Urdu Speech Emotion Recognition: A systematic Literature Review

SOONH TAJ*

Department of Computer Science, Sukkur IBA University, soonhtaj.phdcss22@iba-suk.edu.pk

GHULAM MUJTABA*

Department of Computer Science, Sukkur IBA University, mujtaba@iba-suk.edu.pk

SHER MUHAMMAD DAUDPOTA

Department of Computer Science, Sukkur IBA University, sher@iba-suk.edu.pk

MUHAMMAD HUSSAIN MUGHAL

Department of Computer Science, Sukkur IBA University, muhammad.hussain@iba-suk.edu.pk

Research on Speech Emotion Recognition is becoming more mature day by day, and a lot of research is being carried out on Speech Emotion Recognition in resource-rich languages like English, German, French and Chinese. Urdu is among the top 10 languages spoken worldwide. Despite its importance, few studies have worked on Urdu Speech emotion as Urdu is recognized as a resource-poor language. The Urdu language lacks publicly available datasets, and for this reason, few researchers have worked on Urdu Speech Emotion Recognition. To the best knowledge, no review has been found on Urdu Speech Emotion recognition. This study is the first systematic literature review on Urdu Speech Emotion Recognition, and the primary goal of this study is to provide a detailed analysis of the literature on Urdu Speech Emotion Recognition which includes the datasets, features, pre-processing, approaches, performance metrics, validation methods used for Urdu speech emotion recognition. This study also highlights the challenges and future directions for Urdu Speech Emotion Recognition.

CCS CONCEPTS: • **Computing methodologies → Artificial intelligence; • Computing methodologies → Machine learning → Machine learning algorithms,**

Additional Keywords and Phrases: Speech Emotion Recognition, Urdu Speech, Low Resource Language, Machine Learning, Deep Learning.

1 INTRODUCTION

This paper presents a systematic literature review on Urdu Speech Emotion Recognition. The Introduction section makes a general discussion on Speech Emotion Recognition. The subsequent parts of the Introduction are arranged as follows: Speech Emotion Recognition is briefly discussed in section 1.1, the working model of Speech Emotion Recognition is discussed in section 1.2, the need for Urdu speech emotion recognition is discussed in section 1.3, research motivation is discussed in section 1.4, in section 1.5 comparison is made with existing reviews.

1.1 Speech Emotion Recognition (SER)

Speech is the most natural way to express emotions. Speech Emotion Recognition is a set of tools and techniques for extracting, detecting, and classifying subjective information, such as opinions and feelings from speech or voice data. Signal Processing, Natural Language processing, Machine Learning and Deep Learning Methods are utilized for Speech Emotion Recognition.

*Corresponding author

2375-4699/2023/1-ART1 $15.00
http://dx.doi.org/10.1145/3595377

This field of speech emotion recognition is old; back in the 1920s, a celluloid toy named 'Radio Rex' was developed. This toy had the capability of automatically detecting acoustic energy released by the vowel 'Rex,' i.e., 500Hz [59]. Many researchers then worked in this field to create intelligent and interactive robotic and automated systems that can understand human speech and emotions and behave or serve accordingly. To make Human-Robot interaction smooth, researchers have emphasized using Speech Emotion Recognition [64].

There are many application areas where we find Speech Emotion Recognition very useful. The SER system can be employed as a diagnostic tool for therapists to treat patients of depression based on their emotional and mental state [29]. SER can be used in call centre applications to analyze the agent's customer behaviour using audio call data [50]. Also, the E-commerce and service industry can reap the benefits from Speech Emotion Recognition by sending alerts to customer service and managers about the caller's state of mind towards products and services [60]. SER has also proved beneficial in e-learning systems where stress and frustration can be detected in users' emotions to determine whether learning or studying is conducive or not to provide appropriate countermeasures in e-learning systems [37]. Not only limited to these application areas, Speech Emotion Recognition is also used in computer video games [71]. A unique application of SER is found in detecting fear in audio–video surveillance systems [21]. Also, in the near future, we will have a smart transportation system in which the system will take the steering control for self-driving upon detection of the unhealthy emotion of the driver [55].

1.2 Speech Emotion Recognition (SER) System Working Model

The speech Emotion Recognition system automatically recognizes speech emotions from input speech data. Figure 1 shows the working model of the SER system.

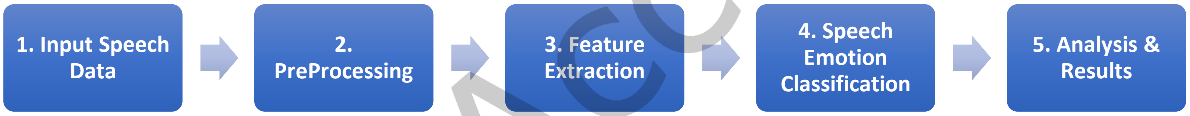

Figure 1: Speech Emotion Recognition System Working Model

1. The first step is to give speech data as input the speech data. Speech data can be voice, call, utterance/ sentence, or spoken dialogue.
2. The next step is to pre-process the given input speech data. In this step, the speech signal is processed, and noise is removed; it can be divided into speech segments, which will help in further processing and feature extraction task. This step enhances the quality of speech data by removing noise and segmenting speech by only keeping relevant and useful segments for speech emotion recognition.
3. After pre-processing of speech data, various speech features, which can be prosodic or acoustic in nature, are extracted from speech data. Most of the time, features such as MFCCs, pitch, energy, and intensity are extracted as these are proven to be helpful for speech emotion recognition.
4. The fourth step in SER is to classify the audio speech into emotions based on the features extracted from the audio speech signal.
5. The last step is to evaluate the performance of the SER system by using various performance metrics like Accuracy, Precision and Recall are commonly used metrics to assess the performance of the SER system.

1.3 Need for Urdu Speech Emotion Recognition

Currently, researchers are focusing on developing interactive systems that can automatically detect emotion from the speech or voice of customers or users and respond quickly in a real environment. For this task to make a speech emotion recognition system as intelligent as it can understand speech in any language, researchers are trying to work on multi-lingual speech emotion recognition.

Much work has been done on speech emotion recognition in famous languages like English, German, Japanese, Chinese languages and many others famous languages. Few authors have contributed to the Urdu Language as there is a lack of availability of datasets [62].

According to rank Source (2022 edition of Ethnologue, a language reference published by SIL International), Urdu is among the top 10 spoken languages [77]. Urdu is the top 10 most spoken languages by number of native speakers (ns) in the world, and there are about 231.3 million native speakers (ns) of the Urdu language in the world [77].

1.4 Research Motivation

Researchers are keen to apply the latest deep learning methods for Speech Emotion Recognition of audio data in different languages, especially resource-rich languages like English, German, French and chinse. Urdu is among the top 10 languages spoken in the world, and there are more than 231 million native speakers of the Urdu language. It is necessary to focus on this language for speech-emotion recognition. This systematic review is the first and foremost Urdu Speech Emotion Recognition review. This systematic review will serve future researchers interested in Urdu Speech Emotion Recognition.

1.5 Comparison with the Existing Literature Reviews

Urdu Speech Emotion Recognition is currently in its initial stages compared to resource-rich languages like English. Few authors have worked on Urdu Speech Emotion Recognition. This directly impacts the availability of reviews or surveys on Urdu Speech Emotion Recognition. From the extensive search, no review study was found on Urdu Speech Emotion Recognition as Urdu is a resource-poor language. Few technical studies exist in the existing Urdu Speech Emotion Recognition literature. This is the first and foremost Systematic Literature review for Urdu Speech Emotion Recognition.

Table 1 shows the summary of related reviews along with their limitation is given, and in Table 2 Comparison of this systematic review is made against existing general reviews found on Speech Emotion Recognition.

Table 1: Related Reviews and Surveys

Ref	Year	Study Type	Objective	Limitations
[53]	2018	Systematic Literature Review	This systematic literature review provided research analysis in SER from the year 2006 to 2017. The objective of this review is to provide detailed analysis of databases, features, and classification techniques. This review discuses speech databases, datasets, Approaches, Features, Classification, and Perf. / Results and Future Work.	This review does not cover the Emotional Model, Experiment Type, Evaluation, Performance Metrics, Validation, Strengths, Weakness/Limitations Implementation Tools of given literature on SER
[67]	2018	Review	The objective of this review is to provide a brief synopsis of the last 2 decades of SER. This review synopsis the Datasets, Approaches, Pre-processing, Features, Classification, Performance, Future Work and Challenges.	This review does not cover Databases, Emotional Models, Modality, Experiment Type, Evaluation, Performance Metrics, Validation, Strengths, Weaknesses/Limitations, and Implementation Tools of the given literature on SER.
[69]	2018	Review	This review aims to provide comprehension for SER, specifically for Databases, Datasets, Features, Classification and Perf. /Results.	This review does not cover Emotional models, Modality, Experiment Type, Approach, Evaluation, Performance Metrics, Validation, Strengths, Weaknesses/Limitations, Future Work, Challenges, and Implementation Tools of the given literature on SER.

[38]	2019	Review	This study aims to provide an overview of Deep Learning Techniques used for SER. This review covers Databases, Datasets, Approaches, Features, Classification, Pref./Results, Strengths, and Future Work.	This review does not cover the Emotional Model, Modality, Experiment Type, Evaluation, Performance Metrics, Validation, and Weak—/Limitations and Implementation Tools of the given literature on SER.
[4]	2020	Survey	This review aims to provide a detailed survey of the current literature on SER. This review brings to light the Emotional Model, Databases, Pre-processing, Features, Modality, Classification, Performance and Challenges.	This review does not cover Experiment Type, Evaluation, Performance Metrics, Validation, Strengths, Weaknesses/Limitations, Future Work, and Implementation Tools of the given literature on SER.
[2]	2021	Review	This survey aims to elaborate on using deep learning techniques for SER. This survey focuses on Databases, Datasets, Approaches, Features, Classification, Performance, Future Work and Challenges.	This review does not cover the emotional Model, Modality, Pre-processing, Evaluation, Performance Metrics, Validation, Strengths, and Weakness/Limitations and Implementation Tools of the given literature on SER.
[27]	2021	Survey	This review explores the SER. techniques for the natural environment in terms of the speaker, text, language and recording environment. Moreover, this review explains Databases, Emotion Models, Datasets, Approaches, Pre-processing, Features, Classification, Performance, Performance Metrics, Strengths, Weaknesses/Limitations, and challenges.	This review does not cover the Modality, Experiment Type, Evaluation, Validation, Implementation Tools, and Future Work of the literature on SER.
[34]	2021	State-of-the-art Review	The objective of this review is to provide an in-depth analysis of Deep Learning approaches that are being utilized for SER. Further, this review presents Databases, Datasets, Pre-processing, Features, Classification, Performance, Performance Metrics, Strengths, Weaknesses/Limitations, Development Tools, Future work, and Challenges.	This review particularly focuses on Deep Learning approaches and does not cover the Emotional Model, Modality, Experiment Type, Evaluation, and Validation of the given literature on SER.
[43]	2021	State-of-the-art Review	The objective of this study is to summarize current research on Deep Multi-modal Emotion Recognition on Human Speech. This review includes in-depth analysis of Datasets, Modality, Approaches, Features, Classification, Performance, Evaluation, Performance Metrics, Validation and Challenges.	This review does not cover the Databases, Emotional Model, Experiment Type, Pre-processing, Strengths, Weakness/Limitations, Implementation Tools, and future Work of given literature on SER.
[73]	2021	Comprehensive Survey	The objective of this review is to identify and synthesize literature on SER. This review provides detailed analysis of Databases, Emotional Model, Datasets, Approaches, Pre-processing, Features, Classification, Performance, Strengths, Future Work and Challenges.	This review does not cover the Modality, Experiment Type, Evaluation, Performance Metrics, Implementation Tools, and Limitations of given literature on SER.

Table 2: Comparison of the most relevant reviews of speech emotion recognition published between 2018 and 2022 and our proposed review.

Meaning of acronyms: D.B. - Databases, Emo. Mod. - Emotional Model, D.S. - Dataset, Exp. Type – Experiment Type, Appr. – Approach), Prep – Pre-processing, Feat. – Features, Classf. – Classification, Perf. – Performance, Eval. – Evaluation, Pref. Metrics / Performance Metrics, Valid. – Validation, Str. – Strengths, Weak. /Lim. – Weaknesses/ Limitations, Impl. Tools – Implementation Tools, F.W. – Future Work, Chall. - Challenges

Ref	Year	Study Scope																	
		D.B.	Emo. Mod.	D.S.	Modality	Exp. Type	Appr.	Prep.	Feat.	Classf.	Perf.	Eval.	Pref. Metrics	Valid.	Str.	Weak. /Lim.	Impl. Tools	F W	Chall.
[53]	2018	✓	×	✓	×	×	✓	×	✓	✓	✓	×	×	×	×	×	×	✓	×
[67]	2018	×	×	✓	×	×	✓	✓	✓	✓	✓	×	×	×	×	×	×	✓	✓
[69]	2018	✓	×	✓	×	×	×	×	✓	✓	✓	×	×	×	×	×	×	×	×
[38]	2019	✓	✓	✓	×	×	✓	×	✓	✓	✓	×	×	×	✓	×	×	✓	×
[4]	2020	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×	×	×	×	×	×	×	✓
[2]	2021	✓	×	✓	×	×	✓	×	✓	✓	✓	×	×	×	×	×	×	✓	✓
[27]	2021	✓	✓	✓	×	×	✓	✓	✓	✓	✓	×	✓	×	✓	✓	×	×	✓
[34]	2021	✓	×	✓	×	×	✓	✓	✓	✓	✓	×	✓	×	✓	✓	✓	✓	✓
[43]	2021	×	×	✓	✓	×	✓	×	✓	✓	✓	✓	✓	✓	×	×	×	×	✓
[73]	2021	✓	✓	✓	×	×	✓	✓	✓	✓	✓	×	×	×	✓	×	×	✓	✓
This Study	2022	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2 METHODOLOGY

The systematic literature review is the standard protocol-based approach for doing research based on evidence-based results that are easy to understand and reproduce. The Systematic Literature review produces quality research and reporting and facilitates readers by providing quality literature with formally synthesized research results [13]. The approach used in this systematic literature review is specially designed for the computer engineering field, and the main contributors to this effort are Kitchenham and Charter [40]. This systematic literature review is completed by following the given 3 phases, namely 1) the planning Phase, 2) Conducting Phase 3) Reporting Phase. The following section discusses these phases in detail.

