I. J. Mathematical Sciences and Computing, **2021, 1, 26-35**
Published Online February 2021 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmsc.2021.01.04

A System to Predict Emotion from Bengali Speech

Prashengit Dhar
Cox's Bazar City College, Bangladesh
E-mail: nixon.dhar@gmail.com

Sunanda Guha
Missouri State University, USA
E-mail: sg75s@missouristate.edu

Received: 06 December 2020; Accepted: 20 January 2021; Published: 08 February 2021

Abstract: Predicting human emotion from speech is now important research topic. One's mental state can be understood by emotion. The proposed research work is emotion recognition from human speech. Proposed system plays significant role in recognizing emotion while someone is talking. It has a great use for smart home environment. One can understand the emotion of other who is in home or may be in other place. University, service center or hospital can get a valuable decision support system with this emotion prediction system. Features like-MFCC (Mel-Frequency Cepstral Coefficients) and LPC are extracted from audio sample signal. Audios are collected by recording speeches. A test also applied by combining self-collected dataset and popular Ravdees dataset. Self-collected dataset is named as ABEG. MFCC and LPC features are used in this study to train and test for predicting emotion. This study is made on angry, happy and neutral emotion classes. Different machine learning algorithms are applied here and result is compared with each other. Logistic regression performs well as compared to other ML algorithm.

Index Terms: Speech recognition, Bengali speech, MFCC, LPC, XgBoost

1. Introduction

Social media as well as our daily activities such as talking to a bot, using smart devices etc. are now producing excessive amount of multimedia contents. These multimedia contents include images, audio, video etc. Analyzing these multimedia contents to extract information, has attracted many researchers recently. To enhance the genuineness of human-computer interaction, emotion detection from the utterance has become very popular nowadays [1]. Emotion can be expressed both verbally and non-verbally. Verbally means detection of emotion from the speech whereas non-verbal forms include physical appearance, facial gesture, change of prosodic parameters as well as alteration in the spectral energy distribution [2]. They have also shown that emotions can be detected from verbal and non-verbal medium. Emotion detection from utterance introduces emotion sensitive HCI (Human Computer interfaces). Our conversation changes after sensing any particular emotion in a person we are talking with. If we sense a person is sad from his utterance, then we behave in a certain way. Thus, emotion detection is mandatory to develop proper HCI. Also, emotion can be detected from the intonations of one's speech. Emotion detection from human speech has become a crucial problem and being applied in many applications. The applications are detection of student state in tutoring systems [3], automatic distressed phone call identification [4] etc. In [3], the researchers investigated the emotion in human-human tutoring conversation based on 8 prosodic features and found that most of the online tutoring systems are unable to understand human emotional state. Thus, understanding human emotion would be very effective for the online tutoring. In past, it has been shown in the research that we can detect emotion from the composition of prosodic features which are tone, spectra, rate of speaking as well as stress distribution [2], [6-7].

For, Bangladesh perspective, very few researches have been done. However, few accomplished researches on speech recognition for Bengali (Bangladesh) language, but emotion recognition from Bengali language speech is a very new and trending topic. This research is focused on recognizing emotion from speech/voice sample.

2. Literature Review

Speech emotion prediction is now a hot topic in signal processing based research. Various researches have been done on speech emotion recognition [29]. Several researchers are working to extract more efficient features from signal

for recognizing emotion. Deep learning is now using widely for recognition tasks. A CNN model is developed by W. Zheng to recognize emotion. S Mirsamadi [10] used automatically extracted relevant features. It adopts local attention by DNN/CNN to concentrate on particular signal region that were showing more emotional salient.

