Review Article

A Review of the Advancement in Speech Emotion Recognition for Indo-Aryan and Dravidian Languages

Syeda Tamanna Alam Monisha and Sadia Sultana

Department of Computer Science and Engineering, Shahjalal University of Science and Technology, Sylhet, Bangladesh

Correspondence should be addressed to Syeda Tamanna Alam Monisha; alammonisha@gmail.com and Sadia Sultana; sadia-cse@sust.edu

Received 3 October 2022; Revised 10 November 2022; Accepted 21 November 2022; Published 1 December 2022

Copyright © 2022 Syeda Tamanna Alam Monisha and Sadia Sultana. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Speech emotion recognition (SER) has grown to be one of the most trending research topics in computational linguistics in the last two decades. Speech being the primary communication medium, understanding the emotional state of humans from speech and responding accordingly have made the speech emotion recognition system an essential part of the human-computer interaction (HCI) field. Although there are a few review works carried out for SER, none of them discuss the development of SER system for the Indo-Aryan or Dravidian language families. This paper focuses on some studies carried out for the development of an automatic SER system for Indo-Aryan and Dravidian languages. Besides, it presents a brief study of the prominent databases available for SER experiments. Some remarkable research works on the identification of emotion from the speech signal in the last two decades have also been discussed in this paper.

1. Introduction

Living in a society, we humans communicate with each other to share our thoughts, feelings, ideas, and different types of information. We use different communication mediums, like text messages, emails, audio, and videos, to express ourselves to others. Besides this, people nowadays use a variety of emojis along with text messages to represent their feelings more precisely. However, without any doubt, among all the communication forms, speech is the most natural and easiest one to express ourselves.

In recent years, contact with computing devices has become more chatty. Dialogue systems like Siri, Alexa, Cortana, and many more have more widely penetrated the consumer market than before [1]. Thus, to make them more like a human conversational partner, it is important to recognize human emotions from the user’s voice signals. Understanding the emotional state of a speaker is important for perceiving the exact meaning of what he or she says. Therefore, research studies on automatic speech emotion recognition, which is the task of predicting the emotional state of humans from speech signals have emerged in recent years as it enhances human-computer interaction (HCI) systems and makes them more natural. Moreover, with the world being digitized day by day, speech emotion recognition has found increasing applications in our daily lives. Call centers, e-tutoring, surveillance systems, psychological treatments, robotics, and online marketing are just some of them.

It has been seen from cross-lingual speech emotion recognition studies that models trained with a language corpus do not perform well when tested on a different language corpus compared to the monolingual recognition rate [2, 3]. However, it will be interesting to find out whether those models perform better for different languages from the same groups. To carry out this study, the first thing is to find out the available resources of the target language group. Hence, in this study, we aimed to investigate recent advancements in SER for the Indo-Aryan and Dravidian language families. Indo-Aryan and Dravidian languages are spoken by 800 million and 250 million people, all over the world, respectively [4, 5]. The speakers of Indo-Aryan languages are mostly from Bangladesh, India, Nepal, Sri Lanka, and Pakistan, and speakers of Dravidian languages...
are mainly from southern India. Having a large number of speakers, yet most of them are low-resource languages. So far, there is no review work that highlights the SER experiments for the Indo-Aryan or Dravidian language groups. Therefore, this study presents a brief review of some work done for the development of SER for languages of the Indo-Aryan and Dravidian families.

The remaining part of the paper is organized as follows: Section 2 gives a brief overview of a speech emotion recognition system with different types of emotional speech corpora, features, and classification algorithms utilized for the development of an SER system. Trends in speech emotion recognition research have been discussed in Section 3. Section 4 discusses some research works on SER in different languages in the last two decades. In Section 5, the advancement of SER works in Indo-Aryan and Dravidian languages is shown, and lastly, the study is concluded in Section 6.

2. Overview of Speech Emotion Recognition System

A speech emotion recognition (SER) system analyzes human speech and makes predictions about the emotion reflected by the speech. The system that recognizes the emotion from speech may be dependent or independent of the speaker and gender. Comparatively, the recognition accuracy of a speaker-dependent system is higher than that of a speaker-independent system, but the disadvantage of this strategy is that the system only responds appropriately to the person who trained the system. As reflected in Figure 1, the first requirement for building an SER system is a suitable speech dataset having different emotional states. For this purpose, raw speech data are collected from speakers in a variety of ways. Based on the generation of the corpus, emotional speech databases may be natural [6–10], acted [11–13], or elicited [14, 15]. Table 1 summarizes some prominent databases used for SER.