2.1 Planning Phase

The Planning phase is the initial phase of the systematic literature review, which consists of initial planning about problem identification, Research Objectives and Research Questions, Search Strategy, Inclusion and Exclusion Criteria, and Quality Assessment Criteria. All the steps followed during the planning phase are mentioned below in this section.

2.1.1 Problem Identification

From the literature review comparison, no study particularly focuses on the Urdu Speech Emotion Recognition review. Many Review papers are already published on Speech Emotion Recognition, but their focus is general. As discussed earlier, the Urdu language is under-studied as a resource-poor language. No systematic review or survey paper has been found in the literature on Urdu Speech Emotion Recognition. Therefore, it is necessary to contribute to the literature on the Urdu Language.

2.1.2 Review Objectives

This systematic literature review is carried out to achieve the following objectives.

1. To explore the existing publicly available datasets that have been used to recognize emotions from Urdu voice data. (RQ1)
2. To explore various Pre-processing techniques that have been used for Urdu SER. (RQ2)
3. To explore various Feature extraction techniques/features used to classify Urdu voice/audio data. (RQ3)
4. To investigate the approaches that have been used in Urdu SER. (RQ4)
5. To investigate the type of Machine Learning (ML) or Deep Learning (DL) approaches used to classify Urdu voice data. (RQ5)
6. To investigate the performance of existing ML or DL classifiers that have been used for Urdu SER. (RQ6)
7. To discuss the validation techniques that have been used for Urdu SER. (RQ7)
8. To investigate the various performance metrics that have been used for the evaluation of the Urdu speech emotion recognition model. (RQ8)
9. To discuss the implementation tools that have been used for Urdu SER. (RQ9)
10. To discuss the challenges and future directions for Urdu SER research. (RQ10)

2.1.3 Research Questions

Following are the research questions designed to be addressed for the Systematic literature review.

- RQ1: What datasets exist to recognize emotions from Urdu Voice Data?
- RQ2: How Urdu voice data has been pre-processed for emotion recognition?
- RQ3: What feature extraction techniques/features have been used for Urdu speech emotion recognition?
- RQ4: What approaches have been used for Urdu speech emotion recognition?
- RQ5: What machine learning or deep learning classifiers have been used for Urdu speech emotion recognition?
- RQ6: What is the performance of existing ML or DL algorithms for Urdu speech emotion recognition?
- RQ7: What validation techniques have been used for Urdu SER?
- RQ8: What performance metrics have been used to evaluate the Urdu speech emotion recognition model?
- RQ9: What implementation tools have been used for Urdu SER?
- RQ10: What are the challenges and future directions for Urdu SER research?

2.1.4 Search Strategy

This review includes studies focusing mainly on the Speech Emotion Recognition of Urdu Audio. Thus, various search keywords are designed accordingly to search related literature. Famous, credible, and high-quality databases, namely WoS (Web of Science) and Scopus are used to retrieve search results. Search Query executed on 17 March at 3:20 PM.

Table 3 shows planning about keywords categorized into different groups and the final search query.

Table 3: Search Keywords and Search Query

Keywords Group	Keywords
Group 1: Keywords that are the focus of the study	Urdu
Group 2: Keywords related to the domain of study	Audio, Speech, Call, Phone, Voice
Group 3: Keywords related to techniques applied	Machine Learning, Deep Learning, Artificial Neural Networks, ANN, CNN, LSTM, RNN., Recurrent Neural Networks, Autoencoder, Transformer, Convolutional Neural Networks
Group 4: Keywords related to approaches used	Emotion, Emotion Recognition, Opinion Mining, Sentiment Analysis, Speech Recognition, Speech Emotion Recognition.
Group 5: Publication Years	2000 to 2022
Group 6: Document Type	Journal and Conference Articles
Group 7: Languages	English
Final Query	(Group 1) AND (Group 2) AND (Group 3) AND (Group 4) AND (Group 5) AND (Group 6) AND (Group 7)

2.1.5 Inclusion & Exclusion criteria

Given Table 4 & Table 5 shows the inclusion and exclusion criteria that are designed for the selection of studies for systematic review purpose.

Table 4: Inclusion Criteria

Inclusion Criteria
1. The included study must have voice data for speech emotion classification purposes.
2. The included study must have voice data in the Urdu Language.
3. The included study must employ either ML or DL to classify Urdu Voice data.
4. The included study must be published between 2000 to 2022.
5. The included study must be either published in Conference or Journal

Table 5: Exclusion Criteria

Exclusion Criteria
1. The study used Urdu voice data only for machine translation, but not for emotion classification will be excluded.
2. Informal studies (not published in well-known journals or conferences) will be excluded.
3. Studies published in a language other than English will be excluded.
4. Studies that are irrelevant to the research questions will be excluded.
5. The study whose full text is unavailable will be excluded.

2.1.6 Quality Assessment Criteria

Quality Assessment Criteria are defined in the planning phase to select final studies. The systematic literature review will add studies that meet the quality assessment criteria. Below given Table 6 shows the quality assessment checklist used as quality assessment criteria for shortlisting the studies for systematic literature review.

Table 6: Quality Assessment checklist

Item #	QA Question	Score	Description
QA1	Are research objectives clearly stated?	0	No, objectives are not stated.
		0.5	Partially, objectives are stated but not clear.
		1	Yes, objectives are clearly stated.
QA2	Is research methodology well defined?	0	No, the research methodology is not well-defined. It needs to go through references.
		0.5	Partially, the methodology is defined but does not mention specific steps for methodology.
		1	Yes, the methodology is well-defined.
QA3	Is there enough information available for the dataset (s) used?	0	No, there is insufficient information for the dataset (s) used.
		0.5	Partially, dataset (s) are given, but the information is incomplete.
		1	Yes, there is enough information about the dataset (s) used.
QA4	Are the pre-processing techniques clearly described with the justification of selection?	0	No, pre-processing techniques are not clearly described.
		0.5	Partially, pre-processing techniques are described, but selection justification is missing.
		1	Yes, pre-processing techniques are clearly described with the justification of selection.
QA5	Are the features used for the SER task mentioned clearly and in detail?	0	No, features used for the SER task are not mentioned in a clear and detailed manner.
		0.5	Partially, features are mentioned but not in a clear and detailed manner.
		1	Yes, features used for the SER task are mentioned in a clear and detailed manner.
QA6	Does the study provide a detailed description of the approach (classifier/ techniques)?	0	No, a description of the approach is not given.
		0.5	Partially, a description of the approach is given, but necessary details of the classifier/ technique are missing.
		1	Yes, the study provides a detailed description of the approach (classifier/ technique).
QA7	Does the study present a detailed evaluation of the proposed approach?	0	No, the study does not present a detailed evaluation of the proposed approach.
		0.5	Partially, the study presents the evaluation of the proposed approach but unclearly.
		1	Yes, the present detailed evaluation of the proposed approach.
QA8	Does the study carry out the comparison of the proposed approach with existing baseline approaches?	0	No, the study does not carry out the comparison of the proposed approach with existing baseline approaches.
		0.5	Partially, a comparison of the proposed approach is made with existing baseline approaches, but it is unclear
		1	Yes, the study carries out the comparison with the existing baseline approach in a clear way.
QA9	Does the study properly interpret and discuss the finding or results, and does the study's conclusion reflect the research findings?	0	No, the study does not properly interpret and discuss the findings or results, and the study's conclusion does not reflect the research findings.
		0.5	Partially, results and findings are mentioned but unclearly, and the study's conclusion does not reflect the research findings.
		1	Yes, the study properly interprets and discusses the findings or results, and the study's conclusion does not reflect the research findings.

2.1.7 Data Extraction Strategy

The data extraction strategy is defined according to the research questions. The data extraction strategy is comprised of 1) Datasets for Urdu SER, 2) Pre-processing techniques, 3) Feature extraction/ Features for Urdu SER, 4) Approaches, 5) Machine Learning and Deep Learning Classifier, 6) Performance of classifiers,

7) Validation techniques, 8) Performance Metrics, 9) Implementation tools, 10) Limitations and Future directions.

2.2 Conducting Phase

The conducting phase of the systematic literature review consists of the execution of search query, screening and selection of studies, Quality Assessment, Data Extraction and Synthesis.

2.2.1 Execution of Search Query

The search query was applied to two major bibliographic databases, i.e., Scopus and WoS. A total of 69 papers were retrieved. The endnote tool saved retrieved studies for further screening and citation management.

2.2.2 Screening & Selection of Studies

The total studies collected from different bibliographic databases were passed to the initial screening process, which will serve to select primary studies. Standard guidelines of PRISMA given in [56] are followed in the screening and selecting of primary studies. Following Figure 2 shows the screening and selection process of retrieved studies.

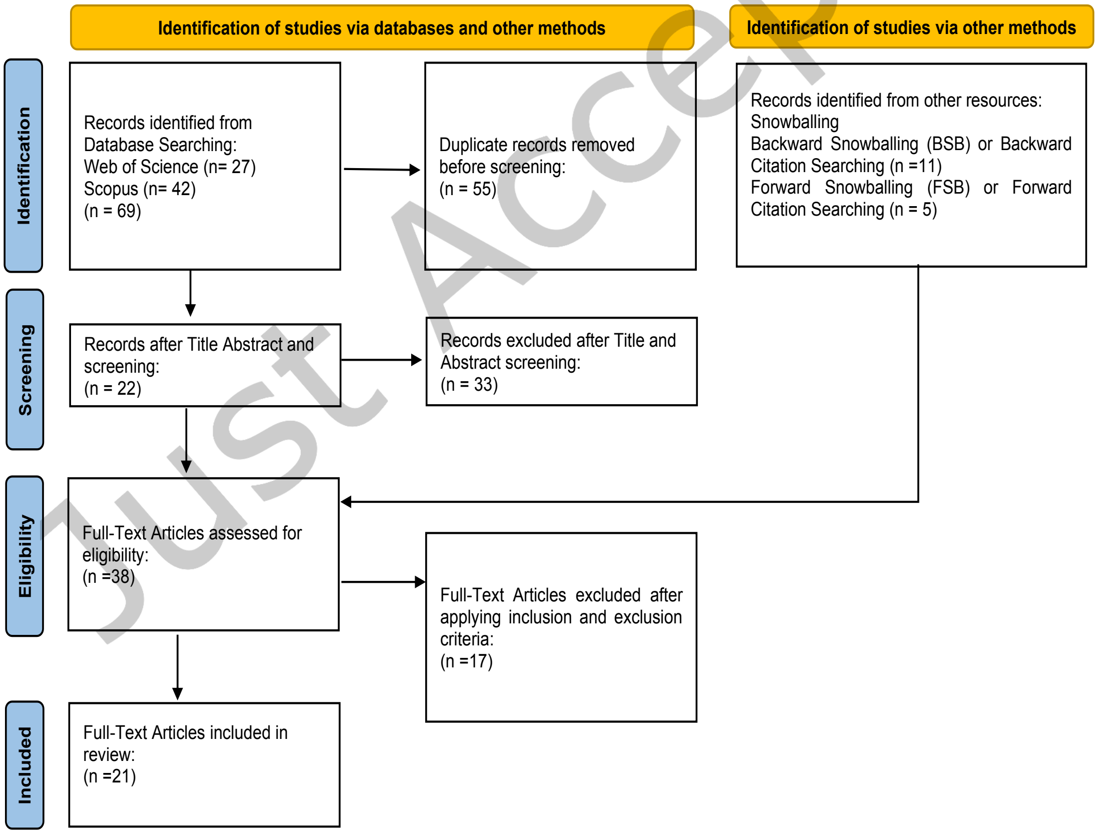
Identification of studies via databases and other methods
Identification of studies via other methods
Identification
Records identified from Database Searching: Web of Science (n= 27) Scopus (n= 42) (n = 69)
Duplicate records removed before screening: (n = 55)
Records identified from other resources: Snowballing Backward Snowballing (BSB) or Backward Citation Searching (n =11) Forward Snowballing (FSB) or Forward Citation Searching (n = 5)
Screening
Records after Title Abstract and screening: (n = 22)
Records excluded after Title and Abstract screening: (n = 33)
Eligibility
Full-Text Articles assessed for eligibility: (n =38)
Full-Text Articles excluded after applying inclusion and exclusion criteria: (n =17)
Included
Full-Text Articles included in review: (n =21)

Figure 2: PRISMA 2020 flow diagram for new systematic reviews, which included searches of databases and other sources.

