An Urdu speech *corpus* for emotion recognition

Awais Asghar[1,2], Sarmad Sohaib[3], Saman Iftikhar[4,5], Muhammad Shafi[6] and Kiran Fatima[7]

[1] Sino-Pak Center for Artificial Intelligence, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology, Haripur, Pakistan
[2] Department of Electrical Engineering, University of Engineering and Technology, Taxila, Punjab, Pakistan
[3] Department of Electrical and Electronic Engineering, University of Jeddah, Jeddah, Saudi Arabia
[4] Faculty of Computer Studies, Arab Open University, Riyadh, Saudi Arabia
[5] Department of Computing, School of Electrical Engineering and Computer Science, National University of Science and Technology, Islamabad, Pakistan
[6] Faculty of Computing and Information Technology, Sohar University, Sohar, Oman
[7] TAFE, New South Wales, Australia

Corresponding author
Awais Asghar,
awais.asghar@spcai.paf-iast.edu.pk

ABSTRACT

Emotion recognition from acoustic signals plays a vital role in the field of audio and speech processing. Speech interfaces offer humans an informal and comfortable means to communicate with machines. Emotion recognition from speech signals has a variety of applications in the area of human computer interaction (HCI) and human behavior analysis. In this work, we develop the first emotional speech database of the Urdu language. We also develop the system to classify five different emotions: sadness, happiness, neutral, disgust, and anger using different machine learning algorithms. The Mel Frequency Cepstrum Coefficient (MFCC), Linear Prediction Coefficient (LPC), energy, spectral flux, spectral centroid, spectral roll-off, and zero-crossing were used as speech descriptors. The classification tests were performed on the emotional speech *corpus* collected from 20 different subjects. To evaluate the quality of speech emotions, subjective listing tests were conducted. The recognition of correctly classified emotions in the complete Urdu emotional speech *corpus* was 66.5% with K-nearest neighbors. It was found that the disgust emotion has a lower recognition rate as compared to the other emotions. Removing the disgust emotion significantly improves the performance of the classifier to 76.5%.

INTRODUCTION

Emotion recognition is a vital aspect towards complete human-machine interaction since effective communications of information is fundamental to human-machine interaction. Emotion recognition is also a vital part of automatic human behavior analysis such as assessing candidates' suitability for a job, assessing emotional intelligence, and lie detection, *etc*. There are many ways in which machines can recognize emotions such as face recognition, gestures, eye movements, body language, and electrocardiogram (ECG)
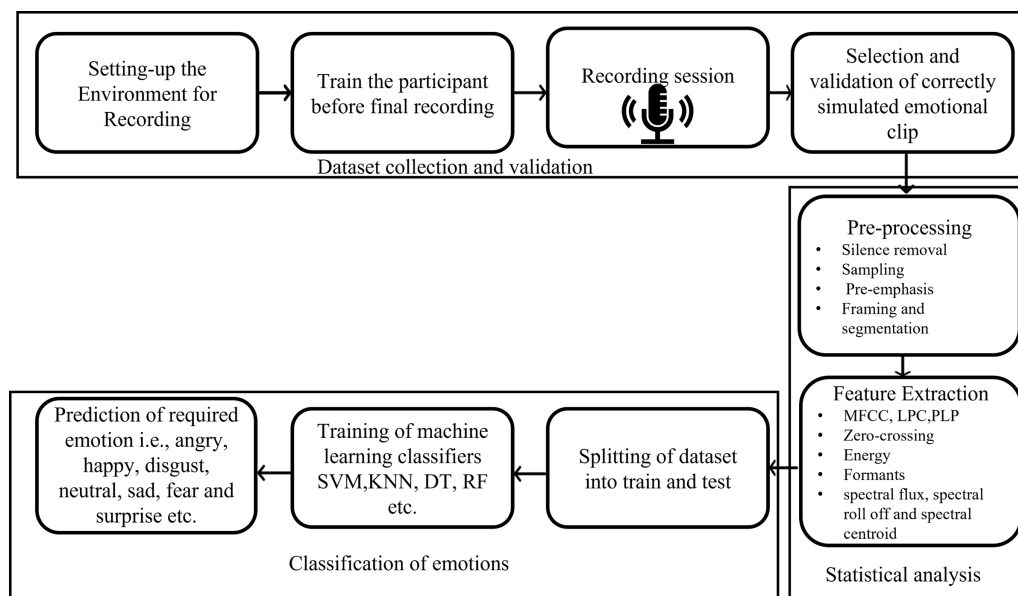
Figure 1 Emotion recognition pipeline.

signals (*Soleymani et al., 2016*). Among all these, speech is an easy and effective form of interaction. Hence, the literature in emotion detection research is focused on the interpretation of emotions from human speech (*Dahake, Shaw & Malathi, 2016*).
There are several applications of emotional understanding such as E-learning where the tutor can change the presentation style when a learner is feeling uninterested, angry, or interested. Similarly, in medical sciences, virtual assessment of the patients' health is possible by listening to his/her voice. In the robot-human communication, the robots can be trained to communicate with human-based emotional states. The cellular services, multimedia devices and call centers have vast area of application related to emotion recognition where devices can detect the human behavior (frustration and annoyance etc.) of end user and react accordingly. Usually, the emotion recognition from the speech is performed by collecting datasets (training, testing and validation), performing statistical analysis (extraction of the features that are associated to the different emotional states), and to classify the emotions from the acoustic signals, as illustrated in Fig. 1 (*Yadav & Aggarwal, 2015*).

Extensive literature is available in the field of human emotions recognition for different languages such as English, French, German, and Malayalam in the last few years (*Ververidis & Kotropoulos, 2003*). For these languages, the developed emotional speech datasets comprises the collection, careful annotation, noise filtering, and validation of speech samples. However, such databases need to be developed for other global languages. The Urdu language has more than 11 million speakers worldwide as a native language and 105 million second-language speakers in the world (*BBC, 2022*). However, speech emotion recognition (SER) from the Urdu language needs further research (*Qasim et al., 2016*; *Kaminska, Sapinski & Anbarjafari, 2017*) and significant improvements such as noise filtering, careful annotation and validation of samples in the development of the

Urdu language emotional dataset. Owing to this lack of consideration in Urdu language dataset collection, Urdu emotional speech database with noise filtering, careful annotation, and sample validation features is realized in this study. The emotion recognition performance is predominantly affected by the pre-processing, feature extraction, and algorithms used to classify the speech into various emotions. In this study, K-nearest neighbour (k-NN), Random Forest (RF), and multiclass Support Vector Machine (SVM) with the linear kernel are used to validate the efficiency of the feature sets.

The remainder of this article is organized as follows. "Background and Related Work" describes the related work and background of research. "Dataset Collection" provides an overview of Urdu emotional speech *corpus* collection, assignments of labels, and Urdu utterances selected for the recording. "Pre-Processing" explores the pre-processing. "Feature Extraction" provides details of feature extraction, and ML algorithms. "Results and Discussions" presents the classification results. Finally, "Conclusion" concludes the paper with future directions.