Vocal emotions from mandarin speechs are classified by Sun and Jiang using hiddden markov model[28]. Two HMM model are used to train and evalaute. Toktam Zoughi et al. adopted the gender-aware deep Boltzmann machine (GADBM) for DNN as pre- training, as a result Boltzmann may exploit the additional facts to increase the prediction accuracy possibly [11]. W. Han proposed temporal classification adopting RNN model to classify emotional state labels. Outcome proved better an effective as compared with other algorithm also [12]. John employed the temporal information's eGeMAPS features and fed into EmNet. The EmNet was prepared as classifier [13]. It achieves an accuracy of 88.9% with the EMO - DB dataset. Z. Zhao made a combination of BLSTM (attention-based) with RNN and FCNs for recognizing emotion from speech and presented more accuracy on predictions comparing with other ML algorithms [14]. Using TDD-Statistics Pooling of attention based TDNN-LSTM model, M.Sarma proposed a speech recognition model [15]. TDD-Statistics Pooling was chooses for improved accuracy. Finally got 70.6% accuracy. One dimensional LSTM is also utilized by J.Zhao for speech recognition, besides 2D LSTM is also used to extract relevant features [16].Finally applying deep leaning model, it results in 52.14% accuracy with speaker independent. Voiced segment selection model used by Yu Gu. [17]. It was adopted to make correct segmentation of audio sample. VSS method works by treating the voice signal segment as like as texture image processing. the Log-Gabor filter was used for classifying sound as voiced and unvoiced. Wootaek Lim created a DNN based on time Distributed layer [18]. Basically it is constructed by combining the CNN and an another circulation NN (Neural Network). Feature information are gathered by using CNN and LSTM together. EmoDB dataset is used for seven emotion classes. Pavitra Patel proposed a PCA based feature extraction method [19]. PCA method were used for extracting reduced feature of loudness, resonance peak and pitch. PCA generally reduces data dimension. Classification was done by Boosted-GMM model with efficictive performance. Htwe and Phyo presented a method to classify noisy speech [27]. Noise is added to the speech signal and then classified using deep learning.

3. Methodology

Features from speech is extracted first. MFCC (Mel-Frequency Cepstral Coefficients) and LPC (Linear prediction coefficients) features are considered in this study. MFCC and LPC both are known as strong feature in classifying speech. Both features from speech are then combined and sent for classification. Fig.1 shows flow of the proposed system

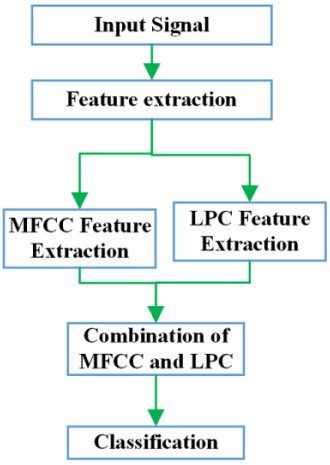

Fig.1. proposed methodology

4. Features

Features are the most important thing in classification related task. Appropriate feature selection leads to better classification. Features are those which represents an input (signal/image etc.) in terms of various properties. In this study, self-recording audio speeches are considered. 301 audio speeches are considered for angry, happy and neutral. Audio data contains both male and female audio speeches.

A. MFCC

Mel-Frequency Cepstral Coefficients (MFCC) considers those coefficients which seize the envelope of the short time spectrum of power. MFCC is computed in a way that the input audio signal is processed into short frames to make

sure the stationarity signal. A periodogram is set to identify the frequencies exist in each frame. periodogram bins are merged by a filter named- Mel filter bank which sums up the energy. It provides an estimation of existing energy in several frequency regions. This step is made for resonating with the way in which cochlea of human works. While perceiving sound, human does not follow a linear scale, rather than it processed in a way of transforming spectral energies into log scale as well as to transform the pattern of non-linear frequency to linear scale which facilitates direct inference.

B. Linear prediction coefficients(LPC)

Linear prediction coefficients (LPC) basically selects vocal tract of human [5]. LPC provides strong and robust feature from speech. It calculates the speech signal by approximating the formants, getting rid of its effects from the speech signal and estimate the concentration and frequency of the left behind residue. The result depicts each sample of the audio signal as a direct incorporation of preceding samples. The coefficients of the difference equation characterize the formants. Thus LPC needs to estimate these coefficients [20]. LPC is an influential speech analysis method and due to its characteristics, it is also known as formant estimation method [21]. Formant frequencies are thosee frequencies where the resonant crests happen. Thus, with this procedure, the positions of the formants in a speech signal are predictable by computing the linear predictive coefficients above a sliding window and finding the crests in the spectrum of the subsequent linear prediction filter [21]. LPC is helpful for encoding the high quality speech at low bit rate [22,23,24].

