Phase-based Cepstral features for Automatic Speech Emotion Recognition of Low Resource Indian languages

CHINMAY CHAKRABORTY, Electronics and Communication Engineering, Birla Institute of Technology Mesra, India
TUSAR KANTI DASH*, Electronics and Communications Engineering, C V Raman Global University, Bhubaneswar, India
GANAPATI PANDA, Electronics and Communications Engineering, C V Raman Global University, Bhubaneswar, India
SANDEEP SINGH SOLANKI, Electronics and Communication Engineering, Birla Institute of Technology Mesra, India

Automatic speech emotion recognition (SER) is a crucial task in communication-based systems, where feature extraction plays an important role. Recently, a lot of SER models have been developed and implemented successfully in English and other western languages. However, the performance of the traditional Indian languages in SER is not up to the mark. This problem of SER in low-resource Indian languages mainly the Bengali language is dealt with in this paper. In the first step, the relevant phase-based information from the speech signal is extracted in the form of phase-based cepstral features (PBCC) using cepstral, and statistical analysis. Several pre-processing techniques are combined with features extraction and gradient boosting machine-based classifier in the proposed SER model. Finally, the evaluation and comparison of simulation results on speaker-dependent, speaker-independent tests are performed using multiple language datasets, and independent test sets. It is observed that the proposed PBCC features-based model is performing well with an average of 96% emotion recognition efficiency as compared to standard methods.

Additional Key Words and Phrases: Speech Emotion Recognition, Low Resource Indian languages, Cepstral features, Phase-based features, LGBM

1 INTRODUCTION

Speech is one of the key parameters for conveying emotions, and it plays an important role in human-machine interactions. Emotions have an impact on both the voice and the linguistic substance of communication [25]. SER has applications in human-computer interaction, robotics, mobile services, call centers, computer gaming, and the psychological assessment of subjects [2]. However, SER is quite challenging because of the acoustic variability, different contents of sentences spoken, types of speakers, speaking styles, as well as the speaking rates, age of the speakers, and different types of speech features used [4]. A lot of development is done in the field of SER for the English language by technical giants such as Amazon, Apple, Google, IBM, and Microsoft.

Authors' addresses: Chinmay Chakraborty, Electronics and Communication Engineering, Birla Institute of Technology Mesra, India, cchakrabarty@bitmesra.ac.in; Tusar Kanti Dash*, Electronics and Communications Engineering, C V Raman Global University, Bhubaneswar, India, tusarkantidash@gmail.com; Ganapati Panda, Electronics and Communications Engineering, C V Raman Global University, Bhubaneswar, India, ganapati.panda@gmail.com; Sandeep Singh Solanki, Electronics and Communication Engineering, Birla Institute of Technology Mesra, India, sssolanki@bitmesra.ac.in.

2375-4699/2022/9-ART $15.00
https://doi.org/10.1145/3563944

However, very less research for SER is done for most of the Indian languages which include 22 official languages and 1652 mother tongues [37].

Most of the speech processing algorithms work on the magnitude and power spectrum while the phase information remains unaltered. But recently it is observed that phase information can be useful in human speech perception, speech intelligibility, and signal enhancement similar to the power spectrum [10, 29]. To measure the importance of phase in automatic speech recognition, several listening tests have been conducted under different phase uncertainty and signal-to-noise ratio conditions. It is observed that a small amount of phase error or uncertainty does not affect the recognition rate much and the phase is also dependent on the noise level [36]. It is reported in [26] that the phase information can be very useful in music processing, beat tracking, speech watermarking, the detection of synthetic speech, speech coding, and psycho acoustics. The importance of the analytic phase in speaker verification is studied using the instantaneous frequency cosine coefficients features and MFCC. It is observed that this combination has shown a performance improvement of 37% over the MFCC alone [43]. The perceptual significance of phase information in speech is studied in [34] and it is observed that that phase can be successfully used for signal quality improvements and improved quality synthesis.

The combination of relative phase information, Mel-Frequency Cepstral Coefficient (MFCC), and modified group delay is used for differentiating human speech and spoofed speech [45]. An improved SER system is proposed using phase-based features by employing fisher kernels, generative Gaussian mixture model, and a linear kernel classifier. This system demonstrates superior performance for differentiating whispered and non-whispered speech GeWEC dataset [11]. By using high dimensional magnitude, five types of phase-based features, and multi-layer perceptron, spoofing speech detection is performed in [46]. The phase-based features have improved the performance of the model by achieving an equal error rate of 0.29% for known spoofing types. Multi-language interpretation plays a crucial role in many online applications and automatic sign language recognition systems. The Scale Invariant Features Transform technique is used to extract robust features [6, 40].

By using the multilingual Transformer method, a low-resource Indian language recognition system is designed for Gujarati, Tamil, and Telugu languages which provides 6% -11% improvements in character error rate [35]. Time Delay Neural Network-based ASR system is designed using combined acoustic modeling and language-specific information in Tamil, Telugu, and Gujarati languages with the word error rate of 16.07%, 17.14%, 17.69% respectively [12]. By using MFCCs, and shifted delta cepstral (SDC), an improved speaker and language identification problem is dealt with for Eastern and Northeastern Indian languages [5].

It is observed from the brief literature review that SER of low-resource Indian languages is a challenging and very less explored task. Similarly, very few SER systems have designed by utilizing the phase information of the signal. The phase-based information can be explored as suitable relevant features. In this paper, these research gaps are utilized for designing a Bengali emotion recognition system from the speech signal. The major research contributions are:

- Extraction of phase-based information in features and correlating them with the emotions of the low-resource Indian languages.
- Application of the cepstral feature extraction techniques on the phase of the signal and develop a new speech-based feature called phase-based cepstral features.
- Conduction of simulation-based experiments on six standard datasets and verification the accuracy of the proposed features along with the gradient machine-based classification techniques.
- Evaluation and comparison of simulation results on speaker-dependent, speaker-independent tests, multiple language datasets, and independent test sets.

The remaining parts of the paper are organized into four sections based on the research objectives. Section I deals with the introduction, literature review, motivations, and objectives of the investigation. The details of the materials and methods employed are dealt with in section II. Section III contains an analysis of results, and

contributions in terms of research findings. The outcome of the research, limitations, and future research scope are presented in section IV.

2 MATERIAL AND METHOD

2.1 Dataset

Two datsets are used for the implementation of SER on low Resource Indian Language.

