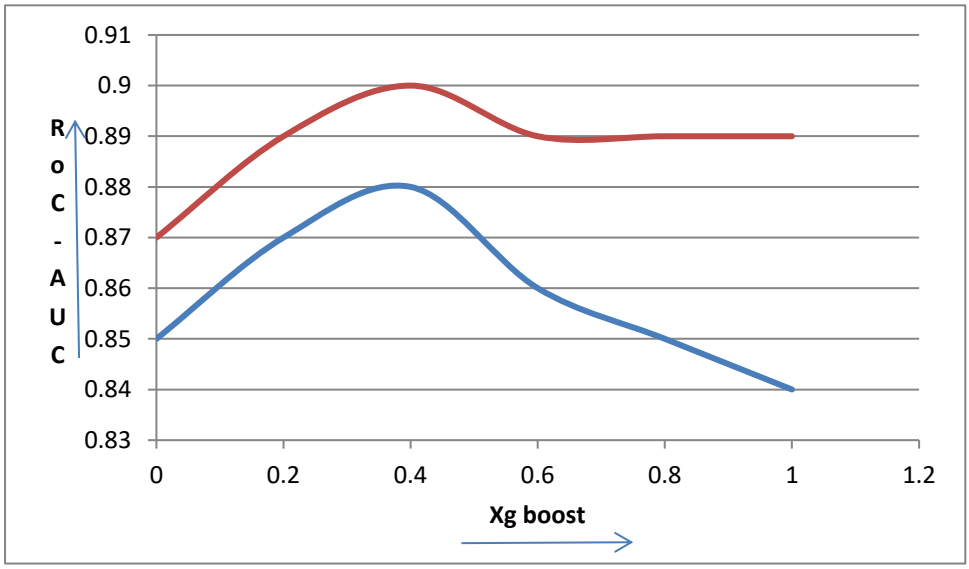

Figure 8: Detection of a Particular Phrase in the client call phase:

Figure 8 shows the detection of a particular phrase in client call phase.

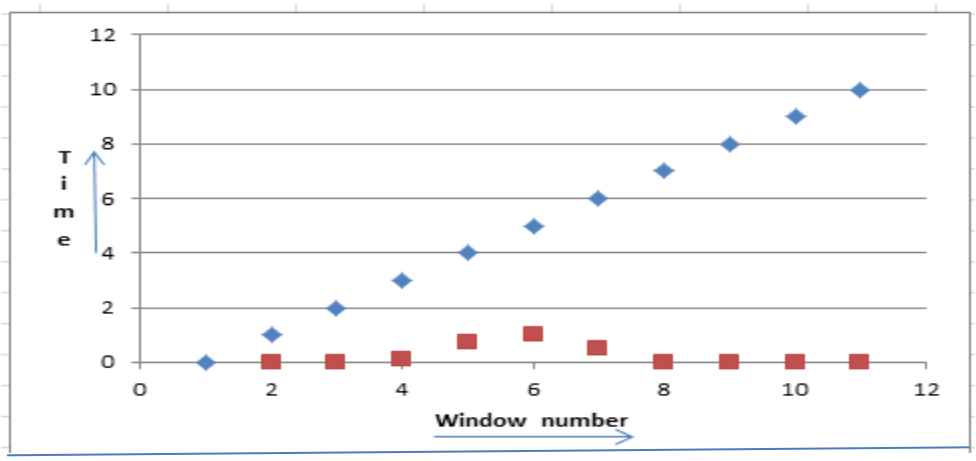

Fig 9: Particular phase in call recognizer system

In excess of 300 further files were marked to determine if the required phrase is pronounced within the first 10 seconds. The prediction accuracy of these files was 87%. The XG-Boost results for detecting the Particular Phrase in the client call phase are shown in **Fig. 9.** This graph visualizes the occurrence of emotions of human beings with associated accuracy.

6. CONCLUSIONS

Appropriate features have been extracted from speech signal units. Machine learning techniques have been deployed to improve performance and emotional registering in a call recognition system, to build on the relatively few previous studies reported for applying multiple classifier systems (MCS). The clustering and classification process is performed in the present study via a K-means algorithm. Details of the Machine learning techniques for the detection of the Tone of Voice of call recognition system are described using the XG Boost Model for Feature extraction and Detection of a Particular Phrase in the client call phase. The present results show that XG Boost is a robust method for data processing and may be further employed to extend studies to a greater number of emotions in conjunction with other machine learning algorithms for call recognition systems.

REFERENCES

[1] Ali Bashashati.**,** Mehrdad Fatourechi, Rabab K Ward and Gary E Birch.2007. A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals. *J. Neural Eng*, 4(2): R32-57 .