BACKGROUND AND RELATED WORK

In the field of natural language processing (NLP) and automatic speech recognition (ASR), several speech corpora have been developed for various languages (*Douglas-Cowie et al., 2003*; *Dimitrios Ververidis, 2019*). Many successful proposals have been proposed in the emotion classification for resource rich languages such as Italian (*Giovannella et al., 2009*), Polish (*Staroniewicz & Majewski, 2009*), German (*Grimm, Kroschel & Narayanan, 2008*), English (*Livingstone & Russo, 2018*), and French (*Gournay, Lahaie & Lefebvre, 2018*). However, emotion recognition in the Urdu language is still a target research area and there is a sufficient opportunity for the improvement. Due to the insufficiency of the emotion recognition techniques for the Urdu language, emotion recognition systems for other languages are summarised below, followed by such systems for the Urdu language.

Livingstone & Russo (2018) and *Zhang, Provost & Essi (2016)* presented a multimodal English language emotional speech and song *corpus* in *Livingstone & Russo (2018)*, *Zhang, Provost & Essi (2016)*. The dataset is collected from 24 professional actors by simulating two neutral statements, that is, "Dogs are sitting behind the door" and "Kids are talking by the door". Seven emotions are selected for the speech whereas five for the song, respectively. Every emotion is simulated with two levels of intensity that is strong and neutral. To validate the dataset, 247 untrained individual opinions are taken on each emotion. *Kaminska, Sapinski & Anbarjafari (2017)* developed an emotion recognition framework for the Polish language, where the dataset is recorded in two different forms of emotional speech that is spontaneous and acted speech. Spontaneous speech samples are collected from live TV shows and programs such as news and reality shows. The acted speech samples are recorded from eight native speakers of both genders (four males and four females) where they uttered 240 sentences in six different emotions. The validation of the dataset is endorsed by the subjective listening test. An accuracy of 72% is achieved in emotion recognition. Statistical analysis is also performed to validate the *corpus*. A pool of the features including Perceptual Linear Prediction (PLP), Bark

Frequency Cepstral Coefficient (BFCC), and Human Factor Cepstral Coefficients (HFCC) is used to classify the emotions. The achieved accuracy of this experiment for natural and acted speech is 81% and 60% respectively. *Lyakso et al. (2015)* developed the first emotional speech *corpus* of children in the Russian language and named as the EmoChildRu. It was comprised of audio samples of 120 children simulated in three different emotions including the comfort, discomfort, and neutral. The basic emotions of anger, sadness, and fear are expressed as discomfort. *Leila et al. (2019)* achieved an accuracy of 83% in recognition of seven basic emotions on the German EmoDB database after applying feature selection and speaker normalization techniques. The Mel Frequency Capstrum Cofficient (MFCC) and Modulation Spectral Features (MSFs) methods were used for feature extraction. *Kumar & Iqbal (2019)* and *Khalil et al. (2019)* discussed different classifiers such as k-NN, SVM, convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM) and some feature extraction techniques in *Kumar & Iqbal (2019)*, *Khalil et al. (2019)* and *Zhao, Mao & Chen (2019)*, respectively. *Pengcheng & Zhao (2019)* proposed an emotion recognition system for the Chinese language, where denoising auto-encoder and sparse autoencoder are used for feature extraction whereas the wavelet kernel sparse SVM classifier is used for the classification. *Tripathi & Beigi (2018)* have used RNN with three hidden layers to recognize emotion for the IEMOCAP database with an accuracy of 71.04%. This study used only four emotions that is happiness, sadness, neutral, and anger. *Tang, Zeng & Li (2018)* recognized seven basic emotions from the *corpus* named as emotional sensitivity assistance system for people with disabilities (EmotAsS) (*Simone et al., 2017*) and achieved an accuracy of 45.12% with RNN, CNN and ResNet. *Sarma et al. (2018)* and *Eskimez, Duan & Heinzelman (2018)* used the IEMOCAP dataset for sentiments recognition, where classification is carried out using the LSTM and CNN. An accuracy of 70.06% and 47% is achieved for LSTM and CNN, respectively. *Latif et al. (2018)* presented a cross-lingual recognition system: Urdu *vs* Western language. A recognition accuracy of 83.04% was achieved for the Urdu dataset when other languages are used in training set on four basic emotions. SVM, logistics regression, and random forest are used for classification. *Panagiotis et al. (2017)* proposed a system with RNN and ResNet that gives recognition rates of 78.7% on the French language based remote collaborative and affective (RECOLA) dataset. The details of the RECOLA are explained by *Fabien et al. (2013)*. *Mao et al. (2017)* introduced an Emotion-discriminative and Domain-invariant Feature Learning Method (EDFLM) in *Mao et al. (2017)*. It provided a good emotion recognition rate on the INTERSPEECH 2009 challenge and the Emo-DB database. *Fayek, Lech & Cavedon (2017)* and *Mirsamadi, Barsoum & Zhang (2017)* both use the IEMOCAP dataset with RNN and CNN obtained 64.78% and 63.5% of accuracy, respectively. *Mirsamadi, Barsoum & Zhang (2017)* used both Low-Level Descriptors (LLDs) and High-Level Statistical Functions (HSFs) as input to SVM in order to differentiate emotions. *Rajisha, Sunija & Riyas (2016)* performed analysis on the Malayalam language to differentiate different sentiments. MFCC, energy, and pitch are used for features extraction. The four basic emotions (happiness, sadness, neutral, and anger) are classified by SVM and artificial neural network (ANN). *Yadav & Aggarwal (2015)* achieved an 85% accuracy to recognize

four emotions with ANN. *Sinith et al. (2015)* tested the SVM with two classification strategies that is one against one, and one against all in *Sinith et al. (2015)*. The SVM gives a higher performance on Berlin emotional database as compared to Malayalam emotional database with a feature set of MFCC, energy, and pitch. *Abbas, Khan & Bashir (2015)* performed a classification of emotions for Urdu language (*Abbas, Khan & Bashir, 2015*) where J48 and Decision tree are tested, achieving an accuracy of 48% with four basic emotions. *Fayek, Lech & Cavedon (2015)* achieved an emotions recognition rate on eINTERFACE and SAVEE database in *Fayek, Lech & Cavedon (2015)* which was 60.53% and 59.7%, respectively. The Polish language emotion speech dataset obtained 70% accuracy with k-NN.

Table 1 presents a summary of the emotion recognition techniques from the literature. *Rauf et al. (2015)* proposed a speaker-independent Urdu language speech recognition system where the dataset comprises the utterances for district names of Pakistan. A total of 139 district names are recognized in major Urdu language accents such as Punjabi, Sindhi, Balouchi, and Pashto. *Ali et al. (2013)* presented an Emotions-Pak *corpus*, where only one utterance "In seven hours it will happen" is recorded in Urdu and other provincial languages of Pakistan. In this *corpus*, four emotions are obtained in a given sentence. To evaluate the performance of recorded emotions, results from the prosodic feature set and subjective listening were compared. *Andleeb, Haider & Abbas (2017)* performed the classification of the special and normal children's speech emotions in Urdu language. A total of 11 different feature extraction techniques including MFCC, Linear Prediction Coefficient (LPC), and PLP are used to classify the special and normal children's speech. The dataset was recorded using 200 special and 200 normal children in four different emotions on the selected utterance "I have to play" in Urdu. *Abbas, Zehra & Arif (2013)* presented a system that recognized the emotions in the provisional languages of Pakistan, where only one utterance was simulated in Pakistani languages for four basic emotions. The achieved accuracy was 75% where Multi-layer Perceptron (MLP), and Naive Bayes were used as classifiers.