2.1.1 Dataset-1 (BanglaSER). BanglaSER is a speech emotion recognition dataset based on the Bengali language. It consists of speech samples collected from 34 participants in the age group of 19 to 47 years old including 17 male and 17 female nonprofessional speakers [8]. It contains 1467 speech samples of five basic human emotions such as: angry, happy, neutral, sad, and surprise. For each emotional state, three trials are undertaken. As a result, the total number of recordings is 1467, which includes 3 statements 3 repetitions 4 emotional states (angry, happy, sad, and surprise) 34 participating speakers equals to 1224 recordings. Additionally, 243 recordings are collected from 27 speakers with 3 statements 3 repetitions of 1 neutral emotional state. The total number of spoken words is 11 with phonemes including 23 vowels, 24 constants. The file formats of the speech samples is wav with the sampling rate of 44.1 kHz and of duration 1 hour 29 min.

2.1.2 Dataset-2 (SUBESCO). This dataset is taken form SUST Bangla Emotional Speech Corpus (SUBESCO) containing speech samples from 20 professional speakers including 10 male and 10 female participants (ages 21 to 35) [38]. Five males and five females took part in each of the two phases of audio recording. For each of the ten sentences, seven emotional states are recorded such as: angry, disgust, fear, happy, neutral, sad, and surprise. For each emotional expression, five trials have been used. As a result, the total number of utterances is 7000, with 10 phrases, 5 repetitions, 7 emotions, and 20 speakers. Each sentence was kept at a constant duration of 4 seconds, with only the silences removed and the full words preserved. The total number of words are 50 including 7 vowels, 23 consonants, 6 diphthongs and 1 nasalization. The audio file format is wav with a sampling rate of 48KHz and the total duration of the dataset is 7 hours 40 min 40 sec.

2.2 Phase Based Cepstral Features

The basic steps for the calculation of the MFCC features are application of pre-emphasis filter on speech signal, frame blocking and windowing, conversion of the signal from time domain to frequency domain using the Fourier transform, application of the Mel filter bank on the power spectrum, determining the logarithm and application of the Discrete cosine transform [27]. However, in these calculations the magnitude of the signal is considered and the phase is discarded. The phase contains important information related to the emotions of the signal. The average values of the phase of the FFT of the speech signals with respect to different emotions of the BanglaSER dataset are shown in Figure 1.

It can be observed that there exists a some difference in phase overally and it can be gainfully employed for derivation of phase-based features. Following this, the phase based cepstral coefficients (PBCC) are designed by incorporating the phase information and the processing steps are explained below.

2.2.1 Pre-emphasis filter. It is used to boost the speech signal strength at the higher frequencies to balance the signal loss in the high-frequency region. The loss due to the lip radiation and the glottal source effects which is approximately −6 dB/octave slope decay as compared to the original speech spectrum [33]. The transfer function of a typical time domain finite impulse response filter used for pre-emphasis is given below:

$$H(z) = 1 - b \cdot z^{-1} \tag{1}$$

where, the slope of the filter is b and its value is taken as 0.9.

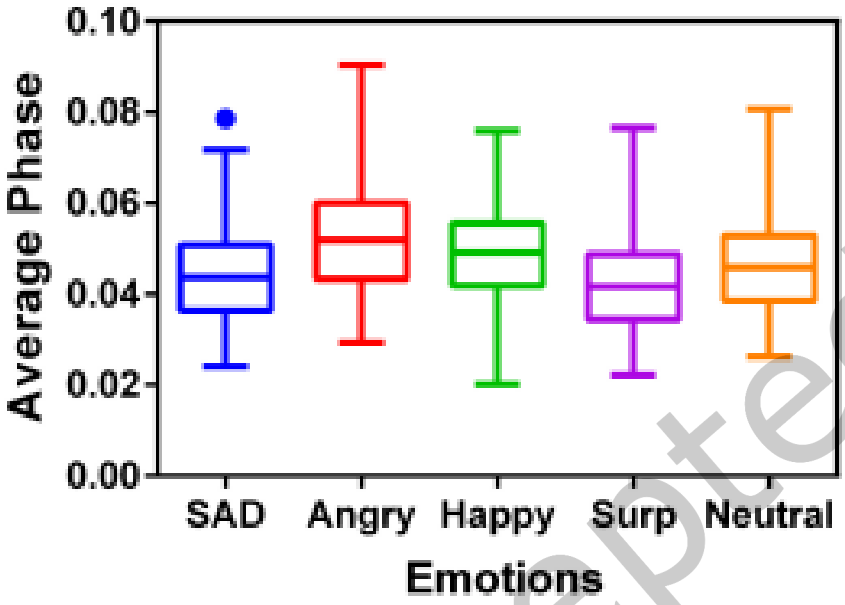

Fig. 1. Relationship between the phase and emotions

2.2.2 Framing and windowing: To extract stable acoustic features from speech signal which is quasi stationary in nature, segmentation is performed. The duration of each segments is usually 20-25 ms with 50% overlapping. The overlapping analysis is required to ensure that the signal is roughly centered in frames. After framing, the Hamming windowing is performed to boost the harmonics, smooth the edges, and lessen the edge effect of the discrete Fourier transform [32]. The hamming window is expressed as

$$w(n) = \begin{cases} 0.54 - 0.46\ \cos\left(\frac{2\pi n}{L-1}\right) & 0 \le n \le L-1 \\ 0 & otherwise \end{cases} \tag{2}$$

Where L is the duration of the window.

2.2.3 Conversion from time domain to frequency domain. After framingand windowing, the signal is converted from time domain to frequency domain using the discrete Fourier transform.

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-\frac{j2\pi nk}{N}} \quad 0 \le k \le N-1 \tag{3}$$

Where N denotes the number of samples in the windowed frame.

The phase of the signal X(k) is calculated as

$$\angle X(k) = \tan^{-1}\left(\frac{imaginary\ (X(k))}{real\ (X(k))}\right) \times \frac{\pi}{180^0} \tag{4}$$

It is expressed in radian. The power spectrum of the phase is calculated by taking the square on $\angle X(k)$.

2.2.4 Application of the Mel filter bank. The human auditory system does not respond to the frequency of the tone. The perception based frequency unit is Mel and the Mel filter bank converts the signal from the linear frequency to the perceptual domain. The conversion from the linear frequency to the Mel scale as:

$$f_m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{5}$$

After conversion to the perceptual domain, the triangular filter shaper is used. The mel auditory filter bank is plotted using the MATLAB platform in Figure 2 for the linear frequency in the range 0 to 3 kHz [22].