Once the data are collected, the raw speech data go through some preprocessing techniques such as noise reduction, silence removal, framing, windowing, and normalization for enhancing the speech signal [34]. After the preprocessing of raw data, the system opts for the feature extraction phase, which analyzes speech signals and obtains different speech characteristics. Any machine learning model’s success is largely dependent on its features. Selecting the right features could result in a more effective trained model, whereas choosing the wrong ones would significantly impede training. The selection of the proper signal features is crucial for better performance in recognizing the emotion of speech. From the beginning of SER research, various arrangements of speech features known as acoustic features like Mel Frequency Cepstral Coefficient (MFCC), pitch, zero crossing rate (ZCR), energy, and linear predictive cepstral coefficients (LPCC) have been used [35]. In various studies, nonspeech characteristics called nonacoustic features have also been integrated with the acoustic ones for the identification of emotion [36, 37]. Gestures, facial images, videos, and linguistic features are some of them.

After the feature selection process, a classifying algorithm is implemented to recognize the speech emotion. For the recognition of emotion from voice signals, many classifying algorithms have been used by researchers. A variety of supervised and unsupervised machine learning models have been employed for this purpose. Hidden Markov model (HMM), support vector machine (SVM), Gaussian mixture model (GMM), K-nearest neighbor (KNN), artificial neural network (ANN), and decision tree (DT) are some of them. In recent years, along with the traditional classification methods, several deep learning techniques are also being utilized for the classification process and have shown promising results. Convolutional neural network (CNN), long short-term memory (LSTM), deep CNN, and recurrent neural network (RNN) are the commonly used ones. In many SER studies, multiple classifiers are integrated to enhance the recognition rate. Authors Zhu et al. [38] combined two classifiers, deep belief network (DBN) and support vector machine (SVM), to classify the emotions of anger, fear, happiness, sadness, neutrality, and surprise in the Chinese Academy of Sciences emotional speech database. They used MFCC, pitch, formant, short-term ZCR, and short-term energy as features and achieved a mean accuracy of 95.8%, which is better than using SVM or DBN individually.

3. Speech Emotion Recognition Trends

The very first approach for determining the emotional state of a person from his/her speech was made in the late 1970s by Williamson [39]. Williamson provided a speech analyzer for the determination of an individual’s underlying emotion by analyzing pitch or frequency changes in the speech pattern. Later on, in 1996, Dellaert et al. [40] published the first research paper on the topic and introduced statistical pattern recognition techniques in speech emotion recognition. Authors Dellaert et al. [40] implemented K-nearest neighbors (KNN), Kernel regression (KR), and Maximum Likelihood Bayes’ (MLB) classifier using pitch characteristic of the utterances for the recognition of four different emotions, happiness, fear, anger, and sadness. Along with MLB and nearest neighbor (NN), Kang et al. [41] implemented the hidden Markov model (HMM), where HMM performs the best with 89.1% accuracy for recognizing happiness, sadness, anger, fear, boredom, and neutral emotions utilizing pitch and energy features. Onward, HMM has been largely used by researchers for speech emotion recognition showing satisfactory results [42–45]. SVM, GMM, and decision tree (DT) are some more traditional machine learning models which have been reliably used over the years for the same purpose [45–50]. In the 2000s, neural network (NN) has also been widely used for speech emotion recognition studies [51–54]. Indeed, in the earlier approaches, the use of conventional machine learning algorithms was widespread for recognizing the underlying emotion in human speech.

However, in the last decade, the trend of using conventional machine learning models for the recognition of emotion from human speech has moved towards deep learning models. Therefore, deep learning approaches have become more
Table 1: Summary of some commonly used databases for speech emotion recognition.

