Published in AJSE, Vol: 24, Issue:3
Received on 30th October 2024
Revised on 25th May 2026
Published on 31st July 2026

AIUB JOURNAL OF SCIENCE AND ENGINEERING
ISSN: 1608 - 3679 (print) 2520 - 4890 (Online)
https://doi.org/10.53799/xsdr1t51

Investigating the Impact of Cross-lingual Acoustic-Phonetic Similarities on Multilingual Speech Emotion Recognition

Syeda Tamanna Alam Monisha [1, *], Sadia Sultana [2]

Abstract

Speech emotion recognition (SER) is a crucial aspect of effective human-computer interaction applications. However, due to diverse phonetic structures and patterns, it is difficult to accurately recognize emotions in speech across different languages. In order to enhance the performance of emotion recognition systems in diverse linguistic circumstances, this study explores how cross-linguistic acoustic-phonetic similarities impact multilingual speech emotion recognition. The study involves a comprehensive analysis of emotion-labeled speech data for Bangla, Hindi, Odia, English, and German languages, with a focus on identifying acoustic-phonetic features that exhibit similarity across languages. We have carried out cross-lingual and multilingual experiments for the five languages. SVM, Random forest, MLP, XGBoost, and CNN classifiers were used for the experiments. A randomized search method with five-fold cross-validation was implemented for hyperparameter tuning, and the machine learning models were evaluated using the parameters obtained. Based on the experimental results, we have found that, machine learning models trained with the languages of a family are capable of recognizing emotions of other languages of the same family better than those of a different one.

Keywords: Cross-lingual, Speech Emotion Recognition, Multilingual, Phonetic similarity

1. Introduction

Speech is undoubtedly the simplest and most natural way of communication, conveying both information and emotions. Emotion plays a crucial role in interpersonal communication, enabling individuals to express their feelings and establish emotional connections. Depending on how it is delivered, the same sentence might have significantly different contextual meanings. In order to understand what a person is trying to say, it is necessary to understand the context. Speech Emotion Recognition (SER) is the field of study focused on understanding and detecting emotions conveyed through human speech. It uses different computational approaches like speech signal processing, feature extraction, and machine learning algorithms to analyze human speech data and predict the underlying emotional state of the speech. Over the past 20 years, human-computer interaction (HCI) has had a remarkable impact on our daily lives. The use of computing devices has permeated daily life, which is evident in how heavily everyone is dependent on computing. For a more human-like HCI system, it is crucial to understand human emotions. A speech emotion recognition (SER) system is capable of predicting the underlying emotional state of human speech. In previous studies, the six basic emotions (happiness, sadness, anger, fear, disgust, and surprise) proposed Ekman [3] in 1992, many times added by a neutral option, were identified by different SER systems. After a good number of high-performance monolingual SER systems, attention has been drawn towards multilingual SER over the last decade. Speech emotion detection systems are often created for a single language, which limits their ability to be generalized to other languages. However, due to globalization, there is a growing need for systems that can understand and interpret emotions in speech across multiple languages. There are many contexts where people of diverse languages meet. For example, emotional recognition plays a crucial role in the mental healthcare sector. As mental healthcare sectors get people from different places around the globe, a multilingual SER system is effective there as it enables more precise diagnosis, individualized mental health care, and treatments for non-native speakers to improve patients' mental health. Multilingual SER can enhance automated call centers, multinational business communications, and customer support by recognizing emotions regardless of the language. Moreover, multilingual SER, independent of spoken language, can support security and law enforcement by identifying emotions in emergency calls or surveillance recordings. Likewise, there are many more contexts where SER systems may encounter people with different languages. However, models find it challenging to reliably identify emotions in multilingual environments because of linguistic and cultural variations. Therefore, multilingual speech emotion recognition aims to address the generalizability issue by developing models that can discriminate emotions across several languages. Understanding how the performance of emotion detection is impacted for numerous languages can be achieved by looking into cross-lingual

Syeda Tamanna Alam Monisha[1] is with the Department of Computer Science and Engineering, Varendra University, Bangladesh. (*corresponding author, email: alammonisha@gmail.com)

Dr. Sadia Sultana[2] is with the Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Bangladesh. (email: sadia-cse@sust.edu)

acoustic-phonetic similarities, which refer to how similar speech elements are across different languages. Understanding and identifying emotions in multilingual contexts can result in more effective human-computer interactions, enhancing cross-cultural communication. In recent times, researchers are more interested in developing a generic SER system that can be implemented for multiple languages. Multilingual SER systems mainly benefit low resource languages due to the lack of availability of proper resources to develop a satisfactory monolingual SER system for the specific language. For the purpose, cross-lingual recognition is also important to observe how the systems perform to an unknown low resource language. Although a good number of experiments have been carried out for multilingual and cross-lingual SER, none of them studied the phonetic similarities among different language speech for SER purposes. While the authors previously explored a DCNN-based multilingual SER framework in [14], the present study focuses on conventional machine learning models and language-family-oriented acoustic analysis. In this research work, we aim to analyze the acoustic-phonetic similarities among multiple language speech and study the impact of phonetic similarities in multilingual speech emotion recognition. This study aims to find how speech characteristics, which may vary across languages, affect the performance of a multilingual speech emotion recognition system. Moreover, research in the speech emotion recognition field has mainly concentrated on resource-rich languages like English, German, Chinese, and French. The vast diversity of emotions conveyed in other languages and cultures has been largely ignored in the field of emotion research. Therefore, in this research, we chose the Bangla, Hindi, and Odia languages of the Indo-Aryan family for similarity measurement. Besides, we used the English and German languages of the Germanic group to see the differences in speech behavior among different groups of languages.

The remaining part of the paper is organized as follows: Recent studies that are relevant are covered in Section 2. Section 3 presents the methodology. Section 4 gives a brief analysis of the speech features, Section 5 shows the experimental results, while section 6 provides a discussion and finally section 7 draws a conclusion.