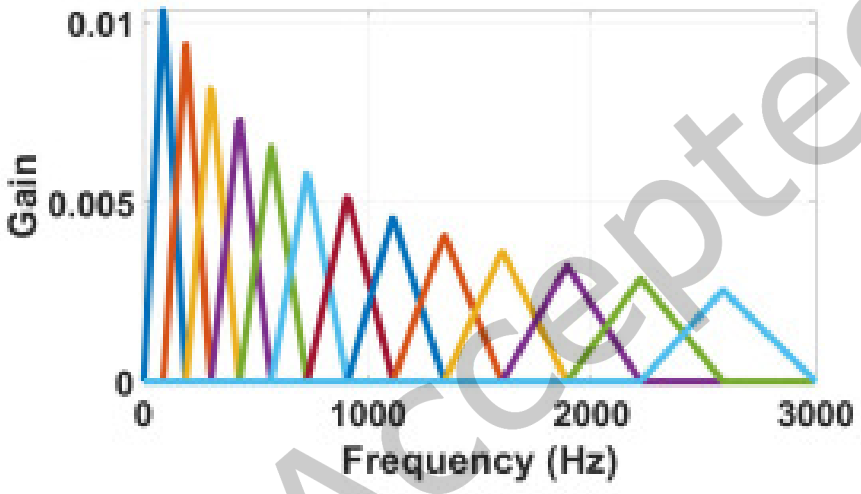

Fig. 2. Mel Filterbank representation

Here, the Mel phase spectrum is computed as

$$p(m) = \sum_{k=0}^{N-1} \left[|\angle X(k)|^2 \cdot H_m(k)\right] \quad 0 \le m \le M-1 \tag{6}$$

Where, the weight associated with the kth phase energy spectrum bin of the mth output Mel band is Hm(k) and m denotes the triangular Mel filter. M is the total number of Mel coefficients to be calculated. Hm(k) is calculated as:

$$H_m(k) = \begin{cases} 0 & , \; if \; k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & , \; if \; f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & , \; if \; f(m) \le k \le f(m+1) \\ 0 & , \; if \; k > f(m+1) \end{cases} \tag{7}$$

2.2.5 Discrete cosine Transform. Discrete cosine Transform is performed along with the log operation to extract PBCC coefficients (c(n)) at the end. It is similar to the inverse Fourier transform to extract the speaker vocal tract

information.

$$c(n) = \sum_{m=0}^{M-1} \log_{10}(p(m)) \cos\left(\frac{\pi\, n\, (m-0.5)}{M}\right) \qquad 0 \le n \le C-1 \tag{8}$$

C is the number of PBCC coefficients which is generally taken as 13. The zeroth coefficient is removed because it contains average log-energy. The energy is replaced by the frame-wise short time energy (STE) in the proposed approach.

2.2.6 Sample-wise statistical features. At the frame and sample levels, several speech features are extracted using PBCC, STE, and statistical parameters. The statistical features are used to extract the relevant data information from the non-stationary nature of the input signal [1]. Different combinations of the statistical features are used in speech recognition to extract the distributions of information apart from the simple average. The different statistical parameters used are: mean, median, root-mean-square, maximum, minimum, 1st and 3rd quartile, interquartile range, standard deviation, skewness, and kurtosis [7]. One complete recording of a single user in each category is called one sample, while the subset of the sample is called a frame. The frame and sample level features are explained in detail in Table 1 taking into consideration that each sample has 'n' frames. The features are identified by their serial numbers, which range from f1 to f91.

Table 1. Feature Details

Type of features	frame level	Sample level
Phase-based Cepstral features	$n \times 12$	mean (f1 - f12) median (f13 - f24) maximum (f25 - f36) minimum (f37 - f48) standard deviation (f49 - f60) rms (f61-f72) kurtosis (f73-f84)
STE	$n \times 1$	mean (f85) median (f86) maximum (f87) minimum (f88) standard deviation (f89) rms (f90) kurtosis (f91)
Total number of features		f1-f91

2.3 Light Gradient Boosting Machine for Classification

Gradient boosting machines are one of the effective ensembles of decision tree algorithms that are used for achieving higher classification and prediction performance. Light Gradient Boosting Machine (LGBM) is the faster and improved version of the gradient boosting machines with very less computational complexities. It is proposed by Microsoft in 2017 [16]. It uses gradient-based one-side sampling and exclusive feature bundling for precise gain estimation and reduction of the number of data processing. It combines several mutually exclusive features into dense form and also removes the zero feature values [21] . If the input data set is exceedingly vast

and it is difficult to extract features from the data set, deep learning and CNN-based methods are suitable. This is so that the deep learning model, which is used for prediction and classification purposes, can automatically extract the necessary characteristics from the raw data. The deep learning-based approach might not always be more effective than the straightforward, time-tested approach. Therefore, finding and using a simple solution is a difficult problem. The speech dataset used in the current study challenge is of medium size, and the audio features to be extracted from speech signals are recognized and chosen [41]. In recent years, LGBM has been employed for several applications such as speech recognition [48], speaker recognition systems [47], and speech enhancement [15]. Because of all successful implementations of LGBM in speech processing, it has been selected in the current work.

3 EXPERIMENTAL SETUP

The proposed PBCC and LGBM-based method is implemented using the following steps as shown in Figure 3. For simulation-based experiments, primarily two Bengali datasets (BanglaSER and SUBESCO) are used. These two datasets are used for speaker-dependent, speaker-independent as well as independent tests. Additionally, four more datasets are used from different languages for checking the effectiveness of the PBCC features. After the collection of datasets, all the speech samples are passed through the speech pre-processing stage before sample-wise feature extraction. These datasets pre-processing steps include standard voice activity detection, dynamic level control, and band pass filtering [9]. For removing the silence from the speech signal the voice activity detection is performed. The dynamic level controller is used for boosting low signal levels as well as lowering peak levels [13]. For speech naturalness and proper speaker identification the frequency range of 100 Hz to 3400 Hz is significant [23]. In the proposed implementation, the speech signal is re-sampled to 8 kHz and passed through a band pass filter with a range of 100Hz-3400Hz. After dataset pre-processing and labeling are completed, the PBCC features are extracted frame-wise and sorted sample-wise. From each speech sample 91-dimensional feature vector is extracted and stored. For reducing the outliers from the extracted feature sets, an effective robust scaler technique is used [42]. These feature sets are now applied to the LGBM classifier for training, validation, and testing. The stratified k-fold cross-validation scheme is used. For performance evaluation, standard evaluation measures such as Accuracy (AC), F-1 Score(F-1), Precision (PR), Recall (RE), Confusion matrix, and area under the curve (AUC) are used [19].

4 ANALYSIS OF RESULTS

On both, datasets, speaker-dependent and speaker-independent experiments are carried out to determine the effectiveness of the proposed PBCC feature-based LGBM classifier model (PBLG). Each experiment is divided into two parts training and testing using a 5-fold stratified cross-validation scheme. The proposed model is evaluated for both the speaker-dependent and speaker-independent SER for checking the robustness of the model.

