Intelligent Systems with Applications 20 (2023) 200266

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Review

# Speech emotion recognition using machine learning — A systematic review

Samaneh Madanian[a,*], Talen Chen[a], Olayinka Adeleye[a], John Michael Templeton[b], Christian Poellabauer[c], Dave Parry[d], Sandra L. Schneider[e]

[a] *Department of Computer Science and Software Engineering, Auckland University of Technology (AUT), Auckland, New Zealand*
[b] *University of South Florida, Department of Computer Science and Engineering, Tampa, FL, USA*
[c] *Florida International University, School of Computing and Information Sciences, Miami, FL, USA*
[d] *School of IT, Media and Communications, Murdoch University, Perth 6150, Australia*
[e] *St Mary's College, Department of Communicative Sciences & Disorders, Notre Dame, IN, USA*

## ARTICLE INFO

*Keywords:*
Speech emotion recognition
Machine learning
Speaker-independent experiment
Classification
Audio emotion recognition

## ABSTRACT

Speech emotion recognition (SER) as a Machine Learning (ML) problem continues to garner a significant amount of research interest, especially in the affective computing domain. This is due to its increasing potential, algorithmic advancements, and applications in real-world scenarios. Human speech contains para-linguistic information that can be represented using quantitative features such as *pitch*, *intensity*, and *Mel-Frequency Cepstral Coefficients* (MFCC). SER is commonly achieved following three key steps: *data processing*, *feature selection/extraction*, and *classification* based on the underlying emotional features. The nature of these steps, coupled with the distinct features of human speech, underpin the use of ML methods for SER implementation. Recent research works in affective computing employed various ML methods for SER tasks; however, only a few of them capture the underlying techniques and methods that can be used to facilitate the three core steps of SER implementation. In addition, the challenges associated with these steps, and the state-of-the-art approaches used in tackling them are either ignored or sparsely discussed in these works. In this paper, we present a systematic review of research that addressed SER tasks from ML perspectives over the last decade, with emphasis on the three SER implementation steps. Different challenges, including the issue of low-classification-accuracy of Speaker-Independent experiments, and solutions associated with them, are discussed in detail. The review also provides guidelines for SER evaluation with a focus on common baselines, and metrics available for experimentation. This paper is expected to serve as a comprehensive guideline for SER researchers to design SER solutions using ML techniques, motivate possible improvements of existing SER models, or trigger novel techniques to enhance SER performance.

## 1. Introduction

Speech emotion recognition (SER), a sub-discipline of affective computing (Picard, 2000), has been around for more than two decades and has led to a considerable amount of published works (Akçay & Oğuz, 2020; Gadhe & Deshmukh, 2015). SER involves recognizing the emotional aspects of speech irrespective of the semantic content (Lech et al., 2020). A typical SER system can be considered as a collection of methodologies that isolate, extract and classify speech signals to detect emotions embedded in them (Akçay & Oğuz, 2020). The use cases of SER in real-world applications are countless, some of which have demonstrated that the inclusion of emotional attributes in *human-machine* interactions can significantly improve the interaction experiences of users (Mustafa et al., 2018). For example, a SER system can evaluate call centre agents' performance by detecting customer emotions such as anger or happiness. This information can support companies in improving service quality or providing targeted training which leads to improving customer satisfaction and call centre efficiency (Mekruksavanich et al., 2020).

SER has become an important building block for many smart service systems in areas such as healthcare, smart homes, and smart entertainment (Zhu et al., 2017). Emergency call centres can use speech emotion analysis to identify hazardous or life-threatening circumstances (Ahmad et al., 2016). SER could also be used by an interactive voice

[*] Corresponding author.
*E-mail addresses:* sam.madanian@aut.ac.nz (S. Madanian), Talen.Chen@aut.ac.nz (T. Chen), oadeleye@aut.ac.nz (O. Adeleye), jtemplet@usf.edu (J.M. Templeton), cpoellab@fiu.edu (C. Poellabauer), david.parry@murdoch.edu.au (D. Parry), sandyl.schneider@gmail.com (S.L. Schneider).

https://doi.org/10.1016/j.iswa.2023.200266
Received 4 September 2022; Received in revised form 27 July 2023; Accepted 8 August 2023
Available online 14 August 2023
2667-3053/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

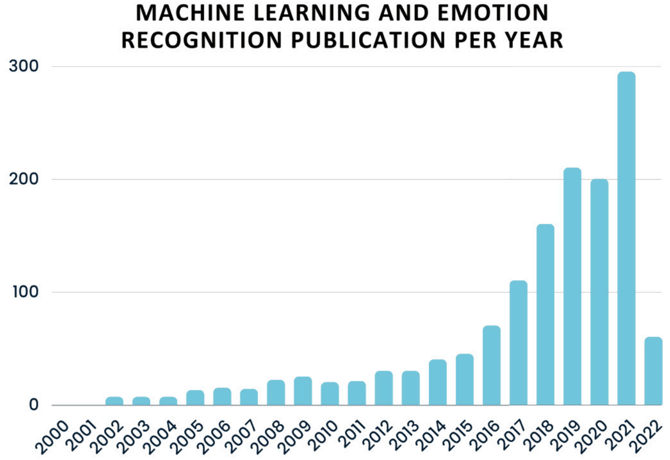

**Fig. 1.** Number of publications per year in the application of ML for emotion recognition.

response system in a car to prevent accidents due to fatigued drivers (Zhou et al., 2016). In clinical settings, SER could promote tele-mental health (Madanian et al., 2022) or use to support mental health diagnosis (Rawat & Mishra, 2015), such as detecting signs of potential suicidal ideation (France et al., 2000). For online education services, SER is a valuable tool, as it allows teachers to assess the degree to which students have mastered new skills by analyzing the emotional content of their responses. This can be used to fine-tune the teaching plan and optimize the learning experience (Zhou et al., 2016).

One of the more daunting tasks of SER is to identify and extract information from speech that is most suitable for computational identification and discrimination of emotion. While human speech contains an abundance of information, including both linguistic and para-linguistic features, it is the para-linguistic features that will be the focus of this research. Linguistic features refer to the qualitative patterns in human articulation, like content and context, while para-linguistic features quantitatively describe the variations in the pronunciation of the linguistic patterns (Anagnostopoulos et al., 2015; Zhao et al., 2019b). These include the *prosodic* features, like pitch and intensity, and *spectral* features, such as Linear Predictor Coefficient (LPC) and Mel-Frequency Cepstral Coefficients (MFCC) (Alu et al., 2017). Moreover, the speech signals can also be represented by more visually-direct forms, such as time-frequency *spectrograms*.

Several SER studies have investigated the connection between human emotions and prosodic/spectral acoustic parameters in speech. More recently, the advancement in digital signal processing, improvements in human-machine interactions (Costantini et al., 2021), and rapid advances in Machine Learning (ML) (Zhang et al., 2020) have significantly increased the use of ML techniques for identifications of emotions. This increase is evident when searching for “machine learning” and “emotion recognition” in scientific databases such as IEEEXplore (Fig. 1) on which we extracted the number of publications per year (search date was 13/06/2022). Based on this interest and emphasize on ML and emotion recognition, and the increased number of studies that have been conducted, it is needed to focus on ML. These studies mainly accomplished SER tasks by using ML pipeline methods that include isolation of the speech signal, dimensionality reduction, speech features extraction, and emotion classification based on the underlying speech features. The main aim is to leverage ML to better understand its users and communicate with them more effectively while improving users’ interaction with technology (Czerwinski et al., 2021).

The distinct features of speech, spectrograms and other attributes of human speech underpin the use of ML methods for SER tasks. Traditionally, ML involves the process of learning patterns and calculating feature parameters from raw data (such as speech, images, ECG and videos). These features are used to train a model that learns to provide the desired output label in either a prediction and/or classification task. Testing numerous distinct features, integrating diverse features into a common feature vector, or using alternative feature selection strategies can provide some insights into which features provide the most efficient clustering of data into classes. In addition, recent ML methods, such as graphs/deep neural networks, provide more elegant ways of bypassing the challenge of an optimal feature selection (Lech et al., 2020).

For SER, the speech features and spectrograms provide a discriminative representation of different emotions in speech, which could be extracted from the audio data to train the ML model. Existing ML-based SER studies have analyzed the acoustic speech parameters and established correlations between the parameters and a speaker’s emotions. Majority of the studies applied standard classifiers such as *Support Vector Machine* (SVM) (Jain et al., 2020; Milton et al., 2013; Bhavan et al., 2019; Kuchibhotla et al., 2014; Kerkeni et al., 2018), *Gaussian Mixture Model* (GMM) (Vondra & Vích, 2009), *K-Nearest Neighbour* (KNN) (Umamaheswari & Akila, 2019), *Recurrent Neural Networks* (RNN) (Yadav et al., 2021), and *Neural Networks* (NN) (Li et al., 2019). The ML approaches used in these works generally follow three key steps: *data pre-processing/speech signal isolation*, *feature extraction/selection*, and *classification* of emotions from audio signals (Yogesh et al., 2017b).

The inherent challenges of recognizing a speaker’s emotional states from the speech are driven by a variety of factors (El Ayadi et al., 2011). First, it is not clear which speech features are most effective in discriminating distinct emotional states. The acoustic variability introduced by the existence of different sentences, speakers, speaking styles and speaking rates adds additional layers of difficulty as these factors could have a direct impact on the retrieved speech features (Swain et al., 2018; El Ayadi et al., 2011). The dependency of certain emotional expressions of the speaker and the speaker’s culture, dialect, and environment could also affect the SER performance. Second, there may be emotion overlaps or multiple emotions perceived in the same utterance, making it difficult to determine the boundaries between each distinct emotional state. Even though many research efforts in SER have explored diverse ML approaches with various combinations of speech features, the majority of these works did not describe the techniques or methods used in carrying out the three core steps (i.e., data pre-processing, feature extraction, and emotion classification) of the SER task. Moreover, the challenges associated with these methods such as the pervasive *low-classification-accuracy* issue of *Speaker-Independent* SER systems, and potential solutions are either not addressed or sparsely discussed.

To assist in understanding the highly diverse use of ML algorithms and their available methods and techniques, we conducted a systematic review of ML-based speech-emotion recognition systems. The aim was to provide a review of techniques, strategies, and methods used in facilitating the three core steps of an ML-based SER, and to address the challenges associated with ML-based SER tasks. We acknowledge that more recently (past 3 to 4 years) are more relevant and interesting in terms of using novel ML algorithms and techniques; however, we selected studies from 2010 onward to demonstrate the evolution of the field and provide the necessary foundation required for the recent developments. In addition, we analyze existing solutions that address the speaker dependency issues and the *low-classification-accuracy* issue of *Speaker-Independent* SER system. The structure of this paper is as follows: Section 2 describes the research background and motivation for this review, Section 3 provides the research methodology, Section 4 presents the research findings and Section 5 presents the ML processes.

## 2. Background and motivations

Speech is the fastest and the most natural mode of communication between humans (Latif et al., 2021; El Ayadi et al., 2011). The speech captures formal features of linguistic expressions (i.e., phonology, morphology, syntax and semantics) along with the emotional states of the human. It carries affective information associated with emotional ex-

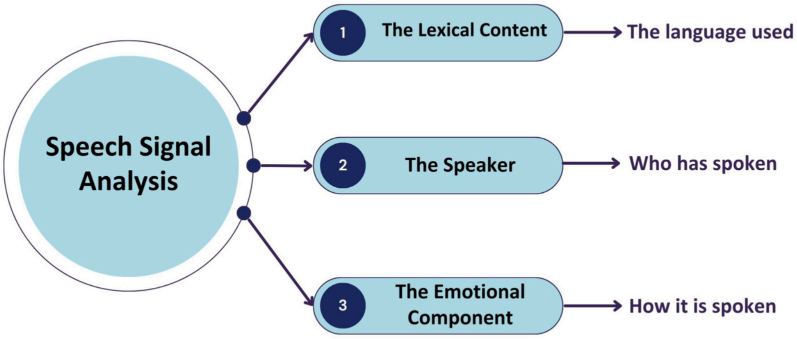

**Fig. 2.** Speech Signal Analysis.

pressions through linguistic (explicit) and para-linguistic cues, which can be extracted using speech processing techniques (Anagnostopoulos et al., 2015; Liu et al., 2018; Latif et al., 2021). Furthermore, speech signal analysis can provide insight into *what has been spoken* (the lexical content), *who has spoken* (the speaker), *the language used*, and *how it is spoken* (the emotional component) (Fahad et al., 2021) - see Fig. 2.

This emotional component is important for understanding human decisions and can reveal the mental state of an individual (Harati et al., 2018; Miner et al., 2020; Mitsuyoshi et al., 2017; Minardi, 2013). These facts motivate the use of speech signals as an effective means of *human-machine* interactions. Although several other modalities including *facial expression* (Ali et al., 2015; Kerkeni et al., 2019), text (Madanian et al., 2023b) and *physiological signals*, such as *heart rate, respiration, skin conductivity, and neural membrane* (Kerkeni et al., 2019; Swain et al., 2018), can also be explored to recognize human emotional states, certain inherent attributes of speech signal make it a more effective source for affective computing. For instance, unlike physiological/biological signals, speech signals can be acquired more readily and economically. Furthermore, it is more difficult to identify emotions from the text as a result of the syntactic and semantic ambiguities, and not all emotions can be detected from facial expressions (Koolagudi et al., 2018). In the case of using heart rate, it may be confounded by physical activity (Madanian et al., 2022).

The para-linguistic content of speech provides extremely valuable acoustic features that can be used to encode the emotional state of the speaker and have been explored using different ML approaches (Latif et al., 2021; Konar & Chakraborty, 2015; Kuchibhotla et al., 2014). These features are highly similar in all languages; therefore, a generic classification model can be used across languages (Fahad et al., 2021). Additionally, emotion detection from speech can also help avoid the privacy concerns associated with facial emotion detection, which could be more appealing from Patients' perspectives on using digital health tools (Madanian et al., 2023a).

To build a *generalizable* speech recognition system, one needs to recognize specific features of a speech signal (e.g., the emotional aspect), while nullifying other speech aspects (e.g., linguistic and cultural information). Specifically, the SER system must be able to accommodate different speakers with different languages (i.e., *Speaker/Language Independence*). For this reason, some recent SER research focused on addressing the challenge of the speaker and language dependency in SER implementations (Kalhor & Bakhtiari, 2021; Sun & Wen, 2015; Liu et al., 2019; Kim et al., 2009; Albornoz & Milone, 2015; Fahad et al., 2021; Abdelwahab & Busso, 2017; Deng et al., 2014).

Many of the recent SER research results leveraged ML methods and techniques to tackle different challenges in SER implementation. Although almost all of these works describe the SER process in three steps (Fig. 3) — *data pre-processing*, *feature extraction*, and *classification* of emotions from audio signals — many studies focus on a single strategy or method for performing the ML tasks.

For example, Alu et al. (2017) only used the $MFCC$ of speech features and a single-classifier architecture ($Convolutional Neural Network-CNN$) for emotion classification. The use of other audio features or ensembled-model architectures (e.g., ensembled-model based on CNN and SVM) were not mentioned. Tashev et al. (2017) used GMM and NN as feature extractors, but very little was mentioned on feature selection and extraction techniques. Niu et al. (2017) performed the SER task with an ensembled-model (based on CNN and RNN) and speech spectrograms. The use of low-level descriptors ($LLDs$) was not considered. Jalal et al. (2019) compared several single-models (e.g., CNN with attention model, $RNN-LSTM$, and $SVM$) performance in the SER task with some of the $LLDs$ and spectrograms. However, the data pre-processing techniques were not mentioned. Some other studies mostly focused only on a very specific step in the ML process or very specific ML algorithms. For example, Swain et al. (2018) concentrated mostly on emotion features and classifiers, or Lieskovská et al. (2021) and Schoneveld et al. (2021) focused on deep learning approaches. Furthermore, some of the studies only focused on addressing a particular challenge in SER. For instance, the data insufficiency and imbalance issues were solved by a particular data augmentation technique in Niu et al. (2017). Shih et al. (2017) address the negative impact of speaker difference in audio data using a Histogram Equalization technique. And Kerkeni et al. (2019) reduce the detrimental influence of audio signal trends using a zero-crossing rate detection strategy. Nevertheless, very few studies provide a comprehensive view of the current problems and solutions in the SER task.

We acknowledge some existing SER surveys over the last decade; such as Fahad et al. (2021); Yadav et al. (2021); Schuller et al. (2011); Latif et al. (2021); Akçay and Oğuz (2020); Mustafa et al. (2018); Swain et al. (2018); Anagnostopoulos et al. (2015); Schuller (2018); El Ayadi et al. (2011); Koolagudi and Rao (2012). Table 1 shows a comparison of these surveys with our review. These works mostly capture the key components of SER, including *database*, *data-processing*, *feature extraction*, *selection*, and *classification*. For example, Fahad et al. (2021) discussed both conventional and deep-learning techniques for SER. The authors presented three categories of databases including acted, evoked, and natural databases, and outlined their pros and cons. Details of challenges and approaches for solving issues related to natural environments such as speaker/language/textual dependencies were described. Acoustic and non-acoustic classes of speech emotion features were also discussed. They explain various techniques suitable for extracting emotion features. Common evaluation metrics were captured in the paper, but no information was provided on evaluation strategies, baselines, and data pre-processing. From databases to evaluation strategies, Schuller et al. (2011) present an end-to-end survey. The authors address SER tasks as both a classification and regression problem. They extensively cover SER components such as features, and feature selection methods. Their survey illustrates how to evaluate SER models using several strategies. Their work also captures the challenges surrounding the robustness of SER in noisy and cross-corpus environments, although it does not go into detail on the methodologies.

