Automatic Evaluation Software for Contact Centre Agents' voice Handling Performance

K.K.A. Nipuni N. Perera, Y.H.P.P. Priyadarshana, K.I.H. Gunathunga, Lochandaka Ranathunga, P.M. Karunarathne, T.M Thanthriwatta

Faculty of Information Technology, University of Moratuwa, Sri Lanka

Abstract - Evaluating the contact center agent's voice handling skills is important in order to enhance the customer satisfaction, through employee performance. Conventionally, supervisors in call center monitoring divisions listen to contact center agents' conversations and then assign scores for the voice of the agent which end up in most cases bias and as erroneous evaluations.
In order to minimize the above mentioned problems, this paper discusses a system, we implemented to evaluate contact center agents' voices automatically avoiding human error and bias decisions. This software application considers speech rate, voice intensity level and emotional state of the contact center agent for their voice handling evaluations. According to the conducted research, it can be seen that the accuracy of evaluating the giving of instructions clearly using speech rate is 72%, evaluating listener comfortable voice intensity of agent is 83% and accuracy of emotion recognition of the agent's is 83% in the implemented system.

Index Terms – Speech rate, voice intensity, emotion recognition, call centre, contact centre

I. INTRODUCTION

Contact center agents play vital role for enhancing customer satisfaction of a business. As a result, a voice of a contact center agent has a significant effect on customer satisfaction. In order to maintain the quality of a contact center agent's voice, regular evaluation on their voice is done by their supervisors. Most of the time, these performance evaluations are affected by human errors and bias perception. According to research on contact center performance, a friendly agent that responds to the customer like a valued customer, an agent who clearly instructs the customers and an agent who empathy with clients' concerns are the contributions towards a positive customer service phone call in a contact center[1].This project introduces a software which checks whether the agent has provided clear instructions, whether the agent spoke with the customer in a comforting level of voice and also the emotional state of the agent. Thereby the software evaluates agents' voice handling by assigning a score for each criterion.

Rate of speech is the determinant factor of a contact center agent's voice when evaluating his/her clear interaction with the customer. Speech rate can be defined as the number of words spoken per minute (wpm).Average speech rate range of a normal speaker in a normal conversation is approximately 150-200 wpm, but listeners can understand speech faster than the narrator speaks[2].Normally listeners can process information 275-300 wpm[3]. Therefore there is an excess capacity of call center agent to deliver information efficiently to the customer by increasing normal speech rate within allocated call time while speaking in clear and comfortable level for the listener. After considering all the above findings and after measuring the speech rate of the contact center agent in an actual contact center, 245wpm is defined as the contact center agent's speech rate that is comfortable for listener. This approximate estimation is used to evaluate the contact center agents' interaction with customer.

Intensity level of the contact center agent's voice is important to decide whether the contact center agent spoke with the customer inside the customer comfortable level of voice. Vocal intensity is a measure of the radiated power per unit area. Intensity is directly related to the amplitude of the signal and it can be measured by taking the Root Mean Square (RMS) amplitude of the signal. Vocal intensity level of a normal conversation is around 40 dB. This value is used for the evaluation of the contact center agents' voice intensity level.

Analyzing contact center agents' emotions through voice is used to recognize whether the contact center agent has treated the customer in a friendly manner and felt empathy for customers' concerns. This method is also known as speech emotion recognition which is an advance research area in the current academic field, but there are only few researches on automatic speech emotion recognition of contact center agents. Automatic speech emotion detection systems identifies the speaker's emotions like happy, angry, sad, neutral, disgust by extracting different features of speaker's voice [4].In order to determine emotional status, different researches have used different features which can be categorized as follows.

- Prosody feature – pitch, energy, zero crossing rate
- Quality feature – Formant frequencies, spectral features
- Derived feature – Mel-Frequency Ceptral Coeffients (MFCC), Linear Predictive Coding Coeffients (LPCC), Mel-Energy spectrum Dynamic Coeffients (MEDC).

Accuracy of the emotion detection directly depends on type of the features that is extracted. In order to increase the accuracy, researchers combine different features for robust emotion detection. After extracting features, next step is to classify the features according to the state of emotion. Many researches have explored several classification methods for emotion detection, such as Neural Network (NN), Gaussian Mixture Model (GMM), Hidden Markov model (HMM), Maximum likelihood Bayesian classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN) and Support vector machines (SVM)[5].