2. Related Works

The field of speech emotion recognition has gained a lot of attention over the last two decades. After the first experiment done for speech emotion recognition in 1978 [21], numerous researches have been conducted on speech emotion recognition in monolingual, cross-lingual, and multilingual contexts. Compared to the number of studies for monolingual speech emotion recognition, the attempts made for cross-lingual and multilingual recognition are few. In 2003, Hozjan and Kačič [7] presented an artificial neural network (ANN) architecture using both short-term low-level and long-term high-level features. Authors used the InterFace corpus, consisting of emotional speech of French, Slovenian, Spanish, and English languages. The average recognition rate for multilingual speech for the system was 37.19%. Swain et al. [19] in 2015 presented an architecture that calculates MFCC, log power, Delta MFCC, Double delta MFCC, LFPC, and LPCC as feature vectors. They implemented HMM and SVM to classify the emotions of Sambalpuri and Cuttacki, and two Odisha native languages. The highest accuracy was obtained using only MFCC as a feature vector. Authors got 82.14% and 78.81% recognition rate for SVM and HMM, respectively, for Sambalpuri language. In 2019, Li and Akagi [11] proposed a multilingual SER system using a three-layer model incorporating acoustic features, semantic primitives, and emotion dimensions. The proposed model using logistic model trees (LMT) was tested on English, German, Chinese, and Japanese emotional speech databases. The system achieved a weighted average precision (WAP) of 81.72% to classify happy, sad, angry, and neutral emotions. Heracleous and Yoneyama [6] proposed a two-pass classification method based multilingual SER system which first identifies the language spoken and then predicts the emotion. The experiment was conducted on the English IEMOCAP, German FAU Aibo, German Emo-DB, and a Japanese speech corpus. The authors implemented fully connected DNN and CNN classifiers along with i-vector features to experiment. The unweighted average recall (UAR) for the system was 85.87% and 84.75% using CNN and DNN, respectively for the Japanese speech corpus. Lee [10] observed that normalization and regularization frameworks improve the generalizability in multilingual speech emotion recognition even merging two highly heterogeneous languages. Author has done comparative analysis with the IEMOCAP and the JTES databases for English and Japanese languages respectively and obtained a maximum of 59.49% unweighted average recall on the IEMOCAP dataset. In 2020, [5] presented two scenarios for English, Spanish, and Italian emotion recognition combining language features and emotion features. The study was conducted on the EmoFilm multilingual emotional speech corpus considering five emotions, happiness, anger, sadness, disgust, and fear. In the first scenario, a single extremely randomized trees (ERT) classifier and in the second one, a stacked generalization ensemble (SGE) with two parallel ERTs for emotion and language features, respectively, topped with one more ERT meta-classifier for the final prediction was used. Authors got the best result from the first method with UAR of 73.3%. Goel and Beigi [4] have conducted cross-lingual cross-corpus study using EmoDB, SAVEE, EMOVO, MASC, and IEMOCAP datasets with SVC classifier and obtained accuracies of 32%, 51%, and 65% on EMOVO, SAVEE, and EmoDB datasets respective while training was done with IEMOCAP. A language independent CNN and BiLSTM based multilingual SER architecture was proposed by Yadav and Vishwakarma [22] in 2020. For validation, the authors used EmoDB and RAVDESS databases for German and English languages, respectively. Using MFCC as feature vector, authors achieved 94% and 73% average accuracy for recognizing emotion from German and English language speech,

respectively. In 2021, Scotti et al. [15] presented a CNN architecture combining textual and deep acoustic features. The proposed architecture was trained and tested on IEMOCAP, EmoFilm, SES, and AESI speech corpora for English, Itaian, Spanish, and Greek language speech. The system obtained an average weighted accuracy (WA) of 78.5% for recognizing four emotions, happy, angry, sad, and neutral. Zehra et al. [24] experimenting with SAVEE, EmoDB, EMOVO, and Urdu corpus using MFCCs, ZCR, pitch, energy, and chroma features showed that ensemble learning approach of Random Forest, Decision Tree, and Sequential Minimal Optimization better performs than the traditional machine learning models both on cross-corpus and within-corpus data. In [1], authors presented a transformer model using data augmentation and fusion of acoustic features. The proposed multilingual SER model was evaluated on EMOVO, BAVED, SAVEE, and EMO-DB databases, among which the highest recognition rate of 95.2% was obtained for the BAVED speech corpus. Lately in 2022, Wang et al. [20] presented a multilingual fusion model which combines handcraft features and DNN-extracted features for feature vector. The model makes use of RNN with an improved local attention mechanism and was evaluated on English IEMOCAP, SAVEE databases and Chinese CASIA, CHEAVD2.0 databases. The highest accuracy obtained by the model was 81.8% for the SAVEE corpus. Yanting Yu, in 2022, showed that combining data mining could improve the performance of multilingual speech emotion recognition [23]. For the experiment, the author developed a model with BiLSTM, which uses data mining technology for effective analysis of speech emotion. The first cross-lingual study involving Bangla with English language was reported in Sultana et al. [16] for a Deep CNN model combining with BiLSTM. Another study Sultana and Rahman [17] showed acoustic feature analysis and presented an optimized feature set for Bangla and English datasets SUBESCO and RAVDESS, respectively. From the studies it can be observed that, due to cultural differences and linguistic nuances the performance measures differ across languages. Authors Kamaruddin et al. [8] also concluded from their study that speech emotion recognition accuracy is influenced by culture.

To the best of our knowledge, no study has specifically explored the acoustic feature similarities among speech signals of languages belonging to the same language family for emotion recognition purposes. Therefore, in this study, we tried to find out the acoustic phonetic similarities across different language speech of the same family and observed their impact on the performance of multilingual emotion recognition in spite of the cultural differences.

3. Methodology

3.1. Emotional Speech Databases

In this study, we have used SUBESCO (SUST Bangla Emotional Speech Corpus) [18], IITKGP-SEHSC (Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus) [9], SITB-OSED (SIT Bhubaneswar-Odia Speech Emotion Database) [13], RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [12], and EmoDB [2] emotional speech databases for Bangla, Hindi, Odia, English, and German languages, respectively. The five languages belong to two language families, Bangla, Hindi, and Odia represent the Indo-Aryan family, whereas English and German belong to the Germanic family. All the five datasets are acted and maintain a balanced distribution of male and female speakers. 7000 Bangla speech data were obtained from the SUBESCO corpus, which contains recordings from twenty professional speakers expressing seven emotional states: happiness, anger, fear, surprise, disgust, sadness, and neutral. For Hindi, the IITKGP-SEHSC corpus was employed, comprising 12000 emotional utterances produced by ten professional radio artists and covering eight emotion categories, namely happy, sad, anger, fear, sarcasm, disgust, surprise, and neutral. Odia language data were collected from the SITB-OSED corpus, which includes 12110 speech samples from five major dialects—Cuttacki, Sambalpuri, Berhampuri, Phulbani, and Baleswari—recorded by twenty native professional speakers. The corpus has six emotional classes: happiness, anger, sadness, surprise, fear, and disgust. For the Germanic language family, English speech samples were extracted from the RAVDESS database. Since the objective of this work is limited to speech-based emotion analysis, only the speech recordings were used, while the song recordings were excluded. The selected subset consists of 1440 emotional utterances representing happiness, anger, sadness, fear, and neutral. German emotional speech data were obtained from the EmoDB corpus, which contains 535 recordings from ten professional actors portraying emotions such as happiness, anger, sadness, boredom, anxiety, disgust, and neutral. A detailed overview of the datasets is presented in Table 1. It should be noted that while some databases provide many takes for a single utterance, others do not. In this case, the models would have been better trained for databases with several takes than for those with a single take. Thus, we had to compress the datasets by taking only one take for each utterance, as indicated in Table 1.