Koolagudi and Rao (2012) presented an SER survey that covers databases, features, and classifiers. They discussed the benefits of us-

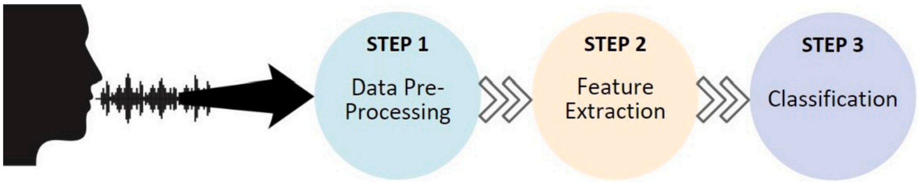

**Fig. 3.** Speech Signal Analysis.

**Table 1**
Comparison of the coverage of our work with existing SER surveys over the last decade.

| Survey Paper | Speaker Independent | Feature Extraction & Selection | Emotion Classifications | Data Processing | Evaluations Strategies |
|---|---|---|---|---|---|
| Fahad et al. (2021) | ✓ | ✓ | ✓ | ✓ | ✗ |
| Yadav et al. (2021) | ✗ | ✗ | ✓ | ✓ | ✗ |
| Latif et al. (2021) | ✗ | ✓ | ✓ | ✗ | ✓ |
| Akçay and Oğuz (2020) | ✗ | ✗ | ✓ | ✓ | ✗ |
| Mustafa et al. (2018) | ✗ | ✗ | ✓ | ✓ | ✗ |
| Swain et al. (2018) | ✗ | ✗ | ✓ | ✓ | ✗ |
| Schuller (2018) | ✗ | ✗ | ✓ | ✗ | ✗ |
| Anagnostopoulos et al. (2015) | ✗ | ✓ | ✓ | ✗ | ✗ |
| El Ayadi et al. (2011) | ✗ | ✓ | ✓ | ✓ | ✗ |
| Koolagudi and Rao (2012) | ✗ | ✗ | ✓ | ✗ | ✗ |
| ***Our Paper*** | ✓ | ✓ | ✓ | ✓ | ✓ |

ing different databases including elicited, acted, and natural databases. They also covered different classification techniques and categories of features including prosodic, excitation, and vocal-tract features. However, they do not capture the techniques to deal with the speaker dependency issues and do not discuss emerging ML techniques used in SER. The survey presented in El Ayadi et al. (2011) addressed speech emotion classification with a focus on three key components of SER including feature selection, speech classification schemes, and preparation of speech emotion databases. The authors discussed issues associated with the feature extraction step in SER, and presented different categories of features including *global and local* and *continuous and voice quality* features and explained the influence of each category on SER performance. However, this survey does not cover speaker-independent challenges and SER evaluation techniques. Similarly, Swain et al. (2018) presented an SER survey that captures components such as databases, features, and different classifiers for speech emotion recognition in several languages. Deep learning, hybrid, and fusion techniques for emotion classification were also discussed in their research. However, their survey did not discuss challenges and the current state-of-the-art approaches for achieving speaker independence, feature selection, evaluation, and several key techniques for speech data pre-processing.

The survey conducted by Akçay and Oğuz (2020) captured essential components of SER including databases, features, emotional models, preprocessing, supporting modalities, and emotion classification. Two methods of modelling emotions and their trade-offs were also discussed. The first approach is based on discrete models, which are based on the six categories of independent emotions: sadness, happiness, fear, anger, disgust, and surprise. These emotions are mostly observed in human daily activities and are only experienced for short periods of time. However, discrete emotions are not able to capture some of the more complex emotional states observed in human communications. Dimensional models, on the other hand, make use of a small number of latent dimensions to characterize human emotions such as *valence*, *arousal*, *control*, and *power*. In a systematic way, these emotions are analogous to one another. The authors also discuss various ML-based and emerging deep learning classifiers such as 3D-CNN and auto-encoders. Although this survey does provide various techniques for speech data preprocessing, it does not include evaluation strategies, speaker adaptation and language-independent techniques, and feature selection.

Our review builds on these prior studies (Torres-Carrión et al., 2018; Grant & Booth, 2009; Kitchenham et al., 2010) by consolidating and bridging the knowledge gap with respect to the state-of-the-art in SER. Unlike the previous studies, this work followed a three-step systematic review, which comprehensively captured existing and emerging methods, techniques, and strategies used for carrying out the main steps in SER implementation. This approach enables us to provide fine-grained information missed in the prior surveys. Moreover, this work provides specific information regarding SER evaluation, and speaker-independent experimentation concerns, specifically, the reason behind the low-classification-accuracy issue associated with speaker-independent SER during the evaluation process in the vast majority of the studies. We further provide a summary of reviewed papers used in this work under the following headings: *Paper Title*, *Signal Pre-processing*, *Feature Extraction and Feature Selection*, *Emotion Classes*, *Machine Learning Algorithms*, *Evaluation Criteria*, and *Classification Results* (see Appendix A).

## 3. Research methodology

To summarize and integrate the available knowledge in the field of ML and SER, we adapted a systematic literature review research methodology based on Xiao and Watson (2019) guidelines to form a more comprehensive knowledge base. This systematic review was conducted in $PubMed$, $IEEE$, $Scopus$, and $GoogleScholar$ databases. The following search queries were constructed to identify the relevant research articles: *‘Automatic’* $AND$ *(‘Emotion Detection’ OR ‘Emotion Recognition’)* $AND$ *(‘Audio’ OR ‘Speech’)* $AND$ *( ’AI’ OR ‘Artificial Intelligence’* $OR$ *‘ML’* $OR$ *‘Machine Learning’* $OR$ *‘Neural Network*’* $OR$ *‘NN’ )*. The search results were limited to English, full text, with publication years ranging from 2011 to 2021 only to capture ML techniques’ trends and evolutions in SER. From the initial 263 collected articles, 95 met the exclusion criteria (Table 2).

We used thematic analysis (Braun & Clarke, 2022) to extract information from the 95 studies, using *NVivo*, a qualitative data analysis software. Based on deductive and inductive reasoning (Yi, 2018) different categories were created to cover different stages of ML for SER (Table 3). Each category provides a reflection on the importance of the stage and ML techniques and methods on the overall performance of SER. The procedure of our systematic review is presented in Fig. 4.

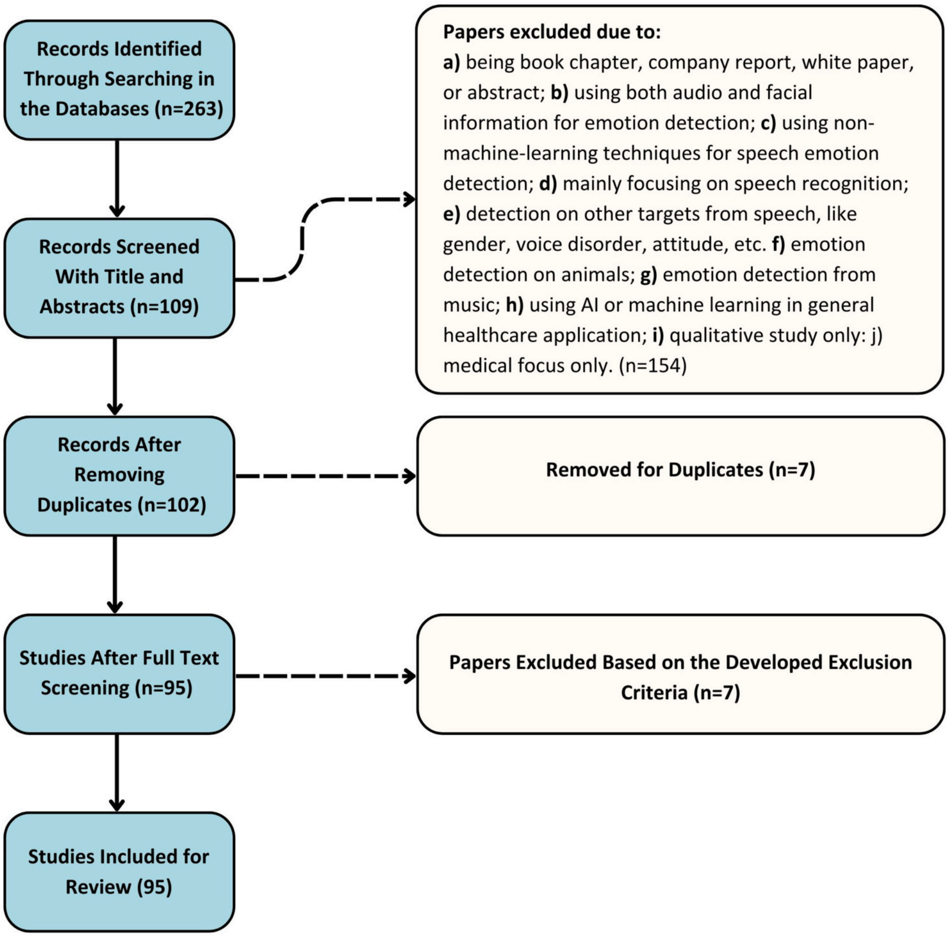
Records Identified Through Searching in the Databases (n=263)
Records Screened With Title and Abstracts (n=109)
Papers excluded due to:
a) being book chapter, company report, white paper, or abstract; b) using both audio and facial information for emotion detection; c) using non-machine-learning techniques for speech emotion detection; d) mainly focusing on speech recognition; e) detection on other targets from speech, like gender, voice disorder, attitude, etc. f) emotion detection on animals; g) emotion detection from music; h) using AI or machine learning in general healthcare application; i) qualitative study only: j) medical focus only. (n=154)
Records After Removing Duplicates (n=102)
Removed for Duplicates (n=7)
Studies After Full Text Screening (n=95)
Papers Excluded Based on the Developed Exclusion Criteria (n=7)
Studies Included for Review (95)

**Fig. 4.** Study selection process.

**Table 2**
Exclusion criteria used for literature selection.

| | Exclusion Criteria |
|---|---|
| **1** | Book chapters, company reports, white papers, or abstracts |
| **2** | Studies in emotion detection or recognition using both audio and facial information |
| **3** | Studies in speech or audio emotion detection/recognition using non-machine-learning techniques |
| **4** | Studies with the main focus on speech recognition |
| **5** | Studies with the detection o other targets from speech, like gender, voice disorder, attitude etc. |
| **6** | Studies with emotion detection on animals |
| **7** | Studies with emotion detection from music |
| **8** | Studies with Artificial Intelligence or ML in general healthcare application |
| **9** | Studies with qualitative study only |
| **10** | Studies with a medical focus only |

**Table 3**
ML Aspects of SER.

| | ML Aspects of SER |
|---|---|
| **1** | Emotion datasets |
| **2** | Data pre-processing |
| **3** | Sampling, quantisation, and pre-emphasis |
| **4** | Audio features |
| **5** | Feature extraction |
| **6** | Emotion classification |
| **7** | Evaluation criteria |
| **8** | Speaker-Independent SER |

## 4. Findings and discussion

Our findings are based on the outcome of our systematic review in four scientific databases (Fig. 5) and presented in SER three steps of data pre-processing, feature extraction, and emotion classification. The findings include the utilised methods or techniques in each of the steps and some special methods.

### *4.1. Data pre-processing*

This step is the foundation in speech processing-based applications (e.g., SER) which can have a significant impact on the subsequent feature extraction and classification steps. If this step is undertaken effectively, it can improve the overall recognition performance (Semwal et al., 2017). However, before starting the pre-processing the speech emotion datasets should be prepared. The majority of the studies used the available speech-emotion datasets. The common datasets are presented in Table 4. Other utilised datasets were: MSP-Improv (Neumann & Vu, 2019), eNTERFACE (Xie et al., 2019), Marathi (Darekar & Dhande, 2018), and TESS (Mekruksavanich et al., 2020).

One of the major goals of the pre-processing step in SER implementation is to remove unwanted noises from audio signals. Audio signals get disturbed by the noise signals at varying decibels such as 0 db, 5 db, and 10 db (Niveditha & Ashok, 2019). Such noise can negatively affect the SER performance. It is worth noting that some pre-processing techniques are employed to facilitate feature extraction processes like feature normalization, which could help reduce the effects of variations of speakers on the recognition process (Akçay & Oğuz, 2020).

There are two routes of pre-processing for audio data with respect to expected features to be extracted (see Fig. 6). First, is the traditional route for $LLDs$ extraction. This includes common audio pre-processing

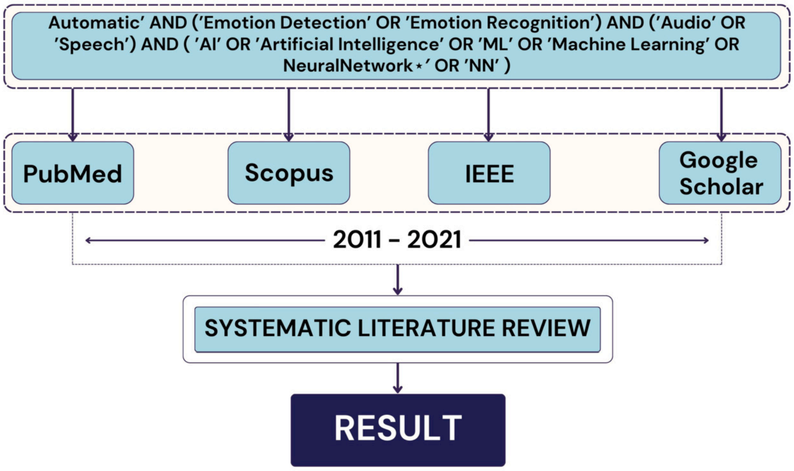
Automatic' AND ('Emotion Detection' OR 'Emotion Recognition') AND ('Audio' OR 'Speech') AND ( 'AI' OR 'Artificial Intelligence' OR 'ML' OR 'Machine Learning' OR NeuralNetwork*' OR 'NN' )
PubMed
Scopus
IEEE
Google Scholar
2011 – 2021
SYSTEMATIC LITERATURE REVIEW
RESULT

**Fig. 5.** SER results based on SLR.

**Table 4**
Speech emotion databases.

| Dataset Name | Sample of Reference Studies | Modality | Actors | Language | Emotion Categories |
|---|---|---|---|---|---|
| IEMOCAP | Niu et al. (2017), Pandharipande et al. (2018), Zhao et al. (2019a), Ramet et al. (2018), Jalal et al. (2019), Neumann and Vu (2019), Jiang et al. (2019), Suganya and Charles (2019) | Audio-Visual | 5 Male and 5 Female | English | Anger, Happiness, Excitement, Sadness, Frustration, Fear, Surprise, Neutral and other |
| SAVEE | Yogesh et al. (2017a), Sivanagaraja et al. (2017), Liu et al. (2018), Mekruksavanich et al. (2020) | Audio-Visual | 4 Male | English | Angry, Disgust, Sad, Fear, Happy, Surprise, and Neutral |
| RAVDESS | Darekar and Dhande (2018), Mekruksavanich et al. (2020), Jalal et al. (2019) | Audio-Visual | 12 Male and 12 Female | English | Happy, Sad, Angry, Fearful, Surprise, Disgust, Calm, and Neutral |
| Emo-DB | Pandharipande et al. (2018), Zhang et al. (2018a), Niu et al. (2017), Zhou et al. (2016), Suganya and Charles (2019), Meng et al. (2019) | Speech | 5 Male and 5 Female | German | Angry, Boredom, Disgusting, Fear, Happy, Sad and Neutral |
| CASIA | Wen et al. (2017), Xie et al. (2019), Liu et al. (2018), Mao et al. (2019), Ke et al. (2018) | Speech | 2 Males and 2 Female | Chinese | Anger, Fear, Happy, Neutral, Sad and Surprise |

methods such as *sampling*, *pre-emphasis*, *windowing and framing*, and *denoising*. Tools like *OpenSMILE* (Eyben et al., 2010) are commonly used for data transformation and feature extraction. The other route captures methods and techniques that transform audio signals into *spectrograms* and *wavelets*, which are used for the SER task. Besides the framing, windowing and pre-emphasising, additional approaches, which are based on Fourier Transform and Wavelet-transform techniques, are used to transform audio segments from time-frequency domain signals into spectrograms and wavelets (Vasquez-Correa et al., 2016; Fahad et al., 2021). A spectrogram represents the strength or loudness of a signal over time at different frequencies in a particular waveform (Badshah et al., 2019). Wavelets give a short-term multi-resolution analysis of time, energy, and frequencies in a speech signal (Schuller et al., 2011).

Some studies use supporting methods in the data pre-processing step to deal with particular challenges while implementing ML-based SER systems (see Table 5). An example of these methods includes the use of a data augmentation strategy for generating more audio data. A detailed description of the pre-processing methods and other supporting methods used during SER audio data processing is provided in the next section.

### 4.2. Sampling, quantisation, and pre-emphasis

*Sampling* is the process of converting the raw audio signal into a digital signal (Niveditha & Ashok, 2019). The sampling rate, the primary parameter of sampling, refers to how many samples per second are extracted from the raw and continuous signal to create the digital and discrete signal. Another application of sampling in SER (e.g., down-sampling technique) is usually used to provide data augmentation (Fayek et al., 2015) and decompose the signal to reduce the influence of perturbations, like jitters, and flutter (Sivanagaraja et al., 2017). The *quantisation* happens with sampling in digitising the raw audio signal and is used to quantize the sampling values to the nearest levels (Niveditha & Ashok, 2019). The pre-emphasis is deployed to augment the effects of high and low frequency to reduce the noise of audio signals. A higher frequency's amplitude will be increased whereas a lower frequency's amplitude is decreased, as the former usually contains relatively more important information compared to the latter, which is blended with noise (Gunawan et al., 2018).

#### 4.2.1. Framing

Framing is the process of partitioning input speech signals into fixed-length segments; where each segment is referred to as a *frame* and thus helps to visualise each frame as an isolated feature vector. Frameshift/overlapping is a key concept in framing, which describes overlapping portions between current and previous frames. The frame size and the overlap ratio are determined during the pre-processing to help differentiate the starting points of each frame and the duration of the frames (Deng et al., 2013; Basu et al., 2017). The overlap rate is employed to reduce information loss and soften the transition from one frame to the next (Ozseven, 2018).