DATASET COLLECTION

Our emotional speech *corpus* comprises 2,500 emotion samples of Urdu speech. There are 20 speakers of both genders (10 males and 10 females) aging between 20 to 40 years. Each speaker utters five times. Every time a speaker utters five different Urdu utterances in five different emotions such as happy, sad, angry, disgust, and neutral. The selected utterances are everyday human-human interaction utterances and easy to understand in all five emotions. The utterances were recorded in the university lab using the Blue Yeti desktop microphone as recording equipment. After collection, the recorded emotional speech utterances were listened by a psychologist and a group of students (10–15) to verify the originality of simulated emotions. The speech utterances which were repeatedly mismatched with the assigned labels were discarded from the emotional *corpus*. A large number of samples were discarded from the disgust emotion which was also highlighted in the Results and Discussion sections. For this reason, the samples per emotion were not balance. The fully filtered emotional speech dataset was then fed to the emotion

Table 1 Summary of literature on emotion recognition from different languages.

Papers with year	Dataset used	Emotions recognized	Technique used	Achieved accuracy
Leila et al. (2019)	Berlin EmoDB	Anger, disgust, fear, joy, neutral, surprise and sadness	SVM and multivariate linear regression (MLR)	83%
Kumar & Iqbal (2019)	EmoDB dataset	Neutral anger and sad	Deep belief network (DBN) and Stacked encoder	65%
Pengcheng & Zhao (2019)	Chinese emotional speech dataset	Anger, scared, happiness, sadness, neutral and surprise	Wavelet-kernel sparse SVM	80.95%
Tripathi & Beigi (2018)	IEMOCAP dataset	Anger, happiness, sadness, and neutral	RNN with 3 layers	71.04%
Tang, Zeng & Li (2018)	EmotAss dataset	Anger, happiness, neutral and sadness	CNN and RNN with ResNEt	45.12%
Sarma et al. (2018)	IEMOCAP dataset	Anger, happiness, neutral, sadness, surprise, fear and disgust	LSTM	70.6%
Eskimez, Duan & Heinzelman (2018)	IEMOCAP dataset	Neutral, sadness, frustration and anger	CNN	47%
Latif et al. (2018)	Urdu language emotional speech dataset	Anger, happiness, neutral and sadness	SVM, logistic regression and random Forest	83.4%
Panagiotis et al. (2017)	Spontaneous emotional RECOLA and AVEC dataset	Happiness, sadness, anger and neutral	CNN and ResNEt	78.7%
Mao et al. (2017)	INTERSPEECH 2009, ABC and EmoDB	Happiness, sadness, neutral, fear, surprise, disgust and anger	Emotion discriminative and domain invariant feature learning method (EDFLM)	65.62%
Fayek, Lech & Cavedon (2017)	IEMOCAP dataset	Neutral, happiness, sadness, anger and silence	RNN and CNN	64.78%
Kaminska, Sapinski & Anbarjafari (2017)	Acted and spontaneous Polish language dataset	Sadness, happiness, anger, neutral, joy, fear, and surprise	SVM and k-NN	81%
Mirsamadi, Barsoum & Zhang (2017)	IEMOCAP dataset	Neutral, anger, sadness, and happiness	Recurrent Neural Network RNN and SVM	63.5%
Zhang et al. (2017)	Chinese emotional speech dataset	Sadness, joy, anger, neutral fear, and surprise	SVM and Deep learning	84.54%
Zhu et al. (2017)	Chinese emotional speech dataset	Sadness, joy, anger, neutral fear, and surprise	Combination of SVM and Deep learning	95.8%
Rajisha, Sunija & Riyas (2016)	Malayalam language emotional speech dataset	Neutral, anger, happiness and sad	ANN and SVM	78.2%
Yadav & Aggarwal (2015)	English emotion speech dataset	Sadness, happiness, anger and neutral	Artificial Neural Network ANN	85%
Qasim et al. (2016)	District name of Pakistan dataset		SVM and GMM	71%
Sinith et al. (2015)	SAVEE and Malayalam emotional speech dataset	Anger, happiness, neutral and sadness	Support vector machine	75%

Papers with year	Dataset used	Emotions recognized	Technique used	Achieved accuracy
Abbas, Khan & Bashir (2015)	Urdu language emotional speech dataset	Anger, sadness, happiness and neutral	Decision tree and J48	40%
Fayek, Lech & Cavedon (2015)	ENTERFACE and SAVEE dataset	Boredom, disgust, sadness, joy, anger and neutral	Deep neural network DNN	60.53%
Abbas, Zehra & Arif (2013)	Provisional language of Pakistan emotional speech dataset	Comfort, happiness, sadness and neutral	Multilayer perceptron (MLP), Naïve Bayes and SMO	75%
Kami´nska & Pelikant (2012)	Polish emotional speech dataset	Sadness, happiness, anger and neutral	k-NN	70%

Selection of speakers
• 20 speakers (10 males, 10 females) hired for recording

Setting-up environment
• Instruction regarding Urdu utterances and emotion elicitation.
• Setting-up of microphone and background noise.

Recording session
• Speaker ask to elicit selected utterances into emotions.
• Requested to repeat it.
• Instant feedback on simulated emotions.

Selection of speech clips
• Psychologists are asked to check the originality of simulated emotion, choose correctly elicited clips .

Initial processing of audio clips
The recorded clips are edited
• Silence part
• Peaks are normalized
• Each clip is saved in (.wav) audio format.

Validation of dataset

Figure 2 Flow chart of Urdu emotional speech dataset creation and validation.

recognition system. The complete process of the emotional speech dataset is outlined in Fig. 2.

Description of audio speech clips

The Urdu emotional speech dataset contains a total of 2,500 audio clips that was simulated by 20 speakers of both genders. Each speaker uttered 125 emotional speech clips that include five emotional states that were angry, happy, neutral, disgust, and sad on five commonly used Urdu language utterances. The full constructed data recording

Table 2 Number of emotions per sample.

Emotion	Number of samples
Angry	500
Disgust	400
Happy	500
Neutral	450
Sad	450

includes the number of clips per speaker = angry (5) × utterances (5) × repetition (5) = 125; for 20 speakers, the total number of audio clips became 125 × 20 = 2,500. In the validation stage, 200 samples, which were not correctly uttered, were filtered out. The distribution of remaining 2,300 audio clips/emotional speech samples is provided in Table 2.

Recording environment

The utterances were recorded in a noise-free lab room in absence of the background noise to achieve good quality. The speakers were asked to sit in front of a microphone, and they may move their bodies freely to express a particular emotion. Further, the speakers were asked to speak in the direction of a microphone to capture the full intensity of voice. The distance between the subject and recording equipment is kept at 25 cm.

Acted or real emotion

A fully developed emotion appears occasionally in the real-life. From the real-life speech samples, it is almost impossible to differentiate between some basic emotions (*Burkhardt et al., 2005*). Hence the literature prefers the acted emotions. There are a few factors to be considered while collecting acted speech. (I): All speakers should act the same verbal content in order to allow the comparability across emotions and speakers. (II): The quality of the recorded voice assumed to be good enough, minimizing background noise; otherwise spectral measurements would not be possible. (III) a reasonable number of speakers should perform all emotions to obtain generalization over the target emotions.

Choice of emotions and speakers

To compare the selection of emotions with early research (*Yadav & Aggarwal, 2015*; *Giovannella et al., 2009*; *Grimm, Kroschel & Narayanan, 2008*), the same emotions were used, such as: happy, sad, angry, disgust, and neutral. These emotions attract more attention and used in the human daily life. These selected emotions are easy to understand by the speakers as well as the listeners. It is important to note that we have not involved trained actors in performing emotional expression. All the speakers were students and faculty members of the department. However, the speakers were aware and trained before the actual recording of the emotions.