3.2. Data Preprocessing

As the datasets used in this research work were developed in studio settingd, there is minimal to no noise in the speech signals. The speech signals were sampled at a sampling rate of 10KHz. The librosa framework was utilized to read and sample the signals. Since there is not much information in the silent parts of speech signals, signal trimming was done under 10dB for silence removal. For feature extraction, each speech signal was divided into segments of 25ms with 10ms overlap. In addition, StandardScaler was used for feature standardization in order to normalize the retrieved acoustic characteristics for every dataset enabling the models to focus more effectively on the underlying acoustic and phonetic patterns.

3.3. Feature Extraction

In this research, we have taken acoustic features only and extracted pitch, intensity, zero-crossing rate (ZCR), formants

Table 1
Characteristics of the emotional speech datasets

Database	Language	Family	Actors	Original Size	Used Size	Emotion wise no. of utterances
SUBESCO	Bangla	Indo-Aryan	20	7000	1400	happiness(200), anger(200), fear(200), sadness(200), surprise(200), disgust(200), neutral(200)
IITKGP-SEHSC	Hindi	Indo-Aryan	10	12000	638	happy(83), disgust(90), surprise(75), sad(75), anger(75), sarcastic(75), fear(90), neutral(75)
SITB-OSED	Odia	Indo-Aryan	20	12110	2352	happy(385), surprise(391), sadness(400), anger(378), disgust(406), fear(392)
RAVDESS	English	Germanic	24	7356	528	sad(96), happy(96), angry(96), fearful(96), neutral(144)
EmoDB	German	Germanic	10	535	474	happiness(61), anxiety(62), boredom(76), sadness(58), disgust(43), anger(97), neutral(77)

(F1, F2 and F3), MFCCs (13), and chromas (12) for further analysis. We have taken mean and standard deviation of pitch and zero-crossing rate, mean, minimum and maximum of intensity, mean and median of the formants, and mean values of MFCCs and chromas. Therefore, in total 38 features were extracted for the study.

3.4. ML Models for Emotion Recognition

For the recognition of emotion from speech signals, we have used four machine learning models: Support vector machine (SVM), Random forest (RF), Multilayer perceptron (MLP), and Extreme gradient boosting (XGBoost). These are some of the most commonly used classification algorithms for emotion detection purposes. SVM is a classifier that can handle high-dimensional data and effectively divide data into multiple classes. It performs well with small to medium-sized datasets and, using a kernel method, can handle both linearly and non-linearly separable data. The RF approach uses ensemble learning to build numerous decision trees during training and combines the results to produce predictions. It is more accurate than a single decision tree and less prone to overfitting. MLP is an artificial neural network comprising multiple layers of interconnected neurons. It is used for both classification and regression problems and is well-known for learning complex patterns in data. XGBoost is an advanced ensemble learning technique combining gradient boosting and regularization and often performs better than other machine learning algorithms. To find the best parameters for each setup, Randomized Search approach with 5-fold cross-validation was used to tune the models. Tables 2 to 5 display the hyperparameter ranges for the four models.

Table 2
Hyperparameter ranges for SVM.

Parameter	Range
Kernel	linear, poly, rbf
C	0.1, 1, 10, 100
Gamma	1, 0.1, 0.01, 0.001, 0.0001

Table 3
Hyperparameter ranges for MLP.

Parameter	Range
Number of hidden layers	3
Number of neurons in i^{th} hidden layer	(150,100,50), (120,80,40), (100,50,30)
Max number of iterations	50, 100, 150
Activation	identity, logistic, tanh, relu
Solver	sgd, adam
Alpha	0.0001, 0.05
Learning rate	constant, adaptive

3.5. Deep Learning Model

In addition to the traditional models, a Convolutional Neural Network (CNN) model was incorporated as a deep learning baseline to validate the robustness of the findings obtained from traditional ML approaches. The objective is to observe if the emotional patterns across languages observed using handcrafted acoustic features remain consistent when evaluated using a deep learning architecture capable of learning higher-level feature representations automatically.

Table 4
Hyperparameter ranges for RF.

Parameter	Range
Number of trees	20, 50, 100 to 2000, +200
Max number of features	sqrt, log2, None
Max depth	10 to 20
Min samples split	2, 6, 10
Min samples leaf	1, 3, 4
Criterion	gini, entropy, log loss
Bootstrap	True, False

Table 5
Hyperparameter ranges for XGBoost.

Parameter	Range
Max depth	3 to 12
Alpha	0, 0.001, 0.01, 0.1
Subsample	0.25, 0.5, 0.75, 1
Learning rate	0.01 to 10, +0.5
Number of trees	10, 25, 40

Table 6
Hyperparameter ranges for CNN.