Although several studies did not apply overlapping during framing and windowing (Lalitha et al., 2019; Muthusamy et al., 2015), Sivanagaraja et al. (2017) noted that overlapping could be better than non-overlapping, especially for SER. Framing is usually carried out to address other specific challenges during speech signal processing. In SER, continuous speech signals restrain the use of processing techniques such as Discrete Wavelets Transform (DWT) and Discrete Fourier Trans-

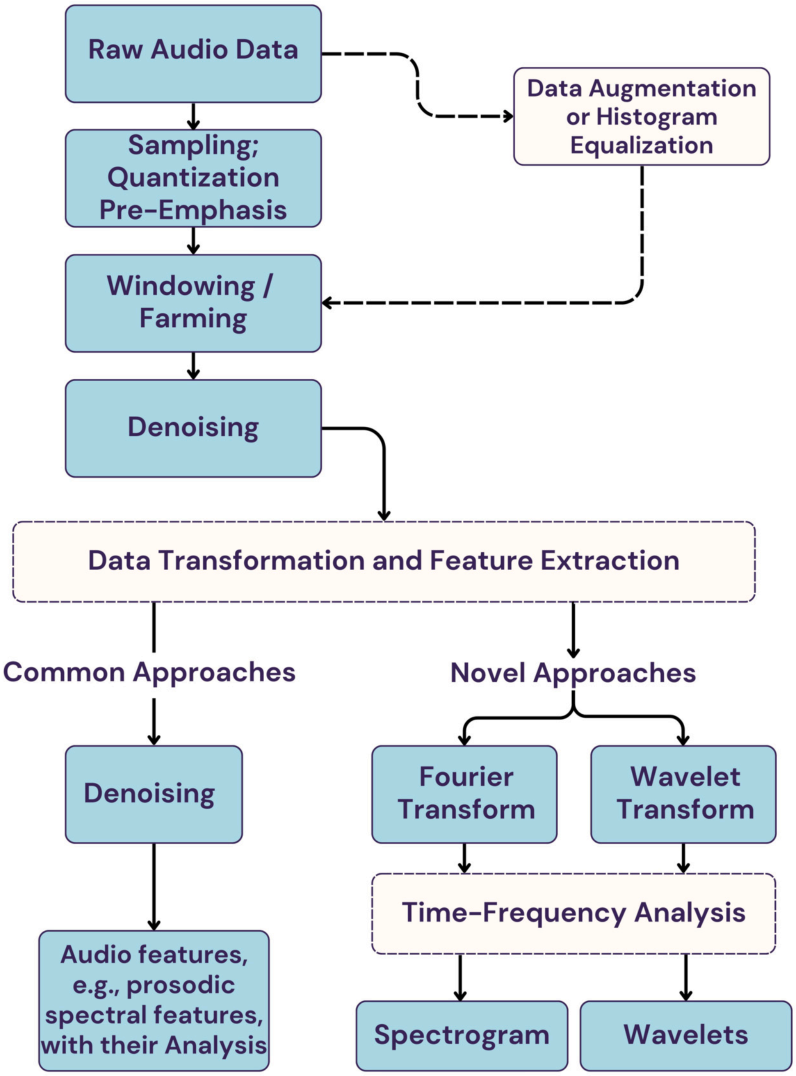
Raw Audio Data
Data Augmentation or Histogram Equalization
Sampling; Quantization Pre-Emphasis
Windowing / Farming
Denoising
Data Transformation and Feature Extraction
Common Approaches
Novel Approaches
Denoising
Fourier Transform
Wavelet Transform
Time-Frequency Analysis
Audio features, e.g., prosodic spectral features, with their Analysis
Spectrogram
Wavelets

**Fig. 6.** Data pre-processing and feature extraction routes.

**Table 5**
Common SER challenges and solutions.

| Challenges | Solutions |
|---|---|
| Silence and noise removal | $AFFT$ and adaptive filtering;<br>The algorithm proposed by Saha et al. (2005);<br>Google $WebRTC\ VAD$. |
| Audio data insufficiency | Data augmentation |
| Data imbalance | Data augmentation; Skew-robust technique. |
| Speaker differences in audio data | Cross-Speaker Histogram Equalisation ($CSHE$) |
| Signal trend issues | Removing signal trend by zero-crossing rate detection method. |
| Perturbations in audio signal | Reducing the impact of perturbations by using down-sampling to decompose the signal. |
| Solve the non-stationary and non-linearity issues | $EMD-TKEO$ technique<br>*(Empirical Model Decomposition - Teager-Kaiser Energy Operator)* |
| Redundant and irrelevant information in LLD and HSF features | Feature selection methods, e.g.,<br>$PCA$, $LDA$, correlation analysis and Fisher Criterion |
| Speaker-independent challenges | Attention model; $LFLB$ (*Local Feature Learning Block*); Enhanced*Gaussian Mixture Model* audio features; 3 Dilated $CNNs$ for feature extraction on 3 transformations (based on delta) of the log- spectrum; A combination loss function of *centre-loss* and *softmax* to train neural network |

form (DFT) during feature extraction. As a result, fixed-size frames are suitable for classifiers (e.g. NN-based classifiers) while preserving emotion information in speech. Furthermore, because the speech signals are non-stationary, emotion can vary in the course of a speech, however, speech stays invariant for a brief amount of time ranging from 10 to 30 ms. This quasi-stationary condition can be approximated and local characteristics retrieved by framing the speech signal (Akçay & Oğuz, 2020). The recommended frame size is $10-20$ ms, as advised by Rabiner (1978).

### 4.2.2. Windowing

Following the framing phase of the speech signal, the next step is to apply a *window function* to the frames. The function is used to mitigate the effects of spectral leakages generated by discontinuities

**Table 6**
Windowing Approaches for SER.

| | Windowing Approach |
|---|---|
| 1 | Hamming Window: the most common approach in SER applications which removes the discontinuities and decreases ripples in each frame (Lokesh & Devi, 2019; Lalitha et al., 2019). Hamming window size of 40 ms could lead to better emotion classification performance (Zhang et al., 2018b). |
| 2 | Rectangular Window: this approach is considered the simplest window that was used in Badshah et al. (2019) for feature extraction spectrograms. |
| 3 | Triangular Window: This method was used only in a very few research studies. Niveditha and Ashok (2019) and Jain et al. (2011) applied a series of triangular windows to find the weighted spectral components in their SER applications. |
| 4 | Hann Window: It is a window function used to perform Hann smoothing (Essenwanger, 1986). It is used in signal processing to select a subset of a series of samples in order to perform Fourier transform or other calculations. |

at the signal's edges during the Fast Fourier Transform (FFT) (Akçay & Oğuz, 2020). After the speech signal is framed, each frame is then windowed to eliminate discontinuities at the beginning and end of the frame and adjust spectral infiltration in the signal and intersection due to overlaps (Ozseven, 2018). Lokesh and Devi (2019) points out that the windowing and framing approaches are used to remove the Gaussian white noise in the audio signal. The identified windowing approaches in our systematic review are discussed in Table 6.

Other forms of windowing techniques such as *Bartlett, Kaiser and Gaussian* windowing functions (Pereira et al., 2016) exist but are rarely used in SER. Based on the investigation of Pereira et al. (2016) on the impact of different windowing functions on SER performance, Hamming and Hanning's windowing approaches performed better than the other techniques while *Bartlett* and *Rectangular* performance were poor. There are few studies discussing the influence of the size of the window on performance's SER. Among 32 studies that mentioned their windowing details, 22 of them use a window size of 25 ms to process the frames. The overlapping/slide size is usually $1/3$ or $1/2$ of the frame size (Rajisha et al., 2016) and the studies that choose a 25 ms window size all set the overlapping size as 6, 10, or 15 ms.

### 4.2.3. *Noise and silence removing*

Noise and silence should be removed to decrease the negative impact of unwanted data on speech signals. The audio signals used for training the ML model are usually recorded in a normal environment with noise backgrounds. Therefore it is necessary to isolate the speech from the noise (Kannadaguli & Bhat, 2019). Moreover, the removal of silence in speech can help some deep learning models (e.g., attention model) to focus on specific parts (Mu et al., 2017). Four studies incorporated silence and/or noise removal steps in their data pre-processing phase. Different techniques were used for this removal with no evidence showing the comparison of these techniques in their contribution to SER performance.

Hussain et al. (2017) used FFT and adaptive filtering for noise and silence removal. An algorithm proposed by Saha et al. (2005) was also used by Kannadaguli and Bhat (2019) to do the same task. A WebRTC Voice Activity Detector (VAD) developed by Google is used by Harár et al. (2017); Mu et al. (2017) to remove the silent parts in the speech sample.

### 4.2.4. *Fourier transforms*

To analyze speech signals more effectively, the frequency domain acoustic features can be used (Fahad et al., 2021). However, raw speech signals are in time-amplitude waves in which the frequency information is not contained, whereas spectrograms require frequency information. Hence, the $FT$ is applied to transfer the raw audio wave from the time domain to the time-frequency domain to gain the frequency information. The Fourier Transform $(FT)$ is famous for transforming the signal into a particular domain to gain its amplitude, phase, and frequency. $FT$ is commonly employed in DNN-based SER applications to extract deep features from the spectrograms (or *Mel-spectrograms*) of the audio signal which requires frequency information. The spectrograms (or *Mel-spectrograms*) can be derived from the frame transformed by $FT$ and the determination of the spectrogram segment size should be considered. The size of the spectrogram segment, in other words, the number of frames contained in a segment, should be long enough to contain sufficient emotional information. Although the specific standard for a segment length $L$ is still an open question in this research area (Zheng et al., 2015), studies such as Kim and Provost (2013) and Provost (2013) pointed out that segments with $SegmentLengthL$ greater than 250 ms could contain enough emotional information.

The common FT-based transformation techniques used in existing SER applications are presented in Table 7.

**Table 7**
Fourier Transforms for SER.

| | Approach for Fourier Transforms |
|---|---|
| 1 | Fast Fourier Transform (FFT): To convert time domain speech signals to frequency domain $FFT$ is used (Basu et al., 2017; Zheng et al., 2015; Rawat & Mishra, 2015; Aouani & Ben Ayed, 2018). FFT generates the frequency spectrum of each frame of a speech signal to obtain the spectrum features in a particular frame. |
| 2 | Discrete Fourier Transform (DFT): DFT converts a finite sequence of equally-spaced samples of a function into a same-length sequence of equally-spaced samples of the Discrete-Time Fourier Transforms (DTFT). DFT is mostly useful to transform temporal speech samples into spectra (Zhu et al., 2017). This allows better compression (Nasreen et al., 2016). DTFT is usually used for frequency analysis of discrete signals, however, since discrete-time signals are continuous, DTFT is not suitable for numerical systems. DFT is usually derived from DTFT when conducting frequency analysis. |
| 3 | Discrete Cosine Transform (DCT): DCT is a reversible frequency domain transformation (Ramakrishnan et al., 2015). It collects cosine waves with varying frequencies to represent data. Due to DCT's symmetrical elongation features, the application of DCT involves less interruption when compared to DFT (Sonmez & Varol, 2019; Kerkeni et al., 2019). |
| 4 | Short-Time Fourier Transform (STFT): FT only provides amplitude with frequency functions, not the time instants at which the frequencies occurred. To overcome this limitation, STFT is a more advanced type of FT which is used to overcome the restriction of the traditional FT (Kerkeni et al., 2019; Zheng et al., 2015). It captures both temporal and spectral information (Fahad et al., 2021). |

### 4.2.5. *Wavelet transform*

As Fourier Transform cannot simultaneously capture time and frequency components of the speech signal, it has some limitations. Emotional speech varies considerably with time more than neutral speech (Fahad et al., 2021), therefore details and rapid component changes in an emotional speech might be difficult to track with a fixed window length of STFT. To address this issue Wavelet Transform (WT) is often used. WT decomposes the signal into high-frequency and low-frequency components which provides better time resolution for high-frequency components and better frequency resolution for lower-frequency components. WT has been used for Time-Frequency analysis and to model emotional speech (He et al., 2013; Wang et al., 2020). Wavelets are a class of functions that are localized both in the time and frequency domains. Each wavelet in the classical setting is a variation of a single wavelet $\psi$ known as the mother wavelet (Ha et al., 2021). The following are different variations of Wavelet Transform:

- **Discrete Wavelet Transform (DWT):** is a wavelet transform with discretely sampled wavelets. A fundamental benefit it has over Fourier Transforms is temporal resolution; it captures both frequency and time information. A group of discrete wavelets can be generated by scaling and translating the mother wavelet into discrete increments (Ha et al., 2021). The time-frequency multiresolution capability of DWT was explored in Palo and Mohanty (2015) which led to better emotion classification.

- **Continuous Wavelet Transform (CWT):** is a non-numerical transformation approach that provides an over complete representation of a signal by letting the translation and scale parameter of the wavelets vary continuously.
- **Wavelet Packet Transform (WPT):** is an extension of DWT and is a generalisation of wavelet decomposition with a wider variety of signal processing capabilities. A wavelet packet is a collection of wavelet functions that have been linearly concatenated (He et al., 2013). As a result, wavelet packets inherit the orthonormality and time-frequency localisation properties of their respective wavelet functions. Instead of only decomposing the low-frequency representation, the entire time-frequency plane can be subdivided into distinct time-frequency components. Wavelet packet analysis has the advantage of allowing the combination of several levels of decomposition to produce the best time-frequency representations of the original signal.

Some SER studies have been leveraging the characteristics of Wavelet Transform to enhance the performance of their application. For instance, Shegokar and Sircar (2016) proposed a speaker-independent SER using features selected based on CWT and prosodic coefficients. Huang et al. (2015) proposed a Wavelet packet filter-bank-based acoustic feature extraction approach for SER that considerably improved emotion recognition performance over the traditional MFCC feature.

#### 4.2.6. Data augmentation

Two common problems for the speech database for SER are: (1) the amount of audio data is not large enough, especially for training a deep learning classifier; (2) the data in different emotion classes is often unbalanced. These two problems would considerably influence emotion classification performance. To address these issues, a data augmentation strategy is being used. Data Augmentation is designed to generate more data to increase the amount of data and to balance the data in different emotion classes. This method can decrease models' over-fitting which leads to enhancing the model generalization. However, among all the included studies in this review, only very few explain their data augmentation strategies and those achieved higher classification performance.

Data Augmentation Algorithm Based on Retinal Imaging Principle (DAARIP) is proposed in Niu et al. (2017). It compared the classification performance between using the raw data and using the augmented data to train the same deep learning architecture. The result demonstrated a significant increase in classification accuracy. $DAARIP$ simulated the relationship between focal distance and image size in human retina. It receives the spectrograms of the speech signal and generates different sizes of images by changing the focal distance.

In another study, the log Mel-spectrogram extracted from an utterance was divided into a certain number of overlapping segments to augment data (Zhao et al., 2019a). Fayek et al. (2015) used a re-sampling strategy to generate more audio data at different sampling frequencies. Although other methods such as adding Gaussian noise, and Vocal-Tract Length Perturbation (VLTP) were explored in their experiments, re-sampling proved to be more successful. In their research, they used the augmented data for training the $DNN$ which significantly improved its generalisation ability. Zhang et al. (2018a) created more audio data by setting the overlapping size to a relatively small number, while in Ramet et al. (2018), random white noises were added to augment data. Other methods such as RNN to utilize past speech frames (Meng et al., 2019) and Multi-Task Learning (Singh et al., 2019) were also suggested.

#### 4.2.7. Cross-speaker histogram equalisation

Another challenge in SER by ML is the differences (e.g., emotion ambiguity) in speech data caused by speaker differences (e.g., gender and age). This challenge is considered to contribute to the low SER performance (Shih et al., 2017). A method called $(CSHE)$ is experimented in Shih et al. (2017) to normalise the training data and the result shows it considerably increases the SER performance by neutral network and $(SVM)$ classifiers.

#### 4.2.8. Signal trend removal

The existence of a signal trend in the speech could be detrimental to the SER task. Great error could be caused by the signal trend in the correlation analysis of time domain (raw audio wave), in the spectral analysis of frequency domain (spectrograms), and in the authentication of low-frequency spectrum (Kerkeni et al., 2019). The signal trend removal is based on the zero-crossing rate detection method (Li & Li, 2011). This method mentioned in Kerkeni et al. (2019) includes two steps. The first step is calculating the signal trend by setting a particular threshold. It could be shown as the two formulae:

$$\frac{R_{IMF_i}}{R_{IMF_1}} < 0.01 \ (i = 2, 3, 4, \ldots n) \tag{1}$$

The $IMF$ is the Intrinsic Model Function derived by $EMD$ (introduced in Section 4.2.1) and $R$ is the *zero-crossing rate*.

$$T(t) = \sum_i IMF \tag{2}$$

$T(t)$ is the signal trend.

The second step is the subtraction of the signal trend from the original speech signal:

$$S(t) = X(t) - T(t) \tag{3}$$

$X(t)$ is the original speech signal. $S(t)$ is the derived signal by removing the trend.

### 4.3. Audio features and feature extraction

This step happens on the speech signals from the pre-processing step and will determine the inputs for training the ML model to make emotion classification. Extraction of the necessary information from the speech signal is a vital step in the SER system (Kerkeni et al., 2019), and has a direct impact on the system performance. Feature extraction is necessary to obtain characteristics from speech signals that are indicative of the emotional content of the speech (Semwal et al., 2017). Feature extraction also aims to remove redundant and unnecessary features leading to enhancing the models' accuracy. The essential consideration for the feature extraction step includes the types of audio features appropriate for the SER task and how to extract them.

Based on our findings, the audio features are mainly categorised into traditional (or hand-crafted) and advanced features. Traditional features include the prosodic, spectral, and acoustic features (Alu et al., 2017), and their statistical features (Zhu et al., 2017). Prosodic features, such as speech intensity, pitch, energy, zero crossing rate, refer to the musical aspect of speech (Alva et al., 2015) and are considered as the main indicator of a speaker's emotional (Ortony et al., 1990; Fei et al., 2016). Spectral features (i.e., the speech signal's short-time representation (Fei et al., 2016)), are calculated by the vocal tract system (Lalitha et al., 2019). MFCC and Linear Predictor Coefficient (LPC) are two common spectral features in the studies. The prosodic and spectral features are the most common traditional features used in SER studies. Nevertheless, some studies (e.g., Alu et al., 2017; Liu et al., 2018) use acoustic or quality features as a supplement.

The mentioned features are Low-Level Descriptors $(LLD)$ which reflect the audio characteristics in the word or frame level. These LLD descriptors are used to derive time, spectral and cepstral domain, for short speech frames of the speech signal (Semwal et al., 2017). On the other hand, to recognise the temporal variations and outlines of the speech, a series of high-level statistical features $(HSF)$ are applied on the LLD at the utterance level (Mirsamadi et al., 2017), such as maximum, minimum, or variance of pitch and $MFCC$. These $HSF$ are dynamic and thus assumed to contain emotional contents more than

**Table 8**
Audio features.

| Categories | Audio features |
|---|---|
| *Prosodic* **features** (Zhu et al., 2017; Anagnostopoulos et al., 2015) (Zhao et al., 2019c) (Alu et al., 2017) | Intensity, pitch (fundamental frequency, energy, zero crossing rate, Harmonics-to-noise Ratio (HNR), shimmer, jitter, speech rate, normalized amplitude quotient, spectral tilt, spectral balance |
| *Spectral* **features** (Zhu et al., 2017; Anagnostopoulos et al., 2015) (Alu et al., 2017) (Fei et al., 2016; Zhao et al., 2019c) | Linear Predictor Coefficient(LPC); Linear Predictor Cepstral Coefficient; Log-frequency Power Coefficient (LFPC); Line Spectrum Pair (LSP); Perceptual Linear Prediction Cepstral Coefficients (PLP); MFCC, One-Sided Auto-correlation Linear Prediction Coefficient (OSALPC); One-Sided Auto-correlation Linear Prediction Cepstral Coefficient (OSALPCC); Zero Crossing Amplitude Peak (ZCPA) |
| *Acoustic features* (Alu et al., 2017; Liu et al., 2018) | Formant frequency, glottis parameters (breath sound, brightness, and throat sound) |
| *Statistical feature* (Koolagudi et al., 2018; Shih et al., 2017) (Huang et al., 2018) | Maximum, minimum, mean, standard deviation, variance, *kurtosis*, *skewness*, relative position, range, Mean Square Error of linear regression. |
| *Deep features* (Fahad et al., 2021; Khalil et al., 2019) | The features extracted by deep neural network |
| *Features based on EMD and TKEO* (Kerkeni et al., 2019) | The features extracted based on TKEO and EMD, like Mel Frequency Cepstral Coefficients based on the Reconstructed Signal (SMFCC); Energy Cepstral Coefficients (ECC); Frequency-weighted Energy Cepstral Coefficients (EFCC); Modulation Spectral (MS); Frequency Modulation (MF) |

the static $LLD$. The usage of $LLD$ and $HSF$ for the traditional SER by $ML$ method usually includes a large number of combinations.