2.2.3 Quality Assessment of Studies

The quality of 21 studies was assessed against the quality assessment criteria (QAC). Nine Quality Assessment (QA) questions had been defined to determine the quality of research studies and to provide quantitative scoring for the final selection or elimination of studies that don't fulfil the QAC. A set of close-ended questions was designed for quality assessment given in Table 6. The scoring for answering each Quality Assessment question was defined in a given range. The answer to each question can be Yes (Y) = 1 score or Partly (P) = 0.5 scores or No (N) = 0 score. Selected Studies were passed through QAC, which were awarded a quality score. The threshold score for QA was given 6. Table 7 shows the score review studies got when passing through the QAC.

Table 7: Quality Assessment Score

Sr. No	Study Ref	QA1	QA2	QA3	QA4	QA5	QA6	QA7	QA8	QA9	Score
1	[72]	1	1	1	1	1	1	1	0	1	8
2	[8]	1	0.5	1	0	1	1	1	0	1	6.5
3	[41]	1	1	1	1	1	1	1	1	1	9
4	[5]	1	1	1	1	1	1	1	0	1	8
5	[7]	1	1	1	0	1	1	1	0	1	7
6	[6]	1	1	1	0	1	1	1	0	1	7
7	[39]	1	1	1	1	1	1	1	0	1	8
8	[35]	1	1	1	1	1	1	1	1	1	9
9	[47]	1	1	1	0	1	1	1	0	1	7
10	[46]	1	1	1	0	1	1	1	1	1	8
11	[70]	1	1	1	0	1	1	1	0	1	7
12	[9]	1	1	1	1	1	1	1	1	1	9
13	[20]	1	1	0.5	1	1	1	1	0	1	7.5
14	[28]	1	1	1	1	1	1	1	0	1	8
15	[36]	1	1	1	0	1	1	1	0	1	7
16	[45]	1	1	1	1	1	1	1	1	1	9
17	[48]	1	1	1	1	1	1	1	1	1	9
18	[66]	1	1	1	0	1	1	1	1	1	8
19	[74]	1	1	1	1	1	1	1	1	1	9
20	[75]	1	1	1	1	1	1	1	1	1	9
21	[11]	1	1	1	1	1	1	1	0	1	8

From the results, all 21 studies scored 6 and above, so all 21 studies were considered resources for systematic literature review.

2.2.4 Data Extraction and Synthesis

After finalizing the studies for systematic literature review, data extraction and data synthesis from these studies are performed, which will help in reporting phase of the review. The data extracted from 21 primary studies was tabulated and used for synthesis. The data extraction was comprised of 1) Datasets for Urdu SER, 2) Pre-processing techniques, 3) Feature extraction/ Features for Urdu SER, 4) Approaches, 5) Machine Learning and Deep Learning Classifier, 6) Performance of classifiers, 7) Validation techniques, 8) Performance Metrics, 9) Implementation tools, 10) Limitations and Future directions.

2.3 Reporting Phase

A total of 21 studies were selected for systematic literature review. Data gathered from the data extraction and synthesis phase addresses the research questions.

2.3.1 *Addressing the Research Questions*

Section 3 is about review findings; it addresses all the research questions. RQ1 is discussed in section 0, RQ2 is discussed in section 2.5, RQ3 is discussed in section 2.6, RQ4 is discussed in section 2.7, RQ5 is discussed in section 2.8, RQ6 and RQ7 are discussed in section 2.9, RQ8 is discussed in section 2.10, RQ9 is discussed in section 2.11, RQ10 is discussed in section 2.12.Review of Literature

In this section, all the selected primary studies are critically analyzed from given aspects, i.e., 1) Datasets for Urdu SER, 2) Pre-processing techniques, 3) Feature extraction/ Features for Urdu SER, 4) Approaches, 5) Machine Learning and Deep Learning Classifiers, 6) Performance of classifiers, 7) Validation techniques, 8) Performance Metrics, 9) Implementation tools, 10) Limitations and Future directions.

2.4 Datasets for Urdu SER (RQ1)

This section discusses the characteristics of Urdu speech datasets. Urdu speech datasets have different characteristics like 1) Corpus Type, 2) Databases, 3) Emotional Model, 4) Experiment Type, and 5) Modality. Figure 3 visualizes the taxonomy for the characteristics of Urdu Speech Datasets.

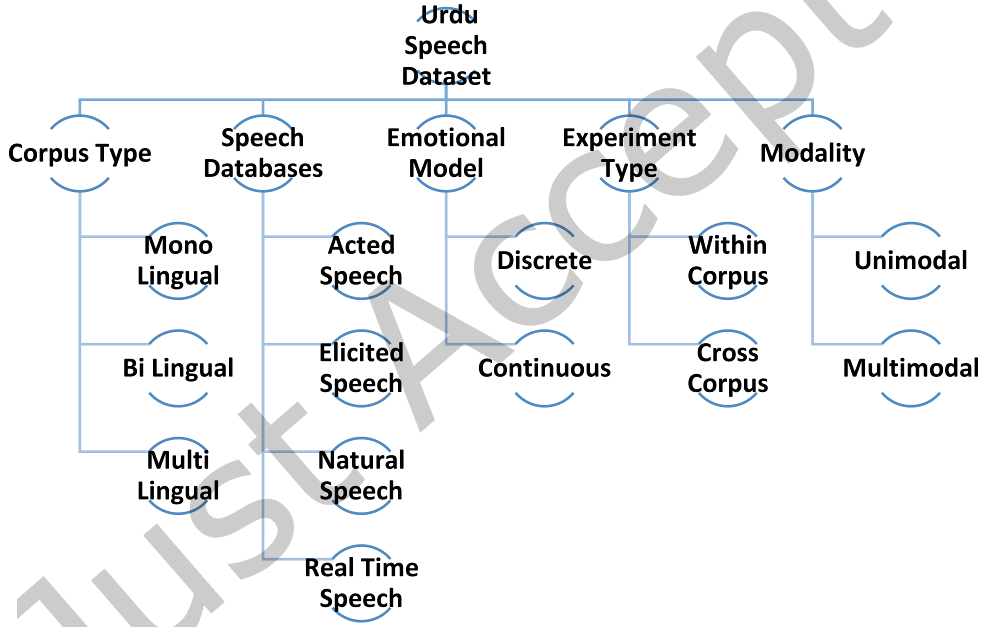

Figure 3: Urdu Speech Dataset CharacteristicSpeech Corpus Type

Generally, Urdu Speech Emotion Recognition datasets are Monolingual Speech Corpus means constituting only one language. Sometimes researchers create a customized corpus that is a mixture of two language speech corpora called Bilingual Speech Corpus. Some datasets are a mixture of more than two languages' speech corpora that we call Multilingual Speech Corpus.

2.4.1 *Speech Database*

It is necessary to decide what kind of database will be used to develop a speech-emotion recognition system. Here by the database, we mean the speech corpora. The speech corpora for emotion recognition are categorized into 3 classes. The categorization of speech databases is based on gathering strategies for developing speech corpora. Speech corpora can be developed using simulated or acted way, and speech corpora can be elicited (Induced) or natural (Spontaneous). The categorization of Speech Emotion Databases found in literature, along with strengths and weaknesses, is summarized and analyzed in Table 8.

1. Acted Speech
   Professional actors or artists are crowd-sourced for the development of acted speech corpora. Professional actors or artists are asked to utter a scripted speech with different emotions. This type of speech is also known as simulated speech.
2. Elicited Speech
   For the development of elicited speech, corpora participants or subjects are put in a situation where they have to express their real emotional attitude using speech. This type of speech is also known as induced speech, as it is developed using artificially created induced situations. With this approach, close to real natural speech can be recorded with control over the lexical and emotional content.
3. Natural Speech
   Natural or spontaneous speech contains realistic and natural emotions. Here speaker expresses the emotions more naturally. T.V. talk shows, interviews, radio shows, call centres, and YouTube video data are the best ways to gather natural speech. Natural speech may contain imbalanced emotional categories.
4. Real-Time Speech
   Processing real-time speech is the ultimate success criterion to be achieved in speech emotion recognition. When the speech emotion recognition system becomes robust enough to recognize emotions from real-time speech, there would be no delay for speech emotion recognition for each single second speech emotion recognition will serve to find emotions for real-time speech. From the systematic literature review, it is found that no research covers Urdu speech emotion recognition for real-time speech.

Table 8: Speech Databases Comparison

Sr. No	Speech Database	Study Ref	Strengths	Weakness
1	Acted Speech	[8], [41], [5], [7], [6], [39], [35]	• Lab Environment is used to record the speech, so acted speech is noise free. • All the emotions are present in acted speech. • It is balanced in nature. • Easy to be analyzed for SER task. • It is easy to process acted speech. • Features can be easily extracted.	• Lack of naturalness • Costly because the lab environment and actors or mature artists are needed to develop the speech database. • Context may be absent.
2	Elicited Speech	[72]	• It is close to natural speech. • Context is available. • Having a medium level of difficulty with SER task	• Participants may not be fully expressive if they know they are being recorded. • All emotions may not be present.
3	Natural Speech	[47], [46], [9], [20], [28], [36], [48], [66]	• Rich source of real emotions. • It can be found easily without paying the extra cost of actors and lab environment setup. • Tv talk shows, radio transmissions, and online reviews provide a natural speech. • Context is present	• It isn't easy for a model to analyze natural speech. • Copyright and privacy issues should be considered. • Special permissions are required to use natural speech available online. • Ethical issues to be considered. • All emotions may not be present. • Emotion categories are imbalanced. • Presence of background noise.
4	Real-Time Speech	No Study Found	• Real-time speech is full of natural emotions and can be utilized for real-time SER systems, which take speech in real-time and analyze this speech without any delay	• Presence of background noise. • All emotions may not be present. • The difficulty level is high for real-time speech to perform SER tasks on it.

2.4.2 Emotional Model

To implement the SER system, it is necessary to develop its emotional model first. For this, we need to define emotions. It is estimated that emotions have more than 90 plus definitions. Authors in [61] defined emotion as "An emotion is not simply a feeling state. Emotion is a complex chain of loosely connected events that begins with a stimulus and includes feelings, psychological changes, impulses to action and specific, goal-directed behaviour. That is to say, feelings do not happen in isolation. They are responses to significant situations in an individual's life, and often they motivate actions." Based on several definitions, two emotional models are considered widely for speech emotion recognition tasks. These are 1) the Discrete Emotional Model and 2) the Dimensional Emotional Model.

1. Discrete Emotional Model
   Most of the research on SER uses a discrete emotional model, constituting the following six emotions identified by Ekman [24], i.e., happiness, sadness, surprise, fear, anger and disgust. These emotions are now universally recognized. This model is also known as the basic emotion set or the big six emotional states.
2. Dimensional Emotional Model
   The dimensional or continuous emotional model consists of two dimensions, i.e., 1) Valence (positive/pleasurable or negative/unpleasurable or it indicates the pleasantness of the voice ranging from unpleasant (E.g., sad, fear) to pleasant (E.g., happy, calm)), 2) Arousal(engaged or not engaged or it is the level of reaction to stimuli and range from inactive (E.g., sleepy, sad) to active (E.g., anger, surprise)) [65],[32].

2.4.3 Experiment Type

Studies selected for the literature review are mainly based on two types of experiments on Urdu Speech data sets, i.e., within-corpus experiments and cross-corpus experiments. In within-corpus experiments, authors have trained and evaluated on the same speech dataset. And in cross-corpus experiments, authors have trained the model on different datasets and evaluated the model by giving a new dataset. The reason for performing cross-corpus experiments is to check the generalizability capability of the SER model for unseen data.

1. Within-Corpus SER
   Within-corpus is the traditional approach for SER in which the same speech corpus is used for training and testing the classifier for emotion detection. Within the corpus training scheme, the SER model is trained on a single corpus, and its performance is evaluated in-domain test set (training data from the same corpus). The main drawback of this approach is that it cannot be generalized for environments with multiple languages, which is the need of today's SER system, especially to train robots or HCI (Human Computer Interaction) systems in a way that people across the world can easily interact with them irrespective of their language. Here in Figure 4, the model for within-Corpus SER is shown. According to this model, Speech Corpus C1 is given to the model for training; pre-processing is performed on training samples from Speech Corpus C1. After pre-processing, features are extracted, and the model is trained based on feature vectors for speech emotion recognition tasks. Once the model is trained, its performance is evaluated by giving test samples from the same dataset or speech corpus.

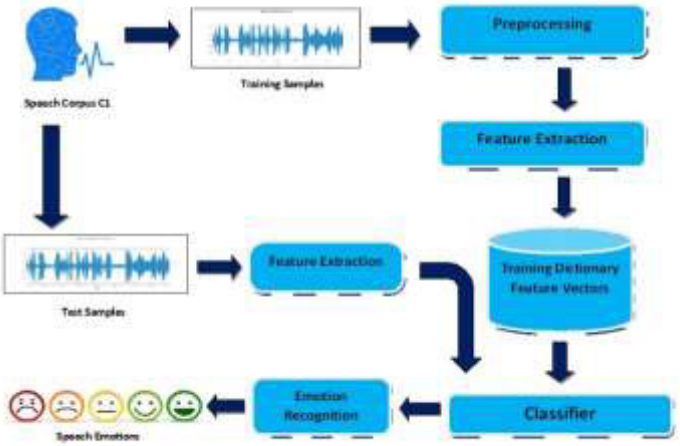

Figure 4: Within-Corpus SER

2. Cross-Corpus SER

   Diversity is observed in human speech when expressed in different languages regarding emotional expression. This diversity in human speech leads to the need for a generalized SER model, which we may call as Cross-Corpus SER model. Cross-Corpus SER model is created by combining several emotional speech corpora within the training. A new speech corpus is given to the model to check its ability to recognize emotions. Here in Figure 5, the model for Cross-Corpus SER is shown. Cross-corpus speech emotion recognition is utilized to build a classifier that can be generalized for applications and acoustic input. It is highly useful for developing practical emotion recognition systems that can adopt language differences by acquiring training from previous languages on which it is trained. Research has shown cross-corpus emotion recognition to be challenging for several reasons, like differences in signal level, type of emotion elicitation, data scarcity, etc. It is tested on out-of-domain corpora means corpora that are not part of the training set to check their performance in a mismatched situation.