Based on prior literature on speech emotion detection, this research identifies happy, angry and neutral emotional states of a contact center agent by extracting MFCC and MEDC of agent's voice and classify emotions with the use Support Vector Machine (SVM) multi class classifier.

MFCCs are based on human perception of the speech. It considers that a human ear act as a filter for speech. The human Ear only concentrates on certain frequencies of the speech. These filters are scaled not in a linear way but in Mel-scale. In Mel scale, filters are spaced linearly in low frequencies and logarithmically in high frequencies [6].Frequencies are converted to Mel scale using following formulation.

$$M = 2595 * \log (1+ (f/700))$$

MEDCs are used to represent energy features of the emotional states in this research. Extracting method of MEDC is same as MFCC. Only difference in MEDSC is taking logarithmic mean of energies after filtering process [7].

Support Vector Machine (SVM) classifier constructs machine learning algorithms efficiently and it is widely used for pattern recognition and feature classification in both emotion detection and speech recognition applications. SVM has supervised learning process which contains learning and testing stages. In learning stage SVM model is trained using training data set while in testing stage training data set is used to test new inputs for classification. Since SVM supports small sample learning and capable of dealing with multi-class issues other than two-class problems, this research used SVM as classifier for emotion recognition [8].

II. METHODOLOGY

This research can be explained under 3 components based on criteria's used for analysis of a contact center agent's voice. Such as

1. Checking whether the agent instructed the customer clearly via speech rate.
2. Checking whether the agent spoke in a comfortable loudness level for listener via voice intensity.
3. Checking whether the agent interacted with the customer in a friendly manner via speech emotion recognition.

Technology used to implement the research solution is the MATLAB 2012 software. The above mentioned sub components are from the supervisors of the call center monitoring division who evaluated and recorded calls of contact center agents for appraisal.

A. Evaluation of Speech Rate

Input for the process is recorded audio conversations between customer and the contact center agent which are in mono wav format. Output of the process is score assigned for contact center agent's speech rate out of 10 marks.

Speech rate of the contact center agent is calculated by the following process illustrates in Figure 1. Once the audio file is selected by the supervisor, the application separates the relevant channel from the wav file and then reduces the noise. Once noise reduction is done, application detects voiced and unvoiced parts and takes a segment which has duration of 5 seconds. After that words per second is calculated and then converting it to words per minute. Finally application assigns a score for contact center agent's speech rate.

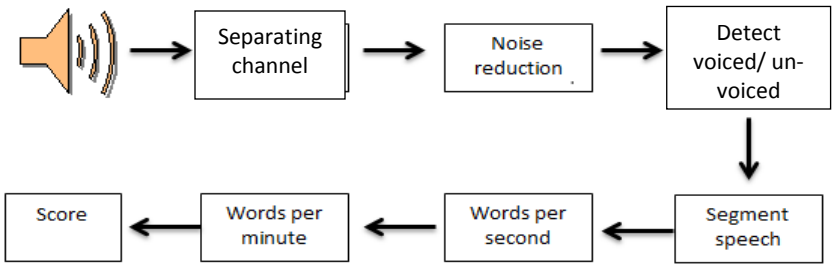

Figure 1: Speech rate calculation process

Speech is rate is calculated after identifying voiced/unvoiced parts. Voiced parts are identified by considering the following measurements.

1. Zero crossing.
2. Magnitude summation.
3. Pitch period value.

After calculating the number of words per minute, the difference of value received and the normal listener comfortable speech rate is obtained. Then a score is assigned for the contact center agent's speech rate .This score is based on the deviation percentage of the speech rate from normal speech rate level of a contact center agent which is comfortable for listeners.

B. Evaluation of Voice Intensity

When evaluating voice intensity of the speaker, as input application uses noise reduced silence removed wav files. Output of evaluating voice intensity is score for contact centre agent out of 5 marks.

In order to assign a score for the contact center agent's vocal intensity level in conversation, Intensity levels are calculated using RMS values. First, the application calculates RMS amplitude normalize value of noise reduced voiced part of the speech and then the application computes logarithmic values of the intensity and thereby assigns a score for intensity level out of 5 marks.