Text material

The utterances used were easy to understand in the emotions, that is, there were no emotional biases involved. The literature suggested two types of text materials that can

Table 3 Chosen Urdu language utterances with English translation.

Sentences in Urdu	English translation
Pakistan kesa hai?	How is Pakistan?
qareeb tareen hospital kahan hai?	Where is the nearest hospital?
kapre fridg pr parey hein	The cloths are lying on the fridge.
tum kahan gaye they?	Where did you go?
kahan ho ajj kal?	Where are you nowadays?

Table 4 Recognition rate of each emotion during validation process.

Emotion	Recognition rate
Angry	96%
Sadness	94%
Neutral	92%
Happy	80%
Disgust	76%

ensure such requirements (*Costantini et al., 2014*), (I): the text material that was emotionally neutral, and (II): normal sentences which are used in everyday life. In the preparation of the database, priority was given to the neutrality of speech material, and thus everyday sentences were used as test utterances. Five sentences were chosen which could easily be interpretable in the above-mentioned emotions. These sentences are given in Table 3.

Recording of data

There was only one session of recording per day with three speakers. All the recordings were completed under the supervision of psychologist and experts, and their opinions on the emotion were also recorded. The collected speech samples were normalized and stored in ".wav" format with sampling frequency 44.1 kHz, and 16 bits per sample. A Blue Yeti desktop microphone was used to record the speech samples. The utterances were recorded in a noise-free lab room in absence of the background noise to achieve the good quality (*Gournay, Lahaie & Lefebvre, 2018*).

Database validation

Based on the opinions of experts and psychologists during the collection stage, the utterances were extracted and initially classified into one of the five discrete emotion categories including happiness, sadness, anger, disgust, and neutral state. A psychologist was asked to listen carefully the randomly presented audio files and indicate which of the emotion is available in the presented files. The psychologist was not allowed to go back to previously presented emotion. Another labelling exercise was carried out where 10 to 15 students were included in the tests. Every student was presented with the acted emotions (.wav audio files) to make a decision about the simulated emotions and check the performance of speakers. Therefore, the speech samples which repeatedly mismatched
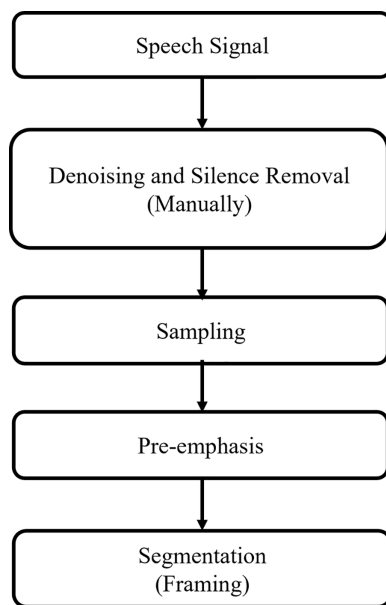
Figure 3 Pre-processing flow of speech signal.

with the labels were discarded from the emotional *corpus*. The fully filtered emotional speech dataset was then fed to the developed emotion recognition system. The recognition rate of each emotion is shown in Table 4.

PRE-PROCESSING

In the emotion recognition system, there can be silence parts and background noise in the spoken utterances. Therefore, the emotional speech signals recordings from the microphone are first pre-processed and made them suitable and noise-free for feature extraction stage. In this study, silence parts and background noise are removed manually. Figure 3 demonstrates the pre-processing steps which are discussed in the subsections.

Pre-emphasis

The high-frequencies were suppressed during the sound production by humans. Therefore, pre-emphasis was applied on the sampled signal to increase the magnitude of higher frequencies, thereby improving the overall signal-to-noise ratio (SNR). The pre-emphasis was implemented as a first order Finite Impulse Response (FIR) filter which is defined as:

$$y(n) = x(n) - a\,x(n-1), \tag{1}$$

where $y(n)$ is the emphasized signal, $x(n)$ is sampled signal and a is the pre-emphasis coefficient, with value raging from 0.9 to 1.0.

Framing

Speech signal is non-stationary by nature and the spectral analysis usually considers the stationary signals. Therefore, framing was used to convert the non-stationary speech signals into stationary signals. During the framing, the speech signal was divided into a

series of the overlapping frames. The frame length was 20 to 30 ms with an overlap of 1/3 of the frame size. Overlapping was used to avoid loss of data due to aliasing.

Hamming window

The sudden change at the onset and offset of frame causes loss of important information. Therefore, Hamming windowing function was applied to all frames. If w(n) is the Hamming window function and y(n) is the input signal frame, then output z(n) is given by equation as:

$$Z(n) = y(n)w(n), \tag{2}$$

where

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi}{N-1}\right), \tag{3}$$

N is number of samples in a frame and z(n) is a final pre-processed signal.

FEATURE EXTRACTION

After all pre-processing, the signal is appropriate for feature extraction. Various statistical values were used in our model to discriminate emotion classes. These statistical values are in the form of vectors known as feature vectors. These feature vectors provide a higher level of representations of audio samples. The extracted features in this study are explained below.

Spectral flux

It is a one-dimensional feature vector against one audio sample. It is a measure of how rapidly the power spectrum of a speech signal varies and is calculated by comparing the power spectrum of two successive frames and computed as the squared difference between the standardized magnitudes of spectra of two consecutive short-term windows and is given by *Alías, Socoró & Sevillano (2016)*

$$\text{Spectral Flux} = (|z(n)| - |z(n-1)|)^2. \tag{4}$$

It is also known as the Euclidean distance among the two standardized spectra.

Spectral centroid

The spectral centroid shows where the centre of gravity of the spectrum of the audio signal is located (*Kamarudin et al., 2014*). It is obtained by taking a weighted average of the frequency components present in the signal. The weighted average is determined by taking Fourier transform of frequencies and their magnitude as weights and calculated as:

$$\text{Spectral Centroid} = \frac{\sum_{n=0}^{N-1} n\, z_t(n)}{\sum_{n=0}^{N-1} z_t(n)}, \tag{5}$$

where $Z_t(n)$ is the magnitude of Fourier transform at frame t and frequency bin n.

Spectral roll off

Spectral roll-off is a feature that is defined as the frequency under which 85% of the signal's spectral energy is accumulated. This measurement gives the centre of mass of energy (higher frequencies) in the spectrum (*Kaur & Kumar, 2017*).

Zero crossing

Zero crossing is a method to classify the voice and non-voice parts of the signal. It is the rate at which speech signals passes through zero level (*Toledo-Pérez, Rodríguez-Reséndiz & Gómez-Loenzo, 2020*). Zero crossing for the signal can be calculated as

$$\text{Zero} - \text{crossing} = \frac{1}{N} \sum_{n=0}^{N} |z(n)| - |z(n-1)|. \tag{6}$$

Energy

Energy is a very basic and fundamental feature in signal processing (*Li & Sun, 2008*). Energy of speech signal is referred to an intensity of a signal and is calculated as

$$\text{Energy} = \frac{1}{N} \sum_{n=0}^{N} |z(n)^2|, \tag{7}$$

For example, energy of the happy and angry is different from sad and neutral.