Linear prediction analysis of speech signal can forecasts any given speech sample at a particular time as a linear weighted accumulation of previous samples. The linear predictive model of speech analysis is given as [22, 25]

$$\hat{s}\,(n) = \sum_{k=1}^{p} a_k s(n-k)$$

where $\hat{s}$ is the predicted sample, s is the speech sample, p is the predictor coefficients.

The prediction error is calculated as [25,26]:

$$e(n)=s(n)-\hat{s}\,(n)$$

Subsequently, each frame of the windowed signal is auto correlated, while the maximum autocorrelation value is the order of the linear prediction analysis. This is followed by the LPC analysis, where each frame of the autocorrelations is transformed into LPC parameters set which entails of the LPC coefficients.

LPC can be derived by following equation.

$$am = log\left[\frac{1-k_m}{1+k_m}\right]$$

The procedure for extracting LPC is shown in fig. 2

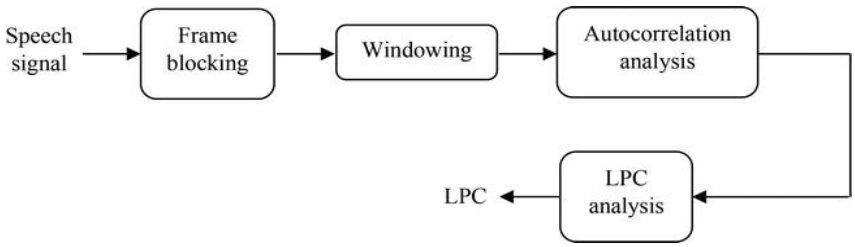

Fig. 2. LPC feature extraction process

5. Training

A. SVM

The SVM algorithm was developed in practice using a kernel. The learning of the hyper plane in linear SVM is done by transforming the fault using some linear algebra, which is out of the scope of this introduction to SVM. A powerful insight is that the linear SVM can be rephrased using the internal product of any 2 given observations, rather than the observations themselves. The internal product between two vectors is the sum of the multiplication of each pair of input values.

B. KNN

The K-Nearest Neighbors algorithm (or k-NN for short) is used for regression and classification. KNN is a non-parametric method. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether KNN is employed for classification or regression. Its purpose is to use a database in which the data points are distributed into several classes to predict the classification of a new sample point. KNN algorithm is one of the simplest classification algorithm and it is one of the most used learning algorithms

C. Adaboost

In adaboost, the output of the opposed learning algorithms ('weak learners') is joined into a weighted add that denotes the ultimate output of the boosted classifier. AdaBoost stands adaptive within the intellect that resulting weak learners are tugged in favor of these instances misclassified by earlier classifiers. AdaBoost is sensitive to clattering information and outliers. In some issues it is less at risk of the overfitting downside than alternative learning algorithms. The individual learners is weak, however as long because the performance of every one is slightly higher than random shot, the ultimate model is tried to converge to a tough learner.

D. Logistic regression

The logistic regression is a predictive analysis. Logistic regression is used to represent data and to clarify the bonding between one dependent binary class and one or more ordinal, nominal, ratio-level or interval independent variables. The principle of Logistic Regression is to explain relationship between features and probability of certain outcome.

E. XGboost

XGBoost is an application of gradient boosting decision trees specially deliberated for speed and performance. It is like optimized boosting method. Gradient boosting works in a way where new models are generated that predict the errors or residuals of preceding models and then merged together to take the final prediction.

6. Performance Analysis

It is mandatory to eavalaute the performance of a model. How well a model works, can be eavlauted by several parameters. Most important and common parametres are precision,recall and F1-score. Precision computes the quantity of positive class predictions that truly belong to the positive class. Recall is the ratio of correctly predicted positive observations to the all observations in actual class. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy.

Table I to table V shows precision, recall and f1-score for KNN, Logistic regression(LR), Adaboost, Support vector Machine (SVM) and Xgboost respectively. Among them logistic regression and SVM performs well, but Logistic regression show better result than SVM. Table VI compares the accuracy among different learning model.