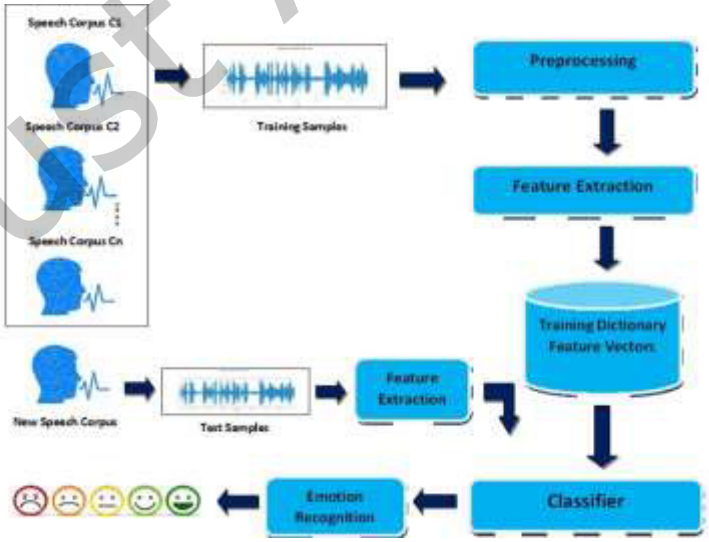

Figure 5: Cross-Corpus SER

Most of the authors for Urdu SER performed both experiments. First, they performed the Within-Corpus experiment for SER, in which the same speech corpus was used to train and evaluate the model. After getting good results in speech emotion recognition, they also trained a model for cross-corpus SER. The model was trained on multiple corpora and evaluated on new speech corpora. For the case of the Urdu Language, the cross-corpus experiment improved its performance as the cross-corpus experiment model was first trained on high-resource languages and evaluated on Urdu [75]. Authors in [5] also suggest using the 1 vs 1 and N vs 1 schemes for cross-corpus experiments.

- 1 vs 1 scheme
  In this scheme, two languages are considered at a time; one language is used for training, and the other language is used for testing the model for speech emotion classification.
- N vs 1 scheme
  Considering the total N languages from which one language is used for training, the remaining languages are used for testing the model for speech emotion classification.

Table *9* shows the strengths and weaknesses of within-corpus and cross-corpus experiment types for speech emotion recognition.

Table 9: Experiment Types for SER

Sr. No	Speech Database	Strengths	Weaknesses
1	Within-Corpus	• Performs well in matched conditions when a test set is given from the same corpus.	• Lack of generalizability when the model is retrained on out-of-domain data. • Poor performance when the model deployed to a real-time setting due to limited generalizability and mismatched conditions means when test set given from different corpus.
2	Cross-Corpus	• It improves the generalization and performance of the model because various speech corpora are combined for training the model.	• Overfitting may occur as the model is well-trained on various speech corpora. • It may display poor performance when given out of domain corpus. • Various speech corpora have distinct emotion classes, so there can be an issue of which emotions to choose and which to discard. Emotions are often mapped into three categories, i.e., positive, negative, and neutral.

2.4.4 Modality

Speech Datasets for emotion recognition can have different modalities integrated into them apart from audio and speech datasets that can have visual data or video and text data. Therefore, this can be taxonomised that speech datasets can be Multi-modal or Unimodal. Multi-modal means datasets with audio video or text modality, whereas unimodal mean datasets with speech data only. Table 10 summarises datasets used for Urdu Speech Emotion recognition and their different characteristics found in the literature.

Table 10: Datasets used for Urdu SER

Meaning of acronyms: Dur - Duration, Spk - Speaker, Exp – Experiment, S.R. – Sampling Rate, M - Male, F – Female, C - Children, NA – Not Available

Dataset Name	Ref	Exp Type	Emotion Model	Corpus Language(s)	Type of speech	Modality (Audio, Video, Text)	Dur	Access	Emotion Classes	Samples	Spk	S.R. kHz
Not specified	[72]	With in-corpus	Discrete	Multi-lingual (English, Mandarin, Urdu, Punjabi, Persian, and Italian)	Acted	Multi-modal Audio, Video	5 Sec	NA	6 Emotion Classes: Anger, Disgust, Fear, Happiness, Sadness, Surprise.	500	8	22.05 kHz
Not Specified	[8]	With in-corpus	Discrete	Multi-lingual (Sindhi, Urdu, Punjabi and Pashto)	Acted	Unimodal Audio	-	NA	4 Emotion Classes: Anger, Sadness, Comfort and Happiness.	404	10 - 5 M 5 F	48 kHz
RML	[41], [5], [35]	With in-corpus & Cross-corpus	Discrete	Multi-lingual (English, Mandarin, Urdu, Punjabi, Persian, Italian)	Acted	Multimodal Audio Video	3-6 sec	Free http://www.rml.ryerson.ca/rml-emotion-database.html	6 Emotion Classes: Anger, Disgust, Fear, Happiness, Sadness, Surprise	720	8 M	22.05 kHz
Not Specified	[7]	With in-corpus	Discrete	Multi-lingual (Urdu, Pashto, Punjabi, Sindhi, and Balochi)	Acted	Unimodal Audio	-	NA	4 Emotion Classes: Anger, Happiness, Neutral and Sad	40	2 - 1 M 1 F	48 kHz
Not Specified	[6]	With in-corpus	Discrete	Monolingual Urdu	Acted	Unimodal Audio	-	NA	4 Emotion Classes: Angry, Happy, Neutral and Sad	24	6 C	48 kHz
Not Specifi	[39]	With in-	Discrete	Monolingual	Acted	Unimodal	-	NA	4 emotion	40	10 C	48 kH

Dataset Name	Ref	Exp Type	Emotion Model	Corpus Language(s)	Type of speech	Modality (Audio, Video, Text)	Dur	Access	Emotion Classes	Samples	Spk	S.R. kHz
ed		corpus		Urdu		Audio			classes: Happy, Sad, Angry, and Neutral			z
URDU [47]	[47], [46], [9], [20], [28], [36], [45], [48], [66], [75]	Within-corpus & Cross-corpus	Discrete	Monolingual Urdu	Natural	Unimodal Audio	-	Free https://github.com/siddiquelatif/URDU-Dataset	4 Emotion Classes: Angry, Happy, Sad, and Neutral	400	38 - 27 M 11 F	44.1 kHz
Urdu-Sindhi Speech Emotion Corpus	[70]	Within-corpus & Cross-corpus	Discrete	Multi-lingual (Urdu-Sindh)	Acted	Unimodal Audio	-	https://zenodo.org/record/3685274#.YsG-ZmBByPk Only feature set available under CC License.	7 Emotion Classes: Anger, Disgust, Happiness, Neutral, Sarcasm, Sadness and Surprise	1,435	2	-
SEMOUR [74]	[74]	Within-corpus	Discrete and dimensional	Monolingual Urdu	Acted	Unimodal Audio	-	http://acoustics-lab.itu.edu.pk/semour/	8 Emotion Classes Anger, Boredom, Disgust, Fearful, Happiness, Neutral, Sadness, Surprise	15,040	8 - 4 M 4 F	48 kHz
Not	[11	With	Discrete	Monolin	Acte	Unimod	-	NA	5	2,500	20	44.

Dataset Name	Ref	Exp Type	Emotion Model	Corpus Language(s)	Type of speech	Modality (Audio, Video, Text)	Dur	Access	Emotion Classes	Samples	Spk	S.R. kHz
Specified [11]	]	in-corpus		gual Urdu	d	al Audio			Emotion Classes Angry, Happy, Neutral, Disgust, and Sad		- 10 M 10 F	1 kHz
BAUM-1 [76]		Cross-corpus	Discrete	Monolingual Turkish	Acted and Natural	Multi-modal Audio, Video	1.82 sec	Free https://archive.ics.uci.edu/ml/datasets/BAUM-1	6 Emotion Classes: Happiness, Surprise, Sadness, Fear, Anger, and Disgust.	1184 total, 521 selected	31 - 18 M 13 F	48 kHz
Berlin Emotional Database (EmoDB) [16]	[41], [9], [46], [47], [20], [36], [48], [66], [75]	Cross-corpus	Discrete	Monolingual German	Acted	Unimodal Audio	1 to 4 sec	Free http://emodb.bilderbar.info/index-1280.html	7 Emotion Classes: Anger, Boredom, Disgust, Fear, Joy, Neutral, and Sadness	535	10 - 5 M 5 F	
SAVEE [33]	[47], [46], [9], [20], [36], [66], [75]	Cross-corpus	Discrete	Monolingual English	Acted	Multimodal Audio Video	-	Free http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/	6 Emotion Classes: Anger, surprise, Disgust, Fear, Happy, Sad, Neutral	480	4 M	44.1 kHz
EMOVO	[47],	Cross-	Discrete	Monolingual	Acted	Unimodal	-	Free http://voice.fub.it/EMOVO	7 Emotion	588	6 - 3	48 kH

Dataset Name	Ref	Exp Type	Emotion Model	Corpus Language(s)	Type of speech	Modality (Audio, Video, Text)	Dur	Access	Emotion Classes	Samples	Spk	S.R. kHz
[22]	[46], [9], [20], [66], [75]	corpus		Italian		Audio			Classes: Anger, Surprise, Disgust, Fear, Joy, Sad, Neutral		M 3 F	z,
RAVDEES [49]	[9], [20], [28], [36], [66],	Cross-corpus	Discrete	Monolingual English	Acted	Multi-modal Audio, Video	-	Open under CC license https://zenodo.org/record/1188976	8 Emotion Classes Anger, Surprise, Disgust, Fear, Happy, Sad, Calm, Neutral	1440	24 - 12 M 12 F	48 kHz
TESS [23]	[20], [36]	Cross-corpus	Discrete	Monolingual English	Acted	Unimodal Audio	-	Free https://tspace.library.utoronto.ca/handle/1807/24487	7 Emotions Classes: Anger, Disgust, Fear, Happiness, Pleasant Surprise, Sadness, and Neutral	2800	2	
NSSED [45]	[45]	within-corpus	Discrete	Monolingual Sindh	Acted	Unimodal Audio	3-10 sec	NA	4 Emotion Classes Happy, Sad, Angry, and Neutral	1231	29 - 16 M 13 F	16 kHz
IEMOCAP [17]	[48]	Cross-corpus	Discrete and Continuous	Monolingual English	Elicited	Multi-modal Audio, Video, Text		Free https://sail.usc.edu/iemocap/iemocap_info.htm	9 Emotion Classes Angry, Excited, Fear, Sad, Surprised,	5531	10 5 M 5 F	16 kHz

Dataset Name	Ref	Exp Type	Emotion Model	Corpus Language(s)	Type of speech	Modality (Audio, Video, Text)	Dur	Access	Emotion Classes	Samples	Spk	S.R. kHz
									Frustrated, Happy, Disappointed, and Neutral Continuous Emotions: Valence, Activation, Dominance			
CREMA-D [18]	[66]	Cross-corpus	Discrete	Monolingual English	Acted	Multi-modal Audio, video	-	Available under Open Database License https://github.com/CheyneyComputerScience/CREMA-D	6 Emotion Classes: Anger, Disgust, Fear, Happy, Neutral, and Sad and four different emotion levels (Low, Medium, High, and Unspecified).	7,442	91	48 kHz
ShEMO [52]	[66]	Cross-corpus	Discrete	Monolingual Persian	Semi Natural	Unimodal Audio	-	Free https://www.kaggle.com/datasets/mansoourehk/shemo-persian-speech-emotion-detection-database	5 Emotion Classes Anger, Fear, Happiness, Sadness and Surprise and Neutral	3000	87 56 M 31 F	44.1 kHz

2.5 Pre-processing Techniques used for Urdu Speech (RQ2)

The input speech used for the speech emotion recognition task is filled with noise, and sometimes there can be the presence of silence in a frame that does not contribute words or any useful information, so this decreases the performance of the SER. model. The first step after collecting input speech is to pre-process it. Pre-processing step plays an important role in the SER task as Pre- Processing helps to improve speech quality by removing noise. It is an important step to be performed before feature extraction. The speech pre-processing starts the noise removal or noise reduction step. After removing noise from the speech, pre-emphasis is performed to balance its frequency spectrum. The general pipeline for pre-processing of the speech signal is shown in Figure 6.