Linear prediction coefficient

The LPC model describes the vocal tract of the humans. In LPC, each sample of the speech signal is expressed as a linear combination of the earlier samples. These coefficients are highly effective representation of the speech signal (*Alim & Rashid, 2018*; *Dave, 2013*). In this analysis, each speech sample is represented by a weighted sum of past speech samples plus an appropriate excitation. The corresponding expression for the LPC model is given as:

$$S_n = \sum_{k-1}^{p} a(k)\, z(n-k) + e(n), \tag{8}$$

where p is the order of LPC, $a(k)$ is the kth coefficient of LPC vector, $z(n-k)$ is the n_{th} speech sample and $e(n)$ is the prediction error. The coefficients $a(k)$ are computed by minimizing the sum of squared differences between the actual speech samples and the linearly predicted ones.

Mel frequency capstrum coefficient

MFCC are the commonly used features in speech recognition systems. It is a short-term power spectrum of an audio signal, which is based on the inverse fast Fourier transform (IFFT) of a log power spectrum on a nonlinear Mel scale of frequency. The Mel scale is a perceived pitch or frequency that is heard by the listener to be equal in distance from one another. Human ear can easily understand the difference between pitch changes at low frequency as compared to high frequency. The incorporation of this scale makes our

Table 5 Feature dimensions.

Features name	Features dimensions
MFCC	13
Mean of MFCC	13
Standard deviation of MFCC	13
LPC	10
Mean of LPC	10
Spectral flux	01
Spectral centroid	01
Spectral rolloff	01
Zero crossing	01
Energy	01
Total feature vector	64

feature vector more closely related to the human hearing system (*Alim & Rashid, 2018*; *Dave, 2013*). Mel scale frequency can be expressed as:

$$f_{mel} = 1125 \ln\left(1 + \frac{f}{700}\right), \tag{9}$$

where f is a linear frequency and f_{mel} is perceived frequency of speech signal. To move back to linear frequency scale from Mel scale perceived frequency we use

$$f = 700\left(e^{\frac{f_{mel}}{1125}} - 1\right) \tag{10}$$

MFCC is implemented using the following steps.

1. Segmented the time-domain speech signal.
2. For each segment, the periodogram estimate of discrete Fourier transformed (DFT) segments is calculated.
3. Applied the Mel scale filter bank on power spectrum, and sum-up the energy for each filter bank.
4. Take the log of Mel scaled energies.
5. Applied the discrete cosine transform (DCT) on a log Mel scaled energies.
6. Keep the first 13 DCT coefficients.

For one audio sample, the total feature vector size is 1×64 as summarized in the Table 5.

RESULTS AND DISCUSSIONS

There are five main blocks in a speech emotion recognition system, that is, emotional speech input, pre-processing, feature extraction, assignment of labels, and classification of the emotions. The complete emotion recognition system is demonstrated in Fig. 4. After feature extraction, each speech sample results in statistical values against every emotion: angry, happy, sad, neutral, and disgust. Each emotion in a speech sample has a unique
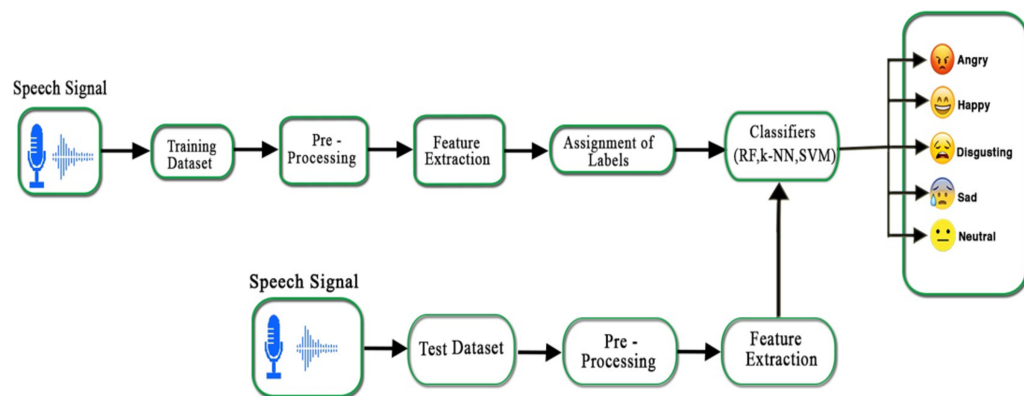
Figure 4 Proposed emotion recognition system for Urdu speech signals.

intensity, pitch, zero-crossing rate, and spectral feature. It is important to classify the emotions from the aforementioned feature vectors.

In this study, we have used three classifiers, that is, SVM, k-NN, and RF to train and test our Urdu speech emotional dataset. The multi-class problem in the SVM is also solved by using one-against-one and one-against-all SVM strategies (*Hassan & Damper, 2010*). These heuristic methods are used to split a multi-class classification problem into multiple binary classification datasets and train a binary classification model on each. The performance of one-against-rest SVM is measured as an average of all binary classifier accuracies. The Urdu speech database is divided into two sets, the training and testing sets, where the training set contains 70% and the testing set contains 30% of the whole dataset. Both sets (training and testing) carry information of each speaker's emotion. During the model training, feature vectors of the training set along with their labels were given to the classifier whereas in testing, the feature vector of the unclassified sample is given to the model. The performance of classifiers was measured on the test data using accuracy, precision, and recall measures.

Finally, the performance of each classifier was compared for each emotion. Our Urdu speech dataset contains five utterances that are simulated in five different emotions *i.e.*, happy, sad, angry, neutral, and disgust. It was observed that 'disgust' is difficult to recognize as compared to the others. It had adverse effects on classification accuracy, while the physiologist also struggled to recognize the disgust emotion. Thus, we divided our data set into two subsets, one with disgust and another without disgust emotion. The classification was implemented in six different ways *i.e.*, females, males, and a complete dataset is subdivided into with and without disgust emotion. In the classification, the emotions angry are labeled as "A", disgust as "D", happy as "H", neutral as "N", and sad as "S".

Table 6 shows the classifiers performance summary with disgust emotion where it can be seen that the k-NN performs better for male and complete datasets. One-*vs*-rest classifier performance is better in the case of the female dataset. Table 7 shows the

Table 6 Comparison of performance of classification algorithms on emotional speech *corpus* with disgust emotions.

ML techniques	For male only			For female only			Complete dataset		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
One-*vs*-rest	69.5%	71%	69%	68.4%	71%	68%	60.6%	62%	61%
One-*vs*-one	70%	71%	70%	65.6%	67%	66%	62.2%	64%	62%
k-NN	73%	73%	72%	66.4%	69%	66%	66.2%	67%	66%
Random Forest	66.5%	67%	66%	58.8%	62%	59%	60.8%	64%	61%

Table 7 Comparison of performance of classification algorithms on emotional speech *corpus* without disgust emotions.

ML techniques	For male only			For female only			Complete dataset		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
One-*vs*-rest	75%	75%	74%	78.5%	81%	79%	70.2%	72%	70%
One-*vs*-one	79.5%	80%	79%	77.5%	78%	79%	70.7%	72%	71%
k-NN	82.5%	84%	83%	76%	76%	76%	76.5%	77%	77%
Random Forest	74%	74%	73%	71%	72%	71%	71.5%	73%	71%

Table 8 Comparison with related work.