Parameter	Range
Number of CNN Layers	2
Kernel size	3
Filters	64, 128
Activation	ReLU
Optimizer	adam
Dense Layers	128, 64
Output Activation	softmax
Loss Function	categorical crossentropy
Batch Size	32
Epochs	100
Validation Split	0.1

4. Feature Analysis

Pitch, intensity, zero-crossing rate (ZCR), formants (F1, F2, and F3), MFCCs, and chromas are the extracted speech features of SUBESCO, IITKGP-SEHSC, SITB-OSED, RAVDESS, and EmoDB databases considered for analysis in this study. Here, SUBESCO, IITKGP-SEHSC, and SITB-OSED are Bangla, Hindi, and Odia language speech databases, respectively, which belong to the Indo-Aryan language family. On the contrary, RAVDESS and EmoDB speech databases are in English and German, respectively, and belong to the Germanic family. Figure 1 and 2 depict the ranges of prominent speech intensity for male and female speakers separately for the five databases. Here, we have shown the results for only anger, happiness, fear, and sad emotions, as these are the only common emotions in all the five language databases. It can be seen from figure 1 and 2 that the intensities of both female and male speakers speech of SUBESCO, IITKGP-SEHSC, and SITB-OSED are higher than the intensities of RAVDESS and EmoDB speech but they are close to each other. The intensities of RAVDESS speech are closer to the ones of EmoDB than the other four. The ranges of the female intensity of Indo-Aryan languages in an average of angry emotion lies between 75-87 dB, happy 70-80 dB, fear 67-80 dB, and sad 70-82 dB, whereas the maximum limit of Germanic languages if 80 dB and the lower limit are below 70 dB. There are some exceptions, such as some intensities of the Hindi and English language speech below the lower limit of their groups. Although English and EmoDB are of the same language family, the reason English speech intensities are a bit lower than German speech may be that English-speaking cultures often value politeness and more moderate speech patterns, which lead to a lower intensity. So, we can say that the intensity ranges of different languages in the same family are quite similar.

Additionally, to quantitatively examine cross-lingual acoustic relationships, similarity analysis was performed using feature-based distance measures on the extracted acoustic representations. Pairwise similarity between languages was evaluated using Euclidean distance computed from the standardized acoustic feature vectors.As shown in Equation 1, the Euclidean distance measures direct distance between feature vectors where smaller distance indicates more similar.

$$d(A, B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \tag{1}$$

Here, where A and B represent the acoustic feature vectors of two languages, and n denotes the total number of extracted features.

Moreover, XGBoost feature importance analysis was used to find the most important features for emotion recognition and investigate whether common audio patterns are shared within the same language family.

As mentioned earlier, due to the difference in database sizes, some having too much data and some less, some exceptions can be seen. For example, the SITB-OSED database for the Odia language contains 2352 recordings, while IITKGP-SEHSC contains only 638. More specifically, if we consider individual emotions, SITB-OSED has 400 utterances for sad emotions, whereas IITKGP-SEHSC has 75 utterances. It is believed that if the size of the databases were balanced, better observations could be reported.

5. Experimental Results

5.1. Evaluation Metric

The performance of each model was reported using the accuracy metric, precision, recall, and F1 score. Mean values of accuracy, precision, recall, and F1 score are reported. The accuracy score in machine learning determines how many of a model's predictions are correct in relation to the total number of predictions.

$$accuracy = \frac{number of correct predictions}{total number of predictions}$$

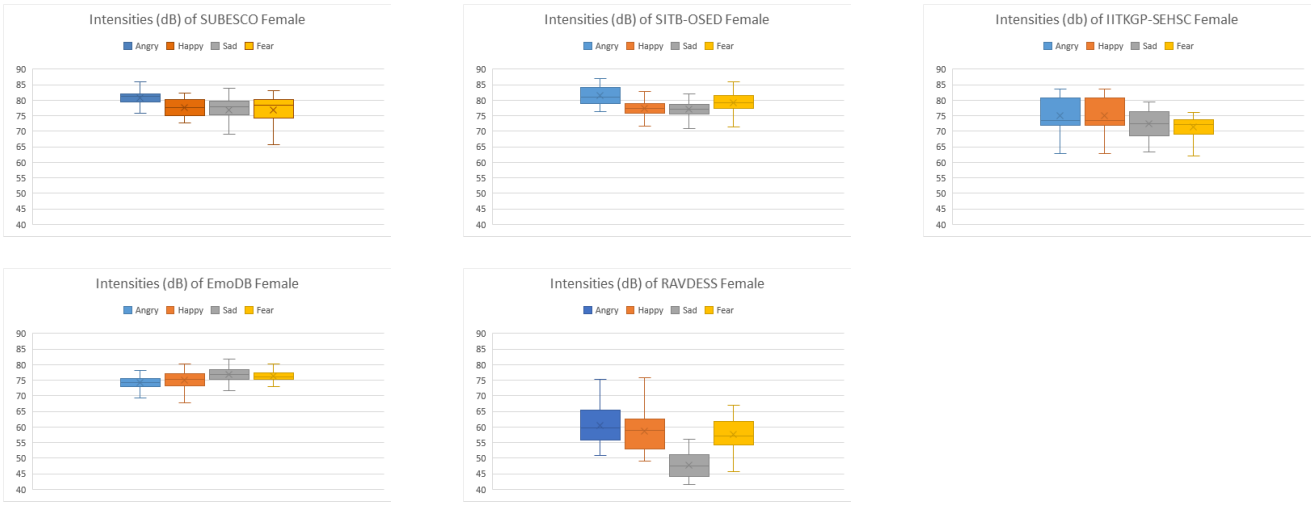

Figure 1: Intensity range for female speakers of five different databases

Figure 2: Intensity range for male speakers of five different databases

The precision, also known as positive predictive value of a model is defined as the ratio of true positive predictions to all positive predictions made by the model.

$$precision = \frac{truepositive}{truepositive + falsepositive}$$

Recall which is also knows as true positive rate or sensitivity is the ratio of true positive predictions to all real positive cases.

$$recall = \frac{truepositive}{truepositive + falsenegative}$$

Lastly, the harmonic mean of precision and recall determines the F1 score.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

5.2. Setup 1: Monolingual Recognition

In this setup we have done monolingual experiment as baseline using same dataset for training and testing purpose. Here, only the four prevalent emotions across all datasets have been taken into consideration, anger, happiness, fear, and sadness. Extracted features of these emotions were used

to train and test the machine learning models. The training-testing ratio was 80:20.

5.3. Setup 2: Cross-lingual Recognition

In setup 2, we have done cross-lingual cross-corpus evaluation to check how the models perform in recognizing the emotions of a different language speech. We have trained the models with one language corpus and used the rest four databases as testing corpus. We have also considered only the four common emotions, happiness, anger, fear, and sadness. From Table 12, it is evident that, when train data is from SUBESCO, IITKGP-SEHSC, SITB-OSED and test data is also from these three, the models better recognize the emotions than test data being from RAVDESS or EmoDB. Same pattern can be observed for RAVDESS and EmoDB also. When training corpus is RAVDESS or EmoDB and test corpus is also any of these two, the accuracy measures are higher than that of test corpus being SUBESCO or IITKGP-SEHSC or SITB-OSED.