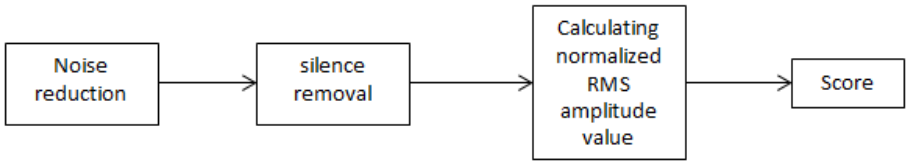

Figure 2: Process of calculating voice intensity

Assuming the voice recorder used in the contact center was having a sensitivity of -70 dBV/Pa which then converted to Volts is 0.000316 V RMS/ Pa, via

$$V = 10^{(dBV/20)}$$

Therefor using this value as the reference value the actual dB of the operator speaking is calculated from using the following equation and the Voltage RMS value of each recording.

$$dB = 20\lg(V\ RMS/0.000316)$$

Considering universal values regarding human hearing levels [9] , and considering the scores given manually the below scoring criteria was as below.

Table I: Relevant score for speaker's voice intensity level.

Intensity Range(dB)	<30	30-34	35-37	38-40	41-47	48-60	61-70	71-75	75<
Score	1	2	3	4	5	4	3	2	1

C. Speech Emotion Detection and Evaluation

Noise reduced and silence removed speech in the training data set and testing data sets are used as inputs. After implementing the only the noise reduced voiced part of speech is to be evaluated. Output of the process is scored for the detected emotion out of 3 marks.

There are three emotional states that application identifies, such as angry, happy and neutral. In emotion detection process, once speech segment is entered, MFCC and MEDC features are extracted. Then extracted coefficients are input to the SVM classifier. In Training stage, coefficients of training data set and classification classes are input to the SVM Classifier in order to develop the model. Once model is built, testing stage get started which provide facility to identify emotion state of any input speech based on model trained. Figure 3 shows overall process of speech emotion detection application [5].

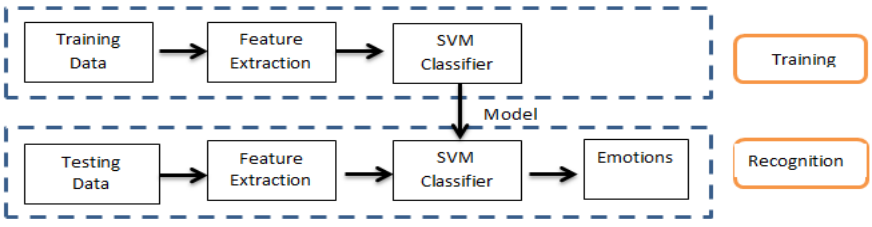

Figure 3: Process of speech emotion detection

As illustrated in Figure 4, First step of the implementation was extracting the MFCC and MEDC feature from the entered wav file.13 MFCC features and MEDC feature were extracted from each audio file.

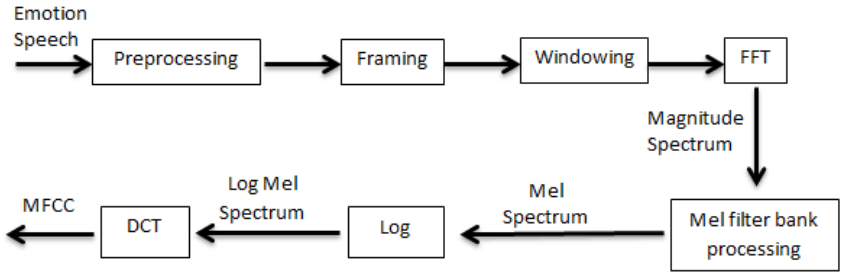

Figure 4: Implementation of MFCC feature extraction

MFCC Feature extraction process contains following stages [5].

1. Preprocessing – in this stage energy at high frequencies in signal is boost by passing signal through FIR (Finite Impulse Response) filter to emphasize high frequencies.
2. Framing – segmenting speech samples obtained from signal into small frames length of 25ms.
3. Windowing – in this stage each frame is windowing to minimize signal discontinuities at start and end of each frame.
4. Fast Fourier Transform (FFT) – each frame of N samples in windowed signal is converted from time domain to frequency domain.
5. Mel filter bank processing – 20 Mel-filters are designed for Mel-processing. From each filter the spectrum are added to get one coefficient each. We considered first 13 coefficients as our features. These frequencies are converted to Mel scale by following below equation.
   M = 2595 * log (1+ (f/700))
6. Taking logarithm – Converting magnitude of Fourier transform from multiplication to addition.
7. Discrete Cosine Transform (DCT) – Shortening the coefficients vector into number of coefficients which will lead to increase accuracy of features.