In addition to traditional features, two categories of advanced features are now available, hypothesized to unveil a richer array of emotional information embedded within the speech. The first type of advanced features is deep features. These features are not obtained manually like the $LLD$ and $HSF$, instead, they are captured automatically by the advanced ML algorithm — deep neural networks. The deep features can be extracted from both raw speech clips and handcrafted features (Zhao et al., 2019b). The second type is the features based on Empirical Model Decomposition ($EMD$) and Teager-Kaiser Energy Operator ($TKEO$) techniques (Kerkeni et al., 2019).

EMD is introduced to solve the stationarity and non-linearity issues in the speech signal. $STFT$ could alleviate the stationarity issue but leave the non-linearity problem unsolved, whereas the EMD method could decompose and analyse the non-linear and/or non-stationary speech signal while preserving the original qualities (Kerkeni et al., 2019). The TKEO technique is used to track the real-time change of the amplitude and frequency of the AM and FM (Amplitude and Frequency Modulation) constituents (Kerkeni et al., 2019). The $EMD$ method could be used to calculate the $IMFs$ (Intrinsic Model Functions), based on which the signal trend can be removed (Section 4.2.8) from the original signal and then the $SMFCC$ feature could be derived from the new signal with $FFT$ and $DCT$ (Discrete Cosine Transform). The combination of $EMD$ and $TKEO$ technique can be used to extract modulation-related features, like Modulation Spectral ($MS$), and the cepstral-based features, like Energy Cepstral Coefficients ($ECC$). The SMFCC and MS features are found to model human auditory perception better than the MFCC does. The cepstral features have similar functionality to the MS while containing different emotional information. The experiment in Kerkeni et al. (2019) showed the combined usage of the two types of features above considerably improved the SER performance.

The comprehensive list of audio categories and the related features is provided in Table 8.

To extract the mentioned features, traditional pre-processing and audio feature extraction for SER by ML is being used. This feature extraction relies on specific tools, that based on our findings, can be categorized into three types: (1) standard $toolkits$ for audio feature extraction, such as $openSMILE$, $PRAAT$, $APARAT$, and $openEAR$ (Yogesh et al., 2017b). The most common toolkit is $openSMILE$, used by 12 out of 17 studies for audio feature extraction. (2) libraries in programing platforms (e.g., Python and MATLAB). For instance, the *PyAudioAnalysis* is used in Giannakopoulos (2015); Manamela et al. (2018) and is a powerful and specialised Python library for extracting audio features (Manamela et al., 2018). (3) lastly, models developed by SER researchers for audio feature extraction for particular purposes.

As discussed in Kerkeni et al. (2019), different feature extraction techniques can be used including Linear Discriminant Analysis (LDA), Principle Component Analysis (PCA), Sequential Forward Selection (SFS), and Recursive Feature elimination (RFE). The developed models for SER feature extraction are numerous. For example, $VGGVox$ is developed by Nagrani et al. (2017) and has been trained with 2000 hours of audio files by over 1200 different speakers to be capable to extract speaker-specific features (Assunção et al., 2019). The extraction of deep features is performed by ($DNN$) automatically. Zhang et al. (2018a) used a simplified Deep CNN model for feature extraction. $DNNs$ is commonly used for extracting deep features include $CNN$ on time-frequency spectrograms (Badshah et al., 2017) or raw audio waves (Bertero & Fung, 2017), $DNN$ on Mel-spectrograms (Fayek et al., 2017) and auto-encoder on spectrograms (Fei et al., 2016). Deep Belief Network was used for classification and feature extraction in Zhu et al. (2017).

The drawing of features from the raw speech signal usually happens on the raw time amplitude waves or spectrograms (or Mel-spectrograms) of the signals. The Mel-spectrogram is derived based on the logarithmic scaling of the spectrogram with Mel-filter banks to make it close to the human's audio perception mechanism (Singh et al., 2019). Therefore, MFCC is used for feature extraction as one of the most effective methods (Rawat & Mishra, 2015).

### 4.3.1. Feature selection

The $LLD$ and $HSF$ features extracted are sometimes not appropriate to be directly used for emotion classification. The huge number of combinations of $LLD$ and $HSF$ features may contain redundant and irrelevant information and thus influence the final SER accuracy (Kerkeni et al., 2019).

To address this issue and reduce the classification cost feature selection method is used that provides selecting the most relevant features for SER task. Different feature selection methods were used in the reviewed studies. Recursive Feature Elimination (RFE) is used by Kerkeni et al. (2019), the combination of *GainRatio*, *InfoGain*, and OneR is used in Getahun and Kebede (2016). Principle Component Analysis (PCA) is used in Ke et al. (2018) and Liu et al. (2018) in which Linear Discriminant Analysis (LDA) is also deployed. Liu et al. (2018) explored a special feature selection method by correlation analysis and Fisher Criterion to perform and the result showed some improvements in final SER performance. To the best of our knowledge, there is no study comparing feature selection methods in terms of their contribution to final SER accuracy. Thus further investigation on comparing the performance of these methods is necessary.

$PCA$ is a famous feature dimensionality reduction method which is based on the analysis of the correlation of the attributes. PCA uses eigenvectors to represent the covariance of samples' vectors in the original dataset, then uses the absolute value of the eigenvectors to evaluate the i-th feature component of samples with regard to its statistical contribution to the feature extraction result and finally drop the not important feature components (Song et al., 2010). PCA focuses more on the global structure of the data rather than the local structure. RFE performs the feature selection using an ML model (e.g., SVM) to evaluate the importance of the features. The features with the least values will be dropped and the rest features are assigned with respective weights, which means RFE selects the features recursively by assessing fewer and fewer features (Kerkeni et al., 2019). On the other hand, LDA decreases the feature dimensionality by measuring the information in different labels or classes. This makes the number of selected features more relevant to the number of classes compared to $PCA$ (Liu et al., 2018). LDA increases the discrimination among diverse classes of features after reducing the feature dimensionality (Arjmandi & Pooyan, 2012). The detailed algorithm of LDA refers to Liu et al. (2018).

The correlation analysis and Fisher Criterion are also great feature selection methods. The correlation analysis firstly uses the partial correction analysis to remove the mutual features that influence each other on representing the same emotional state and then applies Spearman rank correlation coefficient to measure the statistical correlation between two features and thus finally derive the representative features (Liu et al., 2018). Fisher Criterion is a linear dimensionality reduction method and takes the variance between the class of feature components and the variance inside feature components into consideration to evaluate the components' emotionally- discriminative ability. It outperforms $PCA$ in extracting the discriminant embedded information from the emotional features with high dimensionality. The detailed algorithms for correlation analysis and Fisher Criterion refer to Liu et al. (2018).

### *4.4. Emotion classification*

Emotion classification is the final step for SER, which includes the design of the SER model architecture and the selection of the ML algorithms. The model architecture design bridges the feature extraction step and the selection of ML algorithms for SER by suggesting what types of audio features are required and how to use them. The selection of algorithms is the implementation of the architecture design. In this research, we explored the types of ML algorithms for SER models used in past studies. This helps in forming a direction for the design of the SER model, recommendations for model architecture design and ML algorithm selection.

#### *4.4.1. Algorithms for SER development*

ML algorithm selection is a crucial step that influences SER performance. As different datasets and emotion classes were utilised in previous studies, a comparison of SER models based on emotion classification performance is challenging. However, the overall trend of the SER models, including the architecture of the model and the selection of the ML algorithms, used in previous studies could be derived.

The architecture of the ML model deployed in previous studies could be divided into two groups — single-classifier architecture and ensemble model architecture. The former refers to the classifier made from only one ML algorithm, like $SVM$, or $CNN$. In contrast, ensemble models are developed based on a particular combination of multiple algorithms, such as Convolutional Recurrent Neural Network ($CRNN$) that is assembled from $CNN$ and $RNN$. The proportion of the collected studies which stress the ensembled-model architectures in each publication year shows an increasing trend (Fig. 7).

The other two conclusions based on the results of SER algorithms are:

- in all the studies that compare SER performance in ensembled-models and single-classifiers, ensembled-models perform better
- the majority of the algorithms incorporated for developing the classifiers (either single-classifier or ensembled-model architectures) are neural network-related ($ANN$, $DNN$, $CNN$, and $DeepAuto-Encoder$) rather than other algorithms (e.g., $SVM$, and $GMM$).

To sum up, the ensembled-model architecture and neural network-based algorithms are promising considerations to design the SER model. The comprehensive list of utilised algorithms, in past studies, is provided in Appendix A.

#### *4.4.2. SER model design*

Due to the propitious prospect of the ensembled-models, it is necessary to consider the principles for assembling the algorithms to build an ensembled-model. To achieve this, arranging the use of the ensembled-model architectures in previous studies could present a useful suggestion.

The ensembled-model usually divides the whole SER model into the feature-learning component and classification component. The objective is to use $DNN-based$ algorithms (e.g., Deep Belief Network and CNN) as the tool to extract deep features from the audio files (discussed in 4.3) while using the extracted features to make classification with another ML algorithm due to its particular advantages. For example, Zhu et al. (2017); Zhang et al. (2018a); Huang et al. (2018); Jiang et al. (2019); Wen et al. (2017) deployed $DNNs$ (e.g., $DBN$, Deep $Auto-Encoder$ and $DNN$) to extract deep features from audio signals and use $SVM$ as the final classifier to make classification. SVM was selected due to its simplicity and few parameters required consideration. Guo et al. (2018) used a type of $DNN$ and the CNN to extract respective deep features to form a group of fusion features, and then used the Extreme Learning Machine ($ELM$) to make a classification based on these fusion features. As ELM is a simple neural network-based algorithm, it has a higher generalisation ability that can be trained faster and requires fewer data samples for training rather than the DNNs (e.g., $RNN$) do.

Some studies took the contextual information into emotion classification by using the $RNNs$, including Long Short-Term Memory ($LSTM$) and $BiLSTM$, as the final classifier, with the extracted deep features by $CNN$ (Mu et al., 2017; Guo et al., 2018; Luo et al., 2018; Zhao et al., 2019a). A more advanced architecture — $3-Dilated-CNN-with-BiLSTM$ is proposed in Meng et al. (2019) which used three $CNNs$ deep feature extraction from Log Mel-spectrum, delta, and delta of the spectrum of the audio signals. The extracted features fed into the $BiLSTM$ to make $SER$.

Besides the mentioned architecture design and algorithm selection mentioned, there are some methods used by some of the studies to further improve the algorithm for SER. An attention model was used to improve the NN-based algorithms' SER performance (Jalal et al., 2019; Zhang et al., 2018b; Mu et al., 2017; Xie et al., 2019; Ramet et al., 2018; Tao et al., 2018). It helped the model to learn the speaker-independent emotional context information to avoid the bias towards the learned data and boost the emotional-information-mining speed. The detailed implementation of the attention model is discussed in Xie et al. (2019). On the other hand, Shih et al. (2017) used a skew-robust technique to improve the robustness of the ML algorithms (including SVM, MLP, and DNN) against the negative impact of the imbalanced distribution of data samples in different emotion classes. This technique modifies the formula for calculating the sum of cross-entropies between targets and outputs in the candidate algorithms by incorporating the class weights and notably improving their classification accuracy on imbalanced datasets. Sivanagaraja et al. (2017) revised the traditional CNN to deploy a series of 1-D convolution filers to take the multi-scale and multi-frequency transformations from the raw audio signal as the extended inputs to improve its SER performance. The multi-scale and multi-frequency signals are created by the down-sampling technique. These discussed methods and their contributions to solving SER problems are summarised in Table 3.

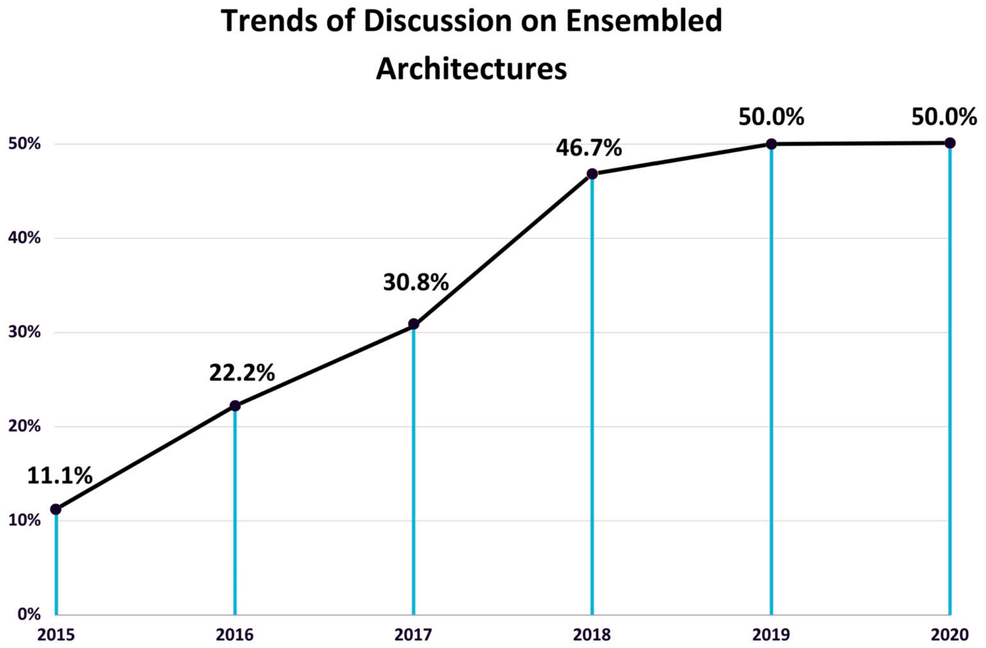

**Fig. 7.** Trends of discussion on ensembled architectures.

### 4.5. Evaluation criteria

The evaluation step is critical to validate the final performance of the developed SER models. Despite several different evaluation criteria used in previous studies, accuracy is the most popular, common and wildly used criteria for evaluating SER's performance (Zhou et al., 2016). Among all the included studies in this review, over 70 percent of the collected studies use it as the only or main validation criterion. The accuracy can be calculated by this formula:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$TP$ and $TN$ stand for true positive and true negative classification, the sum of which is the total number of correctly classified samples. Similarly, $FP$ and $FN$ refer to the number of misclassified samples. Thus, accuracy means the ratio of correct classifications over the whole observation.

According to the different objects that the accuracy is used to measure, there are two types of accuracy used in SER studies — weighted accuracy (WA) and unweighted accuracy ($UA$) or mean class (Chernykh & Prikhodko, 2017; Wen et al., 2017). WA considers the overall accuracy with regard to all the utterances, whereas UA refers to the average score of the different accuracies for different emotions' classification (Zhang et al., 2018b). UA is used as the most important metric to evaluate the SER performance on imbalanced dataset. In another study, Weighted Average Recall (WAR), known as the standard accuracy, and Unweighted Averaged Recall (UAR) were used to evaluate the performance of SER (Zhang et al., 2018a).

Moreover, confusion matrix and related measures, like precision, sensitivity, specificity, recall and F1-score are also used by some studies (e.g., Harár et al., 2017; Niveditha & Ashok, 2019; Lim et al., 2016; Darekar & Dhande, 2018) as the complementary method to accuracy. Trigeorgis et al. (2016) used Concordance Correlation Coefficient (CCC) for arousal-valence distribution to represent the emotion classes. CCC measures the level of agreement in terms of the linear correlation and the bias between two variables (Han et al., 2017b). It is a metric to evaluate the time and value-continuous predictions of emotion (Han et al., 2017a). $CCC$ ranges from -1 to 1. The more it is close to 1, the more a complete concordance appears, and the opposite polar it moves towards, the discordance is more obvious, while 0 stands for no correlation. False Positive Rate (FPR) and False Discovery Rate (FDR) were only used in Darekar and Dhande (2018) and Mannepalli et al. (2018), although in both studies accuracy was the main criteria for performance measurement.

### 4.6. Speaker-independent experiment concern

For the SER task, a primary factor that influences the final performance is whether the experiment is of speaker-dependent or speaker-independent type. Among all the studies, the speaker-dependent experiments gain an accuracy above 70 percent regardless of the databases used and target emotion classes to classify. By contrast, the vast majority of the speaker-independent experiments can only result in an accuracy ranging from 29 to 65 percent. This suggests the challenging task of SER in these cases.

A very few studies gain relatively high classification accuracy in $speaker - independent$ experiments. This suggests there could be some unique techniques or methods used by them that help improve the $SI$ SER performance (Table 3). Zhao et al. (2019b) gained an accuracy above 95 percent in the speaker-independent experiment with regard to SER for 7 emotion classes on $EmoDB$. They used an ensembled-model based on the local feature learning block ($LFLB$) and $RNN - LTSM$. The $LFLB$ is the unique algorithm used in this study. It is similar to CNN and comprises a convolutional layer, a batch normalisation layer, one exponential linear unit layer, and a max-pooling layer, which is designed to learn emotion features - see Zhao et al. (2019b) for LFLB detailed description. Wen et al. (2016) also achieves a high SER performance (over 97 percent of average recognition rate) for 7 emotion classes on BES database by ELM with GMM-enhanced (Gaussian Mixture Model) audio features. Its particularity is the audio features used for SER. The GMM-based clustering method is used in this study to pre-group some special audio features (wavelet packet energy and entropy features) and thus to reduce their intra-class error and to increase their inter-class variance to improve the classification performance. The essence of this method is assuming all the data samples could be located by a mixed-coordinate from various GMM distributions and thus using the GMM-based strategy to cluster the data samples. The centroid of each cluster could be used to calculate the ratio of the data features' means to the corresponding centroid. Lastly, the ratio will be applied to each feature to enhance the data features' discriminative ability. The detailed implementation of this method referenced by

Muthusamy et al. (2015). Meng et al. (2019) also proposed an assembled architecture by 3 Dilated $CNNs$, $BiLTSM$, and Residual Block and achieves nearly 70 percent of accuracy in speaker-independent experiment for 4 emotion states' classification on $IMOCAP$ database. The special technique thereof are the three $CNNs$ used to extract deep features from different forms of spectrograms (mentioned in Section 4.4.2). Moreover, this study also utilised a special loss method for training the $CNN - BiLSTM$ architecture, which could make the extracted deep features more discriminative. This method is a combination usage of soft-max and centre-loss. The former one is responsible for maximising the $inter - class$ distance while the latter one aims at minimising the $intra - class$ error. The centre-loss was uniquely used by this study because it could generate centres of the deep features in each class and penalise the errors between the features and their class centres (Wen et al., 2016). Although there is no further evidence showing the techniques mentioned above help improve the SER performance in SI experiments, they are uniquely used by the studies as fore-mentioned to obtain significant achievement and therefore are valuable for SER researchers to consider applying them to other audio databases or real-conversation experiments.