5.4. Setup 3: Multilingual Recognition

In this setup, we conducted multilingual training and testing to check the impact of acoustic phonetic similarities on multilingual speech emotion recognition. Here we have combined the SUBESCO, IITKGP-SEHSC, and SITB-OSED datasets to one dataset named multilingual Indo-Aryan and RAVDESS and EmoDB to another one named multilingual Germanic to make multilingual datasets. First, we conducted different language family training testing where we have trained the models with multilingual Indo-Aryan language dataset and tested with multilingual Germanic dataset and also have done vice versa. Secondly, we conducted multilingual speech emotion detection for individual language families keeping training and testing databases same where the training testing ration was 80:20.

Table 7
Performance measures for SUBESCO in Setup 1

ML model	Acc.	Pr.	Re.	F1
SVM	0.76	0.77	0.76	0.77
RF	0.79	0.79	0.80	0.79
MLP	0.81	0.79	0.80	0.79
XGB	0.79	0.79	0.78	0.78
CNN	0.71	0.70	0.71	0.70

Table 8
Performance measures for IITKGP-SEHSC in Setup 1

ML model	Acc.	Pr.	Re.	F1
SVM	0.75	0.78	0.73	0.74
RF	0.83	0.86	0.82	0.83
MLP	0.82	0.83	0.82	0.82
XGB	0.79	0.78	0.76	0.77
CNN	0.74	0.75	0.75	0.74

Table 9
Performance measures for SITB-OSED in Setup 1

ML model	Acc.	Pr.	Re.	F1
SVM	0.89	0.88	0.86	0.86
RF	0.90	0.89	0.88	0.88
MLP	0.90	0.91	0.90	0.90
XGB	0.91	0.90	0.89	0.90
CNN	0.86	0.87	0.86	0.86

Table 10
Performance measures for RAVDESS in Setup 1

ML model	Acc.	Pr.	Re.	F1
SVM	0.70	0.71	0.69	0.69
RF	0.72	0.73	0.70	0.71
MLP	0.69	0.70	0.67	0.68
XGB	0.71	0.68	0.67	0.67
CNN	0.68	0.68	0.67	0.67

Table 11
Performance measures for EmoDB in Setup 1

ML model	Acc.	Pr.	Re.	F1
SVM	0.89	0.90	0.84	0.86
RF	0.88	0.87	0.82	0.84
MLP	0.81	0.86	0.81	0.77
XGB	0.86	0.84	0.80	0.81
CNN	0.77	0.75	0.73	0.72

Table 12
Accuracy measures for Setup 2

Train Data	Test Data	ML model	Acc.	Pr.	Re.	F1
SUBESCO	IITKGP-SEHSC	SVM	0.57	0.62	0.57	0.58
		RF	0.63	0.64	0.63	0.63
		MLP	0.55	0.62	0.55	0.57
		XGB	0.59	0.61	0.59	0.60
		CNN	0.50	0.61	0.51	0.55
	SITB-OSED	SVM	0.54	0.58	0.54	0.55
		RF	0.55	0.53	0.55	0.54
		MLP	0.55	0.54	0.55	0.54
		XGB	0.52	0.50	0.52	0.50
		CNN	0.48	0.57	0.49	0.52
	RAVDESS	SVM	0.42	0.46	0.42	0.41
		RF	0.45	0.44	0.44	0.44
		MLP	0.40	0.42	0.40	0.38
		XGB	0.48	0.47	0.48	0.46
		CNN	0.39	0.27	0.38	0.31
	EmoDB	SVM	0.46	0.47	0.45	0.46
		RF	0.39	0.41	0.39	0.39
		MLP	0.37	0.43	0.37	0.38
		XGB	0.37	0.39	0.37	0.37
		CNN	0.40	0.39	0.37	0.38
IITKGP-SEHSC	SUBESCO	SVM	0.54	0.53	0.54	0.52
		RF	0.58	0.59	0.58	0.57
		MLP	0.54	0.55	0.54	0.54
		XGB	0.59	0.58	0.57	0.57
		CNN	0.42	0.42	0.42	0.41
	SITB-OSED	SVM	0.52	0.51	0.52	0.50
		RF	0.54	0.53	0.52	0.52
		MLP	0.52	0.51	0.52	0.52
		XGB	0.53	0.51	0.50	0.51
		CNN	0.41	0.0.47	0.40	0.43
	RAVDESS	SVM	0.40	0.43	0.40	0.40
		RF	0.46	0.42	0.46	0.46
		MLP	0.44	0.45	0.44	0.43
		XGB	0.47	0.48	0.46	0.47
		CNN	0.30	0.15	0.30	0.20
	EmoDB	SVM	0.40	0.43	0.38	0.37
		RF	0.41	0.43	0.41	0.42
		MLP	0.41	0.44	0.41	0.42
		XGB	0.38	0.43	0.38	0.37
		CNN	0.34	0.37	0.33	0.35

Table 12
Accuracy measures for Setup 2 (continued)

Train Data	Test Data	ML model	Acc.	Pr.	Re.	F1
SITB-OSED	SUBESCO	SVM	0.52	0.51	0.50	0.50
		RF	0.58	0.57	0.57	0.57
		MLP	0.53	0.51	0.52	0.52
		XGB	0.55	0.54	0.55	0.54
		CNN	0.45	0.50	0.47	0.48
	IITKGP-SEHSC	SVM	0.52	0.51	0.53	0.51
		RF	0.53	0.51	0.52	0.50
		MLP	0.53	0.52	0.53	0.52
		XGB	0.54	0.54	0.54	0.52
		CNN	0.48	0.44	0.47	0.45
	RAVDESS	SVM	0.43	0.45	0.43	0.42
		RF	0.47	0.48	0.47	0.46
		MLP	0.38	0.46	0.38	0.37
		XGB	0.42	0.46	0.42	0.42
		CNN	0.29	0.37	0.29	0.22
	EmoDB	SVM	0.39	0.41	0.39	0.38
		RF	0.40	0.42	0.40	0.40
		MLP	0.42	0.43	0.42	0.41
		XGB	0.42	0.44	0.42	0.41
		CNN	0.33	0.27	0.31	0.28
RAVDESS	SUBESCO	SVM	0.44	0.45	0.44	0.44
		RF	0.41	0.40	0.41	0.40
		MLP	0.42	0.43	0.41	0.42
		XGB	0.42	0.44	0.42	0.44
		CNN	0.29	0.13	0.29	0.18
	IITKGP-SEHSC	SVM	0.40	0.43	0.40	0.41
		RF	0.42	0.45	0.42	0.42
		MLP	0.42	0.42	0.42	0.41
		XGB	0.45	0.46	0.45	0.45
		CNN	0.24	0.08	0.25	0.12
	SITB-OSED	SVM	0.41	0.41	0.41	0.40
		RF	0.48	0.48	0.48	0.46
		MLP	0.39	0.40	0.41	0.40
		XGB	0.45	0.45	0.45	0.44
		CNN	0.28	0.37	0.29	0.18
	EmoDB	SVM	0.51	0.54	0.50	0.50
		RF	0.58	0.59	0.58	0.58
		MLP	0.53	0.56	0.53	0.54
		XGB	0.55	0.55	0.56	0.55
		CNN	0.48	0.42	0.45	0.43

