International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 9 Issue: 5
DOI: https://doi.org/10.17762/ijritcc.v9i5.5463
***Article Received**: 20 March 2021 **Revised**: 28 March 2021 **Accepted**: 15 April 2021 **Publication**: 31 May 2021*

Voice Feature Extraction for Gender and Emotion Recognition

Dr. Madhu M. Nashipudimath
Department Of Computer Engineering
Pillai College of Engineering
New Panvel – 410 206
madhumn@mes.ac.in

Pooja Pillai
pillaipoopr17ce@student.mes.ac.in

Anupama Subramanian
anupamasub17ce@student.mes.ac.in

Vani Nair
nairvanba17ce@student.mes.ac.in

Sarah Khalife
khalifesarsh17ce@student.mes.ac.in

Abstract— Voice recognition plays a key role in spoken communication that helps to identify the emotions of a person that reflects in the voice. Gender classification through speech is a widely used Human Computer Interaction (HCI) as it is not easy to identify gender by computer. This led to the development of a model for "Voice feature extraction for Emotion and Gender Recognition". The speech signal consists of semantic information, speaker information (gender, age, emotional state), accompanied by noise. Females and males have different voice characteristics due to their acoustical and perceptual differences along with a variety of emotions which convey their own unique perceptions. In order to explore this area, feature extraction requires pre- processing of data, which is necessary for increasing the accuracy. The proposed model followssteps such as data extraction, pre- processing using Voice Activity Detector (VAD), feature extraction using Mel-Frequency Cepstral Coefficient(MFCC), feature reduction by Principal Component Analysis (PCA) and Support Vector Machine (SVM) classifier. The proposed combinationof techniques produced better results which can be useful in the healthcare sector, virtual assistants, security purposes and other fields related to the Human Machine Interaction domain.

Keywords- *Voice feature extraction, Human Machine Interaction, data selection, preprocessing , feature selection and classification.*

I. Introduction

The rapid growth of technology and increasing human demand has made voice recognition systems one of the most desired software programs in various devices. Speech recognition technology converts spoken audio into text and lets users control digital devices by speaking instead of using conventional tools such as keystrokes, buttons, keyboards etc. Some examples of such software are Google voice, digital assistants, car blue-tooth etc. Speech signals contain large amounts of information. Two such pieces of information are gender and emotion which can be distinguished relatively more easily by humans than by computers. At present, the research on voice recognition mainly focuses on the identification of single information, which is not enough to understand the true meaning of speech. Here we intend to use voice feature extraction to identify the gender and emotion of the person using SVM classifier, PCA and MFCC.

II. Literature Survey

A. *Data Selection*

The training of the proposed system is done with the help of predefined datasets. The RAVDESS [2] dataset consisting 4904 files of emotional speech in eight basic emotion categories i.e., angry, disgust, fearful, happy, calm, sad, and neutral is used.

B. *Pre-processing voice signals*

The speech signals obtained from the predefined datasets cannot be fed directly to the feature extraction module. The input signals are initially pre-processed using Voice Activity Detector (VAD) [1] to select active frames and filter out silence frames which do not contain any information. The advantage of using VAD is that even though there may be a long pause at the beginning or end of an utterance, the classifier's actions will not be adversely affected.

C. Feature Extraction

For the purpose of gender and emotion recognition from speech signals, it is important to extract relevant features. In this step the processed speech signals are transformed into a concise but logical representation which is more distinct and reliable than the actual signal.

The short-term power spectrum of sound is described by Mel-frequency cepstrum (MFC), on the basis of a linear cosine transform to log power spectrum with a non-linear Mel scale of frequency. By converting the conventional frequency to Mel Scale, MFCC accounts for human perception for sensitivity at acceptable frequencies. MFC is easy to implement [4] and hence has become a widely used method for speech recognition. The accuracy of the system decreases if the sound samples used have low emotional intensity. It is observed that accuracy can be increased when more datasets are involved.

Linear Prediction Coding (LPC) approximates speech samples as a linear combination of past samples. Then, over a finite interval, a unique set of predictor coefficients can be calculated by minimizing the total of the squared differences between the real speech samples and the linearly predicted samples. An automatic vowel classification system [16] can be presented based on LPC and neural networks. Where traditional linear prediction suffers aliased auto-correlation coefficients LPC gives a very accurate estimate of speech parameters and is comparatively efficient for computation. At the same time, the performance of LPC degrades on the presence of noise in audio signals.