<table>
<thead>
<tr>
<th>S/N</th>
<th>Database</th>
<th>Language</th>
<th>Year</th>
<th>Size of database</th>
<th>Data format</th>
<th>Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RAVDESS [16]</td>
<td>English</td>
<td>2018</td>
<td>7,356 recordings by 24 actors</td>
<td>Audio-visual</td>
<td>Happy, angry, calm, sad, fearful, disgust, surprise, neutral</td>
</tr>
<tr>
<td>3</td>
<td>IEMOCAP [14]</td>
<td>English</td>
<td>2008</td>
<td>12 hours of data by 10 actors</td>
<td>Audio-visual</td>
<td>Happiness, anger, sadness, frustration, neutral</td>
</tr>
<tr>
<td>4</td>
<td>SUBESCO [12]</td>
<td>Bangla</td>
<td>2021</td>
<td>7,000 utterances by 20 native speakers</td>
<td>Audio</td>
<td>Happiness, anger, sadness, fear, surprise, disgust, neutral</td>
</tr>
<tr>
<td>5</td>
<td>FAU AIBO [6]</td>
<td>German</td>
<td>2008</td>
<td>9.2 hours of speech by 51 children talking to robot Aibo</td>
<td>Audio</td>
<td>Angry, emphatic, neutral, positive, rest</td>
</tr>
<tr>
<td>6</td>
<td>BanglaSER [17]</td>
<td>Bangla</td>
<td>2022</td>
<td>1,467 recording by 34 actors</td>
<td>Audio</td>
<td>Angry, happy, neutral, sad, surprise</td>
</tr>
<tr>
<td>7</td>
<td>MELD [7]</td>
<td>English</td>
<td>2018</td>
<td>13,000 utterances from the TV-series friends by multiple speakers</td>
<td>Audio-visual and textual</td>
<td>Anger, fear, joy, surprise, sadness, disgust, neutral</td>
</tr>
<tr>
<td>8</td>
<td>AESDD [18]</td>
<td>Greek</td>
<td>2018</td>
<td>500 utterances from 5 different actors</td>
<td>Audio</td>
<td>Anger, disgust, fear, happiness, sadness</td>
</tr>
<tr>
<td>10</td>
<td>RECOLA [19]</td>
<td>French</td>
<td>2013</td>
<td>3.8 hours of recordings by 46 participants</td>
<td>Audio-visual</td>
<td>Social behaviors (agreement, dominance, engagement, performance, rapport)</td>
</tr>
<tr>
<td>11</td>
<td>CHEAVD [8]</td>
<td>Chinese</td>
<td>2017</td>
<td>140 min emotional segments extracted from talk shows, TV plays, and films by 238 speakers</td>
<td>Audio-visual</td>
<td>Angry, fear, happy, neutral, sad, surprise</td>
</tr>
<tr>
<td>12</td>
<td>LSSED [20]</td>
<td>English</td>
<td>2020</td>
<td>147,025 utterances from 820 subjects</td>
<td>Audio</td>
<td>Happiness, fear, anger, excitement, sadness, boredom, disappointment, disgust, surprise, normal and other</td>
</tr>
<tr>
<td>13</td>
<td>Urdu [2]</td>
<td>Urdu</td>
<td>2018</td>
<td>400 utterances by 38 speakers from Urdu talk shows</td>
<td>Audio</td>
<td>Angry, happy, neutral, sad</td>
</tr>
<tr>
<td>14</td>
<td>Urdu-Sindhi speech emotion corpus [21]</td>
<td>Urdu, Sindhi</td>
<td>2020</td>
<td>1,435 speech recordings</td>
<td>Audio</td>
<td>Happiness, anger, sadness, disgust, surprise, sarcasm, neutral</td>
</tr>
<tr>
<td>15</td>
<td>IITKGP-SEHSC [22]</td>
<td>Hindi</td>
<td>2011</td>
<td>12,000 utterances by 10 professionals</td>
<td>Audio</td>
<td>Happy, anger, fear, disgust, surprise, sad, sarcastic, neutral</td>
</tr>
<tr>
<td>16</td>
<td>KSUEMotions [23]</td>
<td>Arabic</td>
<td>2017</td>
<td>5 hours and 10 minutes of recordings by 23 speakers</td>
<td>Audio</td>
<td>Neutral, happiness, sadness, surprise, anger</td>
</tr>
<tr>
<td>17</td>
<td>Oriya emotional speech database [15]</td>
<td>Oriya</td>
<td>2010</td>
<td>900 emotional utterances for text fragments from various drama scripts of Oriya language by 35 Oriya speakers</td>
<td>Audio</td>
<td>Happiness, sadness, astonish, anger, fear, neutral</td>
</tr>
<tr>
<td>19</td>
<td>CASIA natural emotional audio-visual database [9]</td>
<td>Chinese</td>
<td>2014</td>
<td>2 hours spontaneous emotional segments extracted from 219 speakers from films, TV plays and talk shows</td>
<td>Audio-visual</td>
<td>Happy, angry, disgust, surprise, worried, sad, fear</td>
</tr>
<tr>
<td>20</td>
<td>Egyptian Arabic speech emotion (EYASE) database [25]</td>
<td>Arabic</td>
<td>2020</td>
<td>579 utterances by 3 male and 3 female professional actors from Egyptian TV series</td>
<td>Audio-visual</td>
<td>Angry, happy, neutral, sad</td>
</tr>
</tbody>
</table>
popular, showing promising results. Deep learning algorithms are neural networks with multiple layers. CNN, DCNN, LSTM, BLSTM, and RNN are some widely implemented deep learning techniques for SER [55–57].

In recent times, multitask learning and attention mechanism are also being used for improved performance [58, 59]. For cross-corpus and cross-lingual speech emotion recognition, the transfer learning technique is being widely used [3, 60, 61].