4.1 Speaker-dependent SER

To recognize speech emotion with higher generalization capabilities as well as higher accuracy, extensive speaker-dependent experiments are carried out on both the labeled speech samples. In each dataset, the speech samples are divided into two sets randomly, with the training set containing 80% of the data, and the testing set having the remaining 20%. Then the 91-dimensional PBCC features are extracted and applied to the LGBM classifier. After training the model, the validation accuracy is calculated which is the key indicator of the generalization capabilities of the trained model. If the validation accuracy is higher than the best SER performance is superior. The confusion matrices of these speaker-dependent experiments are listed in Figure 4 and Figure 5 for the BanglaSER, and SUBESCO datasets respectively. Here, the emotions are denoted as anger = ANG, happy = HAP, neutral = NEU, sadness = SAD, surprise = SUR, fear = FEA, disgusting = DIS.

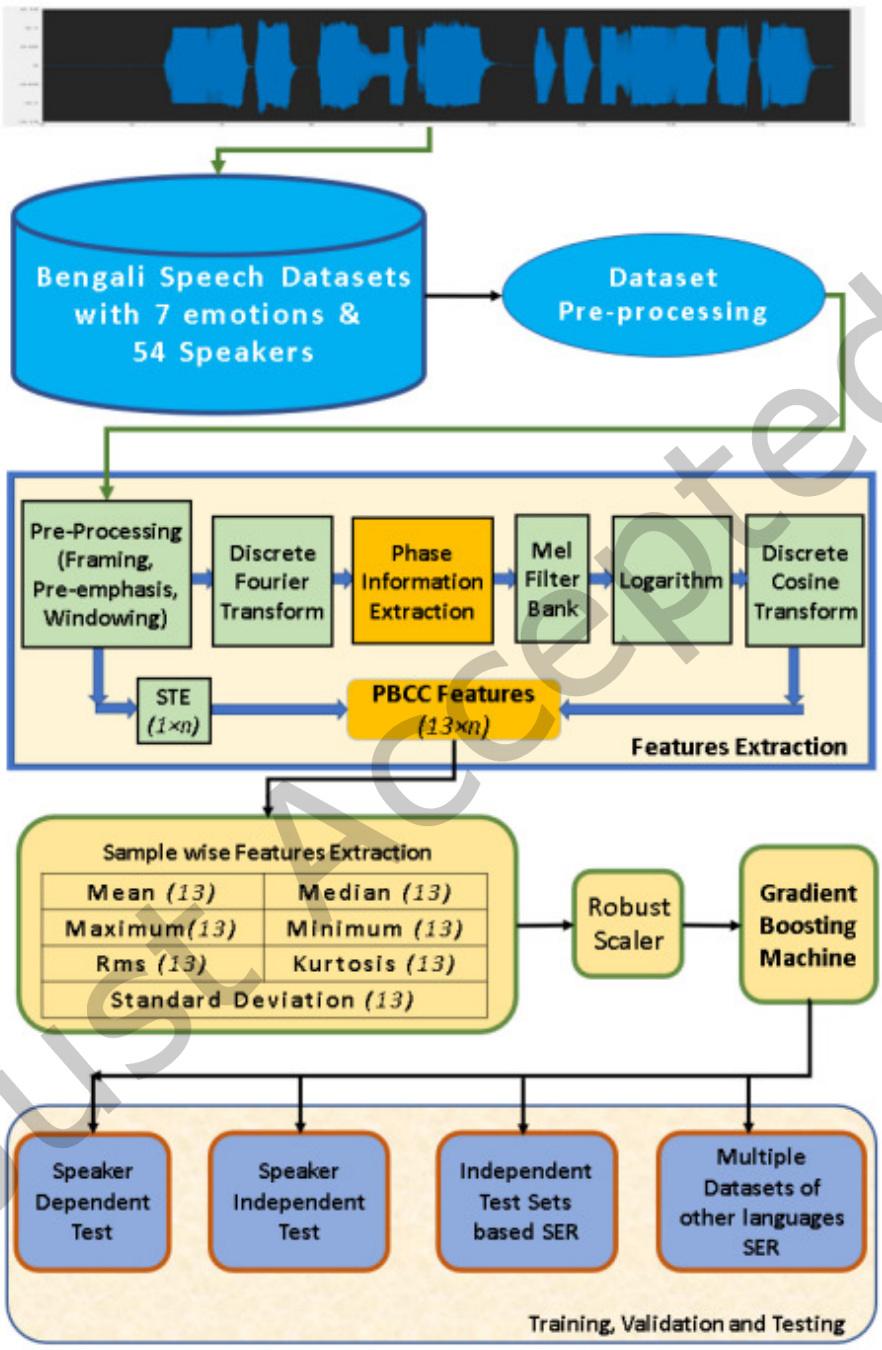

Fig. 3. Block Diagram of the proposed approach

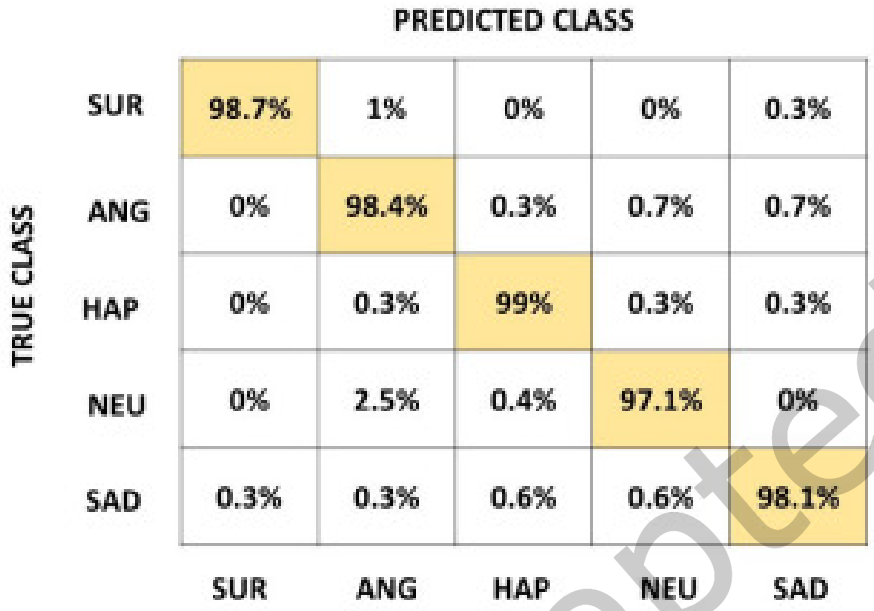

Fig. 4. Confusion matrix of speaker-dependent experiments on BanglaSER dataset based on accuracy

The training and validation are repeated 5 times and the average classification accuracy is reported. By observing the confusion matrices, it is reported that the emotions such as anger and sadness are having the highest detection accuracy in both datasets and the false detection is quite low. It is also observed from evolutionary psychology, that anxiety and anger are two most distinct emotions in humans [31].