Linear Predictive Cepstral Coefficients (LPCC) gives a stable representation of the input speech signal in compressed form as compared to LPC. LPCC are derived from the Fourier transform of the log magnitude spectrum of LPC. The input signal is analyzed by approximating the frequency bands [7] by removing their effects from the signal and approximates the intensity and frequency of the remaining signal. With the help of Discrete Wavelet Transform (DWT), time domain and frequency domain information of the signal can be fetched. DWT decreases the quantity of signals required to recognize the emotions. Different feature extraction techniques like MFCC, pitch, energy, Zero Crossing Rate (ZCR) and DWT are used to extract maximum information of the speech signal and achieve better accuracy [2] with less processing time.

D. Feature Selection

There are many features in a speech signal but not all of them are needed for implementation of the proposed system. Feature selection is required to extract features from audio signals for selection of principal components, as well as to remove redundant and unused information. Principal Component Analysis (PCA) is used to find the principal components out of all available features [1]. PCA is a statistical tool which is used to convert a set of observations of correlated variables to a set of values of linearly uncorrelated variables [13]. It also reduces the processing time since a large set of information requires more processing time.

E. Classification

Emotion and gender recognition is a supervised learning problem. Each pattern used for the training of the classifier carries the correct emotion/gender class label. The most popular approaches for classifications include Bayesian learning, the Linear Discriminant Analysis (LDA), the Support Vector Machine (SVM) [7] which is used as an extension of LDA with a high-dimensional feature space, the multi-layer Neu- ral Network (NN), and the Hidden Markov model (HMM) which captures temporal state transitions. SVM is the most widely used classifier due to its efficiency in classifying high dimensional data where the number of features is greater than number of observations. SVMs have a major benefit over Artificial Neural Network (ANN) that, unlike ANNs, the solution to an SVM is global and exclusive. SVM has a straightforward geometric interpretation and produces a sparse solution.

III. PROPOSED SYSTEM MODEL

Detecting users' emotion and gender accurately, from his/her voice input, requires complex algorithms and intricate deep learning models. To overcome this, we pre-processed the input data meticulously, followed by classification with the help of SVM, which resulted in precise results without the need of complicated Neural Networks. The proposed system shown in Fig. 1 is trained using pre-defined datasets. More emphasis is given to pre-processing of input voice signals contrary to most of the existing systems.

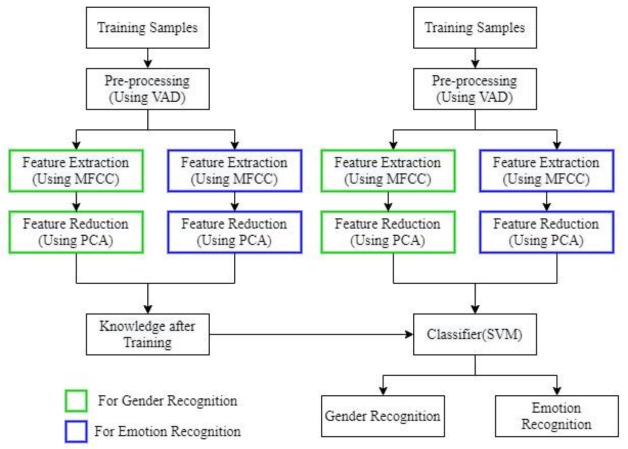
Training Samples
Training Samples
Pre-processing (Using VAD)
Pre-processing (Using VAD)
Feature Extraction (Using MFCC)
Feature Extraction (Using MFCC)
Feature Extraction (Using MFCC)
Feature Extraction (Using MFCC)
Feature Reduction (Using PCA)
Feature Reduction (Using PCA)
Feature Reduction (Using PCA)
Feature Reduction (Using PCA)
Knowledge after Training
Classifier(SVM)
🟩 For Gender Recognition
🟦 For Emotion Recognition
Gender Recognition
Emotion Recognition

Fig. 1. Proposed Model Architecture.