Table 12
Accuracy measures for Setup 2 (continued)

Train Data	Test Data	ML model	Acc.	Pr.	Re.	F1
EmoDB	SUBESCO	SVM	0.46	0.42	0.46	0.42
		RF	0.41	0.40	0.41	0.36
		MLP	0.48	0.47	0.48	0.46
		XGB	0.47	0.47	0.47	0.42
		CNN	0.37	0.39	0.37	0.38
	IITKGP-SEHSC	SVM	0.42	0.33	0.42	0.37
		RF	0.41	0.48	0.41	0.36
		MLP	0.48	0.47	0.48	0.43
		XGB	0.44	0.46	0.44	0.40
		CNN	0.37	0.26	0.37	0.30
	SITB-OSED	SVM	0.43	0.39	0.43	0.39
		RF	0.46	0.43	0.46	0.43
		MLP	0.43	0.41	0.43	0.40
		XGB	0.45	0.43	0.44	0.41
		CNN	0.30	0.24	0.30	0.23
	RAVDESS	SVM	0.52	0.53	0.55	0.54
		RF	0.57	0.54	0.57	0.56
		MLP	0.52	0.53	0.52	0.52
		XGB	0.54	0.51	0.55	0.53
		CNN	0.48	0.45	0.44	0.44

Table 13
Performance measures for Setup 3

Train Data	Test Data	ML model	Acc.	Pr.	Re.	F1
Multilingual Indo-Aryan	Multilingual Germanic	SVM	0.44	0.46	0.44	0.44
		RF	0.45	0.44	0.44	0.43
		MLP	0.43	0.44	0.42	0.43
		XGB	0.47	0.46	0.46	0.46
		CNN	0.34	0.38	0.34	0.32
Multilingual Germanic	Multilingual Indo-Aryan	SVM	0.45	0.44	0.45	0.43
		RF	0.42	0.41	0.42	0.41
		MLP	0.42	0.42	0.43	0.42
		XGB	0.41	0.41	0.41	0.39
		CNN	0.32	0.26	0.32	0.23
Multilingual Indo-Aryan	Same	SVM	0.73	0.73	0.73	0.73
		RF	0.79	0.79	0.79	0.79
		MLP	0.84	0.84	0.84	0.84
		XGB	0.83	0.83	0.83	0.83
		CNN	0.79	0.79	0.79	0.79
Multilingual Germanic	Same	SVM	0.70	0.68	0.70	0.69
		RF	0.71	0.71	0.71	0.69
		MLP	0.69	0.68	0.69	0.68
		XGB	0.71	0.72	0.71	0.71
		CNN	0.69	0.70	0.68	0.68

5.5. Quantitative Similarity Analysis & XGBoost Feature Importance

As shown in Figure 3, the quantitative similarity matrix using Euclidean distance reveals relatively smaller distance values among languages belonging to the same language family compared to languages from different language families, indicating closer acoustic feature distributions across related languages. However, a minor deviation was observed for the Odia–German language pair, which may be attributed to corpus-specific variability and recording-condition differences.

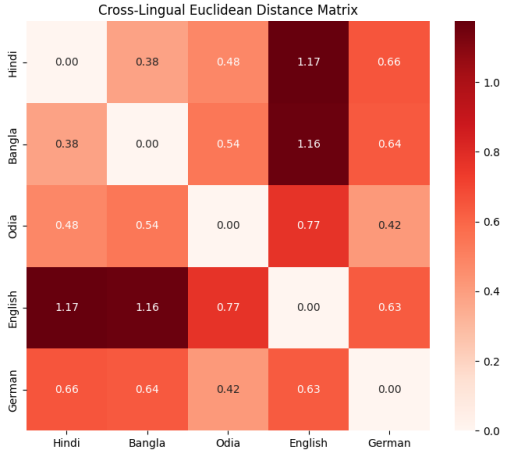

Figure 3: Quantitative Similarity Matrix

Table 14
Emotion wise performance measures while trained with Multilingual Indo-Aryan and tested with Multilingual Germanic Data

ML model	Emotion	Pr.	Re.	F1
SVM	Angry	0.65	0.48	0.56
	Happy	0.36	0.40	0.38
	Sad	0.36	0.47	0.41
	Fear	0.36	0.47	0.41
RF	Angry	0.54	0.61	0.57
	Happy	0.38	0.29	0.33
	Sad	0.45	0.43	0.44
	Fear	0.37	0.41	0.39
MLP	Angry	0.66	0.46	0.54
	Happy	0.38	0.45	0.41
	Sad	0.35	0.44	0.39
	Fear	0.38	0.35	0.37
XGB	Angry	0.62	0.62	0.62
	Happy	0.41	0.32	0.36
	Sad	0.40	0.50	0.45
	Fear	0.39	0.39	0.39
CNN	Angry	0.41	0.47	0.44
	Happy	0.50	0.18	0.26
	Sad	0.32	0.14	0.20
	Fear	0.31	0.61	0.41

Table 15
Emotion wise performance measures while trained with Multilingual Germanic and tested with Multilingual Indo-Aryan Data

ML model	Emotion	Pr.	Re.	F1
SVM	Angry	0.50	0.69	0.58
	Happy	0.37	0.28	0.32
	Sad	0.44	0.49	0.46
	Fear	0.44	0.33	0.38
RF	Angry	0.52	0.81	0.63
	Happy	0.31	0.36	0.33
	Sad	0.42	0.25	0.31
	Fear	0.39	0.27	0.32
MLP	Angry	0.52	0.65	0.58
	Happy	0.52	0.65	0.58
	Sad	0.52	0.65	0.58
	Fear	0.37	0.35	0.36
XGB	Angry	0.47	0.74	0.58
	Happy	0.32	0.38	0.35
	Sad	0.46	0.31	0.37
	Fear	0.39	0.23	0.29
CNN	Angry	0.25	0.37	0.30
	Happy	0.28	0.75	0.41
	Sad	0.35	0.04	0.08
	Fear	0.15	0.12	0.13