Figure 2 depicts an analysis that shows that the use of deep learning techniques like CNN, RNN, LSTM, and DBN has increased over the years, along with traditional machine learning algorithms like SVM, DT, KNN, HMM, and GMM.

4. Survey on Speech Emotion Recognition Research Studies

After the first published research work on speech emotion recognition in 1996, the field of SER has received a great deal of attention over the past 20 years. Moderate progress has been made to create an automatic SER system. Several acoustic and nonacoustic features have been utilized along with different classifying models. Comparatively, the number of SER experiments conducted in English, German, and French languages is higher than that of the research conducted in other languages. One main reason is the availability of established and publicly accessible databases for the mentioned languages. RAVDESS, IEMOCAP, and SAVEE are some prominent emotional speech databases for the English language, EmoDB for Berlin, FAU AIBO for German, and RECOLA for French. The IEMOCAP database was used by researchers in [56, 58, 59, 62–64] for speech emotion recognition. Fayek et al. [56] evaluated deep learning techniques with CNN and LSTM-RNN using the

<table>
<thead>
<tr>
<th>S/N</th>
<th>Database Description</th>
<th>Language</th>
<th>Year</th>
<th>Size of database</th>
<th>Data format</th>
<th>Emotions</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Interface multilingual emotional speech database [26]</td>
<td>English, French, Spanish and Slovenian</td>
<td>2002</td>
<td>In English interface database contains 8,928 utterances, Slovenian 6,080, French 5,600 and Spanish 5,520 by 2 actors</td>
<td>Audio</td>
<td>Joy, disgust, anger, fear, sadness, surprise, neutral</td>
</tr>
<tr>
<td>22</td>
<td>Toronto emotional speech set (TESS) [27]</td>
<td>English</td>
<td>2010</td>
<td>2,800 utterances by 2 actresses</td>
<td>Audio</td>
<td>Happiness, anger, sadness, disgust, pleasant surprise, fear, neutral</td>
</tr>
<tr>
<td>23</td>
<td>ANAD [28]</td>
<td>Arabic</td>
<td>2018</td>
<td>1,384 recordings from Arabic talk shows</td>
<td>Audio</td>
<td>Happy, angry, surprised</td>
</tr>
<tr>
<td>24</td>
<td>Multilingual emotional speech database of north east India (MESDNEI) [29]</td>
<td>Assamese</td>
<td>2009</td>
<td>4,200 utterances of 5 native languages of Assam by 30 speakers</td>
<td>Audio</td>
<td>Sadness, anger, fear, disgust, happiness, surprise, neutral</td>
</tr>
<tr>
<td>25</td>
<td>ShEMO [10]</td>
<td>Persian</td>
<td>2019</td>
<td>Speech data of 3 hours and 25 minutes from online radio plays</td>
<td>Audio</td>
<td>Anger, fear, happiness, sadness, surprise, neutral</td>
</tr>
<tr>
<td>26</td>
<td>SEMOUR*: a scripted emotional speech repository for Urdu [30]</td>
<td>Urdu</td>
<td>2021</td>
<td>27,640 instances recorded by 24 actors</td>
<td>Audio</td>
<td>Anger, happiness, surprise, disgust, sadness boredom, fearful, neutral</td>
</tr>
<tr>
<td>27</td>
<td>Arabic natural corpus [28]</td>
<td>Arabic</td>
<td>2018</td>
<td>1,384 recordings from online Arabic talk shows</td>
<td>Audio</td>
<td>Angry, surprised, happy</td>
</tr>
<tr>
<td>28</td>
<td>EMOVO [31]</td>
<td>Italian</td>
<td>2014</td>
<td>588 utterances of 14 sentences by 6 actors</td>
<td>Audio</td>
<td>Joy, sadness, anger, disgust, fear, surprise, neutral</td>
</tr>
<tr>
<td>29</td>
<td>Punjabi emotional speech database [32]</td>
<td>Punjabi</td>
<td>2021</td>
<td>900 utterances by 15 speakers</td>
<td>Audio</td>
<td>Happy, angry, sad, fear, surprise, neutral</td>
</tr>
<tr>
<td>30</td>
<td>IITKGP-SESC [33]</td>
<td>Telegu</td>
<td>2009</td>
<td>12,000 utterances by 10 professionals</td>
<td>Audio</td>
<td>Happy, anger, sad, disgust, fear, sarcastic, surprise, neutral</td>
</tr>
</tbody>
</table>