The signals are first pre-processed using Voice Activity Detection (VAD) which is used to determine whether the input signals contain speech or not. The next step is feature extraction. Various acoustic features are extracted using Mel- frequency Cepstral Coefficients (MFCC). The next step is to reduce the number of features. This is done using Principal Component Analysis (PCA). The set of correlated features are transformed into a new set of uncorrelated features called as principal components.

Based on the data available after performing the pre- processing steps on the input signal, Support Vector Machine (SVM) classifier is trained to get accurate results. In the following section, the methodology of Data Selection, Data Pre-processing, Feature Extraction and Data Classification to obtain the classification results from the system are discussed in detail.

A. Data Selection

The training of the proposed model is employed using RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset. The database of size 24.8 GB contains a total of 7356 files consisting of voice samples of 24 professional actors ,12 female and 12 males, vocalizing 2 lexically-matched statements. These are in a neutral North American accent. Speech consists of 8 emotions like happy, sad, angry, surprise, calm, disgust and fearful expressions. With additional neutral expression, every expression is produced at 2 levels of emotional intensity- normal and strong.

B. Data Pre-processing using Voice Activity Detection (VAD)

The input signals are pre-processed using Voice Activity Detection (VAD). It is a technique used to detect the presence or absence of speech in an input signal. The unvoiced portions in a signal are removed. It is a binary decision i.e., the output can either be 0 or 1. 0 represents absence of speech whereas 1 represents presence of speech. The function y = VAD(x) is used to express VAD algorithm, wherein the desired target output is:

$$y*:= 0,\ x \text{ is not speech};\ 1,\ x \text{ is speech} \qquad (1)$$

VAD can also be determined as a probability that an input signal contains speech or not. This is called Speech Presence Probability (SPP). The SPP is expressed as the probability which is always in the range 0 to 1. The main objective of calculating SPP is to determine whether the input signal contains speech or not. The SPP is calculated and compared with a threshold value. If the probability is less than the threshold value, then speech is absent in the input signal. A possible definition for the VAD is then

$$VAD(X):= \{0,\ SPP\ (X) < \theta;\ 1,\ SPP\ (X) \geq \theta\} \qquad (2)$$

where θ is a scalar threshold.

C. Feature Extraction using Mel-Frequency Cepstral Coefficient (MFCC) and Principal component analysis (PCA)

1) Mel-Frequency Cepstral Coefficient (MFCC): One of the most popular audio feature extraction methods is the Mel-frequency Cepstral Coefficient (MFCC). It is a technique which takes voice samples as inputs and determines coefficients unique to a particular sample after processing. It provides enough frequency channels to analyze the audio. MFCC involves: framing and windowing, applying the DFT, Mel frequency warping, computing the log of the magnitude, and then applying the inverse DFT. The flow of extracting the MFCC features is shown in Fig 2.

Acoustic features are extracted using MFCC from the output of VAD i.e., the pre-processed speech signals. The Mel scale compares the perceived frequency of a sound to the measured frequency. It scales the frequency to best match what the human auditory system can hear. The following formula is used to convert a frequency measured in Hertz (f) to the Mel scale

$$Mel(f) = 2595 log\ (1 + f/100) \qquad (3)$$

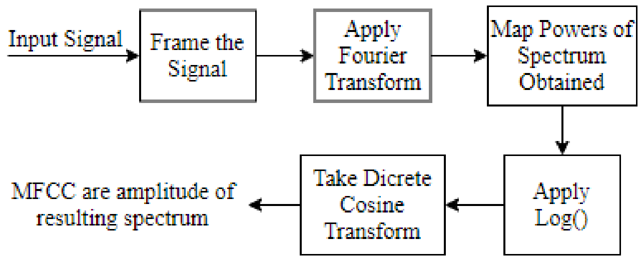
Input Signal
Frame the Signal
Apply Fourier Transform
Map Powers of Spectrum Obtained
Apply Log()
Take Dicrete Cosine Transform
MFCC are amplitude of resulting spectrum

Fig. 2. Flowchart of MFCC Feature Extraction process

2) Principal Component Analysis: Principal Component Analysis (PCA) is applied to the output of MFCC to reduce the size of the signals. It is a statistical technique which is used to reduce the dimension of the data which is then plotted with lesser dimension compared to original data. It is used for dimensionality reduction of the extracted features without any loss of information. One set of variables is transformed into another smaller set using PCA. The newly created variable is difficult to interpret. In many implementations, PCA is used to produce information on true dimensionality of the data set. Consider there are X variables in the data set, among those all X variables will not represent the needed information. PCA transforms a set of correlated variables into a new set of uncorrelated variables called principal components. This reduces the amount of time required to train the classifier as well as the memory space required. Fig 3 illustrates the steps involved in PCA.