## 5. Discussion

This article summarises the main methods for SER since 2010. It can demonstrate the ML trends for SER and forms a comprehensive guideline for developing robust SER models. We aimed to capture all the possible steps in developing a SER system from speech signal analysis to model development and evaluation. This helps with (1) how the audio signals can be transformed into digital signals to be easily recognised by machines (e.g., sampling and pre-emphasis), (2) standard processing methods to prepare the signals for audio features extraction (e.g., framing, windowing and Fourier Transform), (3) types of audio features and their extraction techniques and tools, (4) suggestions for designing the SER models based on the trend and algorithm-advancing techniques based on previous studies, and (5) validation methods of developed ML models. A very recent trend in SER systems included an End-to-End approach in which raw speech signals are used directly for SER development (Tzirakis et al., 2018). However, due to the major method differences between end-to-end systems and the common three-step ML for SER, this study did not consider them in the review.

The overall trend of ML for SER has been evolving fast. In this regard, we explored the most recent trends in the area in this section. This includes different steps of ML such as feature selection and using pre-trained models, selection of ML algorithms, feature embedding and evaluation criteria.

### *5.1. Pre-trained speech feature embeddings*

In addition to handcrafted features and deep features, more recently, studies explore speech representations derived from large-scale pre-trained speech processing models for SER tasks. These self-supervised speech processing models are gaining attention towards SER due to their promises of universal representation of speech (Atmaja & Sasou, 2022). The potentiality of the presence of both linguistic and para-linguistic information in speech representations can lead to investigations of pre-trained speech models in SER domain viable (Atmaja & Sasou, 2022) and these pre-trained models can be used as a feature extractor (Chen & Rudnicky, 2023). Atmaja and Sasou (2022) conducted research with 19 self-supervised pre-trained models as acoustic feature extractors and all performed better than the classical filter-bank. Wav2vec 2.0, WavLM, UniSpeech-SAT and HuBERT are some pre-trained models studied in Atmaja and Sasou (2022). Wav2vec 2.0 is explored in Pepino et al. (2021) as an SER feature extractor. Their speech embedding model was derived from pre-trained wav2vec 2.0 models using simple neural networks, combining the outputs from multiple layers using trainable weights. They have tested with two wav2vec 2.0 models pre-trained on Librispeech, one fine-tuned for speech recognition and the other is not fine-tuned. Results are evaluated on IEMOCAP and RAVDESS databases for four and seven emotion categories respectively. Chen and Rudnicky (2023) fine-tuned wav2vec 2.0 using a proposed new algorithm and evaluated on IEMOCAP and SAVEE datasets to classify four and seven emotion categories. Wav2vec 2.0 models pre-trained on the speech of LibriSpeech without transcriptions were chosen as base models (Chen & Rudnicky, 2023). Several audio-only pre-trained models for SER have been explored in Sharma (2022). They presented a wav2vec 2.0 based multi-lingual multi-task model to identify seven emotions.

While the above experiments investigate discrete emotions, Wagner et al. (2023) studied wav2vec 2.0 and HuBERT pre-trained models on dimensional emotion recognition where emotions are recognized over the dimensions of arousal, valence, and dominance. Utilising a comprehensive testing scheme, they have evaluated the generalisation, robustness, fairness, and efficiency of transformer-based pretrained models on SER. The study demonstrated promising results in valence prediction without using linguistic information explicitly, but implicitly learning from them. Their findings are tally with Chen and Rudnicky (2023) where fine-tuned wav2vec 2.0 model has demonstrated capturing of correlation between linguistic content and emotion labels in SAVEE speech, by surpassing para-linguistic based emotion annotations by human evaluators. Accordingly, recent research efforts on speech processing pre-trained models towards SER may lead SER models that learn both linguistic and para-linguistic information, as illustrated with transformer based speech models (Wagner et al., 2023).

### *5.2. Trends in evaluation and privacy of SER*

In addition to accuracy, SER models are required more trustworthiness in broader deployments conserving the sensitivity of speech signals (Tia). A review study consolidates the trustworthiness of speech-centric ML over different dimensions including privacy, safety and fairness (Tia). Privacy evaluates the models' ability to protect the sensitive information of data owners while safety assesses the ability to produce reliable model outcomes withstanding potential adversarial attacks.

The fairness of models is evaluated on the discrepancy of model performance towards individuals or groups of the population (Gorrostieta et al., 2019; Tia). In general, ML models can unintentionally bias towards certain protected attributes of the population such as race, gender or age (Chen & Rudnicky, 2023; Gorrostieta et al., 2019). Studies evaluate the fairness of models over gender by comparing the accuracy of predictions between male and female groups (Gorrostieta et al., 2019; Wagner et al., 2023). The robustness of SER ML models is defined as their capability of consistent outcomes despite changes to the input signals. This can be quantified by testing SER models against data augmentations without affecting emotion ground truth (Wagner et al., 2023).

Moreover, the computational complexities of models limit practical deployments and applications, despite the promising results in research evaluations. Thereby studies evaluate the efficiency of their computations by focusing on resource optimizations. For instance, Wagner et al. (2023) evaluated the efficiency of their models in terms of optimisation stability, computational complexity, and data efficiency. In optimization stability, they investigated the role of different seeds on convergence time and performance stability of pre-training while computational complexity examined the complexity of the model such as the number of transformer layers. Data efficiency referred to the evaluation of model performance by reducing the amount of training data without sacrificing performance (Chen & Rudnicky, 2023; Wagner et al., 2023).

Besides the technical implementation of SER, privacy protection plays a key role in SER's performance since speech data is stored in central servers in most applications. Speech naturally encompasses certain demographics such as gender, age and language background (Feng et al., 2022; Jaiswal & Provost, 2020). Accordingly, a model trained to recognize emotions by audio even without taking gender as a feature can have gender discriminative capability. Studies evaluate SER model capabilities to minimize the probability of privacy leakage by their model outcomes (Feng et al., 2022; Jaiswal & Provost, 2020). In this regard, Jaiswal and Provost (2020) defined a demographic privacy metric of SER models to measure the model capability to deny the prediction of gender from their SER predictions. The gender leakage probability of an SER model is taken as the gender prediction probability from a new model that is trained from a known set of genders and original SER outcomes. Thereby, the demographic privacy metric is considered as 1-gender leakage probability. On the other hand, Feng et al. (2022) relied on the comparative performance of target and adversarial models to evaluate model performance on information protection.

### *5.3. Emerging techniques in SER*

Recent studies are exploring emerging techniques in SER to enhance their performance, trustworthiness, and computational complexity (Stappen et al., 2020). While studies evaluate model performance over privacy protections, some studies are investigating the protection of speech data from adversarial attacks via Federated learning (FL) in SER (Latif et al., 2020). FL is a distributed ML paradigm that decentralises training models with privacy-sensitive personal data (Tsouvalas et al., 2022). In FL for SER, speech data are kept in user devices and only the model parameters computed locally are transferred to a central server which aggregates updates from multiple participating users to collectively train the model. Therefore, no speech data is transferred between the end user and the central server, but only the initial and updated model parameters do (Latif et al., 2020).

Latif et al. (2020) analyzed the feasibility of FL for SER, training CNN and RNN-based classifiers to recognize four emotion classes based on Mel Filterbank features and different numbers of participating users. However, the study highlights the communication and resource overhead at client ends in FL for SER. Moreover, traditional FL techniques assume the availability of sufficient label data at user ends for training which is challenging in SER applications (Tsouvalas et al., 2022). To address this issue, Tsouvalas et al. (2022) explored semi-supervised learning under FL settings to utilize both labelled and unlabelled data for training on users' devices. They compute pseudo-labels for unlabelled data and only highly confident estimations are considered in the training process considering estimations as target classes similar to ground truth. An attention-based CNN architecture is investigated as an SER classifier under FL (Tsouvalas et al., 2022). To utilize both labelled and unlabelled data at the user end, Amiriparian et al. (2022) also propose a semi-supervised federated learning framework where pseudo labels for unlabelled data are computed based on multiple complementary views. Experiments are carried out with an MLP-based SER model (Feng & Narayanan, 2022).

## 6. Conclusion and future work

SER is becoming an important sub-discipline of human-machine interaction, that could bring benefits to different areas, such as minimizing fatigue driving, assisting some medical diagnosis, and collecting feedback from students in online education. The speech features underlie the use of ML methods to perform SER tasks. Although there are numerous studies in SER by ML method, few of them form a comprehensive guideline of what techniques or methods could be used in the three steps (data pre-processing, feature extraction, and emotion classification) of the SER. Moreover, the common challenges and solutions are rarely presented in an entire view. Particularly, the pervasive low-classification-accuracy issue in the Speaker-Independent experiment is not addressed in previous studies. This paper performed a systematic review of the SER literature to address the common techniques.

This study could create the foundation to understand the detailed occurrences of each specific step of SER systems. This underlies further research efforts through the revision of existing techniques and/or the addition of more useful models to enhance emotion classification performance. The potential enhancement of SER modelling could also be enlightened by the challenges and solutions of SER addressed in this paper. Common challenges in previous speech emotion recognition studies depict problems faced across each step of SER and how these problems impede ML models from learning the audio features effectively. This understanding allows researchers to improve the current techniques in a step-by-step approach which can effectively guide researchers to apply ML methods to detect emotion states from human speech. Furthermore, the challenges and solutions comprehensively addressed in this study can potentially direct the significant improvement of the existing SER models or trigger new and better methods or techniques to solve the current challenges. Particularly, the attention to the underlying strategies to enhance the Speaker-Independent SER performance from previous studies provides a chance to solve the universe issue of low classification accuracy in SI experiments.

This research intended to collect the reported challenges in SER development. For instance, the signal trend issue pointed out by Kerkeni et al. (2019) could influence the correlation analysis of the signal in the time-domain. This suggests that if models are designed to extract features from raw audio signals, there is an underlying factor which is likely to misdirect the ML model to catch the real emotional pattern. On the other hand, the solutions collected from previous studies could be seen as a guideline for solving the challenges or even a trigger for deriving more solutions. This is seen in Muthusamy et al. (2015) where GMM-based clustering, which aims at improving the audio features for training SER models, is motivated by the multiple successful applications of GMM on speech and image processing in earlier studies. Moreover, the data augmentation strategy mentioned in Niu et al. (2017) was created in light of concurrently solving the data insufficiency and data imbalance problems. Therefore, the comprehensive address of both the problems and solutions, via systematic review, could help us in developing more advanced models or techniques, encompassing all issues, thus making considerable improvements in SER.

Finally, this paper distributes a particular focus on Speaker-Independent experiments with regard to several studies that achieve high SER performances in Speaker-Independent tasks. This provides possible solutions for the low performance observed in a vast majority of previous studies. It is not difficult to find that the SER performance is largely impacted by Speaker-Dependent or Speaker-Independent experiments. The accuracy range for Speaker-Dependent experiments is observed to be higher than that of the Speaker-Independent experiments in previously published works. However, some studies still gain high SER accuracy in Speaker-Independent experiments. Except for the common techniques or methods they use, the unique ones in the studies have the highest probability to contribute to this result. Addressing these particular techniques or methods can underpin further research and subsequent applications on other databases. Moreover, these techniques may have the potential to inversely locate specific reasons why Speaker-Independent experiments always lead to lower performance in comparison with Speaker-Dependent experiments. For example, the unique feature-enhancing method used in Muthusamy et al. (2015) indicates the ambiguities in speech features of different people to represent the same emotion. This is because the GMM-based clustering method actually alleviates the ambiguities by adjusting the audio features' intra and inter-class variance.

Nevertheless, this research also has some limitations. Our research results are limited to the searched databases and the articles that met the inclusion and exclusion criteria. Therefore, there might be a potential for some missed topics such as graph networks for SER or Google SpecAugment (Liu et al., 2023) in data augmentation for SER. Therefore, it is recommended that future research focus more on the ML domain advancements and techniques within the past four years and consider the contribution of leading international organisations such as Google and Microsoft. Due to the restricted space and the complicated dimensionality of the SER task, some findings useful for other directions (e.g., real-time application and regression application by the arousal-valence framework) cannot be covered in this paper. Furthermore, as a result of the high generalisation of the search keywords, some studies which further discuss the techniques for solving the extremely specific problems in SER cannot be incorporated. Future research could be conducted based on these limitations.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Table A in Appendix A contains the information of all articles selected for this systematic review.

## Acknowledgements

This research is supported by AUT Faculty of Design and Creative Technology contestable grant 2021 and 2023, and AUT School of Engineering, Computer and Mathematical Sciences emerging researcher fund 2021.

## Appendix A

**Table A**
Summary of Reviewed Papers with Descriptions.

| Paper | Signal Pre-processing, Feature Extraction & Feature Selection | Audio Parameters | Emotion Classes | ML Algorithms | Evaluation Criteria | Classification Result |
|---|---|---|---|---|---|---|
| Alu et al. (2017) | Hamming window (25 ms); PRAAT program, program (frame rate 10 ms) | MFCC (Spectral) | happy, fear, anger, sad, disgust, surprise | CNN | Accuracy | CNN 71.4% |
| Zhou et al. (2016) | Prosodic (intensity, energy); Spectral(LPC, LFPC, LPCC, MFCC); Acoustic formant, glottis) | Sampling: Frequency 16 kHz Framing: Hamming Window (Length 256.5% Overlap) | anger, boredom, disgust, anxiety, happiness, sadness, neural state | Stacked Autoencoder DBN | Accuracy | For speaker independent experiment and the results in the best case: DBN: 39.0% SAE: 29.0% |
| Zhu et al. (2017) | Pre-emphasis; Framing: Window 25 ms with Overlap of 10 ms; Endpoint Detection | Prosodic (pitch, energy, zero-crossing); Spectral (MFCC); Acoustic (formant) | angry, fearful, happy, neutral, sad, surprise, neural state | SVM; DBN; SVM+DBN | Accuracy | For gender-dependent, experiment DBN+SVM: 95.8% DBV: 94.6% SVM: 84.8% |
| Zheng et al. (2015) | Framing: Window 25 ms with sliding at 10 ms each time; Whitening log-transform and processPCA (60 components); Normalisation | Spectral (MFCC, LPC, LPCC); RASTA-PLP; Prosodic (Pitch, Zero Zero crossing) | excitement, frustration, happiness,neutral, surprise | SVM with; hand-crafted acoustic features; DCNN; DCNN with PCA | Accuracy | For speaker-independent experiment DCNN with PCA: 40.02% DCNN: 33.7% SVM: 37.6% |
| Brester et al. (2015) | Praat and (OpenSMILE system) | Prosodic (pitch, intensity), harmonicity; Spectral (MFCC); Acoustic (formant) | For Emo-DB database neutral, anger, fear, joy, sadness, boredom or disgust | MLP; Ensemble of MLPs by MOGA; Ensemble of MLPs by MOGA with FS | F-Score | Ensemble of MLPs Wby MOGA: 82.9% Ensemble of MLPs by MOGA with FS: 81.6% MLP: 80.8% |
| Agrima et al. (2019) | Segmented into the syllabic form (consonant & vowel) | Prosodic (the fundamental frequency, intensity, number of pulses); (Formats) | joy, sadness, and neutral state terms of Ba, Du, and Ki syllabic Funits | Neural networks; SVM; Decision Trees J48 | Accuracy | For Ba unit: Neural network with prosodic and formant features: 75.49% For Du unit: Neural network with prosodic features: 88.40% For Ki unit: SVM with: 80.59% with prosodic and formant features |
| Alex et al. (2018) | Segmented into the syllabic form | Prosodic at the,utterance level; (denoted as group 1) and syllable level (denoted as group 2) | anger, sadness, neutral fear, happiness, boredom, disgust | DNN (group); DNN (group 2); The fused model of DNN (group1) and DNN (group2) | WA | The fused DNNs: 61.68%, DNN (group1), DNN (group2): 58.88% |
| Aouani & Ayed (2018) | Frame segmentation; Pre-emphasis; Windowing; FFT with IFT | Group 1: 39 MFCC, coefficients (12 MFCC + energy, s12 Δ MFCC + energy, 12 ΔΔ MFCC + energy); Group 2: 65 (MFCC features) | anger, disgust, fear, neutral, sad, surprise | SVM;DSVM | Recognition rate | SVM with 65 MFCC features: 73.01% SVM with 39, features: 68.25% |

(continued on next page)