Table I. Result of KNN

	Abeg		
	Precision(%)	recall (%)	f1-score(%)
angry	77	89	83
neutral	97	88	92
happy	78	78	78
macro avg.			**85**
Accuracy			**86**
RAVDEES			
angry	68	86	76
neutral	67	80	73
happy	81	53	64
macro avg.			**71**
Accuracy			**71**
Abeg+RAVDEES			
angry	80	85	83

neutral	88	87	87
happy	75	70	73
macro avg.			**81**
Accuracy			**81**

Table II. Result of LR

	Abeg		
	precision (%)	recall (%)	f1-score(%)
angry	82	95	88
neutral	100	91	95
happy	91	91	91
macro avg.			**91**
Accuracy			**92**
RAVDEES			
angry	81	89	85
neutral	74	80	77
happy	84	73	78
macro avg.			**80**
Accuracy			**81**
Abeg+RAVDEES			
angry	88	83	85
neutral	86	85	85
happy	75	82	78
macro avg.			**83**
Accuracy			**83**

Table III. Result of Adaboost

	Abeg		
	precision(%)	recall (%)	f1-score(%)
angry	83	79	81
neutral	94	88	91
happy	77	87	82
macro avg.			**85**
Accuracy			**86**
RAVDEES			
angry	67	70	69
neutral	80	64	71
happy	60	63	61
macro avg.			**67**
Accuracy			**66**
Abeg+RAVDEES			
angry	74	79	77
neutral	91	79	85
happy	63	66	65
macro avg.			**75**
Accuracy			**75**

Table IV. Result of SVM

	Abeg		
	precision(%)	recall (%)	f1-score(%)
angry	85	89	87
neutral	97	91	94
happy	88	91	89
macro avg.			**90**
Accuracy			**91**
RAVDEES			
angry	84	95	89
neutral	68	68	68
happy	81	71	76
macro avg.			**78**
Accuracy			**80**
Abeg+RAVDEES			
angry	83	83	83
neutral	96	85	90
happy	76	84	80
macro avg.			**84**
Accuracy			**84**

Table V. Result of XgBoost

	Abeg		
	precision (%)	recall (%)	f1-score(%)
angry	88	74	80
neutral	97	91	94
happy	75	91	82
macro avg.			**85**
Accuracy			**87**
RAVDEES			
angry	84	84	84
neutral	78	84	81
happy	79	76	77
macro avg.			**81**
Accuracy			**81**
Abeg+RAVDEES			
angry	89	84	87
neutral	87	90	89
happy	80	84	82
macro avg.			**86**
Accuracy			**86**

Table VI. Comparison of accuracy

	Accuracy (%)		
	Abeg	RAVDEES	Abeg+RAVDEES
LR	92	81	83
KNN	86	71	81
Adaboost	86	66	75
SVM	91	80	84
XgBoost	87	81	86

Confusion matrix and Receiver operating characteristics (ROC) are also measure of performance. Confusion matrix provides well visualization of a classification model. In confusion matrix, column exhibits predicted class and row stands for actual class. Receiver operating characteristics curve shows TPR Vs FPR. Area under curve (AUC) using Logistic Regression model is 0.98 which is nearly 1.For the dataset-Abeg, the confusion matrix and ROC using logistic regression (LR) is shown in table VII is and fig. 3 correspondingly.