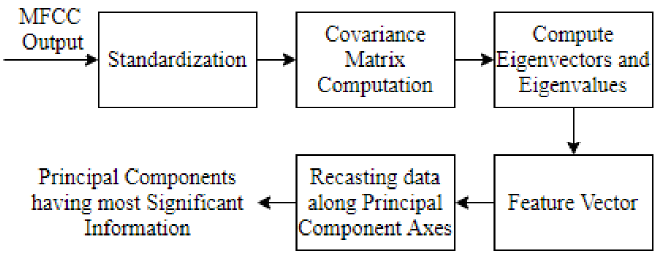
MFCC Output
Standardization
Covariance Matrix Computation
Compute Eigenvectors and Eigenvalues
Feature Vector
Recasting data along Principal Component Axes
Principal Components having most Significant Information

Fig. 3. Steps involved in PCA

D. Classification using Support Vector Machine (SVM)

The classification model of gender and emotion recognition proposed here is based on the SVM classifier. SVM uses binary or multi-class classification. It uses hyper planes in the feature space of high dimensions. This helps to differentiate values based on a particular specification. SVM classification is identical to supervised learning which includes feature extraction to generate desired outputs. The major advantage of SVM is that it is very effortless to train. It is capable of scaling high dimensional data better than neural networks [3].

Based on the knowledge available after pre-processing the speech signals, SVM classifier is used for classification and pattern recognition. Thus, pre-defined datasets and integrated algorithms will be used to classify the gender and emotion of the speech signal. All the pre-processing steps are applied to the speech signal. This pre-processed signal along with the knowledge obtained after training is given as input to the SVM Classifier.

IV. Experimental Evaluation

A. Data Sources

The dataset used for this project is Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). It consists of 7356 files available in three modality formats Audio-only, Audio-Video and Video-only. Audio-only files are present in two forms- Speech and Song. For the implementation of this project, audio speech files are used. The database for the same consists of 1440 audio files recorded by 24 professional actors (12 female, 12 male), articulating two lexically-matched statements in a neutral North American accent. It includes various emotions such as calm, happy, sad, angry, fearful, surprise, and disgust to name a few, wherein each expression is produced at two different levels of emotional intensity (normal and strong) along with an additional neutral expression. This project focuses on more frequently observed emotions like happy, sad, angry, neutral.

B. Principal Components to illustrate variance

On applying MFCC feature extraction, 180 features are extracted. In order to reduce the number of features, principal components are calculated by computing eigenvalues and eigenvectors from the covariance matrix. As shown in Fig 4, to get 95% of variance 55 principal components are required. In the graph plotted below x axis represents the 180 MFCC features extracted from the dataset and y axis represents the cumulative variance for each feature. As observed from the graph, 95% variance is achieved in first 55 features which are considered for further processing.

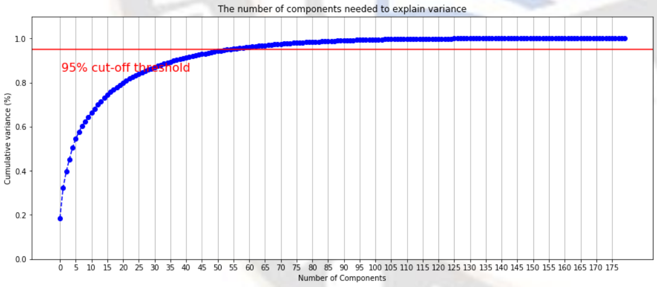

Fig. 4. Plot indicating the number of components needed to illustrate variance

The number of principal components is calculated by combining computational efficiency and performance of the classifier. Two eigenvectors were chosen for illustration purpose as data is plotted using a two-dimensional scatter plot. As shown in Fig 5, the first two components alone contribute around 33% of information in the model.