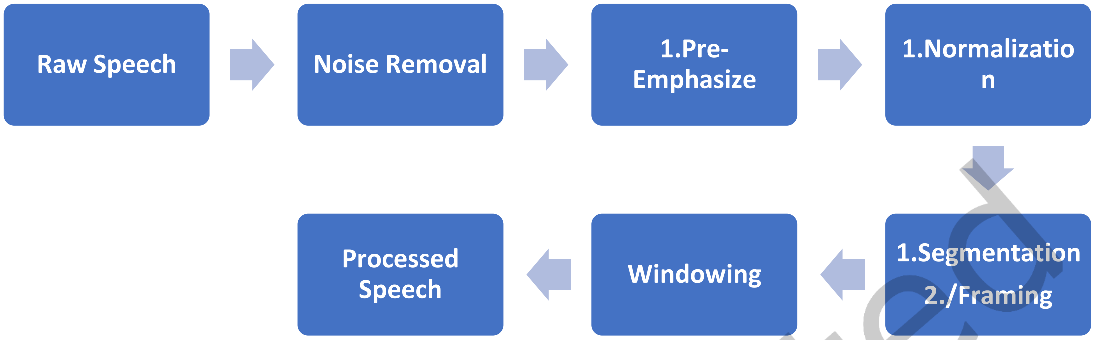

Figure 6: General Pipeline for Speech Pre-processing

Following are some important pre-processing techniques found in the literature.

1. Noise Reduction using thresholding of the wavelet coefficients.
   The Speech data generally contains noise due to the presence of background noise while recording, or it may be the hiss of the recording machine. The presence of noise in speech signals will degrade the overall performance of the SER system. For this reason, performing de-nosing on the speech signal is necessary. A well-known technique for de-nosing the speech signal is to use wavelet transform [3]. Authors in [11] removed noise and silent parts manually. Authors in [72] have used the wavelet transform by thresholding the wavelet coefficients to de-noise the speech signal.
2. Noise Reduction using Wiener Filter
   The noise reduction problem can be solved using the Wiener Filter approach. Literature suggests that the wiener filter is among the optimal approaches for noise reduction [19]. Authors in [5] have used the Wiener Filter approach to reduce the noise present in speech recordings.
3. Pre- Emphasize
   Before feature extraction, it is necessary to apply the pre-emphasized filter on speech signals to balance the frequency spectrum. Normalization can also achieve this [9],[11].
4. Normalization
   In this stage, the maximum value of the speech signal is calculated, and then the whole speech signal sequence is divided by this maximum value [54]. Authors in [75] and [11] have performed normalization on collected speech samples.
5. Segmentation
   Segmentation is the process of partitioning the continuous speech signal into fixed-length segments. In [41], speech signals were segmented using R.T.I. (Relative Time Interval Technique). According to this automatic technique, speech signals were segmented at fixed relative positions, for example, halves, thirds etc. Here authors performed a blind segmentation approach in which no prior delimitation word or syllable boundary is performed. Authors considered that each segment would serve to highlight the emotional primitive that will help in speech emotion recognition. According to the author, speech segmentation helps find segment-level features that are better and more helpful than utterance-level features. Authors also argued that segmentation helps to save time because no

extra computation is required to identify a word or syllable boundaries for feature extraction. Authors also suggest automatic segmentation makes speech emotion recognition more practical and is suitable for real-time speech processing and stream analysis.

6. Framing

   Framing is the method of breaking the continuous speech signal into a series of fixed-length blocks. Framing of speech signals facilitates the block-wise processing of speech signals [44].

7. Windowing

   Windowing is a classical speech signal processing technique that is used to segment speech signals into meaningful speech frames [10]. The windowing technique is applied on the individual frame to minimize the noise, remove discontinuities at the frame boundaries and retain the information in the speech signal. Windowing function Popular windowing techniques like Hamming Window were used by [72],[11].

8. Data Augmentation

   Data augmentation (DA) can be used for an overfitting problem generally found in small datasets. Data augmentation is a technique used to artificially generate data by applying transformations to the existing data to increase the size of the dataset. Authors have used data augmentation to the extracted spectrograms and facial images to deal with the overfitting problem [35]. Also, the author in [45] has used the DA technique to create additional data samples by applying various deformations to the original training dataset. These authors have also discussed different DA methods that they have applied to speech signals, i.e., Time Stretching in which audio data is augmented by varying the length of audio waveforms while keeping other attributes constant, Speed and Pitch in this DA method values of pitch and speed factor are randomly changed to get the augmented audio samples, Noise DA using additive white Gaussian noise (A.W.G.N.) is performed in which noise signal with zero mean value is added to the audio for data augmentation.

2.6 Feature Extraction/ Features used for Urdu SER (RQ3)

Speech signal contains many parameters that reflect emotional characteristics expressed in speech, known as speech features. Speech features can be prosodic, spectral or voice quality, and there are further subclasses. The main Features of Urdu SER are given in Figure 7. This section also briefly discusses the prosodic, spectral and voice quality features and toolkits used in literature for the feature extraction task.

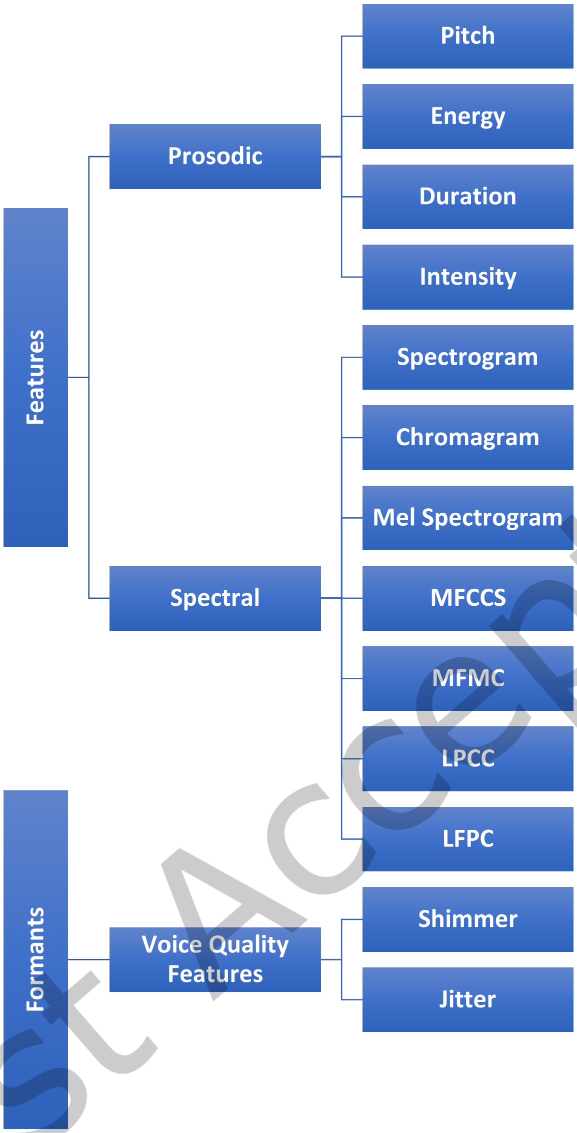

Figure 7: Types of Speech Features

- Prosodic Features
  Prosodic features are the basic features of the speech signal that the human ear perceives. Prosodic features depend on the intonation (which means the rise and fall of the human voice in speaking). The most used prosodic speech features for speech emotion recognition are 1) Pitch, also known as a glottal waveform or fundamental frequency. The pitch depends on the tension of vocal folds and subglottal air pressure. The pitch signal is the result of the vibration of the vocal fold. It contains information about emotions hidden in the speech signal; 2) Energy is also known as power. It is defined as the intensity of voice signal that can be physically detected through the pressure of sounds; 3) Duration, which can be defined as the continuous-time elapsed between the start and end of an emotional sentence, 4) Intensity which can be defined as the amount of energy in a human voice.
- Spectral Features

Spectral features are based on the energy content of different frequency bands in the speech signal. Some most important and widely used speech features are 1) spectrogram, 2) Chromogram, 3) Mel Spectrogram, 4) MFCCs (Mel-frequency Cepstral Coefficient) is among the popular features used for the SER as it mimics the human auditory system for emotion classification, 5) MFMCs (Mel-frequency Magnitude Coefficients), 6) LPC (Linear Predictive Coding), 7) LPCC (Linear Prediction Cepstral Coefficient), 8) LFPC (log frequency power coefficient)

- Voice Quality Features

  Voice Quality Features are used to improve the speech-emotion accuracy of the model. Some commonly used voice quality features are 1) Formants, defined as the concentration of sound energy at a certain frequency in a speech signal. Formants frequencies are important in finding the phonetic content present in the speech signal. Authors in [72] have used formant frequencies as features for speech emotion recognition, 2) Shimmer, and 3) Jitter.

After pre-processing speech signal is subjected to feature extraction to remove irrelevant information and to extract meaningful information that will further serve in speech emotion recognition, here speech signal is divided into segments or frames, which are the meaningful units of information that will serve for speech emotion recognition. Authors in [25], [42] and [30] have emphasized to use of prosodic and spectral features because these features have proved to serve better emotion classification results without the need to use any deep learning technique. In literature, prosodic features like pitch, energy, duration, and intensity are proven to be highly correlated with speech emotions. MFCCs (Mel frequency cepstral coefficients) and the LPCCs (Linear prediction cepstral coefficients) are also widely used speech features for Urdu speech emotion recognition.

Some authors like [46], [70] also emphasize to use of already created feature sets like the Prosody feature set, IS09-Emotion feature set, IS10-Paralinguistics feature set, ComParE feature set and eGeMAPS feature set for speech emotion classification task as these feature sets proved their robustness by combining different features.

- Feature Extraction Toolkits/ Libraries

Most authors have used advanced toolkits/ libraries to extract useful features from speech. Following Table 11 gives a brief overview of Feature Extraction Toolkits/ Libraries used in literature.

Table 11: Feature Extraction Tools used for Urdu SER

Sr. No	Feature Extraction Tools	Purpose	Ref Used
1	PRAAT software [14]	PRAAT is a speech statical analysis software used to analyze, synthesize, and manipulate speech.	[7], [8], [39]
2	OpenSmile toolkit [26]	openSMILE (open-source Speech and Music Interpretation by Large-space Extraction) is an open-source toolkit for feature extraction from speech. It is also used for audio classification and speech emotion recognition tasks.	[46], [47]
3	Librosa Python library [51]	Librsoa is a python library for audio processing and speech analysis. It is to capture information from audio, especially features hidden in audio.	[45], [66], [74]
4	Audio toolbox MATLAB [78]	Audio toolbox is a powerful MATLAB tool containing audio processing and speech analysis algorithms. It is also helpful for feature extraction of speech.	[5], [28]

2.7 Approaches for Urdu SER (RQ4)

This section discusses the main approaches used for Urdu SER. Machine and Deep Learning approaches are generally employed for Urdu Speech Emotion Recognition.

- Machine Learning Approach

Mostly machine learning approach is used for speech-emotion recognition. According to this approach, the algorithm learns the speech emotion classification task from training speech samples. Then it uses these training speech samples to classify emotions from the new speech samples given to it. Figure 8 depicts the working flow of the traditional machine learning approach for speech emotion recognition. Using the machine learning approach, the speech emotion recognition process starts with pre-processing the input speech signal; after pre-processing, the features are extracted from the speech signal and on selected features machine learning algorithm is applied, and speech emotions are recognized.

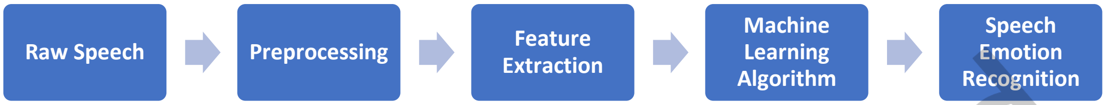

Figure 8: The working flow of the Machine Learning Approach for SER

- Deep Learning Approach

  Deep learning is becoming a suitable approach for SER compared to machine learning. Due to several advantages, deep learning is gaining attention instead of traditional approaches, including its learning and processing capability for raw and unlabeled data, and there is no need for manual feature extraction when using deep learning methods. Figure 9 shows the working flow of the deep learning approach for speech emotion recognition.

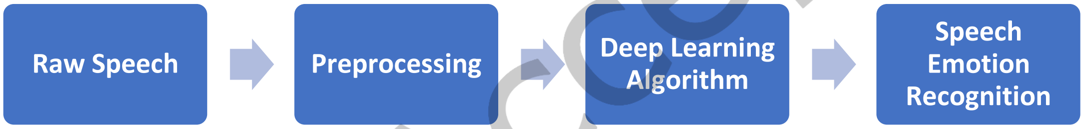

Figure 9: The working flow of the Deep Learning Approach for SER

2.8 Classifiers & their Performance for Urdu SER Research (RQ5-RQ6)

For the development of SER, it is necessary to consider these given factors 1) the good speech emotion dataset that is balanced in nature in terms of emotions and rich in lexical information and has enough speakers, 2) extraction of essential features, 3) choice, of the reliable classifier as speech emotion recognition is considered as a classification problem as there is not any rule of thumb for the choice of the classifier. Every classifier has its own advantages and limitations. This section of the systematic literature review highlights the three factors mentioned above. A summary of current studies on Urdu SER is made according to the datasets, features, approach, classifier, results, and limitations are given in Table 12.

Table 12: Summary of classifiers & their performance

Ref	Dataset(s) used	Features	Appr	Classifiers	Eminent Classifier	Results	Limitations
[72]	Not Specified	Prosodic, MFCCs, Formant.	ML	GMM (Gaussian Mixture Model), KNN (K-Nearest Neighbour), N.N. (Neural Network) and Proposed Classifier model FLDA (Fisher's	FLDA	Accuracy 82.14%	• The dataset is not available for future research. • The model is based on audio-visual modalities, which makes it complex. • Overall model performance can be improved.