MEDC feature extraction process follows Preprocessing ,Framing, Windowing; Fast Fourier Transform (FFT) and Mel filter bank processing steps as same as mentioned above. In order to calculate MEDC mean log energy of each filter is calculated after Mel filter bank processing step. Finally Mel energy spectrum dynamic coefficients are obtained by combining first and second differences of filter energies. Figure 5 shows steps of MEDC feature extraction implementation. For algorithm of MFCC and MEDC feature extractions [7].

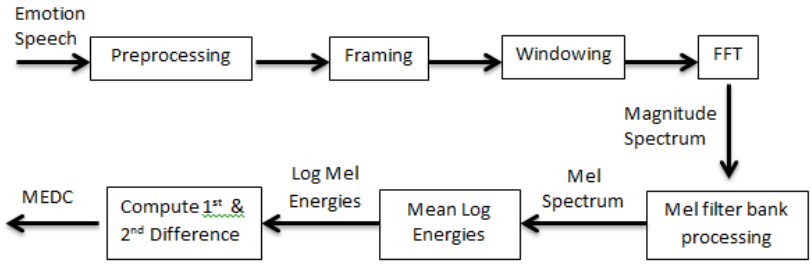

Figure 5: Implementation of MEDC feature extraction

After extracting MFCC and MEDC features, it is necessary to reduce the dimensions of the output vectors since they are going to be inputs into the SVM classifier. Large vectors take long time to be classified. In order to shorten the classifying time by reducing the dimensions of MFCC vector, most represented features were extracted using Principle Component Analysis concept (PCA) [10]. PCA is widely using feature dimension reduction technique. Following steps are followed to implement PCA on MFCC.

1. Obtained feature extraction matrix.
2. Subtracted mean from each data dimension.
3. Calculated covariance matrix.
4. Calculated the eigenvectors and eigenvalues of the covariance matrix.
5. Formed the feature vector.

After that the dimension reduced feature vector is entered to the multi class SVM classifier to classify the 3 emotional states, such as angry, happy and neutral. Multi class SVM is designed with the use of MATLAB SVM toolbox. A non-linear classification is built using Radial Basis Function (RBF) kernel in SVM classifier. SVM classifier is built by following below steps.

1. Constructed the model.
2. Trained constructed model using training dataset.
3. Tested trained model using testing dataset to classify emotions of input speech signal.

Once classification is completed, application assign scores for the emotion mode of the contact centre as follows to evaluate contact centre agents friendliness in speech. Table II shows scores for different emotions.

Table II: Scores assigned for emotional classes

Emotional class	Score
Angry	1
Neutral	2
Happy	3

These are the main 3 modules in this solution. After integrating all the modules, we are providing a friendly customer centric solution where to enhance the customer satisfaction through an effective system.

III. EVALUATION

A. Experiment on accuracy of calculating scores based on speech rate

Accuracy of the calculated score for contact center agent's speech rate is measured by conducting an experiment. 50 audio clips of contact center agents' conversations were used which each has a length of 5 seconds. Each audio clip had a score assigned by contact center monitoring division supervisor. This score given by supervisor is compared with score assigned by the algorithm to measure the accuracy of the algorithm results.

Scores are classified under 3 classes in order to calculate precision, recall and f- measure on accuracy. Score ranges considered are 1-3, 4-7 and 8-10.

Precision, recall and f-measures are calculated for each above mentioned class for measuring accuracy each score class. Table III shows Number of audio clips in 50 audio dataset that belongs to each score range.

Table III: Number of audio clips for each score range

Score Range	Number of audio clips
1-3	15
4-7	15
8-10	20

According to the scenario, definitions of precision and recall are changed as follows.

$$precision\,(A) = \frac{CS(A)}{TS(A)} \times 100$$

Where,

Precision (A) = Precision of score range A
CS (A) = Number of correctly calculated scores within score range A
TS (A) = Total number of calculated scores relate to score range A
Range A can be 1-3, 4-7, 8-10 score ranges.

$$recall\,(A) = \frac{CS(A)}{AS(A)} \times 100$$

Where,

Recall (A) = Recall of score range A
CS (A) = Number of correctly calculated scores within score range A
AS (A) = Actual total number of scores available in testing set which are within score range A
Range A can be 1-3, 4-7, 8-10 score ranges.

$$f-measure = \frac{(2 \times precision \times recall)}{(precision + recall)}$$

Obtained precision, recall and f-measures are shown in Table IV.