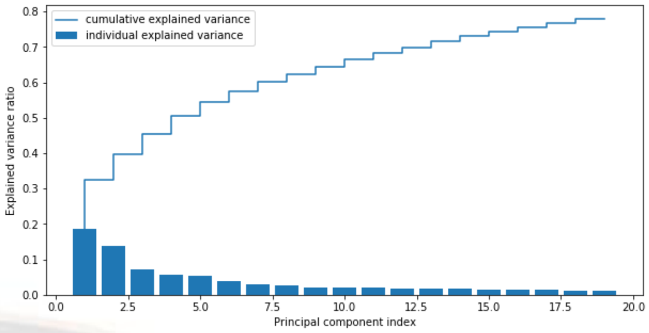

Fig. 5. Plot of 'Explained variance ratio' vs 'Principal component index'

C. Scatterplot for gender

Fig 6 shows the scatter plot for gender of the first two principal components. The x and y axes represent the first two principal components respectively. The data is divided into two clusters, each for male and female represented by distinguishing colors in the figure below depending on the first PC. Few outliers are also observed.

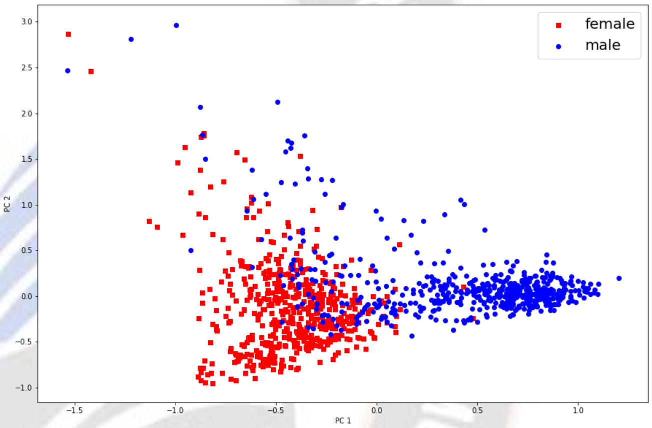

Fig. 6. 2-dimensional PCA scatter plot for Gender

D. Scatterplot for emotion

Fig 7 shows the scatter plot for emotion of the first two principal components. The x and y axes represent the first two principal components respectively. The data is divided into four clusters, each for happy, sad, angry and neutral represented by distinguishing colors in the figure below. Overlapping features are observed for happy, sad and neutral emotions whereas the angry emotion is easily differentiable with few outliers.

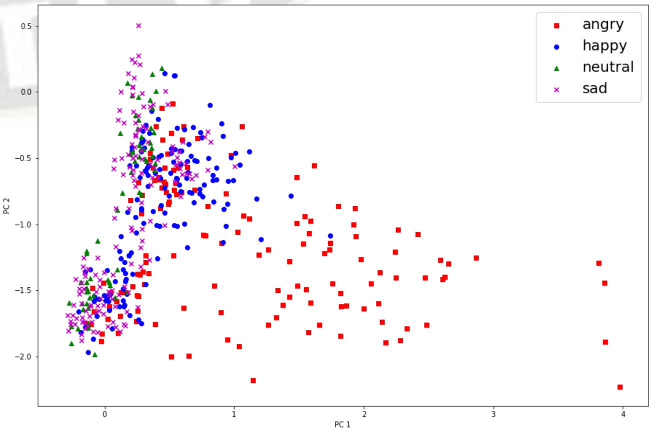

Fig. 7. 2-dimensional PCA scatter plot for Emotion

E. Implementation and Evaluation Measures

The dataset is split into training and testing sets, 75% was used for training the system and 25% for testing. The system is evaluated with the help of various parameters such as Precision, Recall, F1-score and Support which is calculated for each gender and emotion individually. The results of which are displayed in Table I for gender recognition system and in Table II for emotion recognition system.

TABLE I
EVALUATION METRICS FOR GENDER RECOGNITION

Gender	*Precision*	*Recall*	*F1-score*	*Support*
Female	0.98	1.00	0.99	177
Male	1.00	0.98	0.99	183

The accuracy score obtained by the gender recognition system is 98.88% where the train accuracy score is 100% and test accuracy score is 98.88%.