Papers	Languages	Training technique	Features extraction techniques	Emotions	Classifier used	Accuracy
Tripathi & Beigi (2018)	English and German	Speaker dependent	RNN	Anger, happiness, neutral and sadness	RNN with three layers	71.04%
Kaminska, Sapinski & Anbarjafari (2017)	Polish	Speaker dependent independent	MFCC, BFCC, RASTA, energy, formants, LPC and HFCC	Sadness, happiness, anger, neutral, joy, fear and surprise	SVM and k-NN	81%
Rajisha, Sunija & Riyas (2016)	Malayalam	Speaker dependent	MFCC, STE and pitch	Neutral, anger, happiness and sad	ANN and SVM	78%
Ali et al. (2013)	Urdu	Speaker dependent	Duration, intensity, pitch and formants	Anger, sadness, happiness and comfort	Neive Bayes	76%
Abbas, Zehra & Arif (2013)	Urdu	Speaker dependent	Intensity, pitch and formants	Anger, sadness, happiness and comfort	SMO, MLP, J48 and Neive Bayes	75%
Latif et al. (2018)	Urdu	Speaker independent	LLDs low level descriptor	Happiness, sadness, anger and neutral	SVM, logistic regression and RF	83%
Sinith et al. (2015)	English Malayalam and	Speaker dependent	MFCC, pitch and energy	Anger, neutral sadness and happiness	SVM	70%
Our work	Urdu (with disgust emotion)	Speaker dependent	MFCC, LPC, energy, pitch, zero crossing, spectral flux spectral centroid, spectral roll off	Anger, disgust, happiness, sadness and neutral	k-Nearest Neighbours	73%
Our work	Urdu (without disgust emotion)	Speaker dependent	MFCC, LPC, energy, pitch, zero crossing, spectral flux spectral centroid, spectral roll off	Anger, happiness, sadness and neutral	k-Nearest Neighbors	82 .5%
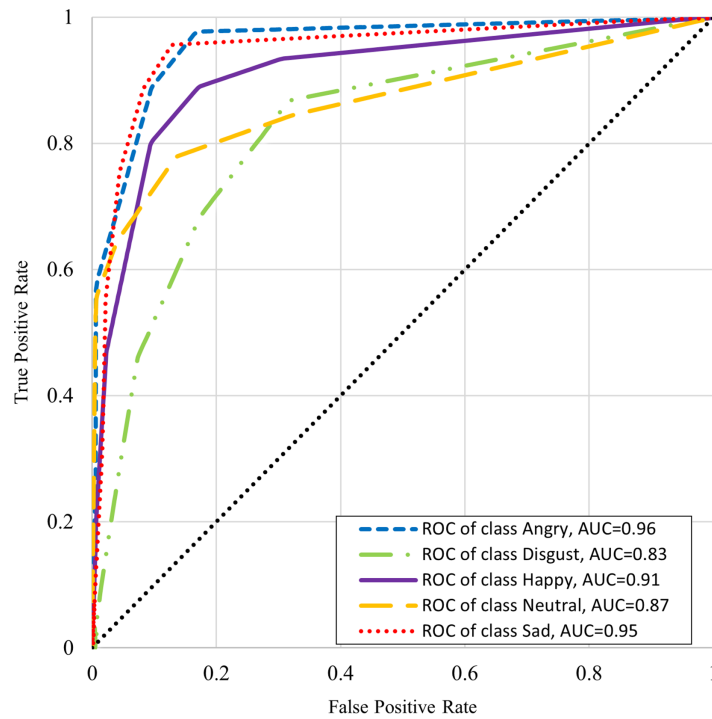
Figure 5 ROC curve of K-NN with disgust emotion.   Full-size DOI: 10.7717/peerj-cs.954/fig-5
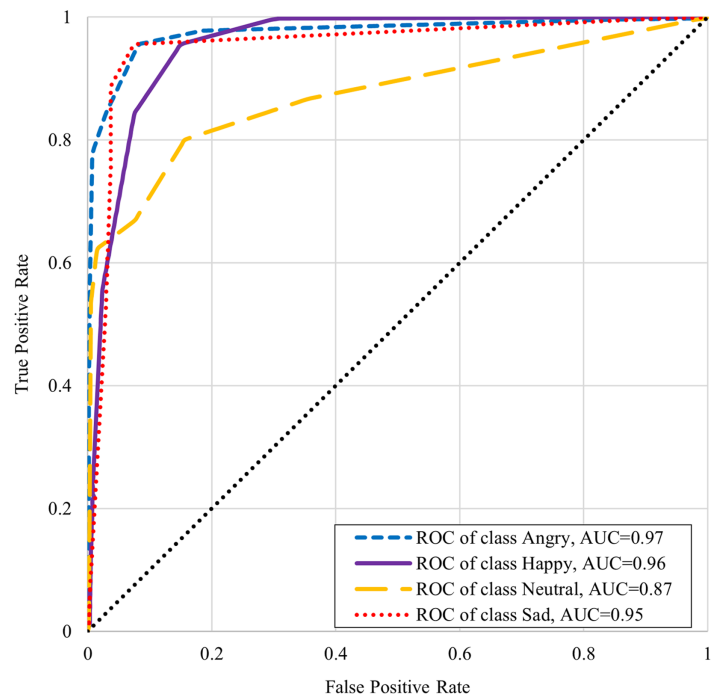
Figure 6 ROC curve of K-NN without disgust emotion.
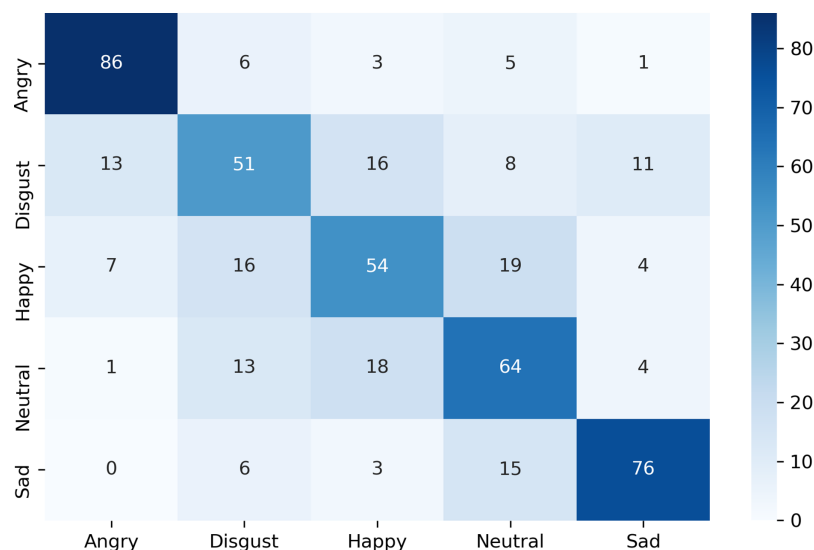
Figure 7 Confusion matrix of k-NN for complete dataset including disgust.

classifiers performance without disgust emotion dataset. It can be observed that the k-NN performs the best for the male and complete dataset here too, whereas onevs- rest classifier performs better in the case of the female dataset in this scenario. The comparison with state-of-the art from literature is presented in Table 8. It is worthwhile to mention here that although one of the benchmarked studies has reported slightly higher accuracy, our work's scope is wide in terms of the number of emotions (with five emotions as compared to four emotions) and the size of the dataset (2,500 samples as compared to 400 samples). The receiver operating characteristic (ROC) curve differentiates between the true positive rate or truly classified samples in opposition to the false positive rate or not truly classified samples. A good classification technique has an upside-down "L" shape curve while others follow diagonals. Figures 5 and 6 show the ROC and area under the curve (AUC) for every emotional state *i.e.* angry, happy, disgust, neutral, and sad. These graphs show that AUC of disgust emotion is less as compared to the rest of emotions. Figure 6 shows that the AUCs of the dataset without disgust emotion are much improved as compared to a dataset with disgust emotion. It is concluded that disgust emotion is difficult to recognize than the rest of the emotions.

