Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-17769-6

1232: HUMAN-CENTRIC MULTIMEDIA ANALYSIS

Semantic speech analysis using machine learning and deep learning techniques: a comprehensive review

Suryakant Tyagi[1] · Sándor Szénási[1,2,3]

Received: 23 December 2022 / Revised: 22 November 2023 / Accepted: 28 November 2023

Abstract

Human cognitive functions such as perception, attention, learning, memory, reasoning, and problem-solving are all significantly influenced by emotion. Emotion has a particularly potent impact on attention, modifying its selectivity in particular and influencing behavior and action motivation. Artificial Emotional Intelligence (AEI) technologies enable computers to understand a user's emotional state and respond appropriately. These systems enable a realistic dialogue between people and machines. The current generation of adaptive user interference technologies is built on techniques from data analytics and machine learning (ML), namely deep learning (DL) artificial neural networks (ANN) from multimodal data, such as videos of facial expressions, stance, and gesture, voice, and bio-physiological data (such as eye movement, ECG, respiration, EEG, FMRT, EMG, eye tracking). In this study, we reviewed existing literature based on ML and data analytics techniques being used to detect emotions in speech. The efficacy of data analytics and ML techniques in this unique area of multimodal data processing and extracting emotions from speech. This study analyzes how emotional chatbots, facial expressions, images, and social media texts can be effective in detecting emotions. PRISMA methodology is used to review the existing survey. Support Vector Machines (SVM), Naïve Bayes (NB), Random Forests (RF), Recurrent Neural Networks (RNN), Logistic Regression (LR), etc., are commonly used ML techniques for emotion extraction purposes. This study provides a new taxonomy about the application of ML in SER. The result shows that Long-Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) are found to be the most useful methodology for this purpose.

Keywords Emotion extraction · Speech · Machine learning · Data analysis · Artificial neural networks · Long-short term memory · Convolutional neural networks

✉ Suryakant Tyagi
suryatyagi10000@gmail.com

Sándor Szénási
szenasi.sandor@nik.uni-obuda.hu

[1] Doctoral School of Applied Informatics and Applied Mathematics, Obuda University, Budapest 1034, Hungary

[2] John von Neumann Faculty of Informatics, Obuda University, Budapest 1034, Hungary

[3] Faculty of Economics and Informatics, J. Selye University, Komárno 94501, Slovakia

Published online: 19 December 2023

1 Introduction

Digital data is constantly growing and becoming more accessible, making it challenging for software tools and technologies to display, store, manage, and analyze it [1]. In real-world social media, such as Twitter and Facebook, Polyglots are much more likely to submit code-switched items by combining two different natural languages [2]. In the data mining (DM) and natural language processing (NLP) sectors, these code-switched texts have generated a plethora of new research areas, including speech recognition, information extraction, language modeling, and lexicon analysis, to mention a few [3]. Emotion identification or sentiment analysis for code-switched texts, which aims to find emotions or sentiments in a piece of mixed-language literature, is one of the most popular research topics [4].

In the past ten years, numerous neural network models have been investigated with the aim of code-switched emotion detection. The current approaches are primarily concerned with building robust neural models with intricate features or architectures [5]. To enhance the code-switched detection model, CNNs and LSTM with the attention mechanism are applied. These techniques, which take characteristics directly from the code-switched text itself, might convey different emotions in either one language or both [6]. The goal of speech emotion recognition (SER) is to identify emotion in speech, regardless of the semantic content. Figure 1 shows the block representation of the SER system. The speech is first given to the ML-based training system then it gets pre-processed with

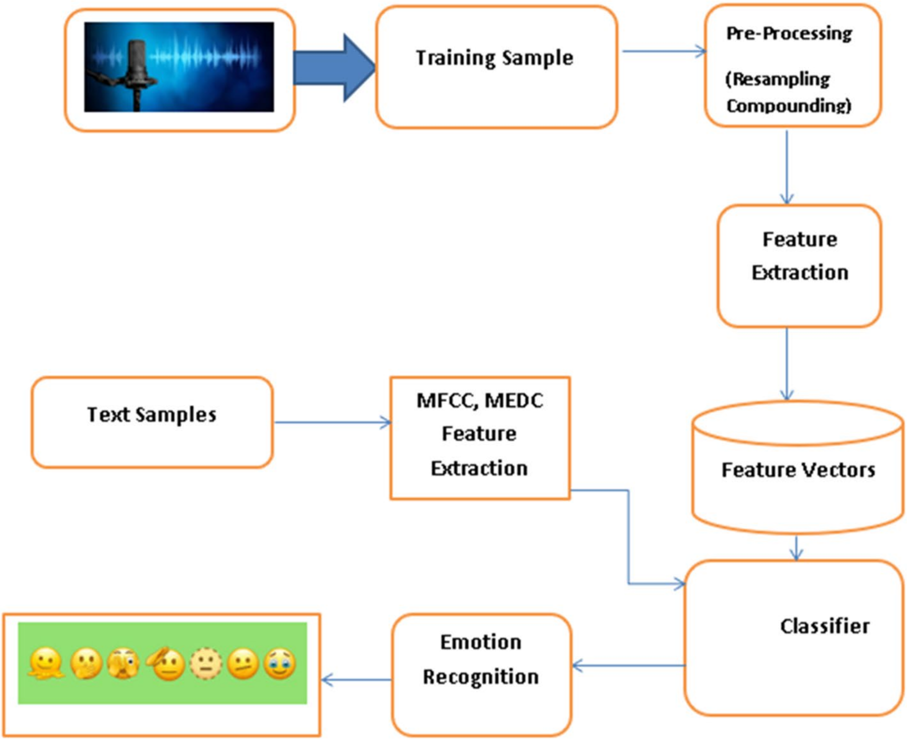

Fig. 1 Block diagram of SER

another pre-processing system. After it undergoes feature extraction to specify the features. Another text sample with Mel frequency cepstral coefficients (MFCC) and (MEDC)-enabled featured extraction goes to the classifier. The classifier differentiates the difference between the two outputs of the feature extractions and then sends a signal to the emotion recognition system. The emotional recognition system finally detects the emotions from the given speech.