Table 16
Emotion wise performance measures for Multilingual Indo-Aryan Data

ML model	Emotion	Pr.	Re.	F1
SVM	Angry	0.72	0.80	0.76
	Happy	0.71	0.66	0.68
	Sad	0.70	0.71	0.71
	Fear	0.80	0.74	0.77
RF	Angry	0.77	0.84	0.80
	Happy	0.77	0.74	0.76
	Sad	0.77	0.80	0.78
	Fear	0.86	0.79	0.82
MLP	Angry	0.87	0.85	0.86
	Happy	0.84	0.82	0.83
	Sad	0.82	0.84	0.83
	Fear	0.84	0.86	0.85
XGB	Angry	0.81	0.86	0.83
	Happy	0.81	0.75	0.78
	Sad	0.79	0.88	0.83
	Fear	0.90	0.83	0.87
CNN	Angry	0.76	0.90	0.83
	Happy	0.79	0.72	0.75
	Sad	0.77	0.81	0.79
	Fear	0.85	0.73	0.78

Table 17
Emotion wise performance measures for Multilingual Germanic Data

ML model	Emotion	Pr.	Re.	F1
SVM	Angry	0.73	0.83	0.78
	Happy	0.65	0.62	0.63
	Sad	0.71	0.69	0.70
	Fear	0.64	0.67	0.65
RF	Angry	0.70	0.92	0.80
	Happy	0.72	0.38	0.50
	Sad	0.79	0.73	0.76
	Fear	0.62	0.81	0.70
MLP	Angry	0.76	0.86	0.81
	Happy	0.62	0.53	0.57
	Sad	0.81	0.70	0.75
	Fear	0.55	0.65	0.60
XGB	Angry	0.79	0.86	0.83
	Happy	0.63	0.56	0.59
	Sad	0.83	0.68	0.75
	Fear	0.56	0.73	0.63
CNN	Angry	0.74	0.87	0.80
	Happy	0.56	0.48	0.52
	Sad	0.88	0.68	0.76
	Fear	0.61	0.69	0.65

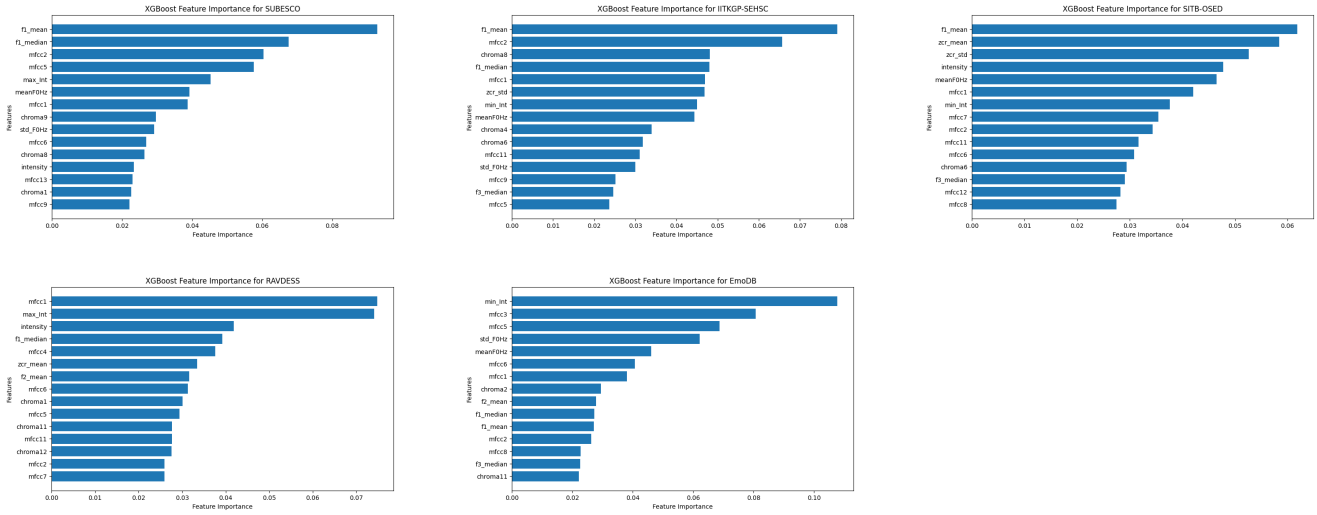

Figure 4: XGBoost feature importance of five different databases

Figure 4 presents the XGBoost feature analysis for the Indo-Aryan and Germanic language datasets. The feature analysis reveals notable similarities among the Indo-Aryan datasets, where f1_mean, mfcc1, and pitch-related features consistently appear among the top contributors, whereas the Germanic datasets demonstrate relatively higher contributions from intensity and MFCC-based features.

6. Discussion

The results of our research indicate that cross-lingual acoustic-phonetic similarities play a vital role in multilingual speech emotion recognition. Comparing Tables 7 to 13, we see that, in most cases the best recognition accuracy was achieved using RF and MLP. From Table 12, cross-lingual speech emotion recognition in Setup 2, we observed that, when models are trained and tested with different languages from the same language family, the emotion recognition accuracy is better than selecting training-testing languages from different families. From the results of Setup 3, we also see that in case of multilingual emotion detection, the models are capable of recognizing emotions of the same family language speech much better than another family language speech. A highest accuracy of 84% and 71% was obtained using multilingual Indo-Aryan (Bangla, Hindi, Odia) and multilingual Germanic (English, German) datasets respectively. On the contrary, while models are trained with multilingual Indo-Aryan and tested with multilingual Germanic, 47% highest accuracy has been obtained and when trained with multilingual Germanic and tested with multilingual Indo-Aryan, 45% highest accuracy has been obtained.

One noteworthy limitation of this study is the relatively small and imbalanced datasets. Expanding the dataset and ensuring more balanced emotion wise utterances for different languages could provide a more comprehensive understanding of cross-lingual emotion recognition. Moreover in this study we have only considered the acoustic-phonetic features and did not take non-speech features into consideration which could have given a more better comprehension. Incorporating non-speech features like facial images, videos, analyzing linguistic features can give better results in multilingual speech emotion recognition.