**Table A** (*continued*)

| Paper | Signal Pre-processing, Feature Extraction & Feature Selection | Audio Parameters | Emotion Classes | ML Algorithms | Evaluation Criteria | Classification Result |
|---|---|---|---|---|---|---|
| Alex et al. (2018) | Segmented into the syllabic form | Prosodic at the, utterance level; (denoted as group 1) and syllable level (denoted as group 2) | anger, sadness, neutral fear, happiness, boredom, disgust | DNN (group); DNN (group 2); The fused model of DNN (group1) and DNN (group2) | WA | The fused DNNs: 61.68%, DNN (group1), DNN (group2): 58.88% |
| Badshah et al., 2017 | Transformed into Spectrograms | Spectrograms (time-frequency), representation of audios | anger, disgust fear, neutral, sadness, boredom | CNN | Accuracy | CNN: 84.3% |
| Bertero et al. (2016) | Filter for CNN: a convolution window of 25 ms (with overlapping of 6 ms) | Raw Audio File | creative/ passionate; criticism/ cynicism; defensiveness/ anxiety; friendly/ warm; hospitality/ anger;-leadership/ charisma;-loneliness/ fulfilment; love/-happiness;-sadness/ sorrow; self-control/ practicality; supremacy/ arrogance; | SVM (L2 regularised) and CNN fo binary Classification across 11 emotion classes | Accuracy | CNN: 62.1% SVM: 54.6% |
| Bertero and Fung (2017) | Filter for CNN: a convolution window of 25 ms with overlapping of 6 ms | Raw Audio File | angry/ Happiness; sadness | CNN and SVM | Accuracy | CNN: 66.1% SVM: 63.0%; For Angry, happy:CNN >SVM For Sad:SVM>CNN |
| Chernykh and Prikhodko (2017) | Framing: Window of 200 ms (with overlapping of 100 ms); Extract audio features by PyAudioAnalysis (Giannakopoulos, 2015) | Acoustic (13 MFCCs);zero crossing rate, energy, entropy of energy; Spectral centroid, spread, entropy, flux, rolloff; 12-dimensional chroma vector, standard deviation of chroma vector | sadness, anger, excitement, neutral | Connectionist Temporal Classification (CTC), which is RNN-based approach | WA; UA | For speaker-independent experiment: CTC: weighted accuracy: 54%, unweighted accuracy: 54% |
| Palo and Mohanty (2015) | Frame segmentation;Hamming window;LP analysis | Time-frequency parameters (pH vectors); Spectral features (LP-VQC vectors) | boredom, angry, sad, happy | PNN with pH PNN with $LP_{VQC}$ | Average classification error | PNN with pH: 10.25% PNN with $LP_{VQC}$:15.14% |
| Assunção et al. (2019) | Frame segmentation (1 to 10 s long); Extract audio features by VGGVox model (Nagrani et al., 2017) PCA or LDA for FS | speaker-specific features from VGGVox model | anger, disgust, fear, happiness; sadness, surprise;the neutral state | KNN (5-nearest neighbour); RF (500-Tree);LMT | Accuracy | For one-speaker's emotion recognition experiment: LMT: 81.1%; KNN 80.1%; RF: 77.3% |
| Chang et al. (2017) | Transformed into Spectrograms (with short time Fourier transform of window size of 1024 samples) | valence, activation | valence/ activation: 5-class;3-class | Multi-task DCGAN; DCGAN; CNN; Multi-task CNN | UA | For 5/3-class: DCGAN: 43.88%/49.80%; Multi-task DCGAN: 43.69%/48.88%; CNN: 38.52%/46.59%; Multi-task CNN: 6.78%/40.5% |
| Chhabra et al. (2016) | Not mentioned | standard deviation of pitch, MFCC, formants, envelope, differentiated envelope, energy, amplitude, and spectrum | fear, disgust, happy, bored, neutral, sad, anger-1.low; 2.-medium; -3.high | Adaptive boost | Accuracy | Adaptive boost: 78.3% |
| Fayek et al. (2015) | Data augmentation for training set (five-fold); Hamming window of 25 ms (with a stride of 15 ms); Transformed into spectrogram by linearly-spaced log Fourier-transform based filter banks; concatenate windows | Not mentioned | ENTERFACE database (anger, disgust, fear, happiness, sadness, surprise) SAVEE database (anger, disgust, fear, happiness, sadness, surprise, neutral) | DNN | Accuracy | For speaker-independent experiment: DNN for ENTERFACE dataset: 60.53% DNN for SAVEE dataset: 59.7% |

**Table A** (*continued*)

| Paper | Signal Pre-processing, Feature Extraction & Feature Selection | Audio Parameters | Emotion Classes | ML Algorithms | Evaluation Criteria | Classification Result |
|---|---|---|---|---|---|---|
| Darekar and Dhande (2018) | PCA for feature-dimension reduction | Cepstral feature, NMF feature, pitch feature | happy, angry, neutral, sad, fear, surprised | NN with PSO-FF; NN with LM; NN with FF; NN with PSO | Accuracy main criterion; Sensitivity; Specificity; Precision; FPR; FNR; NPV; FDR; F1-score; MCC | For benchmark database in accuracy criterion: NN with PSO-FF:88.72%, NN with PSO:87.15%; NN with FF/ LM: 86.81% For Marathi database: NN with PSO-FF: 80.0% |
| Fayek et al. (2016) | Hamming window of 25 ms (with a stride of 10 ms); Log Fourier transform-based filter bank with 40 coefficients distributed on a Mel scale | acoustic features | anger, happiness (including excitement), sadness, neutral | CNN | Error rate; Unweighted Error rate | For speaker-independent experiment: The combined classifier: Error rate 46.44%; Unweighted Error rate 48.96% |
| Deng et al. (2017) | Use openSMILE to frame the raw audio data with window of 25 ms and 10 ms stride to gain the LLD features; Incorporate K-mean pooling to process the raw LLD data; Mean subtraction and standard deviation are performed to each utterance; Down-sampling is used to balance the training data | A frame level of Log-Mel feature vector including 40 log-mel filter bank features, energy and their first deltas; ComParE: including 4 energy-related LLD (e.g., RMS energy, zero-crossing, Sum of auditory spectrum),55 spectral LLD (e.g., MFCC, spectral energy, flux, etc.), and 6 voicing related LLD (e.g., SHS and Viterbi smoothing, etc.) | Group1: anger (1.anger1 2.agner2, 3.otherwise) Group2: anger (1.anger1, 2.anger2) | Experimented classifiers: LSTM (Last Frame), LSTM (Mean Pooling); BLSTM (Mean Pooling); CNN with VGG model; CNN+LSTM with respect to using only Log-Mel feature vector and only ComParE feature vector Baselines from:GMM; SVM; SVM (ComParE) | UAR | Classifiers with the highest performances: For Group1 anger detection: BSTM (Mean Pooling) with Log-Mel LLD: 79.4% BLSTM (Mean Pooling) with ComParEmLLD: 80.1% For Group2 anger detection: CNN+LSTM: 71.0% BLSTM (Mean Pooling) with ComParE LLD: 72.2% |
| Fei et al., 2016 | Pre-emphasis;Framing; Windowing; For DAE: doing wavelet decomposition for each audio frame and derive the Fourier Transform;extend each data frame before and after it to gain continuous data super-frame; normalise the super-frame data | MFCC, PLP, LPCC (for SVM) | anger, fear, happiness, neutral, surprise, sadness | SVM;DAE | Accuracy for classification in each emotion class | For anger: DAE: 86.41% SVM with MFCC: 75.23% SVM with LPCC: 73.51% SVM with PLP: 82.17% |
| Getahun and Kebede (2016) | Channel and format conversion, down-sampling, segmentation, and normalisation by Audacity software; Intensity normalisation by average RMS; Pitch normalisation by average; Framing: window of 25 ms with 10 ms overlapping FS by Weka with regard to Gain Ratio, InfoGain, and OneR | (mean, maximum, minimum, standard deviation, and median of) pitch, energy, MFCC, LPC, LFCC, voice quality features | Anger, Fear, Sadness, Positive | MLP | Weighted average of precision and recall | MLP: 0.721/ 0.724 for precision and recall |
| Fayek et al. (2016) | Framing:Hamming window of 25 ms with a stride of 10 ms; Log Fourier-transform based filter banks with 40 efficient on a Mel scale are collected to generate spectrogram in every frame | Audio features automatically extracted by deep learning | anger, happiness sadness, neutral | DNN (based on feed-forward and recurrent architecture) based on Utterance and Frame each;-Baselines:-DNN+ELM;-SVM;-Replicated-Softmax-Models+SVM; Hierarchical Binary Decision Tree | Accuracy and UAR | For speaker-independent experiment: DNN with Frame-based: 64.78%/ 60.89%; DNN with Utterance-based: 57.74%/ 58.28% Hierarchical Binary Decision Tree: -/ 58.46% Replicated Softmax+SVM: -/57.39% DNN+ELM: 54.3%/ 48.2% |

(*continued on next page*)

**Table A** (*continued*)

| Paper | Signal Pre-processing, Feature Extraction & Feature Selection | Audio Parameters | Emotion Classes | ML Algorithms | Evaluation Criteria | Classification Result |
|---|---|---|---|---|---|---|
| Trigeorgis et al. (2016) | Framing: window length of 3 s with 40 ms overlapping; Preprocess the time-sequences to have zero mean and unit variance; Segmenting the raw waveform to 6 s long sequences | (maximum, minimum, range, mean, and standard deviation) of minimalistic acoustic feature set (eGeMAPS-Eyben et al. (2015)) and LLD | Arousal; Valence | Convolutional recurrent model Baselines: BLSTM (eGeMAPS) BLSTM (ComParE) | Concordance Correlation Coefficient | For Arousal/ Valence classes on testing set (speaker-independent): Convolutional recurrent model: 0.686/ 0.261 BLSTM (ComParE): 0.382/ 0.187 BLSTM (eGeMAPS):0.316/ 0.195 |
| Hussain et al. (2017) | FFT and adaptive filtering are used to remove the noise and Goo | Pitch, intensity, MFCC, frequency contours | anger, disgust, fear, happy, sad and surprise | RBFNN Baselines: FFNN; GRNN;Elman NN | MSE | RBFNN:4.13e-025 GRNN: 0.0056 FFNN: 1.998e+003; Elman NN:1.995e+004 |
| Han et al., 2017a | Framing:window of 25 ms with a stride of 10 ms; Use openSMILE toolkit to gather 13 LLDs; The arithmetic mean and coefficient of variance are calculated based on the LLDs with an analysed window of 8 s (and 40 ms step forward); Speakers' gender and age are balanced | LLDs | Arousal; Valence | Two groups of RNN-SVR, SVR-RNN, RNN-RNN classifiers with training sets of true and pseudo predictions respectively Baselines:SVR; -RNN (2 layers) RNN (4 layers) | Concordance Correlation Coefficient | For Arousal/ Valence classes on testing set (speaker-independent): SVR-RNN (trained with true predictions):0.730/ 0.393 RNN-RNN (trained with pseudo predictions): 0.744/ 0.377 SVR: 0.726/ 0.300RNN (2 layers): 0.738/ 0.278 RNN (4 layers): 0.708/ 0.305 |
| Harár et al. (2017) | Framing: window of 20 ms with no overlapping;Use the Google WebRTC voice activity detector to remove the silent parts of the audio signals; Audio files are standardised by mean value and unit variance | Audio features automatically extracted by deep learning | angry, neutral, sad | CNN | Overall accuracy; precision; f1-score | For the testing segments' overall accuracy: CNN: 77.51% |
| Han et al., 2017b | Dataset is divided and balanced on gender, age, and mother tongue; Framing: window of 25 ms with stride of 10 ms; Use openSMILE toolkit to extract acoustic features; The arithmetic mean and coefficient of variance are computed | LLDs (frame-based or segment-based statistical features), including MFCC and logarithmic energy | Arousal; Valence | Two RNN-BLSTM + RNN-BLSTM classifiers on LLD frame based and functional based features respectively Baselines: RNN-LSTM | Pearson's Correlation Coefficient | For the speaker-independent experiment results in arousal/ valence:RNN-BLSTM + RNN-BLSTM (LLD): 0.729/ 0.309 RNN-BLSTM + RNN-BLSTM (functional): 0.720/ 0.360 RNN-LSTM: 0.350/ 0.199 |
| Guo et al. (2018) | For heuristic features: Framing by window of 265 ms size with 25 ms window shift; Use openSMILE to derive the heuristic features; Calculate the segment-level statistical features For bottleneck features: Use DNN to process the heuristic features | LLDs; Spectrograms | happiness, boredom, neutral, anger, fear, sadness, disgust | CNN-ELM (+ heuristic features)Baselines: DNN-ELM; CNN-BLSTM; CNN-BLSTM(+ heuristic features); CNN-ELM | Precision,-Recall,-F1-Score | For Precision, Recall, and F1-Score:CNN-ELM (+ heuristic: 93.30%, 91.97%, 92.50% |
| Mefiah et al. (2015) | Segment the speech signals and label them; HTK tool is used to do segmentation and alignment; Speech rhythm matrix is calculated based on the segmentation and labels | Speech rhythm matrix | neutral, sad, happy, surprised, questioning and angry | MLP with IM metrics; MLP with PVImetrics; MLP with both metrics | Accuracy | MLP with both metrics: 72.41% MLP with PVI: 63.79% MLP with IM: 55.71% |
| Rajisha et al. (2015) | Framing: Hamming window of 20 to 30 ms wit overlapping of 1/3 or 1/2 or the frame size | MFCC, STE, Pitch | anger, happy, sad, neutral | ANN;SVM | Accuracy | ANN: 88.4%, SVM: 78.2% |
| Lotfidereshgi & Gournay (2017) | Speech signal is separated into orthogonal and complementary components; Then a LP analysis is performed on each frame; PCA is used to reduce feature dimensionality | source and vocal tract components | anger, boredom, disgust, fear, happiness, sadness, neutral | LSM | Recognition rate | LSM: 82.35% |

**Table A** (*continued*)

| Paper | Signal Pre-processing, Feature Extraction & Feature Selection | Audio Parameters | Emotion Classes | ML Algorithms | Evaluation Criteria | Classification Result |
|---|---|---|---|---|---|---|
| Huang et al., 2018 | Use openSMILE to extract acoustic features | 12 functionals (statistical features) of zero-crossing rate, RMS, energy, pitch frequency, HNR, MFCC | angry, happy, sad, neutral | Ladder Network (based on DAE and SVM) Baselines: DAE; SVM (with static acoustic features) | Accuracy | Ladder Network: 0.591% DAE: 0.564% SVM: 0.538% |
| Huang et al. (2017) | Framing: Hamming window of 25 ms with a frame rate of 10 ms.Use Fourier-transform-based filter bank to generate raw speech features.Use openEAR tool to extract audio features | LLDs | Valence (Negative and Positive) | PCASS Baselines:-CT;DAE;MT-;KMM;SHLA-;SSF;SF | Cross-corpus UARs with the training datasets of ABC and Emo-DBy | For Speaker-independent experiment:PCASS: 63.75%/ 61.41%SHLA: 63.36%/ 56.52%KMM: 62.52%/ 58.23%MT: 60.57%/ 61.20%SSF: 59.67%/ 56.86%SF: 59.12%/ 56.35%DAE: 56.20%/ 57.05%CT: 56.03%/ 51.01% |
| Lim et al. (2016) | Transform audio signal into 2D graphs with a frame size of 256 and the overlapping of 50% by STFT; Then the graphs are processed by CNNs, followed by RNN-LSTM networks; | 2D representations of audio signal | neutral, anger, fear, disgust, sadness, boredom, happy | Time distributed CNNs (CNN with RNN-LSTM)Baselines:CNNs;RNN-LSTM | Average precision, recall, and F1-score | Time distributed CNNs: 88.01%/ 86.86%/ 86.65% CNNs: 87.74%/ 86.32%/ 86.06%RNN-LSTM: 79.87%/ 78.83%/ 78.31% |
| Partila et al. (2015) | Preprocess the audio signals by normal operations, such as removing DC element; Pre-emphasis; Segmenting signals into quasi-periodic frames | MFCCs (39), first and second derivative of MFCC, LPC (12), LSP (12), Prosodic features (RMS energy, log-energy, ZCR, MCR, position of maximum, maximum, minimum, HNR) | anger, boredom, disgust, fear, happiness, sadness, neutral state | GMM;KNN (k = 5);ANN | Accuracy | ANN: 90% for MFCC feature |
| Rawat and Mishra (2015) | Remove noise by High Pass Filter; Use framing with hamming window to gain the audio segmentations of equal length; Use FFT to generate the frequency spectrum; Use Mel scale filter bank defines the energy in each frame; Gain the logarithm of the energy;Compute the DCT; Use cepstral mean correction to do compensation for MFCC | MFCC, energy | happy, sad, neutral, disgust, anger | NN | Accuracy for each emotion class | NN: happy with 95%, sad with 94.16%, neutral with 91/58%, disgust with 93.42%, anger with 92.74% of accuracy |
| Rázuri et al. (2015) | Use MATLAB to extract desired audio features | SF, SC, Spectral Rolloff Point, RMS, SCV, ZCR, MFCC, LPC, LPCC, 2DMM, Fraction of Low Energy frame, | anger, disgust, fear, joy, sadness, and surprise | DT (J48); MLP;SVM-P2; BN | Confusion Matrix with more focus on Average Recall | MLP: 96.98% BN: 96.97% SVM-P2: 96.62% DT (J48): 96.22% |
| Wen et al. (2017) | Not mentioned | Prosodic features, Spectral features (LPCC, ZCPA, PLP), HuSWF features with statistical representations (max, min, kurtosis, skewness, median) | EMODB database: anger, anxiety fear, boredom, disgust, happiness, neutral, sadnessSAVEE database: anger, disgust, fear, happiness, sadness, surprise, neutralCASIA: anger, fear, happiness, sadness, surprise, neutral FAU database: anger, emphatic, neutral, positive, rest | RDBN based on DBN and SVM-RBF;-Baselines-:SLDBN-;DLDBN,-TLDBN;-KNN;-SVM | Confusion Matrix; WA UA | In speaker-independent experiment for the four databases: For EMODB in WA: RDBN: 82.32%, LSVM: 81.19%SLDBN: 72.84%KNN: 70.74%DLDBN: 53.85%TLDBN: 24.59% For SAVEE in WA:RDBN: 53.60% LSVM: 46.25% KNN: 43.13% SLDBN: 30.42% TLDBN: 25.00% DLDBN: 20.62% |

(*continued on next page*)