4.2 Speaker-independent SER

Comprehensive speaker-independent experiments are carried out in the same way as the speaker-dependent SER. However, the datasets division is different. In these experiments, the data is split into two groups based on the speakers. For the BanglaSER dataset, the speech samples of the 28 speakers are chosen as the training set, while the samples of the remaining 6 speakers are chosen as the testing set. Similarly, for the SUBESCO dataset, the speech samples of the 16 speakers are chosen as the training set, while the samples of the remaining 4 speakers are chosen as the testing set. The confusion matrix performed on both datasets is shown in Figure 6 and Figure 7 respectively.

From Figures 4, 5, 6 and 7 it can be observed that for the speaker-dependent SER, the average classification accuracy is 98.2% and 98.4% for the BanglaSER, and SUBESCO datasets respectively. Similarly, for the speaker-independent SER, the reported classification accuracy is 97.5% and 95.3% on the BanglaSER, and SUBESCO datasets respectively. It can be observed that the happy emotion is recognized with high recognition accuracies for all the tests. The angry and fear emotions are little bit misrecognized for the SUBESCO and sad and neutral for BanglaSER. After listening to some of these misidentified recordings, it is found that the emotions of anger and fear, as well as neutral and sad, have similar perceptual properties. This is most likely due to differences in speaker culture, environment, and education. Different speakers from varied cultures have different chances of

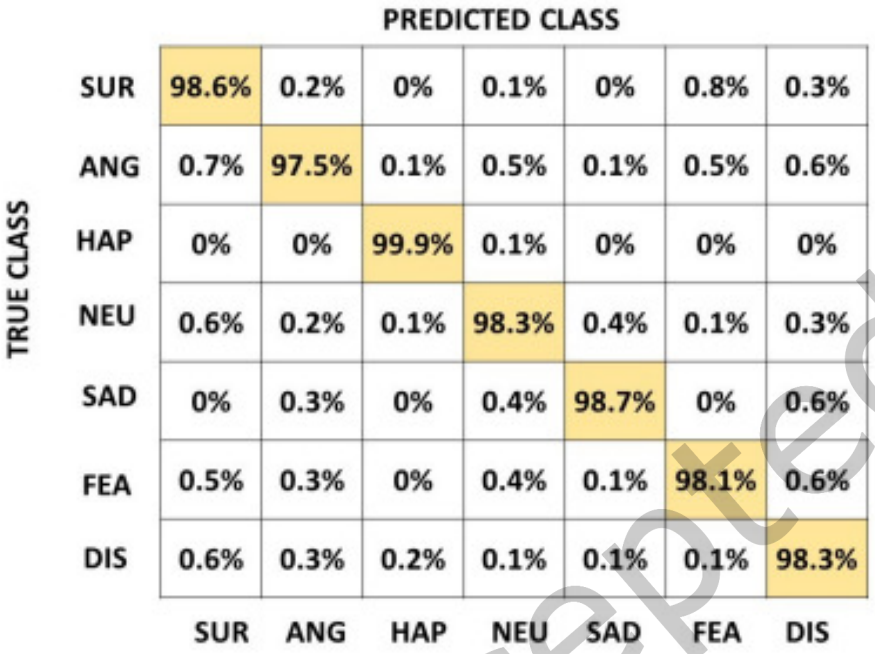

Fig. 5. Confusion matrix of speaker-dependent experiments on SUBESCO dataset based on accuracy

being recognized or misunderstood when they express the same emotion in their speech [49]. With the use of linguistic data, it may also be correctly identified.

4.3 Comparison with Baseline Models

The performance comparison of the proposed PBLG model is done with several standard speech feature extraction techniques including Mel frequency cepstral coefficient (MFCC), Gammatone Cepstrum Coefficient (GTCC), Equivalent rectangular bandwidth scale (ERB) and classical machine learning (ML) based classification methods including support-vector machines (SVM), k-nearest neighbors algorithm (KNN), Naive Bayes(NB), Random Forest (RF). These models (MFCC+SVM [24], MFCC+KNN [18], MFCC+RF [28], GTCC+SVM [39], GTCC+KNN [50], ERB+NB [30], GTCC+RF [3]) are implemented using BanglaSER dataset is presented in Table 2. It is observed that the performance of the proposed method is superior, with a minimum 4% enhancement in accuracy as compared to the standard ML-based models.

The Receiver operator characteristic (ROC) curve is a visual performance indicator showing the relative trade-offs between true and false positives. It provides information about the visualization of the classifier's performance. The ROC curve of the proposed PBLG method for the SUBESCO dataset is shown in Figure 8. It is observed that the PBLG method provides a high true-positive rate and a low false-positive rate with an area under the curve (AUC) of 0.96 as compared to the baseline classifiers.

4.4 Comparison of SER performance on Independent Test Sets

The proposed PBLG SER model is tested for independent tests by training on one dataset and testing on another. As SUBESCO dataset is having 7 emotions, but BanglaSER is having 5 emotions, so for the independent test, the

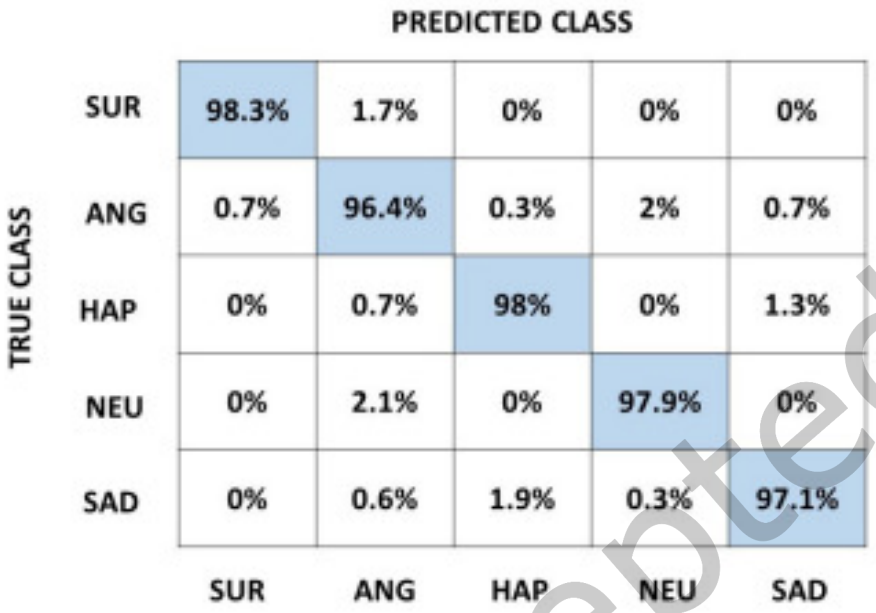

Fig. 6. Confusion matrix of speaker-independent experiments on BanglaSER dataset based on the accuracy