To improve the generalizability of our findings, it is important to expand the dataset used in our study. In this regard, we will expand the database using more languages incorporating a larger number of speakers for each of the emotional classes. It is believed that, expanding the dataset and ensuring more balanced emotion wise utterances for different languages could provide a more comprehensive understanding of cross-lingual emotion recognition. After that along with the current extracted features, our plan is to include more speech features and find the optimal feature combination for emotion detection of multiple languages. Finally, we have future plan to implement deep learning models like Deep CNN (DCNN) and Residual Network (ResNet), incorporating Principal component analysis (PCA) for multilingual speech emotion recognition using the optimal feature combination.

7. Conclusion

This study investigated cross-lingual acoustic–phonetic similarities among different languages for speech emotion recognition. We conducted experiments using datasets consisting of speech samples from five different languages, Bangla, Hindi, Odia, English, and German of two different language family Indo-Aryan and Germanic and employed state-of-the-art machine learning techniques to analyze the data. Analyzing the features of the different databases, we have found that the languages belonging to the same family share a comparable range of speech characteristic values. To evaluate the acoustic phonetic similarity we have conducted cross-lingual cross-corpus study which revealed that when different language training and testing corpus are from

the same language family, recognition rates are better than selecting testing corpus from a different language family. We have also done multilingual experiment and from the results we observed that the multilingual speech emotion recognition accuracy is higher when multiple languages of the same family is used than that of languages being from different linguistic families. This indicates that the acoustic phonetic similarities among different languages of the same family have good impact on multilingual speech emotion recognition.

Data Availability

The data that support the findings of this study are available from the corresponding authors of the respective data sets upon reasonable request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

References

[1] Al-onazi, B.B., Nauman, M.A., Jahangir, R., Malik, M.M., Alkhammash, E.H., Elshewey, A.M., 2022. Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion. Applied Sciences 12, 9188.

[2] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., et al., 2005. A database of german emotional speech., in: Interspeech, pp. 1517–1520.

[3] Ekman, P., 1992. Are there basic emotions? .

[4] Goel, S., Beigi, H., 2020. Cross lingual cross corpus speech emotion recognition. arXiv preprint arXiv:2003.07996 .

[5] Heracleous, P., Mohammad, Y., Yoneyama, A., 2020. Integrating language and emotion features for multilingual speech emotion recognition, in: International Conference on Human-Computer Interaction, Springer. pp. 187–196.

[6] Heracleous, P., Yoneyama, A., 2019. A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme. PloS one 14, e0220386.

[7] Hozjan, V., Kačič, Z., 2003. Context-independent multilingual emotion recognition from speech signals. International journal of speech technology 6, 311–320.

[8] Kamaruddin, N., Wahab, A., Quek, C., 2012. Cultural dependency analysis for understanding speech emotion. Expert Systems with Applications 39, 5115–5133.

[9] Koolagudi, S.G., Reddy, R., Yadav, J., Rao, K.S., 2011. Iitkgp-sehsc: Hindi speech corpus for emotion analysis, in: 2011 International conference on devices and communications (ICDeCom), IEEE. pp. 1–5.

[10] Lee, S.w., 2019. The generalization effect for multilingual speech emotion recognition across heterogeneous languages, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 5881–5885.

[11] Li, X., Akagi, M., 2019. Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model. Speech Communication 110, 1–12.

[12] Livingstone, S.R., Russo, F.A., 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, e0196391.

[13] Maji, B., Swain, M., 2022. Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and bi-gru features. Electronics 11, 1328.

[14] Monisha, S.T.A., Sultana, S., 2025. A deep learning approach toward analyzing the cross-lingual acoustic-phonetic similarities in multilingual speech emotion recognition. Journal of Electrical and Computer Engineering 2025, 4748790.

[15] Scotti, V., Galati, F., Sbattella, L., Tedesco, R., 2021. Combining deep and unsupervised features for multilingual speech emotion recognition, in: International Conference on Pattern Recognition, Springer. pp. 114–128.

[16] Sultana, S., Iqbal, M.Z., Selim, M.R., Rashid, M.M., Rahman, M.S., 2021a. Bangla speech emotion recognition and cross-lingual study using deep cnn and blstm networks. IEEE Access 10, 564–578.

[17] Sultana, S., Rahman, M.S., 2023. Acoustic feature analysis and optimization for bangla speech emotion recognition. Acoustical Science and Technology 44, 157–166.

[18] Sultana, S., Rahman, M.S., Selim, M.R., Iqbal, M.Z., 2021b. Sust bangla emotional speech corpus (subesco): An audio-only emotional speech corpus for bangla. Plos one 16, e0250173.

[19] Swain, M., Sahoo, S., Routray, A., Kabisatpathy, P., Kundu, J.N., 2015. Study of feature combination using hmm and svm for multilingual odiya speech emotion recognition. International Journal of Speech Technology 18, 387–393.

[20] Wang, C., Ren, Y., Zhang, N., Cui, F., Luo, S., 2022. Speech emotion recognition based on multi-feature and multi-lingual fusion. Multimedia Tools and Applications 81, 4897–4907.

[21] Williamson, J.D., 1978. Speech analyzer for analyzing pitch or frequency perturbations in individual speech pattern to determine the emotional state of the person. US Patent 4,093,821.

[22] Yadav, A., Vishwakarma, D.K., 2020. A multilingual framework of cnn and bi-lstm for emotion classification, in: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE. pp. 1–6.

[23] Yu, Y., 2022. Multilingual speech emotion research based on data mining. Advances in Multimedia 2022.

[24] Zehra, W., Javed, A.R., Jalil, Z., Khan, H.U., Gadekallu, T.R., 2021. Cross corpus multi-lingual speech emotion recognition using ensemble learning. Complex & Intelligent Systems 7, 1845–1854.

Syeda Tamanna Alam Monisha is a Lecturer in the Department of Computer Science and Engineering (CSE), Varendra University, Rajshahi, Bangladesh. She received the B.Sc. (Engg.) and M.Sc. degrees from the Department of CSE, Shahjalal University of Science and Technology (SUST), Sylhet. Her research interests include speech analysis, natural language processing, and artificial intelligence.

Dr. Sadia Sultana is a Professor in the Department of Computer Science and Engineering (CSE), Shahjalal University of Science and Technology (SUST), Sylhet, Bangladesh. She received the B.Sc. (Hons.), M.Sc., and Ph.D. degrees from the Department of CSE, SUST. Her research interests include speech analysis, data science, artificial intelligence, and pattern recognition.