Figure 2: Growth of using deep learning techniques for SER. The graph displays the publication number in the last two decades, having machine learning or deep learning terms in the title of SER papers, according to Google Scholar around the time this article was submitted.
IEMOCAP database and achieved 64.78% and 61.71% test accuracy for CNN and LSTM-RNN, respectively. Implementing spectrogram-based self-attentional CNN-BLSTM, Li et al. [58] gained a weighted accuracy of 81.6% and unweighted accuracy of 82.8% for the IEMOCAP dataset for classifying angry, happy, neutral, and sad emotions. Using BLSTM with an attention mechanism, Yu and Kim [59] got a weighted accuracy of 73% and an unweighted accuracy of 68% for the IEMOCAP corpus. Meng et al. [62] used attention mechanism-based dilated CNN with residual block and BiLSTM for both IEMOCAP and Berlin EmoDB and got 74.96% speaker-dependent and 69.32% speaker-independent accuracy for IEMOCAP and 90.78% speaker-dependent and 85.39% speaker-independent for Berlin EmoDB.

A combination of prosodic and modulation spectral features (MSFs) with an SVM classifier was implemented by Wu et al. [65] for the Berlin EmoDB database and the recognition rate was 91.6% for recognizing the emotions in the Berlin EmoDB database. An improved recognition rate of 96.97% was observed by deep convolutional neural network (DCNN) for the Berlin EmoDB database for the recognition of angry, neutral, and sad emotions [66]. For Chinese language authors, Zhang et al. [67] employed SVM and a deep belief network (DBN) with MFCC, pitch, and formant features and got 84.54% mean accuracy by SVM and 94.6% by DBN for the Chinese Academy of Sciences emotional speech database. A higher mean accuracy of 95.8% was achieved for the same Chinese dataset in [38] by combining deep belief network (DBN) with support vector machine (SVM).

Experiments have also been conducted for cross-lingual speech emotion recognition. Sultana et al. [3] showed a cross-lingual study for English and Bangla languages using RAVDESS and SUBESCO datasets, respectively, where the proposed system integrates a deep CNN and a BLSTM network with a TDF layer. Transfer learning was used for the cross-lingual experiment, achieving weighted accuracy of 86.9% for SUBESCO and 82.7% for RAVDESS. Latif et al. [2] used an SVM classifier for cross-lingual emotion recognition for Urdu, German, English, and Italian languages. The authors used SAVEE, EmoDB, EMOVO, and URDU databases for English, German, Italian, and Urdu languages, respectively, for the evaluation of the cross-corpus study. Xiao et al. [68] investigated the cross-lingual study of emotion recognition from speech using the databases EmoDB, DES, and CDESD for German, Danish, and Mandarin languages, respectively. Using CDESD as the training dataset and EmoDB as the testing dataset, the authors achieved the best accuracy of 71.62% for the cross-corpus study with a sequential minimal optimization (SMO) classifier. The IEMOCAP and Recola databases were used for cross-lingual study by Neumann [69] for English and French languages, respectively, where an attentive convolutional neural network (ACNN) was used. 59.32% unweighted average recall was achieved for the IEMOCAP test database while trained on Recola and 61.27% for Recola while training was carried out on the IEMOCAP database. A cross-lingual cross-corpus study was carried out for four languages, German, Italian, English, and Mandarin by Goel and Beigi [70]. Transfer learning and multitask learning techniques were used, providing accuracy of 32%, 51%, and 65% for EMOVO, SAVEE, and EmoDB databases, respectively, using IEMOCAP as the training database.

Apart from using available prominent databases, researchers are also creating emotional speech corpus using acted, elicited, or natural recordings and experimenting with various classification models for the identification of speech emotion. A multilingual database containing 720 utterances by 12 native Burmese and Mandarin speakers was built by Nwe et al. [43]. Using the short-time log frequency power coefficients (LFPC) feature, the authors implemented the HMM classifier, which classifies six emotions, namely, anger, disgust, fear, joy, sadness, and surprise, with an average accuracy of 78% and the best accuracy of 96%.

5. Advancement of Speech Emotion Recognition in Indo-Aryan and Dravidian Languages

Indo-Aryan languages, also known as Indic languages, are the native languages of the Indo-Aryan people, which are a branch of the Indo-Iranian languages in the Indo-European language family. An estimation made at the beginning of the 21st century shows that more than 800 million people, mostly in India, Bangladesh, Sri Lanka, Nepal, and Pakistan, speak Indo-Aryan languages [4]. Hindi, Bangla, Sinhala, Urdu, Punjabi, Assamese, Nepali, Marathi, Odia, Gujarati, Sindhi, Rajasthani, and Chhattisgarhi are some prominent Indo-Aryan languages. Besides, Dravidian or Dravidic languages are spoken by 250 million people primarily in southern India, southwest Pakistan, and north-eastern Sri Lanka [5]. Tamil, Malayalam, Telugu, and Kannada are the most spoken Dravidian languages. Although a lot of work on speech emotion recognition in English, German, Chinese, Mandarin, and French languages has been conducted by researchers, compared to that, the number of experiments in the Indo-Aryan and Dravidian languages is not much. Inadequacy of available resources and variation in the nature of the languages are some reasons for that. However, in the last decade, improvement has been seen in speech emotion recognition research for both language families. Figure 3 shows an analysis of research works done for some of the languages.