The confusion matrix of the complete dataset with and without disgust emotion is shown in Figs. 7 and 8 respectively, where actual and predicted emotions are listed on vertical and horizontal axis, respectively. As can be seen from Fig. 7, the disgust emotion is the most wrongly predicted class which results in reduction of system accuracy. The confusion matrix without the disgust emotion in Fig. 8 shows a reduction in misclassification of the emotion which thereby results in enhanced accuracy of the system.
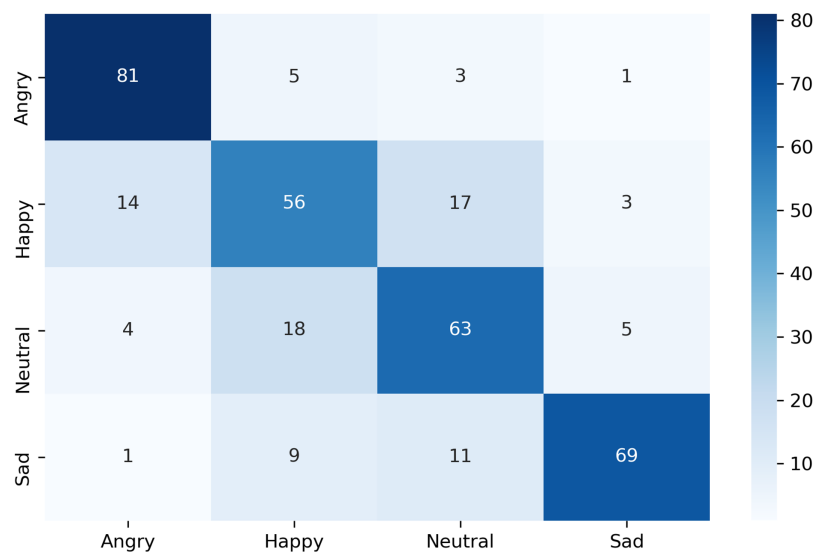
Figure 8 Confusion matrix of k-NN for complete dataset without disgust.

CONCLUSION

This study presented the design and development of emotional speech *corpus* for the Urdu language. For the development of this *corpus*, five sentences in the Urdu language were simulated in five different emotions, that is, happy, sad, angry, disgust, and neutral. The recognition of emotions from Urdu speech signals using different machine learning techniques was carried out. The Urdu emotional speech data of opposite genders obtains different recognition rates. Different feature sets were studied for better classification of emotions, and only those features were adopted that show a good description of the speech signals. The experimental results showed that males have distinct emotions as compared to the female emotions. There was a large difference in the model performance with disgust and without disgust emotion. The maximum overall recognition accuracy achieved with disgust emotion was 72.5% with k-NN, 68.5% with one-against-rest classifier, and 66.2% on k-NN for male, female, and the complete dataset, respectively. For the dataset without disgust emotion, maximum overall recognition accuracy was 82.5% with k-NN, 78.5% with one-against-rest classifier, and the 76.5% on k-NN for male, female, and the complete dataset respectively.

This study could potentially play a vital role in the automatic human behavior analysis for Urdu speakers. Some of the use cases of the proposed study in human behavior analysis are assessing candidates' suitability for a job, assessing emotional intelligence, lie detection, etc. In future, we are devoted to developing a more robust Urdu dataset with more emotions and human behaviors.

Funding
The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Awais Asghar conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Sarmad Sohaib conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, and approved the final draft.
- Saman Iftikhar analyzed the data, prepared figures and/or tables, and approved the final draft.
- Muhammad Shafi analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Kiran Fatima performed the computation work, authored or reviewed drafts of the paper, proof reading and final touch to the paper, and approved the final draft.

Data Availability

The raw data are available in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.954#supplemental-information.

REFERENCES

Abbas SA, Khan A, Bashir N. 2015. Analyzing the impact of prosodic feature (pitch) on learning classifiers for speech emotion corpus. *International Journal of Information Technology and Computer Science* **2**:54–59 DOI 10.5815/ijitcs.2015.02.07.

Abbas AS, Zehra S, Arif A. 2013. Performance evaluation of learning classifiers for speech emotions corpus using combinations of prosodic features. *International Journal of Computer Applications* **76(2)**:35–43 DOI 10.5120/13221-0634.

Ali SA, Zehra S, Khan M, Wahab F. 2013. Development and analysis of speech emotion corpus using prosodic features for cross linguistics. *International Journal of Scientific and Engineering Research* **4(1)**:1–8.

Alías F, Socoró JC, Sevillano X. 2016. A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. *Applied Sciences* **6(5)**:143 DOI 10.3390/app6050143.

Alim SA, Rashid NKA. 2018. Some commonly used speech feature extraction algorithms. In: *From Natural to Artificial Intelligence-Algorithms and Applications*. London: IntechOpen DOI 10.5772/intechopen.80419.

Andleeb M, Haider N, Abbas S. 2017. A novel approach for features extraction towards classifying normal and special children speech emotions in Urdu. *International Journal of Computer Science and Network Security* **17(7)**:188.

BBC. 2022. Languages Urdu: A Guide to Urdu 10 facts about the Urdu language. *Available at http://www.bbc.co.uk/languages/other/urdu/guide/facts.shtml*.

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. 2005. A database of German emotional speech. In: *Ninth European Conference on Speech Communication and Technology*.

Costantini G, Iaderola I, Paoloni A, Todisco M. 2014. EMOVO corpus: an Italian emotional speech database. In: *International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA), 3501–3504.

Dahake P, Shaw K, Malathi P. 2016. Speaker dependent speech emotion recognition using MFCC and support vector machine. In: *International Conference on Automatic Control and Dynamic Optimization Techniques*. IEEE, 1080–1084.

Dave N. 2013. Feature extraction methods LPC, PLP and MFCC in speech recognition. *International Journal for Advance Research in Engineering and Technology* **1(6)**:1–4.

Dimitrios Ververidis C. 2019. A state-of-the-art review on emotional speech databases. *Available at Citeseerx.ist.psu.edu*.

Douglas-Cowie E, Campbell N, Cowie R, Roach P. 2003. Emotional speech: towards a new generation of databases. *Speech Communication* **40(1–2)**:33–60 DOI 10.1016/S0167-6393(02)00070-5.

Eskimez SE, Duan Z, Heinzelman W. 2018. Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5099–5103.

Fabien R, Sonderegger A, Sauer J, Lalanne D. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Piscataway: IEEE, 1–8.

Fayek HM, Lech M, Cavedon L. 2015. Towards real-time speech emotion recognition using deep neural networks. In: *2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS)*. Piscataway: IEEE, 1–5.

Fayek H, Lech M, Cavedon L. 2017. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks* **92**:60–68 DOI 10.1016/j.neunet.2017.02.013.

Giovannella C, Conflitti D, Santoboni R, Paoloni A. 2009. Transmission of vocal emotion: do we have to care about the listener? The case of the Italian speech corpus EMOVO. In: *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*.

Gournay P, Lahaie O, Lefebvre R. 2018. A canadian french emotional speech dataset. In: *Proceedings of the 9th ACM Multimedia Systems Conference on-MMSys 18*.