Table IV: Precision, recall and f-measure of experiment on speech rate score

Score range	Precision	Recall	f-measure
1-3	0.533	0.533	0.533
4-7	0.786	0.733	0.758
8-10	0.809	0.85	0.829

Accuracy is proportion of true results to the total population.

$$Accuracy = \frac{RS}{TS} \times 100$$

Where,

RS = Number of relevant scores generated for given dataset
TS = Size of total dataset

Therefore according to Table IV, Accuracy of assigning scores for speech rate by algorithm is 72%.

B. Experiment on Accuracy of speech emotion detection

In order to evaluate the accuracy of speech emotion detection, 60 emotional audio clips which are labelled were used for the experiment. 20 audio clips were in each emotional class. From each emotional class 80% have been used for training and remaining 20% is used for training dataset.

Table V: Number of samples used for training and testing

Emotional class	Training dataset	Testing dataset	Total
Angry	16	4	60
Happy	16	4	
Neutral	16	4	

After the experiment precision, recall and f-measure for each emotional class have been obtained. For the task of emotion recognition precision and recall are calculated using prior mentioned equations.
Where,

Precision (A) = Precision of emotion class A
CS (A) = Number of correctly recognized emotional speech instances of class A
TS (A) = Total number of instances recognized as class A
Recall (A) = Recall of emotion class A
AS (A) = Total number of instances of class A in testing dataset.
Class A can be angry, happy and neutral emotional classes.

In above equations, Class A stands for emotional classes such as angry, happy and neutral. Table VI shows confusion matrix of recognizing emotional classes.

Table VI: Confusion matrix of recognizing emotional classes

Emotional class	Angry	Happy	Neutral
Angry	75	0	0
Happy	25	100	25
Neutral	0	0	75

As seen from the confusion matrix in Table VI, the emotion of happy after testing got misclassified with angry and neutral emotions. Based on the confusion matrix, the precision, recall and f-measure are shown in table VII.

Table VII: Results of precision, recall and f- measure

Emotional class	Precision	Recall	F-measure
Angry	1	0.75	0.857
Happy	0.67	1	0.802
Neutral	1	0.75	0.857

Accuracy of speech emotion recognition algorithm is 83.33%.

C. Experiment on accuracy of calculating scores based on voice intensity

Accuracy of the calculated score for contact center agent's voice intensity was measured by conducting an experiment. 50 audio clips of contact center agents were used as the dataset. Each audio clip had a score assigned by the contact center monitoring division supervisor. This score given by supervisor was compared with score assigned by the algorithm to measure the accuracy of the algorithm results. It is rarely assigning score 1, 2 and 4 for a contact center agent for voice intensity even the marks are given out of 5. Therefore for the experiment, two ranges are defines as 1-3 and 4-5
Table VIII shows Number of audio clips in 30 audio clips dataset that belongs to each score range.

Table VIII: Number of audio clips in dataset belongs to each score range

Score range	Number of audio clips
1-3	11
4-5	19

.

According to the scenario, equations definition mentioned under measuring speech rate accuracy are used for obtaining precision, recall and f-measures are shown in Table IX.

Table IX: Precision, recall and f-measure of experiment on voice intensity score.

Score range	Precision	Recall	f-measure
1-3	1	0.545	0.7055
4-5	0.791	1	0.883

Therefore according to Table IX, Accuracy of assigning scores for voice intensity by algorithm is 83.33%.

IV. RESULTS

When considering experiments results in the previous chapter, accuracy of assigning scores for contact center agents' speech rate is 72%. The application can accurately assign scores of 8-10 range for audio clip when comparing with other two score ranges. Since 8-10 score range audio clips are having clear speech rate it may lead to be accurately identified at the same time accurately scored.

Accuracy of speech emotion detection algorithm is 83.33% which is in the satisfactory level. Among the three emotional classes, "angry" and "neutral" can be recognized accurately when comparing with the "happy" class. According to the confusion matrix of emotional classes, it can be seen that "happy" emotion is confused with "angry" and "neutral". Due to this reason classifying happy emotional speeches get less accurate and thereby f-measure represents a comparatively low value. Accuracy of classifying happy emotional state can be increased by training the existing classifier thorough increasing number of happy emotion data in training data set.It will help system to identify happy state correctly since classifier is trained using more data.