Table VII. Confusion matrix for LR

	angry	neutral	happy
angry	18	0	1
neutral	2	31	1
happy	2	0	21

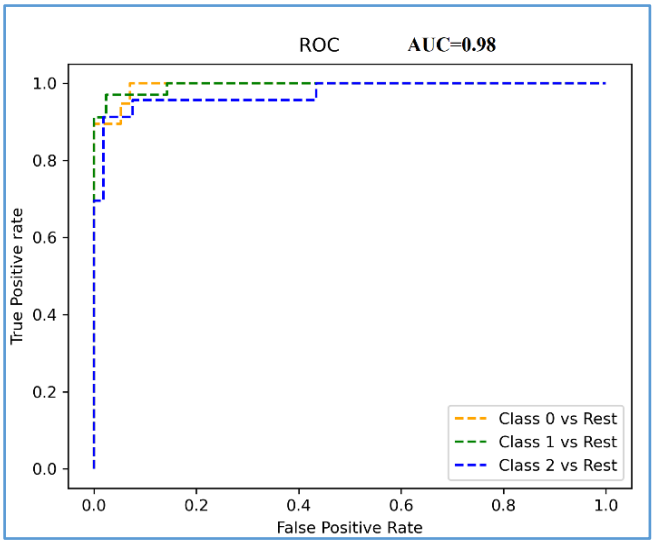

Fig. 3. ROC(Abeg dataset) for Logistic Regression model

In case of RAVDEES dataset, logistic regression(LR) and XgBoost both results in 81%. Receiver operating characteristics curve of XgBoost and LR using Ravdees dataset in depicted in fig. 4 and 5. Combining both Abeg and Ravdees dataset, XgBoost performs better than other with an accuracy of 86%. The ROC curve for combined dataset using XGboost is depicted I fig. 6.