**Table A** (*continued*)

| Paper | Signal Pre-processing, Feature Extraction & Feature Selection | Audio Parameters | Emotion Classes | ML Algorithms | Evaluation Criteria | Classification Result |
|---|---|---|---|---|---|---|
| Wen et al. (2017) | Not mentioned | Prosodic features, Spectral features (LPCC, ZCPA, PLP), HuSWF features with statistical representations (max, min, kurtosis, skewness, median) | EMODB database: anger, anxiety fear, boredom, disgust, happiness, neutral, sadnessSAVEE database: anger, disgust, fear, happiness, sadness, surprise, neutralCASIA: anger, fear, happiness, sadness, surprise, neutral FAU database: anger, emphatic, neutral, positive, rest | RDBN based on DBN and SVM-RBF;-Baselines-:SLDBN-;DLDBN,-TLDBN;-KNN;-SVM | Confusion Matrix; WA UA | In speaker-independent experiment for the four databases:% For CASIA in WA:RDBN: 48.50% LSVM: 42.08% SLDBN: 38.50% KNN: 34.33% DLDBN: 29.50% TLDBN: 18.25% For FAU in UA: RDBN: 42.20% SLDBN: 40.52% LSVM: 37.37% KNN: 35.70% DLDBN: 30.50% TLDBN: 30.10% |
| Muthusamy et al. (2015) | Framing: window width of 32 ms with no overlapping; Frames with low-energy were dropped; The rest speech signals are flattened by a first-order pre-emphasis filter; Glottal wave-forms are formed based on inverse filtering and linear predictive analysis; Feature enhancement by GMM; Feature reduction by SLDA | 30 relative energy features, 30 relative entropy features; glottal waveform | BES database: anger, disgust, fear, neutral, happiness, sadness, boredom SAVEE database: anger, disgust, fear, neutral, happiness, sadness, surprise SES database: neutral, surprise, happiness, sadness, anger | KNN;ELM; | Average Recognition Rate | For BES database in speaker-independent experiment:ELM with enhanced features:97.24%ELM with raw features: 56.61%KNN: 49.12% |
| Muthusamy et al. (2015) | Framing: window width of 32 ms with no overlapping; Frames with low-energy were dropped; The rest speech signals are flattened by a first-order pre-emphasis filter; Glottal wave-forms are formed based on inverse filtering and linear predictive analysis; Feature enhancement by GMM; Feature reduction by SLDA | 30 relative energy features, 30 relative entropy features; glottal waveform | BES database: anger, disgust, fear, neutral, happiness, sadness, boredom SAVEE database: anger, disgust, fear, neutral, happiness, sadness, surprise SES database: neutral, surprise, happiness, sadness, anger | KNN;ELM; | Average Recognition Rate | For SAVEE database in speaker-dependent experiments:ELM with GMM-enhanced features: 97.60%ELM with raw features:58.33%KNN with GMM-enhanced features: 94.27% KNN with raw features:50.31% For SES database:ELM with GM features: 92.79%ELM with raw features:42.14%KNN with GMM features: 84.58%KNN with raw features: 27.25% |
| Niu et al. (2017) | Transform speech files into spectrograms by short time Fourier transform with window size of 512, overlapping of 384; Gain larger images by data augmentation technique; Convert the images to the size of 256*256 | Speech spectrogram | anger, happiness, sadness, neutral, frustration, excitement, surprise, fear | DRCNN with raw data; DRCNN with augmented data | Accuracy | DRCNN with raw data: 41.54%DRCNN with augmented data: 99.25% |
| Mirsamadi et al. (2017) | Extract acoustic features from framed audio files with 25 ms at the rate of 100 frames/s; Use statistical functions to gain statistical features | LLDs (pitch, voicing probability, energy, etc.);HSFs (mean, min, max, etc.) | happy, sad, neutral, angry | RNN-frame-wise;RNN-final frame;RNN-mean pool;RNN-weighted pool with attention | WA; UA | In terms of WA and UA:RNN-weighted pool with attention: 63.5%/ 58.8%RNN-mean pool: 62.7%/ 57.2% |
| Le et al. (2017) | Arousal and valence are computed in each 40 ms of the audio files; Extract log Mel filterbank coefficients for every utterance by a 25 ms window with 10 ms frame shift; Combine every four adjacent frames into the features with 160-dimension and 40 ms span to align with the labels; Apply z-normalisation on the features in each utterance; Label discretisation with K-Means; | valence, arousal | valence, arousal of audio files | CLS-Raw;CLS-DecodedREG-MSE; REG-CCC | CCC | CLS-Raw: 0.682;CLS-Decoded: 0.680REG-CCC: 0.664REG-MSE: 0.652 |

**Table A** (*continued*)

| Paper | Signal Pre-processing, Feature Extraction & Feature Selection | Audio Parameters | Emotion Classes | ML Algorithms | Evaluation Criteria | Classification Result |
|---|---|---|---|---|---|---|
| Koolagudi et al. (2018) | Extract spectral features from the frame of 20 ms in speech signal with a variation less than 20 ms | Prosodic (pitch, intensity, jitter, shimmer);Spectral (MFCCs, formant);Statistical values (minimum, maximum, mean, standard deviation) of prosody | anger, fear, happiness, neutral, sadness | VQ;K-Means;GMM-;ANN | Accuracy | For MDB and IIT-KGP databases: GMM: 84%/ 81% K-Means: 74%/ 71% ANN: 72%/ 79% VQ: 57%/ 62% |
| Zhang et al. (2017) | Augment the data;Framing: Hamming window of 25 ms with 10 ms of overlapping;Use 64 Mel-filter banks from 20 to 8000 Hz to derive the entire log Mel-spectrogram;Apply a context window of 64 frames to the log Mel-spectrogram to obtain 2-D segments with the size of 64*64, among which a frame shift of 30 frames is used to produce overlapping segments | log Mel-spectrogram (3D) segments;utterance features from segment-levelfeatures by DTPM | EMO-DB dataset:anger, joy, sadness, neutral, boredom, disgust, and fearRML dataset:anger, disgust, fear, joy, sadness, surpriseeNETERFACE05 dataset: anger,disgust, fear, joy, sadness, surpriseBAUM-1s dataset:joy, anger, sadness, disgust, fear, surprise,boredom,contempt | DCNN-DTPM;DCNN-Average (These two methods are both based on DCNN and SVM) | WAR | For EMO-DB, RML, eNETERFACE05, and BAUM-1s datasets:DCNN-DTPM: 76.27%/ 62.40%/ 56.08%/ 38.42%DCNN-Average: 72.35%/ 59.46%/ 51.33%/ 36.10% |
| Shih et al. (2017) | Use CSHE to normalise data to eliminate speaker difference in audio files | 16 LLDs: ZCR, RMS energy: pitch frequency (normalised to 500 Hz), HNR, MFCC, delta of LLDs; 12 functionals: mean, standard deviation, kurtosis, etc. | anger, emphatic, neutral, positive, rest | SVM, NN, DNN, and their advanced versions by SR and CSHE+SR respectively | UA | For speaker-independent experiment:NN+CSHE+SR: 45.3%NN+SR: 39.6%NN:30.1% -DNN+CSHE+ -SR: 44.9% -DNN+SR: 39.9% -DNN: 28.6%;-SVM+- CSHE+SR: 42.6% -SVM+SR: 41.1%SVM:29.8% |
| Sivanagaraja et al. (2017) | Speech signals are resampled; Segmenting the speech recording with various lengths with overlapping/ non-overlapping respectively; Use multi-scale decomposition and multi-frequency decomposition to process the speech signals | raw speech signals; down-sampled speech signals; | happiness, anger, fear, sadness | MCNN | accuracy | MCNN with audio signals of 700 ms length and 50% overlapping: 50.28% MCNN with audio signals of 700 ms length and non-overlapping: 46.08% |
| Liu et al. (2018) | FS by correlation analysis and Fisher Criterion | prosodic, quality, and spectrum features of speaker-dependent and speaker-independent groups | surprise, happy, sad, angry, fear, neutral | ELM-DT;SVM-DT-Baselines-:KNN;NN-BP | Average Recognition Rate | ELM-DT with FS: 89.6% SVM-DT with FS: 87.2% ELM-DT without FS: 88.25% SVM-DT without FS: 87.63%;NN-BP with FS: 82.3%KNN with FS: 80.7% |
| Hadjadji et al. (2019) | Not mentioned | MFCC, duration, shimmer, jitter | neutral, joy, anger, sadness | MLP | Intra-speaker Recognition Rate;Inter-speaker Recognition Rate | For intra-speaker: MLP: 98.01%; For inter-speaker with 13 speakers' utterances as the learning base and one/ four/ non speaker in the learning base as the test samples: 92.50%/ 84.25%/ 54.75% |
| Jalal et al. (2019) | Not mentioned | For proposed method:F0, MFCC, energy augmented by delta and delta-delta, which are denoted, For baseline systems:log-spectrogram feature with 128 filter-banks, the eGeMAPS and super-vector features | RAVDESS database:-neutral, calm, happy, sad, angry, fearful, surprise and disgust | CNN with self-attention model;-Baselines-:TCapsNet; -RNN-LSTM; RNN-BLSTM; CNN; SVM | UA | For RAVDESS database:CNN with self-attention model: 95.1% TCapsNet: 68.1 to 69.4; BLSTM: 63.9% BLSTM+CNN: 51.3% SVM: 36.3% CapsuleNet + BLSTM: 35.4% CNN: 34.6% |

(*continued on next page*)

**Table A** (*continued*)

| Paper | Signal Pre-processing, Feature Extraction & Feature Selection | Audio Parameters | Emotion Classes | ML Algorithms | Evaluation Criteria | Classification Result |
|---|---|---|---|---|---|---|
| Ke et al. (2018) | FS by PCA | RMS of energy, ZCR, fundamental frequency, sounding probability, MFCC, the 1st, 2nd, and 3rd formant frequency and bandwidth; Statistical features, such as range, maximum, minimum, etc. | neutral, angry, fear, happy, sad and surprise | SVM with one-to-one method; ANN-BP | Accuracy | ANN: 75.0% SVM: 75.83% |
| Jiang et al. (2019) | Use OpenSMILE to extract LLDs; Use deep speech recognition networks to extract high-level acoustic features | LLDs (IS10 including energy, pitch, jitter, etc., MFCCs, eGeMAPS features); High-level features (SoundNet and VGGish Bottlenecks) | angry, happy, neutral, sad | Proposed method: DNNs-SHLA and SVM architecture | Accuracy | Proposed method: 0.64% |

## References

Abdelwahab, M., & Busso, C. (2017). Ensemble feature selection for domain adaptation in speech emotion recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5000–5004). IEEE.

Ahmad, J., Muhammad, K., Kwon, S-i., Baik, S. W., & Rho, S. (2016). Dempster-Shafer fusion based gender recognition for speech analysis applications. In *2016 international conference on platform technology and service (PlatCon)* (pp. 1–4). IEEE.

Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56–76.

Albornoz, E. M., & Milone, D. H. (2015). Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles. *IEEE Transactions on Affective Computing*, *8*, 43–53.

Ali, H., Hariharan, M., Yaacob, S., & Adom, A. H. (2015). Facial emotion recognition using empirical mode decomposition. *Expert Systems with Applications*, *42*, 1261–1277.

Alu, D., Zoltan, E., & Stoica, I. C. (2017). Voice based emotion recognition with convolutional neural networks for companion robots. *Science and Technology*, *20*, 222–240.

Alva, M. Y., Nachamai, M., & Paulose, J. (2015). A comprehensive survey on features and methods for speech emotion detection. In *2015 IEEE international conference on electrical, computer and communication technologies (ICECCT)* (pp. 1–6). IEEE.

Amiriparian, S., Christ, L., König, A., Meßner, E.-M., Cowen, A., Cambria, E., & Schuller, B. W. (2022). Muse 2022 challenge: Multimodal humour, emotional reactions, and stress. In *Proceedings of the 30th ACM international conference on multimedia MM '22* (pp. 7389–7391). New York, NY, USA: Association for Computing Machinery.

Anagnostopoulos, C.-N., Iliou, T., & Giannoukos, I. (2015). Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artificial Intelligence Review*, *43*, 155–177.

Aouani, H., & Ben Ayed, Y. (2018). Emotion recognition in speech using MFCC with SVM, DSVM and auto-encoder. In *2018 4th international conference on advanced technologies for signal and image processing (ATSIP)* (pp. 1–5).

Arjmandi, M. K., & Pooyan, M. (2012). An optimum algorithm in pathological voice quality assessment using wavelet-packet-based features, linear discriminant analysis and support vector machine. *Biomedical Signal Processing and Control*, *7*, 3–19.

Assunção, G., Menezes, P., & Perdigão, F. (2019). Importance of speaker specific speech features for emotion recognition. In *2019 5th experiment international conference (exp. at'19)* (pp. 266–267). IEEE.

Atmaja, B. T., & Sasou, A. (2022). Evaluating self-supervised speech representations for speech emotion recognition. *IEEE Access*, *10*, 124396–124407. https://doi.org/10.1109/ACCESS.2022.3225198.

Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W. (2017). Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)* (pp. 1–5). IEEE.

Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., Kwon, S., & Baik, S. W. (2019). Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, *78*, 5571–5589.

Basu, S., Chakraborty, J., Bag, A., & Aftabuddin, M. (2017). A review on emotion recognition using speech. In *2017 International conference on inventive communication and computational technologies (ICICCT)* (pp. 109–114). IEEE.

Bertero, D., & Fung, P. (2017). A first look into a convolutional neural network for speech emotion detection. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5115–5119). IEEE.

Bhavan, A., Chauhan, P., Shah, R. R., et al. (2019). Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*, *184*, Article 104886.

Braun, V., & Clarke, V. (2022). *Thematic analysis: A practical guide / Virginia Braun and Victoria Clarke*. London: SAGE Publications Ltd.

Chen, L.-W., & Rudnicky, A. (2023). Exploring Wav2Vec 2.0 fine-tuning for improved speech emotion recognition. arXiv:2110.06309.

Chernykh, V., & Prikhodko, P. (2017). Emotion recognition from speech with recurrent neural networks. arXiv preprint, arXiv:1701.08071.

Yogesh, C. K., Hariharan, M., Ngadiran, R., Adom, A. H., Yaacob, S., Berkai, C., & Polat, K. (2017a). A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert Systems with Applications*, *69*, 149–158. https://doi.org/10.1016/j.eswa.2016.10.035. https://www.sciencedirect.com/science/article/pii/S0957417416305759.

Costantini, G., Parada-Cabaleiro, E., & Casali, D. (2021). Automatic emotion recognition from DEMoS Corpus by machine learning analysis of selected vocal features. In *Biosignals* (pp. 357–364).

Czerwinski, M., Hernandez, J., & McDuff, D. (2021). Building an AI that feels: AI systems with emotional intelligence could learn faster and be more helpful. *IEEE Spectrum*, *58*, 32–38.

Darekar, R. V., & Dhande, A. P. (2018). Emotion recognition from Marathi speech database using adaptive artificial neural network. *Biologically Inspired Cognitive Architectures*, *23*, 35–42.

Deng, J., Zhang, Z., Eyben, F., & Schuller, B. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, *21*, 1068–1072.

Deng, L., Hinton, G., & Kingsbury, B. (2013). New types of deep neural network learning for speech recognition and related applications: An overview. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8599–8603). IEEE.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*, 572–587.

Essenwanger, O.M. (1986). Elements of statistical analysis.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*, 190–202.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on multimedia* (pp. 1459–1462).

Fahad, M. S., Ranjan, A., Yadav, J., & Deepak, A. (2021). A survey of speech emotion recognition in natural environment. *Digital Signal Processing*, *110*, Article 102951.

Fayek, H. M., Lech, M., & Cavedon, L. (2015). Towards real-time speech emotion recognition using deep neural networks. In *2015 9th International conference on signal processing and communication systems (ICSPCS)* (pp. 1–5). IEEE.

Fayek, H. M., Lech, M., & Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, *92*, 60–68.

Fei, W., Ye, X., Sun, Z., Huang, Y., Zhang, X., & Shang, S. (2016). Research on speech emotion recognition based on deep auto-encoder. In *2016 IEEE international conference on cyber technology in automation, control, and intelligent systems (CYBER)* (pp. 308–312). IEEE.

Feng, T., Hashemi, H., Annavaram, M., & Narayanan, S. S. (2022). Enhancing privacy through domain adaptive noise injection for speech emotion recognition. In *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7702–7706).

Feng, T., & Narayanan, S. (2022). Semi-FedSER: Semi-supervised learning for speech emotion recognition on federated learning using multiview pseudo-labeling. In *Proc. interspeech 2022* (pp. 5050–5054). https://www.isca-speech.org/archive/interspeech_2022/feng22_interspeech.html.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, *47*, 829–837.

Gadhe, R. P., & Deshmukh, R. R. (2015). Emotion recognition from isolated Marathi speech using energy and formants. *International Journal of Computer Applications*, *125*.

Getahun, F., & Kebede, M. (2016). Emotion identification from spontaneous communication. In *2016 12th International conference on signal-image technology & Internet-based systems (SITIS)* (pp. 151–158). IEEE.

Giannakopoulos, T. (2015). pyAudioAnalysis: An open-source Python library for audio signal analysis. *PLoS ONE*, *10*, Article e0144610.

Gorrostieta, C., Lotfian, R., Taylor, K., Brutti, R., & Kane, J. (2019). Gender de-biasing in speech emotion recognition. In *Interspeech* (pp. 2823–2827).

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, *26*, 91–108.

Gunawan, T. S., Alghifari, M. F., Morshidi, M. A., & Kartiwi, M. (2018). A review on emotion recognition algorithms using speech analysis. *Indonesian Journal of Electrical Engineering and Informatics*, *6*, 12–20.

Guo, L., Wang, L., Dang, J., Zhang, L., & Guan, H. (2018). A feature fusion method based on extreme learning machine for speech emotion recognition. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 2666–2670). IEEE.

Ha, W., Singh, C., Lanusse, F., Upadhyayula, S., & Yu, B. (2021). Adaptive wavelet distillation from neural networks through interpretations. *Advances in Neural Information Processing Systems*, *34*.

Han, J., Zhang, Z., Ringeval, F., & Schuller, B. (2017a). Prediction-based learning for continuous emotion recognition in speech. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5005–5009). IEEE.

Han, J., Zhang, Z., Ringeval, F., & Schuller, B. (2017b). Reconstruction-error-based learning for continuous emotion recognition in speech. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2367–2371). IEEE.

Harár, P., Burget, R., & Dutta, M. K. (2017). Speech emotion recognition with deep learning. In *2017 4th International conference on signal processing and integrated networks (SPIN)* (pp. 137–140). IEEE.

Harati, S., Crowell, A., Mayberg, H., & Nemati, S. (2018). Depression severity classification from speech emotion. In *2018 40th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 5763–5766). IEEE.

He, L., Lech, M., Zhang, J., Ren, X., & Deng, L. (2013). Study of wavelet packet energy entropy for emotion classification in speech and glottal signals. In *SPIE: Vol. 8878. Fifth international conference on digital image processing (ICDIP 2013)* (pp. 581–586).

Huang, J., Li, Y., Tao, J., Lian, Z., Niu, M., & Yi, J. (2018). Speech emotion recognition using semi-supervised learning with ladder networks. In *2018 First Asian conference on affective computing and intelligent interaction (ACII Asia)* (pp. 1–5). IEEE.

Huang, Y., Wu, A., Zhang, G., & Li, Y. (2015). Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition. *IET Signal Processing*, *9*, 341–348.

Hussain, L., Shafi, I., Saeed, S., Abbas, A., Awan, I. A., Nadeem, S. A., & Rahman, B. (2017). A radial base neural network approach for emotion recognition in human speech. *International Journal of Computer Science and Network Security*, *17*, 52.

Jain, A., Prakash, N., & Agrawal, S. (2011). Evaluation of MFCC for emotion identification in Hindi speech. In *2011 IEEE 3rd international conference on communication software and networks* (pp. 189–193). IEEE.