5.1. Emotional Speech Databases for Indo-Aryan and Dravidian Languages

Some established and validated emotional speech corpora are available for some of the languages. Hindi is the most spoken language among the Indo-Aryan languages in terms of native speakers. The IIITKGP-SESC, Indian Institute of Technology Kharagpur Simulated Emotion Speech Corpus, developed by a team of Indian Institute of Technology Kharagpur in 2009, is the first corpus in Telugu, an Indian language [33]. The corpus contains 12,000 emotional speech utterances in Telugu, with happiness, surprise, anger, disgust, sadness, fear, sarcasm, and neutral emotions expressed by ten speakers.

Afterward, emotions being language independent, Koolagudi et al. [22] felt the necessity of a speech corpus in other Indian languages and created the Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus.
(IITKGP-SEHSC) developed in the Hindi language. The database contains 12,000 utterances of Hindi speech recorded by ten professional FM radio artists in India. Eight emotions, namely, happiness, sadness, surprise, anger, sarcasm, fear, disgust, and neutral, are present in the database.

A publicly available speech emotion corpus is available in the Urdu language containing 400 utterances by 38 speakers from different Urdu talk shows annotated with emotions of anger, happiness, neutral, and sadness [2]. Asghar et al. [71] built a corpus comprising 2,500 emotional speech utterances by 20 speakers with sadness, anger, disgust, happiness, and neutrality emotions.

SUBESCO, SUST Bangla Emotional Speech Corpus, is the largest available emotional speech corpus for the Bangla language consisting of more than 7 hours of speech with 7,000 utterances [12]. Happiness, surprise, anger, sadness, disgust, fear, surprise, and neutrality are the emotional states present in the database.

Mohanty and Swain [15] developed an Oriya emotional speech corpus for the Oriya language having six emotion classes, namely, happiness, anger, fear, sadness, astonishment, and neutrality.

For the Assamese language, there exists an emotional speech corpus containing utterances in five native Assamese languages, namely, Assamese, Karbi, Bodo (or Boro), Missing (or Mishing), and Dimasa [29].

A Punjabi speech database was created by Kaur and Singh [32] consisting of 900 emotional speech utterances by 15 speakers. Happiness, fearful, angry, surprised, sad, and neutral are the six emotions present in the database.

Kannada emotional speech (KES) database developed by Geethashree and Ravi [72] contains acted emotional utterances in the local languages of Karnataka. The database includes the basic emotions of happiness, sadness, anger, and fear, with a neutral state by four native Kannada actors.

A Malayalam elicited emotional speech corpus for recognizing human emotion from the speech was built by Jacob [73]. The database consists of 2,800 speech recordings in the six basic emotions and neutral by ten native educated and urban female Malayalam speakers.

Apart from these corpora, there are many more small speech databases created for emotion recognition purposes in Indo-Aryan and Dravidian languages [74–77].

5.2. Speech Emotion Recognition for Indo-Aryan and Dravidian Languages. Over the last fifteen years, there has been a moderate progress in SER research for languages of the Indo-Aryan and Dravidian families. Although the earlier approaches were traditional machine learning-based, in recent times, state-of-the-art models are being used by researchers with good performance. After the first large Telugu (IITKGP-SESC) [3] and Hindi (IITKGP-SEHSC) [22] emotional speech databases were published in 2009 and 2011, respectively, many experiments have been done for the languages. In 2021, Agarwal and Om [78] used deep neural network with deer hunting optimization algorithm and got the highest accuracy of 93.75% for the IITKGP-SEHSC dataset. The model implemented for the RAVDESS database outperforms the state-of-the-art accuracy giving 97.14% highest recognition rate [78]. Combining DCNN and BLSTM, the model proposed by Sultana et al. [3] obtained state-of-the-art efficiency with 82.7% and 86.9% accuracy for the RAVDESS and SUBESCO databases for English and Bangla languages, respectively.

Swain et al. [93] in 2022 implemented a deep convolutional recurrent neural network-based ensemble classifier for Odia and RAVDESS databases, which provides better results than some state-of-the-art models for the mentioned databases, giving an accuracy rate of 85.31% and 77.54%, respectively. Likewise, conventional approaches, along with deep learning techniques, are also showing good performance for the language families. Table 2 summarizes some experiments on speech emotion recognition for Indo-Aryan and Dravidian languages.