It can be seen that calculating scores for voice intensity also has 83.33% accuracy. Among considered two score ranges, 4-5 score range can be identified accurately and less confusion comparatively. Therefor the algorithm performs well for calculating 4-5 range scores.

V. CONCLUSION

According to the research carried out, it can be seen that it is imperative to have an automated system to evaluate contact center agents voice handling skills besides existing manual system. Since automated system can be used as supportive tool for supervisor's evaluations in order to minimize bias and erroneous supervisions.

Audio clips used for experiments are mono channel and they were not high quality. Due to this reason accuracy level is low. As further work, there are many features that can be added to the implemented application. In present work, contact center agent's performances on how clearly the customer was instructed is only measured by the speech rate of contact center agent. It can be further expanded by considering speakers speech articulation, because integrating both speech rate and speech articulation will give more accurate results on evaluating the contact center agent.

Feature extraction for the emotional speech recognition can be expanded and made more accurate by combining both prosody and derive features for recognition. Prosody features like pitch, loudness and speech rate of agent can be combine with derived features like MFCC and MEDC in order to obtain more accurate results. Combining acoustic features with lexical features of speaker's utterance will also enhance accuracy of speech emotion detection.

In conclusion, it can be seen that by combining both manually conducting process and implemented automatic process is essential to enhance contact center agents' performance since it provides unbiased evaluation on employee performance which enhance employee satisfaction and there by customer satisfaction.

ACKNOWLEDGMENT

The authors want to acknowledge the full support and collaboration received from both academic staff and students of Faculty of Information Technology, University of Moratuwa, Sri Lanka.

REFERENCES

[1] http://www.jabra.com/~/media/Documentation/ContactCenter/Report_Value_of_Voice.pdf
[2] Hand book of Research on Educational Communications and Technology, Association for Educational Communication and Technology, 3nd ed, 2008, pp. 966-968.
[3] http://www.advisortoday.com/archives/article.cfm?articleID=119
[4] A.M. Samantaray," Development of a Real-time Embedded System for Speech Emotion Recognition", National Institute of Technology, Rourkela, May 2014.
[5] Y. Pan, P. Shen, L. Shen," Feature Extraction and Selection in Speech Emotion Recognition", Shanghai Jiao Tong University, China.
[6] F. Zheng, G. Zhang, "Integrating the Energy Information into MFCC", *International Conference on Spoken Language Processing, Beijing, pp. I-389~292, Oct. 16-20.*
[7] Y. Chavhan, M. L. Dhore, P. Yesaware, "Speech Emotion Recognition Using Support Vector Machine", *International Journal of Computer Applications, 2010, Vol. 1, No. 20.*
[8] A. Joshi," Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering, August 2013,Vol. 3.*
[9] www.home.earthlink.net/~dnitzer/4HaasEaton/Decibel.html
[10] L. Yu, K. Zhou, Y. Huang, "A Comparative Study on Support Vector Machines Classifiers for Emotional Speech Recognition", March 2014, Vol. 2, No.1.

AUTHORS

First Author – K.K.A. Nipuni N. Perera, Undergraduate of Faculty of Information Technology, University of Moratuwa, Sri Lanka, nipuninamali@gmail.com
Second Author – Y.H.P.P. Priyadarshana , Undergraduate of Faculty of Information Technology, University of Moratuwa, Sri Lanka, toprasanyapa@gmail.com
Third Author – K.I.H. Gunathunga, Undergraduate of Faculty of Information Technology, University of Moratuwa, Sri Lanka, isuruhasarel@gmail.com
Fourth Author – Dr. Lochandaka Ranathunga, B.Sc. Sp(Hons), M.Sc., PGDip in DEd. (IGNOU), PhD (Malaya), MIPSL, MCSSL , Head of the Department of Information Technology, Faculty of Information Technology, University of Moratuwa, lochandaka@itfac.mrt.ac.lk.
Fifth Author – P.M. Karunarathne, MBA, Msc, B.Sc.Eng. , Head of the Department of Interdisciplinary studies, Faculty of Information Technology, University of Moratuwa, pmkaru@itfac.mrt.ac.lk.
Sixth Author – T.M Thanthriwatta, B.Sc.(Hons), Lecturer at Faculty of Information Technology, University of Moratuwa.

Correspondence Author - K.K.A. Nipuni N.Perera, nipuninamali@gmail.com, +94717563490