TABLE II
EVALUATION METRICS FOR EMOTION RECOGNITION

Emotion	*Precision*	*Recall*	*F1-score*	*Support*
Angry	0.80	0.84	0.82	51
Happy	0.66	0.66	0.66	44
Neutral	0.69	0.76	0.72	33
Sad	0.71	0.60	0.65	40

The accuracy score obtained by the emotion recognition system is 72.02% where the train accuracy score is 100% and test accuracy score is 72.02%.

F. *Evaluation Results*

The output of the system is analyzed using a confusion matrix. Fig 8 represents the confusion matrix as obtained for the gender recognition system and Fig 9 represents the confusion matrix for emotion recognition system.

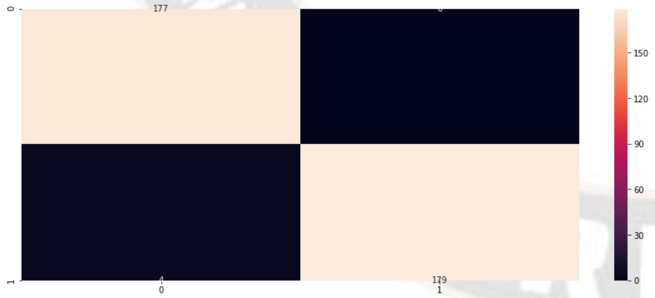

Fig. 8. Confusion matrix for gender recognition system

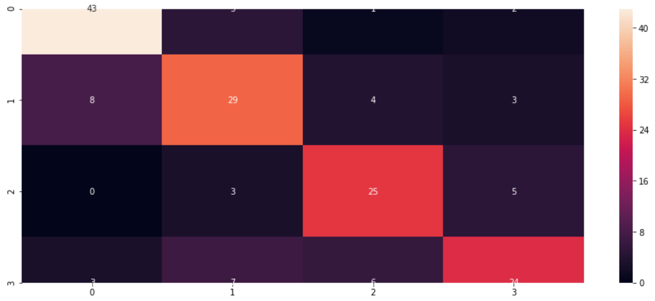

Fig. 9. Confusion matrix for emotion recognition system

CONCLUSION

This proposed system uses 1440 voice samples in .wav format and produces 98.88% accuracy for gender recognition and 72.02% accuracy for emotion recognition. This improvement in accuracy is achieved due to better pre-processing of data with the help of VAD and better feature extraction by using MFCC and PCA. Assorted evaluation parameters are considered while computing the final result. Inclusion of kernel PCA will produce more reliable emotion classification results. There can be further advancements to this project as there are rapid technological advancements and increasing need of human machine interaction systems.

ACKNOWLEDGMENT

It is our privilege to express our sincerest regards to our supervisor Dr. Madhu M. Nashipudimath for the valuable inputs, able guidance, encouragement, whole-hearted cooperation throughout the duration of this work. We deeply express our sincere thanks to our Head of the Department Dr. Sharvari Govilkar and our Principal Dr. Sandeep M. Joshi for encouraging and allowing us to present this work.

REFERENCES

[1] Sharma, Gyanendra & Mala, Shuchi. “Framework for gender recognition using voice”. (2020). 32-37. 10.1109/Confluence47617.2020.9058146.

[2] Koduru Anusha, Hima Bindu Valiveti, and Anil Kumar Budati. ” Feature extraction algorithms to improve the speech emotion recognition rate.” International Journal of Speech Technology 23.1.(2020).

[3] Mr. Sundar Ka, Sadagopan E.Nb, Chandran Mc, Aswin Raja S,” Emotion Recognition Using Support Vector Machine.” (2020).

[4] Gumelar, Agustinus Bimo, et al.” Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks.” 2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH). IEEE, 2019.

[5] Jiang, Wei & Wang, Zheng & Jin, Jesse & Han, Xianfeng & Li, Chunguang.”Speech Emotion Recognition with Heterogeneous Fea- ture Unification of Deep Neural Network. Sensors”. (2019). 19. 2730. 10.3390/s19122730.

[6] Aggarwal, Gaurav, and Rekha Vig. ” Acoustic Methodologies for Clas- sifying Gender and Emotions using Machine Learning Algorithms.” 2019 Amity International Conference on Artificial Intelligence (AICAI). IEEE, 2019.