[2] Antosik-Wójcińska A.Z.2020. Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling, *International Journal of Medical Informatics*, 138, 104131.

[3] Ayadi, M.E., M. S. Kamel and F. Karray, 2011.Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognition* 44, 572-587.

[4] Bezooijen *et al.*, 1983.Recognition of vocal expression of emotion: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology*, 14 (4) 387-406

[5] Bou-Ghazale .S., and J. Hansen, 2000. A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Trans. Speech Audio Process*, 8 (4) 429-442

[6] Chiriacescu .,, 2009.Automatic emotion analysis based on speech, *M.Sc. Thesis, Computer Science, Delft University of Technology,* Netherlands .

[7] Corballis, M. C., 2002.*From Hand to Mouth: The Origins of Language.* Princeton: Princeton University Press.

[8] Dellaert, F., Polzin, Th., and Waibel, A., 1996.Recognizing emotions in speech. ICSLP 96- Proc. 4th *International Conference on Spoken Language Processing, Philadelphia, PA, USA, 3-6 October*.

[9] Elliot, C., and Brzezinski, J., 1998.Autonomous agents as synthetic characters. *AI Magazine*, 19:13-30 .

[10] Emerich, E., and Lupu, A.2009. Apatean, Emotions recognitions by speech and facial expressions analysis.*17th European Signal Processing Conference, Glasgow, Scotland, August 24-28* .

[11] Jang, J-S., 2020.Data clustering and pattern recognition, available at the links for on-line courses at the author's homepage at http://mirlab.org/jang .

[12] MacNeilage, P., 2008.*The Origin of Speech,* Oxford University Press, UK.

[13] Murray, I.R., and Arnott**,** J.L., 1993.Towards a simulation of emotion in synthetic speech: A review of the literature on human vocal emotion, *J. Acoustical Society America*, 93 (2), 1097-1108.

[14] Lee, C.M., Narayanan, S.S., 2005. Towards detecting emotions in spoken dialogs, *IEEE Transactions Speech Audio Processing*, 13 (2) 1-8.

[15] Picard, R.W., *Affective Computing*: MIT Press, Cambridge, USA.

[16] Reynolds, D.A., 2001, Gaussian mixture models, *Technical Report, MIT Lincoln Laboratory, USA.*

[17] Reynolds, D.A., 2008.Gaussian mixture models, *Encyclopedia of Biometric Recognition*, Springer, New York .

[18] Reynolds, D.A., and R. C. Rose, 1995.Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Transactions Speech Audio Processing*, 3, 72-83 .

[19] Shen P .,Z. Changjun and X. Chen, 2011.Automatic speech emotion recognition using support vector machine, *International Conference Electronic and Mechanical Engineering and Information Technology, Harbin, Heilongjiang, China, 12-14 August.*

[20] Scherer,K.R.., R. Banse, H. G. Wallbott and T. Goldbeck, 1991.Vocal cues in emotion encoding and decoding, *Motivation and Emotion*, 15, 123-148 .

[21] Schuller, B.,S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll,2005. Speaker independent speech emotion recognition by ensemble classification, *IEEE International Conference on Multimedia and Expo, ICME, pp. 864–867, Amsterdam, Netherlands, July 6.*

[22] Vaibhav V. Kamble, Ratnadeep R. Deshmukh and Anil R. Karwankar, 2014.Emotion Recognition for Instantaneous Marathi Spoken Words, *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA), November* 14-15, Bhubaneswar, Odisha, India, pp. 335-346 .

[23] Ververidis D.,and C. Kotropoulos, 2006.Emotional speech recognition: resources, features and methods, *Speech Communication*, 48 (9) 1162-1181 .

[24] Vogt, E.T. Andre and J. Wagner, 2008, Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realization, C. Peter and R. Beale (Eds.): *Affect and Emotion in HCI*, *Lecture Notes in Computer Science*, 4868, pp. 75–91.

[25] You, M,C. Chen, J. Bu, J. Liu, and J. Tao, 2006.A hierarchical framework for speech emotion recognition, *IEEE International Symposium on Industrial Electronics, Montreal, Quebec, Canada, July 9-13*, pp. 515–519.

[26] Zhou .J., G.Wang., Y. Yang and P. Chen, 2006.Speech emotion recognition based on rough set and SVM, *5th IEEE International Conference on Cognitive Informatics, 17-19 July, Beijing, China, ICCI* 2006, 53–61 .

[27].https://data.world/chrisryser/call-center-test-data/workspace/file?filename=Call+Center+results.xlsx (2020).