5 Advances in Human-Computer Interaction
<table>
<thead>
<tr>
<th>S/ N</th>
<th>Reference</th>
<th>Database Name</th>
<th>Language</th>
<th>Approach used</th>
<th>Recognized emotions</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Koolagudiet al. [79]</td>
<td>IITKGP-SESC</td>
<td>Telugu</td>
<td>SVM and GMM with energy and pitch parameters</td>
<td>Happy, anger, fear, disgust, sarcastic, sad, neutral, surprise</td>
<td>63.75% average accuracy obtained</td>
</tr>
<tr>
<td>2</td>
<td>Sultana et al. [3]</td>
<td>SUBESCO and RAVDESS</td>
<td>Bangla and English</td>
<td>The system integrates a DCNN and a BLSTM network with a TDF layer</td>
<td>Happy, calm, sad, surprise, fearful, disgust, angry, neutral</td>
<td>For the SUBESCO and RAVDESS datasets, the proposed model has achieved weighted accuracies of 86.9% and 82.7%, respectively</td>
</tr>
<tr>
<td>3</td>
<td>Kumar and Yadav [80]</td>
<td>IITKGP-SEHSC</td>
<td>Hindi</td>
<td>Deep LSTM with GMFCC and DMFCC features</td>
<td>Happy, fear, angry, sad, neutral</td>
<td>The proposed framework gives average accuracy of 91.2% for male speech and 87.6% for female speech</td>
</tr>
<tr>
<td>4</td>
<td>Mohanty and Swain [15]</td>
<td>Oriya emotional speech database</td>
<td>Oriya</td>
<td>Fuzzy K-means</td>
<td>Anger, sadness, astonish, fear, happiness, neutral</td>
<td>65.16% recognition rate by incorporating mean pitch, first two formants, jitter, shimmer, and energy as feature vectors</td>
</tr>
<tr>
<td>5</td>
<td>Samantaray et al. [48]</td>
<td>MESDNEI</td>
<td>Assamese</td>
<td>SVM with dynamic, quality, derived, and prosodic features</td>
<td>Happy, anger, fear, disgust, surprise, sad, neutral</td>
<td>82.26% average accuracy rate for speaker-independent case</td>
</tr>
<tr>
<td>6</td>
<td>Bhavan et al. [81]</td>
<td>EmoDB, RAVDESS and IITKGP-SEHSC</td>
<td>German, English and Hindi</td>
<td>Bagged ensemble of SVM using MFCCs, spectral, and centroids</td>
<td>Happy, sad, calm, angry, surprise, fear, disgust, neutral</td>
<td>Obtained accuracy EmoDB: 92.45%, RAVDESS: 75.69% and IITKGP-SEHSC: 84.11%</td>
</tr>
<tr>
<td>7</td>
<td>Swain et al. [82]</td>
<td>Self-created database using utterances from two native languages of Odisha: Cuttacki and Sambalpuri</td>
<td>Oriya</td>
<td>SVM using MFCC as feature vector</td>
<td>Happiness, fear, anger, disgust, sad, neutral, surprise, neutral</td>
<td>82.14% recognition accuracy for SVM classifier</td>
</tr>
<tr>
<td>8</td>
<td>Zaheer et al. [30]</td>
<td>SEMOUR’</td>
<td>Urdu</td>
<td>Ensemble classifier, CNN combined with VGG-19 model</td>
<td>Anger, disgust, happiness, surprise, boredom, sadness, fearful, neutral</td>
<td>The proposed model achieved 56% speaker-independent recognition rate</td>
</tr>
<tr>
<td>9</td>
<td>Wankhade et al. [47]</td>
<td>Speech emotional database containing dialogues from different bollywood movies</td>
<td>Hindi</td>
<td>SVM classifier with MFCC and MEDC feature set</td>
<td>Angry, happy, sad, neutral</td>
<td>71.66% recognition rate using SVM classifier</td>
</tr>
<tr>
<td>10</td>
<td>Ali et al. [83]</td>
<td>Self-created speech emotional corpus recorded in 5 regional languages of Pakistan</td>
<td>Urdu, Sindhi, Pashto, Punjabi, and Balochi</td>
<td>Learning classifiers (adaboostM1, J48, classification via regression, decision stump) with prosodic features</td>
<td>Happiness, sad, anger, neutral</td>
<td>40% classification accuracy with pitch feature</td>
</tr>
<tr>
<td>11</td>
<td>Ancilin and Milton [84]</td>
<td>Urdu</td>
<td>Urdu</td>
<td>SVM classifier with mel frequency magnitude coefficient (MFMC)</td>
<td>Happy, sad, anger, neutral</td>
<td>95.25% emotion recognition rate using MFMC</td>
</tr>
<tr>
<td>12</td>
<td>Farhad et al. [85]</td>
<td>Urdu</td>
<td>Urdu</td>
<td>Neural network, random forest and meta iterative classifiers with pitch and MFCC features</td>
<td>Happy, sad, angry</td>
<td>With an accuracy of 78.75%, random forest outperforms other classifiers</td>
</tr>
</tbody>
</table>
6. Conclusion