Table 2. Comparison of prediction performance of the proposed method with existing baseline models using Bangla SER dataset

Mehods	Performance Evaluation Measures			
	AC	F-1	PR	RE
MFCC+SVM (Mod_1)	0.84	0.83	0.83	0.82
MFCC+KNN (Mod_2)	0.83	0.81	0.82	0.82
MFCC+RF (Mod_3)	0.94	0.94	0.93	0.93
GTCC+SVM (Mod_4)	0.88	0.87	0.86	0.86
GTCC+KNN (Mod_5)	0.87	0.87	0.86	0.86
ERB+NB (Mod_6)	0.89	0.88	0.89	0.88
GTCC+RF (Mod_7)	0.94	0.92	0.94	0.94
PBLG (proposed)	**0.98**	**0.97**	**0.97**	**0.97**

common five emotions are considered from the SUBSECO dataset which are: angry, happy, sad, and surprise. In the first part of the independent tests, the PBLG model is trained with the SUBESCO dataset, and the trained model is used to predict the emotions of the BanglaSER dataset. In the second instance, the BanlaSER model is used for training and SUBESCO model is used for testing. At the end, a combined Bengali dataset is prepared by including both the SUBESCO and BanglaSER and held -out test is performed. These results are listed in Table 3 and it is observed that the proposed model works quite satisfactorily in the combined dataset, but the performance degrades when the training is done in BanglaSER dataset and tested in SUBESCO dataset. The main reason for

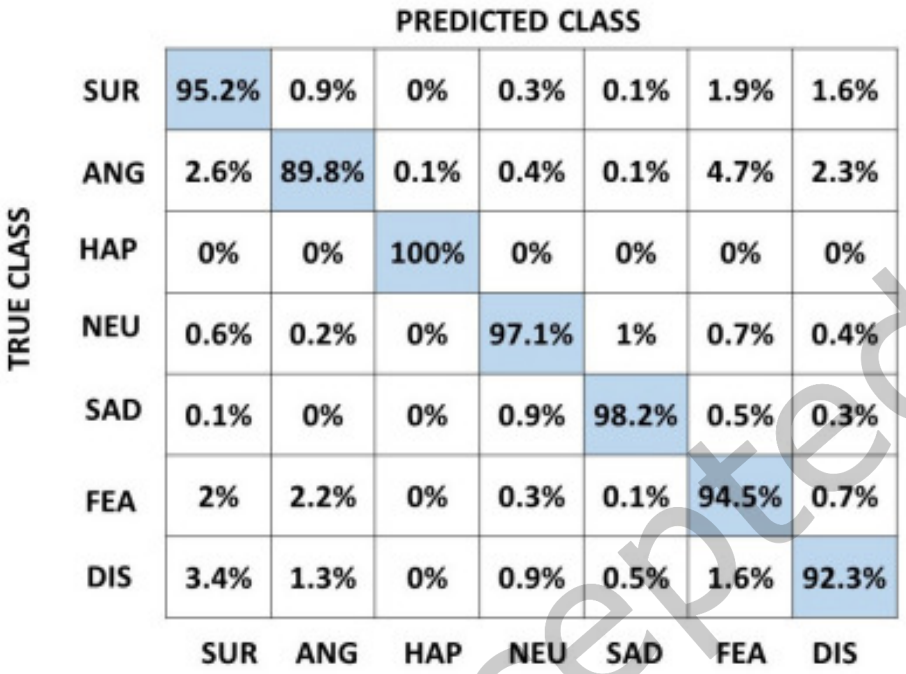

Fig. 7. Confusion matrix of speaker-independent experiments on SUBESCO dataset based on the accuracy

performance degradation in independent tests is the difference in the sentences spoken in both datasets, the availability of fewer number samples for training as well as different speakers used for both datasets.

Table 3. Performance evaluation results of the independent tests

Dataset	Performance Evaluation Measures			
	AC	F-1	PR	RE
SUBESCO(Training)	0.98	0.97	0.97	0.97
BanglaSER (Testing)	0.79	0.78	0.79	0.79
BanglaSER (Training)	0.98	0.98	0.97	0.97
SUBESCO (Testing)	0.75	0.74	0.74	0.74
Combined Dataset (Training)	0.98	0.97	0.97	0.97
Combined Dataset (Testing)	0.96	0.95	0.95	0.95

4.5 Comparison of SER performance on other language datasets

To evaluate the performance of the proposed PBLG model on other language datasets four standard speech emotion datasets are used. These are: Acted Emotional Speech Dynamic Database for Greek language (AESDD) [44], Arabic Natural Audio dataset (ANAD) [17], Canadian French Emotional speech dataset (CFESD) [14], Ryerson

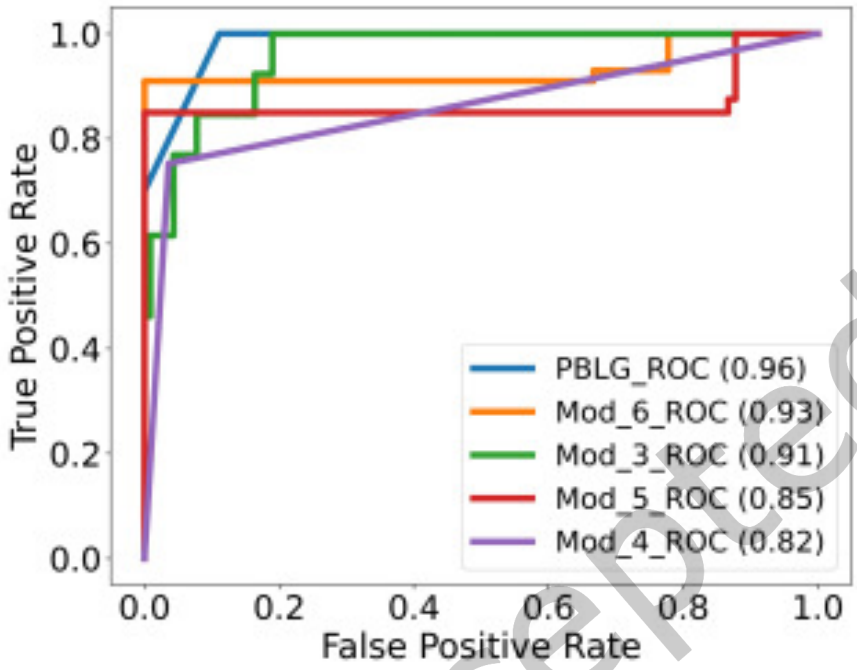

Fig. 8. ROC curves of best 5 classifier models using the SUBESCO dataset

Audio-Visual Database of Emotional Speech and Song (RAVDESS) of English language [20]. The results are shown in Figure 9. It can be observed that the proposed model is demonstrating overall satisfactory performance across multiple other language datasets.