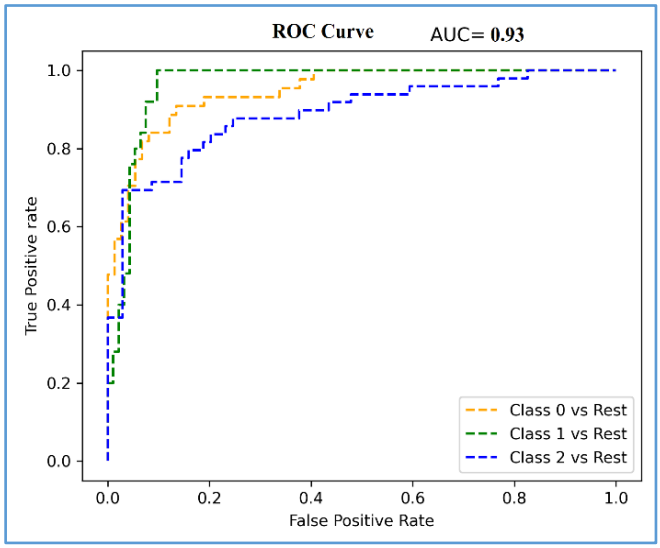

Fig. 4. ROC(Ravdees dataset) for Xgboost model

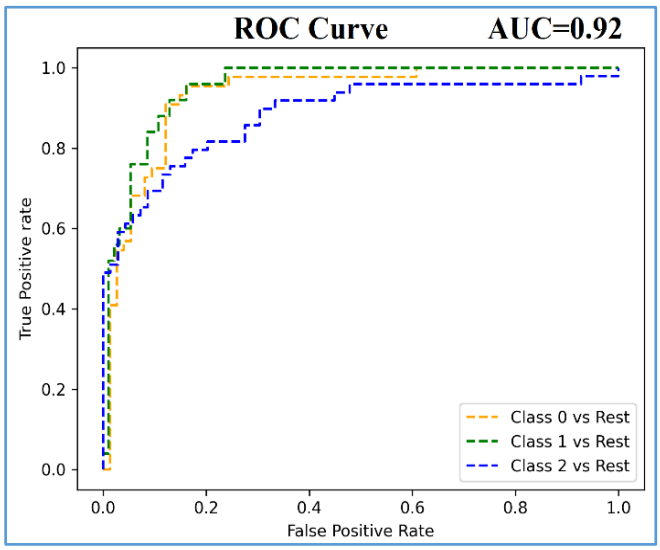

Fig. 5. ROC (Ravdees dataset) for LR

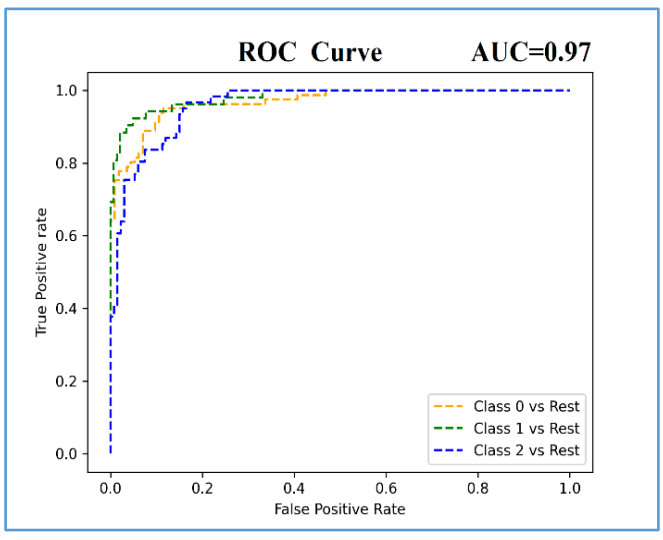

Fig. 6. ROC (Ravdees+Abeg dataset) for Xgboost model

7. Conclusion

MFCC and LPC feature based emotion prediction from Bengali speech is presented in this paper. Comparing various ML model, logistic regression provides best output in this study for the Abeg dataset. MFCC is a powerful feature in signal analysis. Moreover LPC is strong on the basis of computation and accuracy. Proposed method combines MFCC and LPC features. 92% accuracy is achieved for 3 classes of Abeg dataset and 86% accuracy is achieved by XgBoost while Abeg and RAVDEES datasets were combined based on 3 classes. This emotion prediction can help others to understand ones state of mind. In smart home technology, emotion prediction can play a great role. The research is made on for only 3 classes and ita one of the limitation of the research. More classes are yet needed to be added and need to evaluate the performance and efficiency with different feature.

References

[1] Yu, Feng, et al. "Emotion detection from speech to enrich multimedia content." Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg, 2001.
[2] Polzin, T., and Waibel, A., “Emotion-Sensitive HumanComputer Interfaces”, Proceedings of the ISCA-Workshop on Speech and Emotion, 2000.
[3] D. Litman and K. Forbes, “Recognizing emotions from student speech in tutoring dialogues,” in ASRU, 2003
[4] P. Belin, S. Fillion-Bilodeau, and F. Gosselin, “The montreal affective voices: a validated set of nonverbal affect bursts for research on auditory affective processing,” in Behavior Research Methods, 2008
[5] Al-Sarayreh KT, Al-Qutaish RE, Al-Kasasbeh BM. Using the sound recognition techniques to reduce the electricity consumption in highways. Journal of American Science. 2009.
[6] F., Polzin, T., and Waibel, A., “Recognizing Emotion in Speech”, Proceedings of the ICSLP, 1996
[7] Erickson, D., Abramson, A., Maekawa, K., and Kaburagi, T., “Articulatory Characteristics of Emotional Utterances in Spoken English”, Proceedings of the ICSLP, 2000.
[8] Paeschke, A., and Sendlmeier, W. F., “Prosodic Characteristics of Emotional Speech: Measurements of Fundamental Frequency Movements”, Proceedings of the ISCA-Workshop on Speech and Emotion, 2000.
[9] W. Q. Zheng, J. S. Yu, and Y. X. Zou, ‘‘An experimental study of speech emotion recognition based on deep convolutional neural networks,’ in Proc. IEEE Int. Conf. Affect. Comput. Intell. Interact., Sep. 2015, pp. 827–831
[10] S. Mirsamadi, C. Zhang, and E. Barsoum, ‘‘Automatic speech emotion recognition using recurrent neural networks with local attention,’’ in Proc. 42nd IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Mar. 2017, pp. 2227–2231
[11] T. Zoughi and M. M. Homayounpour, ‘‘Gender aware deep Boltzmann Machines for phone recognition,’’ in Proc. Int. Joint Conf. Neural Netw., 2015, pp. 1–5
[12] W. Han, X. Chen, Z. Wang, H. Li, B. Schuller, and H. Ruan, ‘‘Towards temporal modelling of categorical speech emotion recognition,’’ in Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2018, pp. 932–936
[13] J. W. Kim and R. A. Saurous, ‘‘Emotion recognition from human speech using temporal information and deep learning,’’ in Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2018, pp. 937–940
[14] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, ‘‘Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition,’’ in Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2018, pp. 272–276
[15] M. Sarma, D. Povey, N. K. Goel, K. K. Sarma, N. Dehak, and P. Ghahremani, ‘‘Emotion identification from raw speech signals using DNNs,’’ in Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2018, pp. 1–5
[16] J. Zhao, X. Mao, and L. Chen, ‘‘Speech emotion recognition using deep 1D & 2D CNN LSTM networks,’’ Biomed. Signal Process. Control, vol. 47, pp. 312–323, Jan. 2019
[17] Gu Y, Postma E, Lin H X, et al. Speech Emotion Recognition Using Voiced Segment Selection Algorithm:,22nd European Conference on Artificial Intelligence (ECAI 2016), pp. 1682- 1683.
[18] Lim W, Jang D, Lee T. “Speech emotion recognition using convolutional and Recurrent Neural Networks”, Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific
[19] Patel P, Chaudhari A, Kale R, “Emotion Recognition From Speech With Gaussian Mixture Models & Via Boosted GMM”. International Journal of Research In Science & Engineering, 2017.
[20] Agrawal S, Shruti AK, Krishna CR. Prosodic feature based text dependent speaker recognition using machine learning algorithms. International Journal of Engineering Science and Technology. 2010,pp 5150-5157
[21] Gill AS.,” A review on feature extraction techniques for speech processing”, International Journal Of Engineering and Computer Science 2016,pp 18551-18556
[22] Kumar R, Ranjan R, Singh SK, Kala R, Shukla A, Tiwari R.,” Multilingual speaker recognition using neural network”, In Proceedings of the Frontiers of Research on Speech and Music, FRSM. 2009. pp. 1-8
[23] Paulraj MP, Sazali Y, Nazri A, Kumar S. A speech recognition system for Malaysian English pronunciation using neural network. In: Proceedings of the International Conference on Man-Machine Systems (ICoMMS). 2009
[24] Tan CL, Jantan A. Digit recognition using neural networks. Malaysian Journal of Computer Science. 2004,pp 40-54
[25] Agrawal S, Shruti AK, Krishna CR. Prosodic feature based text dependent speaker recognition using machine learning algorithms. International Journal of Engineering Science and Technology. 2010,pp 5150-5157
[26] Al-Sarayreh KT, Al-Qutaish RE, Al-Kasasbeh BM. Using the sound recognition techniques to reduce the electricity consumption in highways. Journal of American Science. 2009,pp 1-12