Ref	Dataset(s) used	Features	Appr	Classifiers	Eminent Classifier	Results	Limitations
[8]	Not Specified	Intensity, Pitch, Formants	ML	Linear Discriminant Analysis) MLP (Multi-Layer Perceptron), Naive Bayes, J48, and S.M.O. (Sequential Minimal Optimization)	J48	Accuracy 75%	• The dataset is not available for future research. • The authors have designed only one carrier sentence for the experimental purpose, which is insufficient for emotion recognition results and classifier performance.
[41]	RML	Pitch, Energy, MFCCs, LPCC	ML	Ensemble Classifiers, i.e., Random Forest (RF) and Kernel Factory (KF)	KF	Accuracy 78.36%	• Authors have focused on just segment-level feature extraction from speech signals, though, which is better than utterance level, but the latest feature extraction techniques emphasised on frame level feature extraction as the speech frames act as atoms of emotions from which we can identify emotional primitives that could be used to deduce the emotion of an utterance or speech signal.
[5]	RML	MFCCs	ML	SC (Standard Classifier), EPC (Emotional Profile Classifier), EC (Ensemble Classifier), and EEPC (Ensemble Emotional Profile Classifier)	EEPC	Within-corpus Accuracy using SVM (Avg baseline Accuracy 70.58%) Cross-corpus Accuracy 1 vs 1 using S.C. (Avg baseline Accuracy 45%) Cross-corpus Accuracy N vs 1 using E.E.P.C. (Avg baseline Accuracy 58.1%)	• The performance of the proposed classifier is very low; it could be further improved.
[7]	Not Specified	Pitch, Intensity & Formant	ML	AdaboostM1, classification via regression, Decision stump, J48 with Pitch, Intensity, Formant, and their combination.	All classifiers performed well.	Accuracy 43%	• The dataset is not available for future research. • Only four basic emotions used in this study can be further extended. • Model performance is too low. • The emotion set included just 4 emotions. This can be further extended to the standard emotion set.
[	Not	Rasta-PLP &	ML	K-star, Logit boost,	All classifiers	Accuracy 100%	• The dataset is not available

Ref	Dataset(s) used	Features	Appr	Classifiers	Eminent Classifier	Results	Limitations
6]	Specified	MFCCs		J48, LAD Tree and Random Forest with MFCCs., Rasta PLP and their combination.	performed well.		for future research. • Authors have just focused on spectral features and have investigated the effect of emotion and spectral features (MFCCs and Rasta PLP) on spoken utterances in the Urdu language, as many studies analysed both spectral and prosodic features for emotion recognition and classifier performance and analysed the effect of both spectral and prosodic features. • The authors have designed only one carrier sentence for the experimental purpose, which is insufficient for emotion recognition results and classifier performance. • The emotion set included just 4 emotions. It can be further extended to the standard emotion set.
[39]	Not Specified	Pitch, Intensity & Formant	ML	ANN (artificial neural networks), N.B. (Naive Bayes), D.S. (Decision Stump), J-48, K-Star with Pitch, Intensity, Formant, and their combination.	ANN	Accuracy 45%	• The dataset is not available for future research. • Authors have just focused on prosodic features of speech, but there is also a significant effect of using spectral features like MFCCs on classifier performance for speech emotion recognition; classification accuracy is also 45% which can be further improved. • The authors have designed only one carrier sentence for the experimental purpose, which is insufficient for emotion recognition results and classifier performance. • The emotion set included just 4 emotions. It can be further extended to the standard emotion set.
[35]	RML BAUM-1s	-	DL, ML	A-V emotion recognition model developed using CNN (Convolutional Neural Network),	-	Within-corpus Accuracy RML 82.97%, BAUM-1s 56.01%, Cross-corpus	• Despite model competitiveness for audio-visual emotion recognition, the model misclassified emotions as fear and

Ref	Dataset(s) used	Features	Appr	Classifiers	Eminent Classifier	Results	Limitations
				DNN (Deep Neural Network), LSTM (Long short-term memory) and RL (Reinforcement Learning)		Accuracy Training RML Testing BAUM-1s 38.00% Training BAUM-1s Testing RML 50.80% Training RML & BAUM-1s Testing RML 78.45%, Testing BAUM-1s 54.53%	sadness. • The proposed architecture is highly complex due to its multi-modal nature, which means audio and video are analysed for speech emotion recognition.
[47]	URDU EMOVO SAVEE EMO-DB	eGeMAPS	ML	SVM (Support Vector Machine)	-	Within-corpus Accuracy URDU 83.40%, EMOVO 74.01%, SAVEE 65.10%, EMO-DB 81.30%. Cross-corpus Accuracy Training - URDU Testing EMOVO 43.75%, SAVEE 50.87%, EMO-DB 55.12%. Training - EMOVO Testing - URDU 45.86%, Training - S.A.V.E.E. Testing - URDU 40.10%, Training - EMO-DB Testing - URDU 57.87%.	• Baseline, within-corpus and cross-corpus accuracies can be further improved by employing deep learning methods.
[46]	URDU EMOVO SAVEE EMO-DB	eGeMAPS	ML	The model implemented using SVM (Support Vector Machine), GAN (Generative Adversarial Networks)	-	Within-corpus Accuracy URDU 83.40%, EMOVO 74.01%, SAVEE 65.10%, EMO-DB 81.30%. Cross-corpus Accuracy Training - URDU Testing EMOVO 61.3%, SAVEE 53.2%, EMO-DB 65.3%. Training - EMOVO Testing - URDU	-

Ref	Dataset(s) used	Features	Appr	Classifiers	Eminent Classifier	Results	Limitations
						53.6%, Training - SAVEE Testing - URDU 58%, Training - EMO-DB Testing - URDU 65.2%. Cross-corpus n vs 1 Testing EMOVO 61.8%, SAVEE 56.7%, EMO-DB 68% URDU 67.3%	
[70]	Urdu-Sindhi Speech Emotion Corpus	The prosody feature set, the IS09-Emotion feature set, the IS10-Paralinguistics feature set, the ComParE feature set, and the eGeMAPS feature set.	ML	Logistic Regression baseline classifier	Logistic Regression with ComParE feature set	Within-corpus Accuracy Sindhi speech 46.81%, Urdu speech 57.14%, Cross corpus Accuracy Training Sindhi speech Testing Urdu speech 17.69% Training Urdu speech Testing Sindhi speech 19.15%	• The authors have used the Logistic regression classifier as a baseline. One can argue that the more powerful machine learning models, such as those based on deep learning, will likely perform better than logistic regression.
[9]	URDU	LPCC, LFPC., MFMC., MFCCs	ML	Multi-class SVM	-	Accuracy 95.25%	• The authors haven't used any feature selection technique as the performance of MFMC can be further improved by using feature selection techniques.
[20]	URDU	MFCCs	DL	LSTM with Meta Learning, LSTM with Transfer Learning and Multi Learning	LSTM with Meta-Learning	Cross-corpus F1 Scores Training - TESS, EMODB, and RAVEDESS Testing - URDU 0.72	• Meta-learning for speech emotion recognition requires complex architecture for hyper-parameter searches to stabilize training. • It requires enough training data to surpass its performance threshold.
[28]	RAVDESS, URDU, Arabic	MFCCs and Pitch	ML	Random Forest, Neural Network, Meta Iterative Classifier	Random Forest	Accuracy URDU 78.75%	• The model gave a good performance in the Urdu Language, but a decline in performance was observed for R.A.V.D.E.S.S. and Arabic Dataset. • The accuracy of the

Ref	Dataset(s) used	Features	Appr	Classifiers	Eminent Classifier	Results	Limitations
							emotion recognition model can be improved by employing different deep learning algorithms.
[36]	URDU	GeMAPS	DL	Multi-layer Perceptron	-	Accuracy URDU 65.4%	• Extreme speaker imbalance in Urdu Dataset due to its spontaneous nature requires whole corpus normalization.
[45]	NSSED, URDU	Mel spectrograms	ML and DL	SVM, LSTM, 1-dimensional Convolution Neural Network (1DCNN)	(1DCNN)	Accuracy URDU 88% Sindhi 91%	• Limited emotion classes in both datasets.
[48]	URDU IEMOCAP	Spectrograms	DL	Neural Network with Pseudo Multilabel (N.N.P.M.)	-	Cross-corpus Accuracy Training - IEMOCAP Testing URDU UA 54.55%, WA 51.31%.	• Model performance can be further improved using the latest deep-learning approaches.
[66]	SAVEE RAVDEES CREMA-D EmoDB EMOVO SheMo URDU	Mel-spectrogram	ML, DL	SVM (Support Vector Machine), Decision Tree based Bagging, CNN (Convolutional Neural Network), C.R.N.N. (Convolutional Recurrent Neural Network)	C.R.N.N. (Convolutional Recurrent Neural Network)	Multi-lingual Accuracy 89.77%	• The authors have just focused on anger identification. This work can be further extended by adding some more emotion classes.
[74]	SEMOUR	MFCCs, chromagram, and Mel-spectrogram	DL	Gaussian Naive Bayes, Logistic Regression, SVM, Decision Tree, ANN, Random Forest, D.N.N. (Deep Neural Network)	D.N.N. (Deep Neural Network)	Accuracy 91%	-
[75]	URDU EMOVO SAVEE EMO-DB	MFCCs, Spectral (Roll off, flux, centroid, bandwidth), Energy (Root-mean-square energy), Raw Signal (Zero crossing rate), Pitch (Fundamental frequency), and Chroma features.	ML	SVM (baseline classifier), Sequential Minimal Optimization (S.M.O.), Fandom Forest (R.F.), J48, Ensemble Classifier (S.M.O., R.F., and J48)	Ensemble Classifier increased accuracy for both cross-corpus and within-corpus compared to the baseline classifier SVM.	Within-corpus Accuracy URDU 96.75%, EMOVO 87.14%, SAVEE 69.31%, EMO-DB 89.75%. Cross-corpus Accuracy Training URDU Testing - EMOVO 58.16%, SAVEE 43.34%, EMO-DB 57.14%. Training - EMOVO Testing - URDU 62.5%,	-

Ref	Dataset(s) used	Features	Appr	Classifiers	Eminent Classifier	Results	Limitations
[11]	Not Specified	MFCCs, LPC, energy, pitch, zero crossing, spectral flux, spectral centroid, spectral roll-off	ML	One-vs-rest One-vs-one k-NN (k-Nearest Neighbors) Random Forest	KNN	Training - SAVEE Testing - URDU 45%, Training - EMO-DB Testing URDU 52.5%. Accuracy Without disgust emotion 82 .5% and with disgust, emotion was 72.5%	• The dataset is not available for future research. • Deep Learning methods can be used to achieve higher accuracy. • The model Doesn't perform well for disgust emotion.

The above summary of classifiers and their performance shows that traditional machine learning algorithms previously used for Urdu SER were not as robust as the latest deep learning algorithms. C.R.N.N. (Convolutional Recurrent Neural Network) used in [66] achieved an emotion recognition accuracy of 89%, D.N.N. (Deep Neural Network) used in [74] achieved an emotion recognition accuracy of 91%, which clearly shows the robustness of deep learning algorithms. Recent research also emphasizes using an ensemble classifier for the Urdu speech emotion recognition task as it achieves the highest Accuracy for Urdu Speech emotion recognition which is 96.75%.

2.9 Performance Metrics for Urdu SER (RQ7)

The performance evaluation of current studies is generally carried out using Accuracy, Precision, Recall, Confusion Matrix and U.A.R. (Unweighted Average Recall). This section gives brief details of these performance evaluation metrics. Accuracy is the easiest to understand and widely used performance evaluation metric for SER tasks.

The recall is defined as the ratio between all the instances correctly classified in the positive class against the total number of actual members of the positive class. F-score is also known as F1-Score or F-measure. It is based on the precision and recall of the model. It is the harmonic means of precision and recall of the classification model. A confusion Matrix is the accurate description of predictions made by the classification model in a matrix form by representing correctly predicted and wrongly predicted instances. U.A.R. (Unweighted Average Recall) is known as balanced accuracy, as it is the average of recall in the positive and negative classes.

Below given Table 13 shows a Comparative analysis of various performance evaluation metrics used in current studies on Urdu SER. Most of the studies used accuracy for performance evaluation as it is the most appropriate metric for emotion classification.

Table 13: Comparison of various performance evaluation metrics used in current studies on Urdu SER

Sr. No	Study Ref	Accuracy	Precision	F-score	Recall	Confusion Matrix	U.A.R.
1	[72]	✓	✗	✗	✗	✓	✗
2	[8]	✓	✗	✗	✗	✗	✗
3	[41]	✓	✗	✓	✗	✗	✗
4	[5]	✓	✗	✗	✗	✗	✗
5	[7]	✓	✗	✗	✗	✗	✗
6	[6]	✓	✓	✗	✗	✗	✗
7	[39]	✓	✓	✗	✗	✗	✗
8	[35]	✓	✗	✗	✗	✓	✗
9	[46]	✗	✗	✗	✗	✗	✓
10	[47]	✓	✗	✗	✗	✗	✗
11	[70]	✓	✗	✗	✗	✓	✓
12	[9]	✓	✗	✗	✓	✓	✗
13	[20]	✗	✗	✓	✗	✗	✗
14	[28]	✓	✓	✗	✓	✗	✗
15	[36]	✗	✗	✗	✗	✗	✓
16	[45]	✓	✓	✓	✓	✓	✗
17	[48]	✓	✗	✗	✗	✗	✗
18	[66]	✓	✗	✗	✗	✓	✗
19	[74]	✓	✓	✓	✓	✗	✗
20	[75]	✓	✓	✗	✓	✗	✗
21	[12]	✓	✓	✗	✓	✓	✗

2.10 Validation Techniques used for Urdu (RQ8)

Validation methods in Urdu speech emotion recognition are used to ensure the performance of the developed SER model and prevent it from overfitting and underfitting problem. Validation methods are applied after model training. Without ensuring the validity of the developed SER model, one can't rely on model predictions for speech emotion recognition. This section discusses validation methods used in the existing literature on Urdu speech emotion recognition.