5 CONCLUSION

In this study, novel phase information-based cepstral coefficients are designed and used along with a gradient boosting machine-based classifier for automatic speech emotion recognition of low-resource Indian language. A total of six standard datasets from five different languages are used for the performance evaluation of the proposed model. The simulation results show that the proposed model performs significantly well across multiple datasets. Although the proposed method provided better results compared to standard approaches, still the performance is low in the case of independent tests. This can be improved further by combining linguistic data information. The model may be tested with other Indian language datasets in the future. The major contributions of the investigation can be summarized.

- Application of various preprocessing techniques before features extraction has enabled the model to extract the relevant PBCC feature and these features along with the gradient boosting machines have demonstrated an average SER of 97% on two Bengali speech datasets.
- The detection accuracy of the proposed model is observed to be robust as it offers consistent detection performance over six standard datesets as well, as in independent test sets.
- Improvement in detection accuracy is observed when compared to the standard methods including different types of cepstral features.

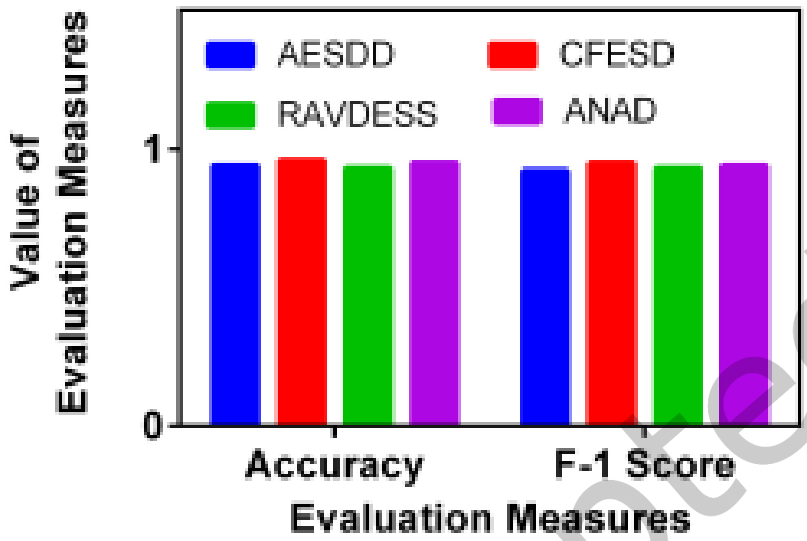

Fig. 9. Comparison of SER performance on other language datasets

REFERENCES

[1] Gaurav Aggarwal, Sarada Prasad Gochhayat, and Latika Singh. 2021. Parameterization techniques for automatic speech recognition system. , 209-250 pages.

[2] Mehmet Berkehan Akçay and Kaya Oğuz. 2020. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication* 116 (2020), 56–76.

[3] Pulung Nurtantio Andono, Guruh Fajar Shidik, Dwi Puji Prabowo, Dewi Pergiwati, and Ricardus Anggi Pramunendar. 2022. Bird Voice Classification Based on Combination Feature Extraction and Reduction Dimension with the K-Nearest Neighbor. *Int. J. Intell. Eng. Syst* 15 (2022), 262–272.

[4] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition* 44, 3 (2011), 572–587.

[5] Joyanta Basu, Soma Khan, Rajib Roy, Tapan Kumar Basu, and Swanirbhar Majumder. 2021. Multilingual speech corpus in low-resource eastern and northeastern Indian languages for speaker and language identification. *Circuits, Systems, and Signal Processing* 40, 10 (2021), 4986–5013.

[6] S Bharathi and T Ananth Kumar. 2020. Translation its Results and Insinuation in Language Learning. *PalArch's Journal of Archaeology of Egypt/Egyptology* 17, 9 (2020), 5081–5090.

[7] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. 2020. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3474–3484.

[8] Rakesh Kumar Das, Nahidul Islam, Md Rayhan Ahmed, Salekul Islam, Swakkhar Shatabda, and AKM Muzahidul Islam. 2022. BanglaSER: A speech emotion recognition dataset for the Bangla language. *Data in Brief* 42 (2022), 108091.

[9] Tusar Kanti Dash, Chinmay Chakraborty, Satyajit Mahapatra, and Ganapati Panda. 2022. Gradient Boosting Machine and Efficient Combination of Features for Speech-Based Detection of COVID-19. *IEEE Journal of Biomedical and Health Informatics* (2022).

[10] Tusar Kanti Dash, Sandeep Singh Solanki, and Ganapati Panda. 2020. Improved phase aware speech enhancement using bio-inspired and ANN techniques. *Analog Integrated Circuits and Signal Processing* 102, 3 (2020), 465–477.

[11] Jun Deng, Xinzhou Xu, Zixing Zhang, Sascha Frühholz, Didier Grandjean, and Björn Schuller. 2017. Fisher kernels on phase-based features for speech emotion recognition. , 195-203 pages.