[7] Jain, Manas & Narayan, Shruthi & Balaji, Pratibha & Bhowmick, Abhijit & Muthu, Rajesh. ” Speech Emotion Recognition using Support Vector Machine”. (2018). 45-55. https://link.springer.com/article/10.1007/s10772-020-09672-4

[8] Poonam Rani, and Ms Geeta. ” Gender and Emotion Recognition Using Voice”. International Journal of Electronics Engineering (ISSN: 0973- 7383). Volume 10 • Issue 2 pp. 165-174 June 2018-Dec 2018.

[9] Hossain, Nazia & Jahan, Rifat & Tunka, Tanjila. “Emotion Detection from Voice Based Classified

DOI: https://doi.org/10.17762/ijritcc.v9i5.5463
***Article Received**: 20 March 2021 **Revised**: 28 March 2021 **Accepted**: 15 April 2021 **Publication**: 31 May 2021*

Frame-Energy Signal Using K-Means Clus- tering". International Journal of Software Engineering & Applications. 9. 37-44. 10.5121/ijsea.2018.9403. (2018).

[10] Kerkeni, Leila & Serrestou, Youssef & Mbarki, Mohamed & Raoof, Kosai & Mahjoub, Mohamed."Speech Emotion Recognition: Methods and Cases Study". 175-182. 10.5220/0006611601750182. (2018).

[11] Alshamsi, Humaid & K e¨puska, Veton & Alshamsi, Hazza. Auto- mated Speech Emotion Recognition on Smart Phones.2018 9th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2018, pp. 44-50, doi: 10.1109/UEM- CON.2018.8796594. (2018).

[12] Wang, Zhong-Qiu & Tashev, Ivan. Learning utterance-level representa- tions for speech emotion and age/gender recognition using deep neural networks. 10.1109/ICASSP.2017.7953138. (2017).

[13] Sengupta, Saptarshi & Yasmin, Ghazaala & Ghosal, Dr.Arijit."Classification of Male and Female Speech Using Perceptual Features". 10.1109/ICCCNT.2017.8204065. (2017).

[14] Pahwa, Anjali & Aggarwal, Gaurav."Speech Feature Extraction for Gender Recognition". International Journal of Image, Graphics and Signal Processing. (2016). 8. 17-25. 10.5815/ijigsp.2016.09.03.

[15] Xavier, Arputha rathina."Basic Analysis on Prosodic Features in Emo- tional Speech". International Journal of Computer Science, Engineering and Applications. 2. 99-107. 10.5121/ijcsea.2012.2410. (2012).

[16] Paulraj, M. P., et al." A speech recognition system for Malaysian English pronunciation using Neural Network." (2009).

[17] Rong, Jia & Li, Gang & Chen, Yi-Ping Phoebe."Acoustic feature selection for automatic emotion recognition from speech". Information Processing & Management. 45. 315-328. 10.1016/j.ipm.2008.09.003. (2009).

[18] Rosenberg, Aaron & Sambur, Marvin."New Techniques for Automatic Speaker Verification. Acoustics, Speech and Signal Processing". IEEE Transactions on. Vol. ASSP-23. 169 - 176. 10.1109/TASSP.1975.1162667. (1975).

[19] Kaggle, [Online], Available: RAVDESS Emotional speech audio. https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio

[20] Global Time, [Online]. Available: China leads in emotion recognition tech, reinforces privacy rules to tackle abuse. https://www.globaltimes.cn/page/202103/1217212.shtml#:~:text=China%20leads%20in%20emotion%20recognition,to%20tackle%20abuse%20%2D%20Global%20Times&text=Global%20Times%20reporter%20based%20in,Shanghai%20and%20its%20surrounding%20areas.

[21] Towards Data Science, [Online]. Available: Speech Emotion Recognition Using RAVDESS Audio Dataset — by Muriel Kosaka. https://towardsdatascience.com/speech-emotion-recognition-using-ravdess-audio-dataset-ce19d162690

[22] Stackexchange,[Online]. Available: Spoken utterance classification on RAVDESS using MFCC. https://datascience.stackexchange.com/questions/80383/spoken-utterance-classification-on-ravdess-using-mfcc

[23] ScienceDirect, [Online]. Available: Confusion Matrix - an overview. https://www.sciencedirect.com/topics/engineering/confusion-matrix