- Train-Test Split validation

  To evaluate the model's performance as an unbiased estimate, it is necessary to evaluate the model on data that the basic way for validation of the model train test split method. According to this method, the dataset is randomly divided into train and test sets. The ideal train-test split ratio is 80:20, which means 80% of the data is used for training, and 20% is used for testing the model. The authors have divided the dataset into three partitions, i.e., Training, Validation, and Testing having a 60:20:20 ratio. This training and Test dataset distribution can vary; some authors prefer 70% for training and 30% for testing the model. Authors in [7], [6], [39], [20], [45], [74], [12], [66] have used Train-Test split validation.

- K fold cross-validation
  In k-fold cross-validation, the dataset used for validation is divided into k subsets approximately of the same size. Each time k-1 subsets of datasets are used for training, and one subset of the dataset is used for testing. K-fold cross-validation is suitable when the dataset is large as it will spend less time validating. Authors in [8], [41], [9], [28] have used a standard practice of 10-fold cross-validation in which the dataset is divided into 10 subsets for validation. This helped to prevent overfitting. Authors in [5] have used 4-fold stratified cross-validation (S.C.V.) cross-validation. In stratified cross-validation, the stratified sampling technique is used instead of random sampling. Stratified sampling method, the population is divided into groups called 'strata' based on a

characteristic or features as they appear in the population. It ensures that the training and test set have the same proportion of characteristics or features of interest as in the original dataset.

- Leave-one-out (LOO) cross-validation.

  In this cross-validation, the approachh dataset is split into training and test set. For n samples of the dataset, one sample from the dataset at a time is taken for testing, and the rest of the samples of the dataset or n-1 samples are used for training of the SER system. The process continues for n times (where n represents the total number of samples of the dataset). LOO is an extreme case of k-fold cross-validation where n=k, or we can say it is exhaustive in nature. LOO cross-validation takes all samples for validation and has a low bias.

  For this reason, the major drawback can be the high variance because, in this validation, we are validating the model against one sample of the dataset at a time if that sample of the dataset is an outlier. Also, it takes a lot of execution time as the validation continues for n times each time one sample is used for testing, and the rest of the samples or n-1 samples are used for training the model. It is computationally infeasible for this reason. When the size of the dataset is, a small LOO validation method is suitable. Authors in [72], [47],[46], [36], [75] have used LOO cross-validation.

2.11 Implementation Tools (RQ9)

This section discusses the implementation tools for Urdu SER. Many open-source frameworks and libraries are available for Urdu SER. Given below, Table 14 shows the details about implementation tools that are used for Urdu Speech emotion recognition.

Table 14: Implementation Tools for Urdu SER

Sr. No	Implementation Tools	Description	Ref Used
1	Weka Data Mining Tool [15]	Weka is a tool with various machine-learning algorithms for data mining tasks.	[8], [6]
2	TensorFlow [1]	TensorFlow is the most popular open-source platform for designing machine learning and deep learning models.	[46], [36], [66]
3	Scikit Learn Toolkit [58]	Scikit Learn is an open-source library for machine learning in python. It provides a diverse range of classification, regression, and clustering algorithms.	[70], [36], [72]
4	PyTorch Toolkit [57]	PyTroch is an open-source machine-learning framework based on the Torch library. It supports both research prototyping and product development.	[48]
5	Keras [31]	Keras is the latest deep-learning framework. That provides a python interface for the development of robust deep learning models.	[66]

2.12 Challenges and Future Directions (RQ10)

This section discusses the challenges and future directions for research on Urdu SER. As the field of Urdu SER. is emerging filed and getting the attention of the latest researchers. As this field is emerging and not mature for low resource language Urdu, for this reason, there are still some challenges that need to be considered, and along with these challenges, there are new research areas as well to be explored in the Future.

2.12.1 Challenges in Urdu SER

1. Difficulty levels for Speech Databases

Speech Emotions are difficult to be analyzed, and when subjected to evaluation, the performance of speech emotion recognition varies according to the degree of naturalness found in speech. It is easy to analyze emotions from speech that is acted in nature. For elicited speech, emotion recognition is at a medium level of difficulty. For natural and real-time speech, the task of emotion recognition is of high difficulty level. Urdu Speech Databases and their difficulty level are illustrated in Figure 10.

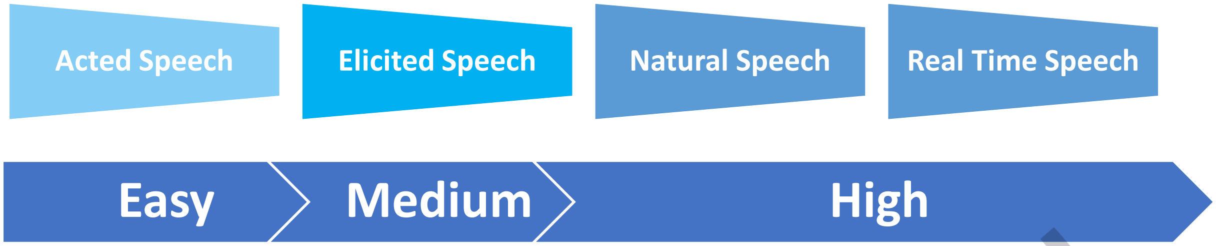

Figure 10: Urdu Speech Databases and their difficulty level

2. Limited Lexical Variability in Speech Databases

   Lexical variability is the quality of a speech database to is rich in lexical content and has a variety of emotional utterances. Current databases are mostly acted in nature, and these databases lack lexical variability, like a few sentences or even just one sentence uttered in different emotional styles. Due to this reason, the SER system lacks generalizability and has not had a good performance. To introduce generalizability in the SER system and make it more robust, it is necessary to train the speech databases, which are rich in lexical variability or have a variety of emotional utterances and are full of lexical content.
3. A limited number of speakers in Speech Databases

   Most literature found on Urdu speech emotion recognition using speech databases with a limited number of speakers. If the same speaker is used for training and testing the SER. model, it may result in poor performance when the model is deployed in real time due to a lack of generality. Therefore, it is necessary to use a speech database with many speakers.
4. Effect of Recording Environment

   Conditions for recording the speech corpora also affect the performance of the SER system. Clear recordings performed in the lab following standard recording conditions help the model easily classify speech emotions. Whereas speech corpora recorded in noisy environments must be well pre-processed before using them for the SER task.
5. Real-time Speech Emotion Recognition

   Studies found in the systematic review used acted speech for speech emotion recognition. There was not any single study found on real-time speech emotion recognition. The speech emotion recognition of real-time speech is a challenging task. Research must contribute to making the SER system robust enough to recognize emotions from real-time speech. Also, finding corpora for the Urdu language in the natural environment is also a challenging task.
6. Lack of robustness in the SER model

   Most SER models are implemented using machine learning methods, and few studies have used deep learning also. It is observed that current studies lack robustness in terms of performance. Still, there is a need to work on the robustness of the SER model to provide good performance results with generalizability ability and perform even on unseen speech data.

2.12.2 Future Directions for Urdu SER

Urdu speech emotion recognition is a growing research field with many future research directions. Some latest future directions are discussed in this section.

1. Active Learning for Urdu SER
2. Deep Learning approaches are proven to be better than traditional machine learning approaches. Deep learning requires large training data, but it is difficult to achieve good performance results due to a lack of data resources. For this problem, an active learning approach can be effectively used for speech emotion recognition tasks even with small training data.
3. Multiple Classifier Methods (MCM)
4. Few studies are using multiple classifier methods (MCM). This approach can be considered as a future direction for Urdu Speech Emotion Recognition, where an ensemble of multiple classifiers can be used to make the final decision for speech emotion recognition.
5. Multi-modal Emotion Recognition
6. Emotion expression is a multi-modal activity in real life. Emotions are not only expressed using speech; the facial expression and linguistic information uttered in speech can be considered for developing the SER system. With multi-modal emotion recognition, robustness can be observed in HRI (Human-Robot Interaction) system. The system can use audio and visual modalities for emotion recognition and perform actions more robustly and accurately.
7. Using Generative Models for Cross-Corpus Speech Emotion Recognition
8. GAN (Generative Adversarial Network) can be used in cross-lingual speech emotion recognition scenarios. The GAN model will help for domain adaption in case of practical applications to generalize multiple languages to learn different language invariant representations without requiring target language data labels [46].
9. Using Meta-Learning for Cross-Corpus Speech Emotion Recognition
10. Meta-Learning can be used to develop a real-time speech-emotion recognition system where the SER model will learn emotion-specific features from training by resource-rich languages and show its generalizability ability for unseen languages like Urdu. In [20], Meta-Learning helps LSTM-based SER models to learn language-independent emotion-specific features for cross-corpus speech emotion recognition to outperform approaches like transfer learning and multitask learning. Authors in [20] empirically proved that meta-learning training is useful in cross-corpus scenarios for generalizing from resource-rich to resource-poor language families. Authors performed Meta-Learning with LSTM implementation trained model on TESS., EMODB RAVEDESS, which are resource-rich languages and tested the model on Urdu, which is resource poor language. Authors achieved good performance results for Urdu speech emotion recognition i.e., 0.72 F1 score.
11. Real-time Urdu Speech Emotion Recognition
12. Current studies found in the literature evaluate the SER model on acted or artificially created speech corpus. Real-time speech is difficult to be evaluated, and emotion recognition becomes difficult for real-time speech. To make speech emotion recognition robust, it is necessary to train the model to efficiently perform speech emotion recognition for real-time speech.
13. Diversifying Urdu Speech Emotion Recognition
14. There are several dialects and accents of the Urdu Language. The Urdu language has 4 dialects, i.e., Urdu, Dakhini, Hyderabadi Urdu and Rekhta [68]. And there are six accents, namely Urdu, Punjabi, Pashto, Saraiki, Balochi, and Sindhi, for the Urdu dialect [63]. Urdu is a language spoken by a large community, so while designing or collecting data for Urdu SER, it is necessary to keep in mind to diversify the Urdu speech corpus. This will make Urdu corpus rich in acoustic information with diversity from different dialects and accents.
15. Text Dependent Speech Emotion Recognition

16. Most of the research on speech emotion recognition is carried out in text-independent way speech signal is processed directly. An alternate approach can be text dependent approach in which speech can be analyzed according to the linguistic content present in the speech. The text-independent approach helps find the aspects hidden behind emotions and helps summarise the speech. It can serve in different applications like call centre data and customer service.
17. Aspect-Based Speech Emotion Recognition ABSER
18. No study was found on aspect-based speech emotion recognition from the literature collected using this systematic literature review. Current studies only classify emotions but do not recognize the aspect or context of speech. Recent advances in SER are to find speech emotions along with aspects. This will help businesses get clearer insights about customer feedback or customer service calls.
19. Work on the Dimensional Emotional Model
20. Few authors like [74] have worked on dimensional or continuous emotional models means emotions can be represented in terms of valence and arousal. This helps to provide more information about an emotional state.
21. Use of Reinforcement Learning for Speech Emotion Recognition
22. Authors in [35] have implied reinforcement learning using REINFORCE algorithm along with deep learning architecture to train the RL agent to learn and explore the most effective methods for the estimation of system confidence values to predict accurate emotions and to improve multi-modal emotion recognition for real-time requirements of HRI (Human-Robot Interaction). The use of Reinforcement learning is also a new research area to be explored further to make a robust Speech Emotion Recognition System that can be effectively used for HRI

3 CONCLUSION

This paper provides a systematic literature review on Urdu Speech Emotion Recognition. This paper is the first systematic literature review on Urdu Speech Emotion Recognition. Few studies have worked on Urdu Speech Emotion Recognition; this review critically appraises the existing literature on Urdu Speech Emotion Recognition. It is observed that despite of lack of datasets, researchers are contributing toward Urdu Speech Emotion Recognition. Authors have also created benchmarks for Urdu SER, guiding future researchers.

Furthermore, due to the diversity found in the Urdu language in terms of dialects and accents, it is necessary to include this diversity when creating datasets for Urdu SER. This systematic literature review investigated datasets, pre-processing techniques, features, approaches, performance metrics and validation methods that have been adopted for Urdu SER. From the systematic review it is found that there are few publicly available datasets of Urdu Speech which are acted and natural collected, but no study found on real-time speech. Most of the work found on Urdu SER. were doing cross-corpus SER. this highlights the generalizability ability of different languages. Most of the studies used Machine Learning approaches. Due to advancements in deep learning, the latest techniques must be explored in the Future to make Urdu SER more robust. Challenges and future directions are also discussed in this systematic literature review.

ACKNOWLEDGEMENT

This work was supported by the National Research Program for Universities (NRPU), Higher Education Commission under Project Ref No. 20-14457/NRPU/R&D/HEC/2021-2020.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, and Michael Isard. 2016. Tensorflow: a system for large-scale machine learning. Savannah, GA, USA, 265–283.