[12] Noor Fathima, Tanvina Patel, C Mahima, and Anuroop Iyengar. 2018. TDNN-based Multilingual Speech Recognition System for Low Resource Indian Languages.. In *Interspeech*. 3197–3201.
[13] Dimitrios Giannoulis, Michael Massberg, and Joshua D Reiss. 2012. Digital dynamic range compressor design—A tutorial and analysis. *Journal of the Audio Engineering Society* 60, 6 (2012), 399–408.
[14] Philippe Gournay, Olivier Lahaie, and Roch Lefebvre. 2018. A canadian french emotional speech dataset. In *Proceedings of the 9th ACM multimedia systems conference*. 399–402.
[15] Monika Gupta, R K Singh, and Sachin Singh. 2022. G-Cocktail: An Algorithm to Address Cocktail Party Problem of Gujarati Language Using Cat Boost. *Wireless Personal Communications* (2022), 1–20.
[16] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).
[17] S Klaylat, Z Osman, R Zantout, and L Hamandi. 2018. Arabic Natural Audio Dataset, v1. *Mendeley Data* (2018).
[18] Rahul B Lanjewar, Swarup Mathurkar, and Nilesh Patel. 2015. Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques. *Procedia computer science* 49 (2015), 50–57.
[19] Jake Lever, Martin Krzywinski, and Naomi Altman. 2016. Points of significance: model selection and overfitting. *Nature methods* 13, 9 (2016), 703–705.
[20] Steven R Livingstone and Frank A Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (2018), e0196391.
[21] Satyajit Mahapatra and Sitanshu Sekhar Sahu. 2022. ANOVA-particle swarm optimization-based feature selection and gradient boosting machine classifier for improved protein-protein interaction prediction. *Proteins: Structure, Function, and Bioinformatics* 90, 2 (2022), 443–454.
[22] MATLAB. [n. d.]. designAuditoryFilterBank. https://in.mathworks.com/help/audio/ref/designauditoryfilterbank.html
[23] G Miet, A Gerrits, and Jean-Christophe Valiere. 2000. Low-band extension of telephone-band speech. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, Vol. 3. 1851–1854.
[24] A Milton, S Sharmy Roy, and S Tamil Selvi. 2013. SVM scheme for speech emotion recognition using MFCC feature. *International Journal of Computer Applications* 69, 9 (2013).
[25] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. 2227–2231.
[26] Pejman Mowlaee, Rahim Saeidi, and Y Stylianou. 2014. Interspeech 2014 special session: Phase importance in speech processing applications. In *Proc. Interspeech*. 1623–1627.
[27] Sugan Nagarajan, Satya Sai Srinivas Nettimi, Lakshmi Sutha Kumar, Malaya Kumar Nath, and Aniruddha Kanhe. 2020. Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales. *Digital Signal Processing* 104 (2020), 102763.
[28] Fatemeh Noroozi, Tomasz Sapiński, Dorota Kamińska, and Gholamreza Anbarjafari. 2017. Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology* 20, 2 (2017), 239–246.
[29] Kuldip K Paliwal and L Alsteris. 2003. Usefulness of phase in speech processing. In *Proc. IPSJ Spoken Language Processing Workshop, Gifu, Japan*. 1–6.
[30] Zhichao Peng, Zhi Zhu, Masashi Unoki, Jianwu Dang, and Masato Akagi. 2017. Speech emotion recognition using multichannel parallel convolutional recurrent neural networks based on gammatone auditory filterbank. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. 1750–1755.
[31] Michael Bang Petersen. 2010. Distinct emotions, distinct domains: Anger, anxiety and perceptions of intentionality. *The Journal of Politics* 72, 2 (2010), 357–365.
[32] Lawrence R Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77, 2 (1989), 257–286.
[33] K Sreenivasa Rao and K E Manjunath. 2017. *Speech recognition using articulatory and excitation source features*. Springer.
[34] Ibon Saratxaga, Inma Hernaez, Michael Pucher, Eva Navas, and Iñaki Sainz. 2012. Perceptual importance of the phase related information in speech. In *Thirteenth Annual Conference of the International Speech Communication Association*.
[35] Vishwas M Shetty and Metilda Sagaya Mary NJ. 2020. Improving the performance of transformer based low resource speech recognition for Indian languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 8279–8283.
[36] Guangji Shi, Maryam Modir Shanechi, and Parham Aarabi. 2006. On the importance of phase in human speech recognition. *IEEE transactions on audio, speech, and language processing* 14, 5 (2006), 1867–1874.
[37] Amitoj Singh, Virender Kadyan, Munish Kumar, and Nancy Bassan. 2020. ASRoIL: a comprehensive survey for automatic speech recognition of Indian languages. *Artificial Intelligence Review* 53, 5 (2020), 3673–3704.
[38] Sadia Sultana, M Shahidur Rahman, M Reza Selim, and M Zafar Iqbal. 2021. SUST Bangla Emotional Speech Corpus (SUBESCO): An audio-only emotional speech corpus for Bangla. *Plos one* 16, 4 (2021), e0250173.

[39] Anuja Thakur and Sanjeev Kumar Dhull. 2022. Language-independent hyperparameter optimization based speech emotion recognition system. *International Journal of Information Technology* (2022), 1–9.
[40] Alaa Tharwat, Tarek Gaber, Aboul Ella Hassanien, Mohamed K Shahin, and Basma Refaat. 2015. Sift-based arabic sign language recognition system. In *Afro-european conference for industrial advancement*. 359–370.
[41] Daniel Sáez Trigueros, Li Meng, and Margaret Hartnett. 2018. Face recognition: From traditional to deep learning methods. *arXiv preprint arXiv:1811.00116* (2018).
[42] Andreas François Vermeulen. 2019. *Industrial Machine Learning: Using Artificial Intelligence as a Transformational Disruptor*. Apress.
[43] Karthika Vijayan, Pappagari Raghavendra Reddy, and K Sri Rama Murty. 2016. Significance of analytic phase of speech signals in speaker verification. *Speech Communication* 81 (2016), 54–71.
[44] Nikolaos Vryzas, Rigas Kotsakis, Aikaterini Liatsou, Charalampos A Dimoulas, and George Kalliris. 2018. Speech emotion recognition for performance interaction. *Journal of the Audio Engineering Society* 66, 6 (2018), 457–467.
[45] Longbiao Wang, Yohei Yoshida, Yuta Kawakami, and Seiichi Nakagawa. 2015. Relative phase information for detecting human speech and spoofed speech. In *Sixteenth Annual Conference of the International Speech Communication Association*.
[46] Xiong Xiao, Xiaohai Tian, Steven Du, Haihua Xu, Engsiong Chng, and Haizhou Li. 2015. Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge.. In *Interspeech*. 2052–2056.
[47] Yuan Xiwen and Zhu Xiaosong. 2021. Speaker Recognition System with Limited Data based on LightGBM and Fusion Features. In *2021 6th International Conference on Computational Intelligence and Applications (ICCIA)*. 160–164.
[48] Jiali Yu, Yuanyuan Qu, Zhongkai Zhang, Qidong Lu, Zhiliang Qin, and Xiaowei Liu. 2021. Speech recognition based on concatenated acoustic feature and lightGBM model. In *Twelfth International Conference on Signal Processing Systems*, Vol. 11719. 181–188.
[49] Jianfeng Zhao, Xia Mao, and Lijiang Chen. 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical signal processing and control* 47 (2019), 312–323.
[50] Changrui Zhu and Wasim Ahmad. 2019. Emotion recognition from speech to improve human-robot interaction. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*. 370–375.