Jain, M., Narayan, S., Balaji, P., Bhowmick, A., Muthu, R. K., et al. (2020). Speech emotion recognition using support vector machine. arXiv preprint, arXiv:2002.07590.

Jaiswal, M., & Provost, E. M. (2020). Privacy enhanced multimodal neural representations for emotion recognition. In *Proceedings of the AAAI conference on artificial intelligence: Vol. 34* (pp. 7985–7993).

Jalal, M. A., Moore, R. K., & Hain, T. (2019). Spatio-temporal context modelling for speech emotion classification. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)* (pp. 853–859). IEEE.

Jiang, W., Wang, Z., Jin, J. S., Han, X., & Li, C. (2019). Speech emotion recognition with heterogeneous feature unification of deep neural network. *Sensors*, *19*, 2730.

Kalhor, E., & Bakhtiari, B. (2021). Speaker independent feature selection for speech emotion recognition: A multi-task approach. *Multimedia Tools and Applications*, *80*, 8127–8146.

Kannadaguli, P., & Bhat, V. (2019). Comparison of artificial neural network and Gaussian mixture model based machine learning techniques using DDMFCC vectors for emotion recognition in Kannada. In *2019 3rd International conference on electronics, materials engineering & nano-technology (IEMENTech)* (pp. 1–6). IEEE.

Ke, X., Zhu, Y., Wen, L., & Zhang, W. (2018). Speech emotion recognition based on SVM and ANN. *International Journal of Machine Learning and Computing*, *8*, 198–202.

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., & Mahjoub, M. A. (2018). Speech emotion recognition: Methods and cases study. In *ICAART (2)* (pp. 175–182).

Kerkeni, L., Serrestou, Y., Raoof, K., Mbarki, M., Mahjoub, M. A., & Cleder, C. (2019). Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Communication*, *114*, 22–35.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, *7*, 117327–117345.

Kim, E. H., Hyun, K. H., Kim, S. H., & Kwak, Y. K. (2009). Improved emotion recognition with a novel speaker-independent feature. *IEEE/ASME Transactions on Mechatronics*, *14*, 317–325.

Kim, Y., & Provost, E. M. (2013). Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 3677–3681). IEEE.

Kitchenham, B. A., Brereton, P., Turner, M., Niazi, M. K., Linkman, S., Pretorius, R., & Budgen, D. (2010). Refining the systematic literature review process—two participant-observer case studies. *Empirical Software Engineering*, *15*, 618–653.

Konar, A., & Chakraborty, A. (2015). *Emotion recognition: A pattern analysis approach*. John Wiley & Sons.

Koolagudi, S. G., Murthy, Y. S., & Bhaskar, S. P. (2018). Choice of a classifier, based on properties of a dataset: Case study-speech emotion recognition. *International Journal of Speech Technology*, *21*, 167–183.

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, *15*, 99–117.

Kuchibhotla, S., Vankayalapati, H. D., Vaddi, R., & Anne, K. R. (2014). A comparative analysis of classifiers in emotion recognition through acoustic features. *International Journal of Speech Technology*, *17*, 401–408.

Lalitha, S., Tripathi, S., & Gupta, D. (2019). Enhanced speech emotion detection using deep neural networks. *International Journal of Speech Technology*, *22*, 497–510.

Latif, S., Khalifa, S., Rana, R., & Jurdak, R. (2020). Poster abstract: Federated learning for speech emotion recognition applications. In *2020 19th ACM/IEEE international conference on information processing in sensor networks (IPSN)* (pp. 341–342).

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. W. (2021). Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing*.

Lech, M., Stolar, M., Best, C., & Bolia, R. (2020). Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding. *Frontiers of Computer Science*, *2*, 14.

Li, X., & Li, X. (2011). Speech emotion recognition using novel HHT-TEO based features. *Journal of Computers*, *6*, 989–998.

Li, Y., Zhao, T., & Kawahara, T. (2019). Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. In *Interspeech* (pp. 2803–2807).

Lieskovská, E., Jakubec, M., Jarina, R., & Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, *10*, 1163.

Lim, W., Jang, D., & Lee, T. (2016). Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA)* (pp. 1–4).

Liu, L.-Y., Liu, W.-Z., & Feng, L. (2023). SDTF-Net: Static and dynamic time–frequency network for speech emotion recognition. *Speech Communication*, *148*, 1–8.

Liu, Z.-T., Wu, M., Cao, W.-H., Mao, J.-W., Xu, J.-P., & Tan, G.-Z. (2018). Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, *273*, 271–280.

Liu, Z.-T., Xiao, P., Li, D.-Y., & Hao, M. (2019). Speaker-independent speech emotion recognition based on CNN-BLSTM and multiple SVMs. In *International conference on intelligent robotics and applications* (pp. 481–491). Springer.

Lokesh, S., & Devi, M. R. (2019). Speech recognition system using enhanced mel frequency cepstral coefficient with windowing and framing method. *Cluster Computing*, *22*, 11669–11679.

Luo, D., Zou, Y., & Huang, D. (2018). Investigation on joint representation learning for robust feature extraction in speech emotion recognition. In *Interspeech* (pp. 152–156).

Madanian, S., Nakarada-Kordic, I., Reay, S., & Chetty, T. (2023a). Patients' perspectives on digital health tools. *PEC Innovation*, *2*, Article 100171. https://doi.org/10.1016/j.pecinn.2023.100171. https://www.sciencedirect.com/science/article/pii/S2772628223000511.

Madanian, S., Parry, D., Adeleye, O., Poellabauer, C., Mirza, F., Mathew, S., & Schneider, S. (2022). Automatic speech emotion recognition using machine learning: Digital transformation of mental health.

Madanian, S., Rasoulipanah, H., & Yu, J. (2023b). Stress detection on social network: Public mental health surveillance: Public mental health surveillance. In *Proceedings of the 2023 Australasian computer science week ACSW '23* (pp. 170–175). New York, NY, USA: Association for Computing Machinery.

Manamela, P. J., Manamela, M. J., Modipa, T. I., Sefara, T. J., & Mokgonyane, T. B. (2018). The automatic recognition of Sepedi speech emotions based on machine learning algorithms. In *2018 International conference on advances in big data, computing and data communication systems (icABCD)* (pp. 1–7). IEEE.

Mannepalli, K., Sastry, P. N., & Suman, M. (2018). Emotion recognition in speech signals using optimization based multi-SVNN classifier. *Journal of King Saud University: Computer and Information Sciences*, *34*, 384–397. https://doi.org/10.1016/j.jksuci.2018.11.012. https://www.sciencedirect.com/science/article/pii/S1319157818307158.

Mao, S., Tao, D., Zhang, G., Ching, P. C., & Lee, T. (2019). Revisiting hidden Markov models for speech emotion recognition. In *ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6715–6719).

Mekruksavanich, S., Jitpattanakul, A., & Hnoohom, N. (2020). Negative emotion recognition using deep learning for Thai language. In *2020 Joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT & NCON)* (pp. 71–74). IEEE.

Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D Log-Mel spectrograms with deep learning network. *IEEE Access*, *7*, 125868–125881.

Milton, A., Roy, S. S., & Selvi, S. T. (2013). SVM scheme for speech emotion recognition using MFCC feature. *International Journal of Computer Applications*, *69*.

Minardi, H. (2013). Emotion recognition by mental health professionals and students. *Nursing Standard*, *27*.

Miner, A. S., Haque, A., Fries, J. A., Fleming, S. L., Wilfley, D. E., Wilson, G. T., Milstein, A., Jurafsky, D., Arnow, B. A., Agras, W. S., et al. (2020). Assessing the accuracy of automatic speech recognition for psychotherapy. *npj Digital Medicine*, *3*, 1–8.

Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2227–2231). IEEE.

Mitsuyoshi, S., Nakamura, M., Omiya, Y., Shinohara, S., Hagiwara, N., & Tokuno, S. (2017). Mental status assessment of disaster relief personnel by vocal affect display based on voice emotion recognition. *Disaster and Military Medicine*, *3*, 1–9.

Mu, Y., Gómez, L. A. H., Montes, A. C., Martínez, C. A., Wang, X., & Gao, H. (2017). Speech emotion recognition using convolutional-recurrent neural networks with attention model. *DEStech Transactions on Computer Science and Engineering*.

Mustafa, M. B., Yusoof, M. A., Don, Z. M., & Malekzadeh, M. (2018). Speech emotion recognition research: An analysis of research focus. *International Journal of Speech Technology*, *21*, 137–156.

Muthusamy, H., Polat, K., & Yaacob, S. (2015). Improved emotion recognition using Gaussian mixture model and extreme learning machine in speech and glottal signals. *Mathematical Problems in Engineering*, *2015*.

Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. arXiv preprint, arXiv:1706.08612.

Nasreen, P. N., Kumar, A. C., & Nabeel, P. A. (2016). Speech analysis for automatic speech recognition. In *Proceedings of international conference on computing, communication and science*.

Neumann, M., & Vu, N. T. (2019). Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7390–7394).

Niu, Y., Zou, D., Niu, Y., He, Z., & Tan, H. (2017). A breakthrough in speech emotion recognition using deep retinal convolution neural networks. arXiv preprint, arXiv:1707.09917.

Niveditha, C., & Ashok, K. (2019). ACNN based speech emotion recognition and noise suppression using modified cuckoo search algorithm. In *2019 2nd International conference on intelligent computing, instrumentation and control technologies (ICICICT), vol. 1* (pp. 79–86). IEEE.

Ortony, A., Clore, G. L., & Collins, A. (1990). *The cognitive structure of emotions*. Cambridge University Press.

Ozseven, T. (2018). Evaluation of the effect of frame size on speech emotion recognition. In *2018 2nd International symposium on multidisciplinary studies and innovative technologies (ISMSIT)* (pp. 1–4). IEEE.

Palo, H. K., & Mohanty, M. N. (2015). Classification of emotional speech of children using probabilistic neural network. *International Journal of Computer and Electrical Engineering*, *5*, 311–317.

Pandharipande, M., Chakraborty, R., Panda, A., & Kopparapu, S. K. (2018). An unsupervised frame selection technique for robust emotion recognition in noisy speech. In *2018 26th European signal processing conference (EUSIPCO)* (pp. 2055–2059).

Pepino, L., Riera, P., & Ferrer, L. (2021). Emotion recognition from speech using Wav2Vec 2.0 embeddings. arXiv:2104.03502.

Pereira, M., Chapaneri, S., & Jayaswal, D. (2016). Analysis of windowing techniques for speech emotion recognition. In *2016 International conference on information communication and embedded systems (ICICES)* (pp. 1–6). IEEE.

Picard, R. W. (2000). *Affective computing*. MIT Press.

Provost, E. M. (2013). Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 3682–3686). IEEE.

Rabiner, L. R. (1978). *Digital processing of speech signals*. Pearson Education India.

Rajisha, T., Sunija, A., & Riyas, K. (2016). Performance analysis of Malayalam language speech emotion recognition system using ANN/SVM. *Procedia Technology*, *24*, 1097–1104.

Ramakrishnan, A. G., Abhiram, B., & Mahadeva Prasanna, S. R. (2015). Voice source characterization using pitch synchronous discrete cosine transform for speaker identification. *The Journal of the Acoustical Society of America*, *137*, EL469–EL475. https://doi.org/10.1121/1.4921679. https://pubs.aip.org/asa/jasa/article-pdf/137/6/EL469/15318468/el469_1_online.pdf.

Ramet, G., Garner, P. N., Baeriswyl, M., & Lazaridis, A. (2018). Context-aware attention mechanism for speech emotion recognition. In *2018 IEEE spoken language technology workshop (SLT)* (pp. 126–131). IEEE.

Rawat, A., & Mishra, P. K. (2015). Emotion recognition through speech using neural network. *International Journal of Advanced Research in Computer Science and Software Engineering*, *5*, 422–428.

Saha, G., Chakroborty, S., & Senapati, S. (2005). A new silence removal and endpoint detection algorithm for speech and speaker recognition applications. In *Proceedings of the NCC* (p. 5). Citeseer volume 2005.

Schoneveld, L., Othmani, A., & Abdelkawy, H. (2021). Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, *146*, 1–7.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, *53*, 1062–1087.

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

Semwal, N., Kumar, A., & Narayanan, S. (2017). Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models. In *2017 IEEE international conference on identity, security and behavior analysis (ISBA)* (pp. 1–6). IEEE.

Sharma, M. (2022). Multi-lingual multi-task speech emotion recognition using Wav2Vec 2.0. In *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6907–6911).

Shegokar, P., & Sircar, P. (2016). Continuous wavelet transform based speech emotion recognition. In *2016 10th International conference on signal processing and communication systems (ICSPCS)* (pp. 1–6). IEEE.

Shih, P.-Y., Chen, C.-P., & Wang, H.-M. (2017). Speech emotion recognition with skew-robust neural networks. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 2751–2755). IEEE.

Singh, C., Kumar, A., Nagar, A., Tripathi, S., & Yenigalla, P. (2019). Emoception: An inception inspired efficient speech emotion recognition network. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)* (pp. 787–791). IEEE.

Sivanagaraja, T., Ho, M. K., Khong, A. W., & Wang, Y. (2017). End-to-end speech emotion recognition using multi-scale convolution networks. In *2017 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC)* (pp. 189–192). IEEE.

Song, F., Guo, Z., & Mei, D. (2010). Feature selection using principal component analysis. In *2010 international conference on system science, engineering design and manufacturing informatization, vol. 1* (pp. 27–30). IEEE.

Sonmez, Y. Ü., & Varol, A. (2019). New trends in speech emotion recognition. In *2019 7th International symposium on digital forensics and security (ISDFS)* (pp. 1–7). IEEE.

Stappen, L., Baird, A., Rizos, G., Tzirakis, P., Du, X., Hafner, F., Schumann, L., Mallol-Ragolta, A., Schuller, B. W., Lefter, I., Cambria, E., & Kompatsiaris, I. (2020). MuSe 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media: Emotional car reviews in-the-wild. In *Proceedings of the 1st international on multimodal sentiment analysis in real-life media challenge and workshop MuSe'20* (pp. 35–44). New York, NY, USA: Association for Computing Machinery.

Suganya, S., & Charles, E. Y. A. (2019). Speech emotion recognition using deep learning on audio recordings. In *2019 19th International conference on advances in ICT for emerging regions (ICTer)* (pp. 1–6).

Sun, Y., & Wen, G. (2015). Emotion recognition using semi-supervised feature selection with speaker normalization. *International Journal of Speech Technology*, *18*, 317–331.

Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: A review. *International Journal of Speech Technology*, *21*, 93–120.

Tao, F., Liu, G., & Zhao, Q. (2018). An ensemble framework of voice-based emotion recognition system for films and TV programs. In *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (pp. 6209–6213). IEEE.

Tashev, I. J., Wang, Z.-Q., & Godin, K. (2017). Speech emotion recognition based on Gaussian mixture models and deep neural networks. In *2017 information theory and applications workshop (ITA)* (pp. 1–4). IEEE.

Torres-Carrión, P. V., González-González, C. S., Aciar, S., & Rodríguez-Morales, G. (2018). Methodology for systematic literature review applied to engineering and education. In *2018 IEEE global engineering education conference (EDUCON)* (pp. 1364–1373). IEEE.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5200–5204). IEEE.

Tsouvalas, V., Ozcelebi, T., & Meratnia, N. (2022). Privacy-preserving speech emotion recognition through semi-supervised federated learning. In *2022 IEEE international conference on pervasive computing and communications workshops and other affiliated events (PerCom workshops* (pp. 359–364).

Tzirakis, P., Zhang, J., & Schuller, B. W. (2018). End-to-end speech emotion recognition using deep neural networks. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5089–5093).

Umamaheswari, J., & Akila, A. (2019). An enhanced human speech emotion recognition using hybrid of PRNN and KNN. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)* (pp. 177–183). IEEE.

Vasquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, J. R., Vargas-Bonilla, J. F., & Noeth, E. (2016). Wavelet-based time-frequency representations for automatic recognition of emotions from speech. In *Speech communication; 12. ITG symposium* (pp. 1–5). VDE.

Vondra, M., & Vích, R. (2009). Recognition of emotions in German speech using Gaussian mixture models. In *Multimodal signals: Cognitive and algorithmic issues* (pp. 256–263). Springer.

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–13. https://doi.org/10.1109/TPAMI.2023.3263585.

Wang, K., Su, G., Liu, L., & Wang, S. (2020). Wavelet packet analysis for speaker-independent emotion recognition. *Neurocomputing*, *398*, 257–264.

Wen, G., Li, H., Huang, J., Li, D., & Xun, E. (2017). Random deep belief networks for recognizing emotions from speech signals. *Computational Intelligence and Neuroscience*, *2017*.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision* (pp. 499–515). Springer.

Xiao, Y., & Watson, M. (2019). Guidance on conducting a systematic literature review. *Journal of Planning Education and Research*, *39*, 93–112. https://doi.org/10.1177/0739456X17723971.

Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., & Schuller, B. (2019). Speech emotion classification using attention-based LSTM. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *27*, 1675–1685.

Yadav, S. P., Zaidi, S., Mishra, A., & Yadav, V. (2021). Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN). *Archives of Computational Methods in Engineering*, 1–18.

Yi, E. (2018). Themes don't just emerge—coding the qualitative data. Medium.

Yogesh, C., Hariharan, M., Ngadiran, R., Adom, A. H., Yaacob, S., Berkai, C., & Polat, K. (2017b). A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal. *Expert Systems with Applications*, *69*, 149–158.

Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, *59*, 103–126.

Zhang, S., Zhang, S., Huang, T., & Gao, W. (2018a). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, *20*, 1576–1590. https://doi.org/10.1109/TMM.2017.2766843.

Zhang, Y., Du, J., Wang, Z., Zhang, J., & Tu, Y. (2018b). Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC)* (pp. 1771–1775). IEEE.

Zhao, H., Xiao, Y., Han, J., & Zhang, Z. (2019a). Compact convolutional recurrent neural networks via binarization for speech emotion recognition. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6690–6694). IEEE.

Zhao, J., Mao, X., & Chen, L. (2019b). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, *47*, 312–323.

Zhao, Z., Bao, Z., Zhao, Y., Zhang, Z., Cummins, N., Ren, Z., & Schuller, B. (2019c). Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. *IEEE Access*, *7*, 97515–97525.

Zheng, W., Yu, J., & Zou, Y. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 International conference on affective computing and intelligent interaction (ACII)* (pp. 827–831). IEEE.

Zhou, X., Guo, J., & Bie, R. (2016). Deep learning based affective model for speech emotion recognition. In *2016 Intl IEEE conferences on ubiquitous intelligence & computing, advanced and trusted computing, scalable computing and communications, cloud and big data computing, Internet of people, and smart world congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)* (pp. 841–846). IEEE.

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