[2] Babak Joze Abbaschian, Daniel Sierra-Sosa, and Adel Elmaghraby. 2021. Deep learning techniques for speech emotion recognition, from databases to models. *Sensors* 21, 4 (2021), 1249.

[3] Rajeev Aggarwal, Jai Karan Singh, Vijay Kumar Gupta, Sanjay Rathore, Mukesh Tiwari, and Anubhuti Khare. 2011. Noise reduction of speech signal using wavelet transform with modified universal threshold. *International Journal of Computer Applications* 20, 5 (2011), 14–19.

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

[5] Enrique Marcelo Albornoz and Diego H Milone. 2015. Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles. *IEEE Transactions on Affective Computing* 8, 1 (2015), 43–53.

[6] Syed Abbas Ali, Najmi Ghani Haider, and Maria Andleeb. 2016. Evaluating the Performance of Learning Classifiers and Effect of Emotions and Spectral Features on Speech Utterances. *International Journal of Computer Science and Information Security (IJCSIS)* 14, 10 (2016).

[7] Syed Abbas Ali, Anas Khan, and Nazia Bashir. 2015. Analyzing the impact of prosodic feature (pitch) on learning classifiers for speech emotion corpus. *International Journal of Information Technology and Computer Science* 2, (2015), 54–59.

[8] Syed Abbas Ali, Sitwat Zehra, and Afsheen Arif. 2013. Performance evaluation of learning classifiers for speech emotions corpus using combinations of prosodic features. *International Journal of Computer Applications* 76, 2 (2013).

[9] J. Ancilin and A. Milton. 2021. Improved speech emotion recognition with Mel frequency magnitude coefficient. *Appl. Acoust.* 179, (August 2021), 10. DOI:https://doi.org/10.1016/j.apacoust.2021.108046

[10] R Aparna. A STUDY ON IMPACT OF VARIOUS WINDOWING TECHNIQUES IN CONTINUOUS SPEECH SIGNAL SEGMENTATION. *International Journal of Applied Engineering Research* 10, 76 , 2015.

[11] Awais Asghar, Sarmad Sohaib, Saman Iftikhar, Muhammad Shafi, and Kiran Fatima. 2022. An Urdu speech corpus for emotion recognition. *PeerJ Computer Science* 8, (2022), e954.

[12] Awais Asghar, Sarmad Sohaib, Saman Iftikhar, Muhammad Shafi, and Kiran Fatima. 2022. An Urdu speech corpus for emotion recognition. *PeerJ Computer Science* 8, (2022), e954.

[13] Margaret Bearman, Calvin D Smith, Angela Carbone, Susan Slade, Chi Baik, Marnie Hughes-Warrington, and David L Neumann. 2012. Systematic review methodology in higher education. *Higher Education Research & Development* 31, 5 (2012), 625–640.

[14] Paul Boersma. 2011. Praat: doing phonetics by computer [Computer program]. *http://www. praat. org/* (2011).

[15] Remco R Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, and David Scuse. 2018. WEKA manual for version 3-9-3. *The University of Waikato, Hamilton, New Zealand* (2018).

[16] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. 1517–1520.

[17] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, (2008), 335–359.

[18] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini %J IEEE transactions on affective computing Verma. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. 5, 4 (2014), 377–390.

[19] Jingdong Chen, Jacob Benesty, Yiteng Huang, Simon %J IEEE Transactions on audio Doclo, and language processing. 2006. New insights into the noise reduction Wiener filter. 14, 4 (2006), 1218–1234.

[20] Suransh Chopra, Puneet Mathur, Ramit Sawhney, and Rajiv Ratn Shah. 2021. Meta-learning for low-resource speech emotion recognition. IEEE, 6259–6263.

[21] Chloé Clavel, Ioana Vasilescu, Laurence Devillers, Gaël Richard, and Thibaut Ehrette. 2008. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication* 50, 6 (2008), 487–503.

[22] Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, and Massimiliano Todisco. 2014. EMOVO corpus: an Italian emotional speech database. European Language Resources Association (ELRA), 3501–3504.

[23] Kate Dupuis and M. Kathleen %J Canadian Acoustics Pichora-Fuller. 2011. Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. 39, 3 (2011), 182–183.

[24] Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion* 98, 45–60 (1999), 16.

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

[26] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. 1459–1462.

[27] Md Shah Fahad, Ashish Ranjan, Jainath Yadav, and Akshay Deepak. 2021. A survey of speech emotion recognition in natural environment. *Digital signal processing* 110, (2021), 102951.

[28] Moomal Farhad, Heba Ismail, Saad Harous, Mohammad Mehedy Masud, and Azam Beg. 2021. Analysis of emotion recognition from cross-lingual speech: Arabic, English, and Urdu. IEEE, 42–47.

[29] Daniel Joseph France, Richard G Shiavi, Stephen Silverman, Marilyn Silverman, and M Wilkes. 2000. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering* 47, 7 (2000), 829–837.

[30] Yuanbo Gao, Baobin Li, Ning Wang, and Tingshao Zhu. 2017. Speech emotion recognition using local and global features. Springer, 3–13.

[31] Antonio Gulli and Sujit Pal. 2017. *Deep learning with Keras*. Packt Publishing Ltd.

[32] Hatice Gunes and Björn Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31, 2 (2013), 120–136.

[33] Philip Jackson and SJUoSG Haq. 2014. Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK* (2014).

[34] Rashid Jahangir, Ying Wah Teh, Faiqa Hanif, and Ghulam Mujtaba. 2021. Deep learning approaches for speech emotion recognition: state of the art and research challenges. *Multimedia Tools and Applications* (2021), 1–68.

[35] Ioannis Kansizoglou, Loukas Bampis, and Antonios Gasteratos. 2019. An active learning paradigm for online audio-visual emotion recognition. *IEEE Transactions on Affective Computing* 13, 2 (2019), 756–768.

[36] Aaron Keesing, Yun Sing Koh, and Michael Witbrock. 2021. Acoustic Features and Neural Representations for Categorical Emotion Recognition from Speech. 3415–3419.

[37] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, and Mohamed Ali Mahjoub. 2017. A review on speech emotion recognition: Case of pedagogical interaction in classroom. IEEE, 1–7.

[38] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, and Thamer Alhussain. 2019. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* 7, (2019), 117327–117345.

[39] Sallar Khan, Syed Abbas Ali, and Jawaria Sallar. 2018. Analysis of children's prosodic features using emotion based utterances in Urdu language. *Engineering, Technology & Applied Science Research* 8, 3 (2018), 2954–2957.

[40] Barbara Kitchenham. 2004. Procedures for performing systematic reviews. *Keele, UK, Keele University* 33, 2004 (2004), 1–26.

[41] Vladimer Kobayashi and Vicente Calag. 2013. Detection of affective states from speech signals using ensembles of classifiers. *IET Intelligent Signal Processing Conference 2013 (ISP 2013)* (January 2013). DOI:https://doi.org/10.1049/cp.2013.2067

[42] Shashidhar G. Koolagudi, Y. V. Murthy, and Siva P. %J International Journal of Speech Technology Bhaskar. 2018. Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition. 21, 1 (2018), 167–183.

[43] Panagiotis Koromilas and Theodoros Giannakopoulos. 2021. Deep multimodal emotion recognition on human speech: A review. *Applied Sciences* 11, 17 (2021), 7962.

[44] Maria Labied, Abdessamad Belangour, Mouad Banane, and Allae Erraissi. 2022. An overview of Automatic Speech Recognition Preprocessing Techniques. IEEE, 804–809.

[45] Muddasar Laghari, Muhammad Junaid Tahir, Abdullah Azeem, Waqar Riaz, and Yi Zhou. 2021. Robust speech emotion recognition for sindhi language based on deep convolutional neural network. IEEE, 543–548.

[46] Siddique Latif, Junaid Qadir, and Muhammad Bilal. 2019. Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. IEEE, 732–737.

[47] Siddique Latif, Adnan Qayyum, Muhammad Usman, and Junaid Qadir. 2018. Cross lingual speech emotion recognition: Urdu vs. western languages. IEEE, 88–93.

[48] J. Li, N. Yan, and L. Wang. 2021. Unsupervised Cross-Lingual Speech Emotion Recognition Using Pseudo Multilabel. 366–373. DOI:https://doi.org/10.1109/ASRU51503.2021.9688171

[49] Steven R Livingstone and Frank A Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one* 13, 5 (2018), e0196391.

[50] Jianhua Ma, Hai Jin, Laurence T Yang, and Jeffrey J-P Tsai. 2006. *Ubiquitous intelligence and computing*. Springer.

[51] Brian McFee, Colin Raffel, Dawen Liang, Daniel P. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. 18–25.

[52] Omid Mohamad Nezami, Paria Jamshid Lou, and Mansoureh Karami. 2019. ShEMO: a large-scale validated database for Persian speech emotion detection. *Language Resources and Evaluation* 53, (2019), 1–16.

[53] Mumtaz Begum Mustafa, Mansoor AM Yusoof, Zuraidah M Don, and Mehdi Malekzadeh. 2018. Speech emotion recognition research: an analysis of research focus. *International Journal of Speech Technology* 21, (2018), 137–156.

[54] Bashar M. Nema and Ahmed A. %J Al-Mustansiriyah Journal of Science Abdul-Kareem. 2017. Preprocessing signal for speech emotion recognition. 28, 3 (2017), 157–165.

[55] Jonathan Shi Khai Ooi, Siti Anom Ahmad, Hafiz Rashidi Harun, Yu Zheng Chong, and Sawal Hamid Md Ali. 2017. A conceptual emotion recognition framework: stress and anger analysis for car accidents. *International journal of vehicle safety* 9, 3 (2017), 181–195.

[56] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, and Sue E. %J Systematic reviews Brennan. 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. 10, 1 (2021), 1–11.

[57] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32, (2019).

[58] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12, (2011), 2825–2830.

[59] Gulnaz Nasir Peerzade, R. R. Deshmukh, and S. D. Waghmare. 2018. A review: Speech emotion recognition. *Int. J. Comput. Sci. Eng* 6, 3 (2018), 400–402.

[60] Valery Petrushin. 1999. Emotion in speech: Recognition and application to call centers. 22.

[61] Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89, 4 (2001), 344–350.

[62] Syed Asif Ahmad Qadri, Teddy Surya Gunawan, Muhammad Fahreza Alghifari, Hasmah Mansor, Mira Kartiwi, and Zuriati Janin. 2019. A critical insight into multi-languages speech emotion databases. *Bulletin of Electrical Engineering and Informatics* 8, 4 (2019), 1312–1323.

[63] Muhammad Qasim, Sohaib Nawaz, Sarmad Hussain, and Tania Habib. 2016. Urdu speech recognition system for district names of Pakistan: Development, challenges and solutions. IEEE, 28–32.

[64] Javier G Rázuri, David Sundgren, Rahim Rahmani, Antonio Moran, Isis Bonet, and Aron Larsson. 2015. Speech emotion recognition in emotional

feedbackfor human-robot interaction. *International Journal of Advanced Research in Artificial Intelligence (IJARAI)* 4, 2 (2015), 20–27.

[65] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[66] A. Saitta and S. Ntalampiras. 2021. Language-agnostic speech anger identification. Ieee, NEW YORK, 249–253. DOI:https://doi.org/10.1109/tsp52935.2021.9522606

[67] Björn W Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM* 61, 5 (2018), 90–99.

[68] Tariq Rahim Soomro and Saqib Muhammad Ghulam. 2019. Current status of urdu on Twitter. *Sukkur IBA Journal of Computing and Mathematical Sciences* 3, 1 (2019), 28–33.

[69] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. 2018. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology* 21, (2018), 93–120.

[70] Z. S. Syed, S. A. Memon, M. S. Shah, and A. S. Syed. 2020. Introducing the Urdu-Sindhi Speech Emotion Corpus: A Novel Dataset of Speech Recordings for Emotion Recognition for Two Low-Resource Languages. *Int. J. Adv. Comput. Sci. Appl.* 11, 4 (April 2020), 805–810.

[71] Mariusz Szwoch and Wioleta Szwoch. 2015. Emotion recognition for affect aware video games. Springer, 227–236.

[72] Yongjin Wang and Ling Guan. 2008. Recognizing human emotional state from audiovisual signals. *IEEE transactions on multimedia* 10, 5 (2008), 936–946.

[73] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. 2021. A comprehensive review of speech emotion recognition systems. *IEEE Access* 9, (2021), 47795–47814.

[74] Nimra Zaheer, Obaid Ullah Ahmad, Ammar Ahmed, Muhammad Shehryar Khan, and Mudassir Shabbir. 2021. SEMOUR: A Scripted Emotional Speech Repository for Urdu. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021).

[75] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu. 2021. Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex Intell. Syst.* 7, 4 (August 2021), 1845–1854. DOI:https://doi.org/10.1007/s40747-020-00250-4

[76] Sara Zhalehpour, Onur Onder, Zahid Akhtar, and Cigdem Eroglu Erdem. 2016. BAUM-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing* 8, 3 (2016), 300–313.

[77] "Yau", Ethnologue, 2022. [Online]. Available: https://www.ethnologue.com/language-of-the-day/2022-06-29. [Accessed: 30- Jun- 2022].

[78] "Audio Toolbox," Mathworks.com. Retrieved July 2, 2022 from https://ch.mathworks.com/products/audio.html