Gournay P, Lahaie O, Lefebvre R. 2018. A Canadian French emotional speech dataset. In: *Proceedings of the 9th ACM Multimedia Systems Conference*. 399–402.

Grimm M, Kroschel K, Narayanan S. 2008. The Vera am Mittag German audio-visual emotional speech database. In: *IEEE International Conference on Multimedia and Expo*. Piscataway: IEEE.

Hassan A, Damper RI. 2010. Multi-class and hierarchical SVMs for emotion recognition. In: *Interspeech*.

Kamarudin N, Al-Haddad SAR, Hashim SJ, Nematollahi MA. 2014. Feature extraction using spectral centroid and Mel frequency cepstral coefficient for Quranic accent automatic identification. In: *2014 IEEE Student Conference on Research and Development*. Piscataway: IEEE, 1–6.

Kaminska D, Sapinski T, Anbarjafari G. 2017. Efficiency of chosen speech descriptors in relation to emotion recognition. *EURASIP Journal on Audio,Speech, and Music Processing* **2017(1)**:165 DOI 10.1186/s13636-017-0100-x.

Kami´nska D, Pelikant A. 2012. Recognition of human emotion from a speech signal based on Plutchik's model. *International Journal of Electronics and Telecommunications* **58(2)**:165–170 DOI 10.2478/v10177-012-0024-4.

Kaur CP, Kumar R. 2017. Study and analysis of feature based automatic music genre classification using Gaussian mixture model. In: *2017 International Conference on Inventive Computing and Informatics (ICICI)*. IEEE, 465–468.

Khalil RA, Jones E, Babar MI, Jan T, Zafar MH, Alhussain T. 2019. Speech emotion recognition using deep learning techniques: a review. *IEEE Access* **7**:117327–117345 DOI 10.1109/ACCESS.2019.2936124.

Kumar AK, Iqbal MLJ. 2019. Machine learning based emotion recognition using speech signal. *International Journal of Engineering and Advanced Technology.* **9(1S5)**:295–301 DOI 10.35940/ijeat.a1068.1291s52019.

Latif S, Qayyum A, Usman M, Qadir J. 2018. Cross lingual speech emotion recognition: Urdu vs. Western languages. In: *2018 International Conference on Frontiers of Information Technology (FIT)*. Piscataway: IEEE, 88–93.

Leila K, Serrestou Y, Mbarki M, Raoof K, Mahjoub MA, Cleder C. 2019. Automatic speech emotion recognition using machine learning. In: *Social Media and Machine Learning*. London: IntechOpen.

Li J, Sun S. 2008. Energy feature extraction of EEG signals and a case study. In: *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. Piscataway: IEEE, 2366–2370.

Livingstone SR, Russo FA. 2018. The Ryerson audio-visual database of emotional speech and song: a dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE* **13(5)**:e0196391 DOI 10.1371/journal.pone.0196391.

Lyakso E, Frolova O, Dmitrieva E, Grigorev A, Kaya H, Salah AA, Karpov A. 2015. EmoChildRu: emotional child Russian speech corpus. In: *Speech and Computer Lecture Notes in Computer Science*. 144–152.

Mao Q, Xu G, Xue W, Gou J, Zhan Y. 2017. Learning emotion discriminative and domain-invariant features for domain adaptation in speech emotion recognition. *Speech Communication* **93**:1–10 DOI 10.1016/j.specom.2017.06.006.

Mirsamadi S, Barsoum E, Zhang C. 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Piscataway: IEEE, 2227–2231.

Panagiotis T, Trigeorgis G, Nicolaou MA, Schuller BW, Zafeiriou S. 2017. End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing* **11**:1301–1309 DOI 10.1109/JSTSP.2017.2764438.

Pengcheng W, Zhao Y. 2019. A novel speech emotion recognition algorithm based on wavelet kernel sparse classifier in stacked deep autoencoder model. *Personal and Ubiquitous Computing* **23(3–4)**:521–529 DOI 10.1007/s00779-019-01246-9.

Qasim M, Nawaz S, Hussain S, Habib T. 2016. Urdu speech recognition system for district names of pakistan: development, challenges and solutions. In: *2016 Conference of the Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques*. Piscataway: IEEE, 28–32.

Rajisha TM, Sunija AP, Riyas KS. 2016. Performance analysis of Malayalam language speech emotion recognition system using ANN/SVM. *Procedia Technology* **24(3)**:1097–1104 DOI 10.1016/j.protcy.2016.05.242.

Rauf S, Hameed A, Habib T, Hussain S. 2015. District names speech corpus for Pakistani languages. In: *International Conference Oriental Held Jointly Conference on Asian Spoken Language Research and Evaluation*.

Sarma M, Ghahremani P, Povey D, Goel NK, Sarma KK, Dehak N. 2018. Emotion identification from raw speech signals using DNNs. In: *Interspeech*. 3097–3101.

Simone H, Sagha H, Cummins N, Schuller BW. 2017. Emotional speech of mentally and physically disabled individuals: introducing the EmotAsS database and first findings. In: *Interspeech*. 3137–3141.

Sinith MS, Aswathi E, Deepa TM, Shameema CP, Rajan S. 2015. Emotion recognition from audio signals using support vector machine. In: *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)*. Piscataway: IEEE, 139–144.

Soleymani M, Asghari-Esfeden S, Fu Y, Pantic M. 2016. Analysis of eeg signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing* **7(1)**:17–28 DOI 10.1109/TAFFC.2015.2436926.

Staroniewicz P, Majewski W. 2009. Polish emotional speech database–recording and preliminary validation. In: *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions Lecture Notes in Computer Science*. 42–49.

Tang D, Zeng J, Li M. 2018. An end-to-end deep learning framework for speech emotion recognition of atypical individuals. In: *Interspeech*. 162–166.

Toledo-Pérez DC, Rodríguez-Reséndiz J, Gómez-Loenzo RA. 2020. A study of computing zero crossing methods and an improved proposal for EMG signals. *IEEE Access* **8**:8783–8790 DOI 10.1109/ACCESS.2020.2964678.

Tripathi S, Beigi H. 2018. Multi-modal emotion recognition on IEMOCAP dataset using deep learning. *ArXiv preprint*. DOI 10.48550/arXiv.1804.05788.

Ververidis D, Kotropoulos C. 2003. A review of emotional speech databases. In: *Proceedings Panhellenic Conference on Informatics*. 560–574.

Yadav P, Aggarwal G. 2015. Speech emotion classification using machine learning. *International Journal of Computer Applications* **118(13)**:44–47 DOI 10.5120/20809-3564.

Zhang B, Provost EM, Essi G. 2016. Cross-corpus acoustic emotion recognition from singing and speaking: a multi-task learning approach. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. Piscataway: IEEE, 5805–5809.

Zhang W, Zhao D, Chai Z, Yang LT, Liu X, Gong F, Yang S. 2017. Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services. *Software: Practice and Experience* **47(8)**:1127–1138 DOI 10.1002/spe.2487.

Zhao J, Mao X, Chen L. 2019. Speech emotion recognition using deep 1D and 2D CNN LSTM networks. *Biomedical Signal Processing and Control* **47**:312–323 DOI 10.1016/j.bspc.2018.08.035.

Zhu L, Chen L, Zhao D, Zhou J, Zhang W. 2017. Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. *Sensors* **17(7)**:1694 DOI 10.3390/s17071694.