Speech emotion recognition being an integral part of HCI, a successful SER system with a healthy level of accuracy is essential for the better performance of a human-computer interaction system. This paper presents a survey on speech emotion recognition research for Indo-Aryan and Dravidian languages. A brief review of 31 research studies, including the development of emotional speech corpora and implemented approaches with utilized features for emotion recognition purposes, has been covered for the mentioned language families. Besides, a thorough study of some standard available emotional speech corpora and research works conducted for the identification of emotional states from human speech for different languages has also been presented in this paper. Therefore, researchers working in this field might find helpful insights about speech emotion recognition in this study.

Data Availability

No data were used to support this study.

Table 2: Continued.

<table>
<thead>
<tr>
<th>S/N</th>
<th>Reference</th>
<th>Database Name</th>
<th>Language</th>
<th>Approach used</th>
<th>Recognized emotions</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>Darekar and Dhande [86]</td>
<td>Marathi database</td>
<td>Marathi</td>
<td>Adaptive ANN combining cepstral, non-negative matrix factorization (NMF) and pitch features</td>
<td>Happy, sad, angry, fear, neutral, surprised</td>
<td>Proposed model obtains 80% accuracy combining the 3 features</td>
</tr>
<tr>
<td>14</td>
<td>Koolagudi et al. [87]</td>
<td>IITKGP-SESC</td>
<td>Telugu</td>
<td>SVM and GMM model with epoch parameters were used</td>
<td>Happy, anger, fear, sadness, disgust, neutral</td>
<td>Average recognition rates are 58% and 61% for SVM and GMM, respectively</td>
</tr>
<tr>
<td>15</td>
<td>Kandali et al. [49]</td>
<td>Self-created acted emotional speech database by 27 speakers</td>
<td>Assamese</td>
<td>GMM classifier with MFCC features</td>
<td>Happy, sad, disgust, fear, angry, surprise, neutral</td>
<td>Highest mean classification score is 76.5%</td>
</tr>
<tr>
<td>16</td>
<td>Dhar and Guha [88]</td>
<td>Abeg: self-collected Bangla emotional speech dataset</td>
<td>Bangla</td>
<td>Logistic regression model with MFCC and LPC features</td>
<td>Happy, angry, neutral</td>
<td>Proposed model achieved 92% accuracy combining MFCC and LPC features</td>
</tr>
<tr>
<td>17</td>
<td>Jacob [89]</td>
<td>Hindi emotional speech database containing 2240 wav files collected from 10 speakers</td>
<td>Hindi</td>
<td>ANN model with jitter and shimmer features</td>
<td>Happy, sad, anger, fear, surprise, disgust, neutral</td>
<td>83.3% overall accuracy obtained combining jitter and shimmer features</td>
</tr>
<tr>
<td>18</td>
<td>Fernandes and Mannepalli [90]</td>
<td>Acted emotional speech database containing 1400 utterances by 10 actors</td>
<td>Tamil</td>
<td>LSTM and BiLSTM with MFCC, MFCC delta, spectral kurtosis, bark spectrum, and spectral skewness features</td>
<td>Happy, anger, sad, fear, boredom, disgust, neutral</td>
<td>84% accuracy rate obtained using LSTM and BiLSTM with dropout layers</td>
</tr>
<tr>
<td>19</td>
<td>Rajisha et al. [91]</td>
<td>Acted emotional dataset created by the authors</td>
<td>Malayalam</td>
<td>ANN and SVM classifier with MFCC, short-time energy, and pitch features</td>
<td>Happy, anger, sad, neutral</td>
<td>88.4% recognition rate obtained using ANN and 78.2% with SVM</td>
</tr>
<tr>
<td>20</td>
<td>Kannadaguli and Bhat [92]</td>
<td>Self-created database containing 2800 emotional recordings</td>
<td>Kannada</td>
<td>Bayesian and HMM model with MFCC feature</td>
<td>Happy, excited, angry, sad</td>
<td>Average emotion error rate of 25.5% for Bayesian and 0.2% for HMM approach</td>
</tr>
</tbody>
</table>

Conflicts of Interest

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

References


Advances in Human-Computer Interaction


Advances in Human-Computer Interaction