[27] Htwe Pa Pa Win, Phyo Thu Thu Khine, " Emotion Recognition System of Noisy Speech in Real World Environment", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.12, No.2, pp. 1-8, 2020.DOI: 10.5815/ijigsp.2020.02.01
[28] Sun Menghana, Jiang Baochena, Yuan Jing, "Vocal Emotion Recognition Based on HMM and GMM for Mandarin speech", I.J. Education and Management Engineering 2012, pp 25-31
[29] J. Sirisha Devi, Srinivas Yarramalle, Siva Prasad Nandyala,"Speaker emotion recognition based on speech features and classification techniques", IJIGSP, vol.6, no.7, pp. 61-77, 2014.DOI: 10.5815/ijigsp.2014.07.08

Authors' Profiles

Prashengit Dhar received his B.Sc. degree in Computer Science and Engineering from University of Science and Technology Chittagong (USTC) and M.Sc. degree in Computer Science and Engineering from Port City International University. Currently he is working as a lecturer in a college. He has published many papers in conferenc98e and journal. His research interests include image processing, pattern recognition and machine learning.

Sunanda Guha received her B.Sc. and M.Sc. degree in Computer Science and Engineering from University of Chittagong. Currently she is studying her Masters in computer science in the Missouri State University. She has published several papers in conference and journal. Her research interests include Machine Learning, Expert Systems, Internet of Things, Big Data and Image processing.

How to cite this paper: Prashengit Dhar, Sunanda Guha," A System to Predict Emotion from Bengali Speech ", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.7, No.1, pp. 26-35, 2021. DOI: 10.5815/ijmsc.2021.01.04