Languages usually have different ways of expressing emotions, which keeps these techniques from progressing. As a result, a successful model ought to be able to more efficiently and effectively mine both monolingual and bilingual data. A parallel translation with a bilingual perspective translates texts using code-mixing into both languages. By doing this, we can prevent information loss and preserve the original contexts as much as we can in both languages [7]. Additionally, the total system can use attention-based Bidirectional LSTMs as the shared encoder under adversarial learning to dynamically and selectively use both the monolingual private and the bilingual shared features in code-mixed texts [5]. A bilingual-view parallel translation, which translates code-mixed texts into both languages, provides the least amount of information loss while preserving the original settings [8]. The task of extracting features from concurrent translation texts in two languages is handled by an adversarial dual-channel encoder [9]. Additionally, the total system can use attention-based Bidirectional LSTM Networks as the shared encoder under adversarial learning to dynamically and selectively use both the monolingual private and the bilingual shared features in code-mixed texts [10].

In recent years, the connection between human and machine communication has grown in significance. Many studies were conducted in the 1950s to teach robots to recognize human voices [11, 12]. To enhance the effectiveness of human-to-machine communication, the statistics of the human voice must be recognized [13]. A range of circumstances, such as educational and therapeutic settings, as well as entertainment and the arts, can benefit from the use of emotional content in speech [14]. Speech signals may now be used to communicate between humans and machines thanks to several technological developments [15]. Speech recognition and speech-to-text (STT) technology have made mobile phones an increasingly popular means of communication [16]. One of the signal recognition fields of study that is expanding the quickest is speech recognition. SER is a new area of study that has the potential to advance a variety of industries, including automatic translation systems and human–machine interfaces [17]. As a result, the study concentrates on a variety of speech extraction traits, emotion databases, and classification strategies. Figure 2 shows the implication of CNN methods in the SER.

Speech contains signals that reveal the speaker, their language, and their emotions in addition to the message. The majority of the speech processing algorithms currently in use

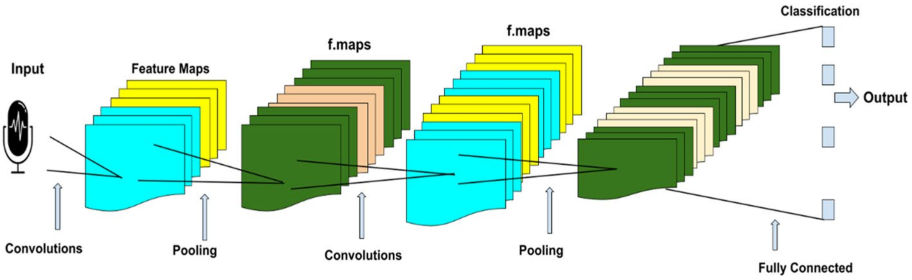

Fig. 2 Application of CNN methodology in SER [158]

work admirably with neutral studio recordings, but when it comes to emotional speech, they struggle [18]. This is because it is challenging to represent and define the emotions that are expressed through spoken language [19]. Communication becomes more natural when it contains emotional content. By using the appropriate semantics, emotions can be used to describe the same idea in a different way [20]. The project's primary goal is to analyze and examine the potential use of ML and DL for emotion-based hazardous speech identification [21]. Naive Bayes, Support Vector Machine, and K-Nearest Neighbor models are three ML methods are used to check the identify emotion label. The Naïve Bayes model demonstrates strong performance with an accuracy of 82.3% in emotion detection. Additionally, the F1-Score metrics for this model stand at 0.89. This achievement is particularly noteworthy in the context of analyzing emotions and identifying them within Tamil song lyrics.

This review was formulated by following the steps of the PRISMA methodology and it is organized into different sections. Section 2 explains the overview and research gap. Section 3 discusses the traditional emotional extraction in speech. Section 4 reviews the previous works of recent years. The outcomes of the application of the PRISMA approach are reported together with the responses to the research questions and the acquired results are discussed in Section 5. In Section 6, conclusions and closing remarks are offered.

2 Overview and research gap

In past years a lot of research has been going on SER. The major focus is on extracting emotions like anger, happiness, sadness, fear, surprise, and disgust from the speech but almost more than 300 emotions apperars in the speech [22]. It is very difficult to find out all emotions from speech and decide whether it is harmful to society. DL techniques are the subset of ML that is widely used in extraction from voice, especially data models created for the special object for pattern recognition and detection making. One of the frequently used approaches is multi-model learning with deeper layers of architecture like RNN, Deep Belief Neiwork, Deep Boltzmann Machine (DBM), CNN, and Auto Encoder [23]. In recent years, emotional extraction in speech has gained much attention, especially after the popularity of social media. The peer-reviewed journals have shown a significant increase in the past few years. Science Direct shows 5006 results on the topic out of which 649 are review articles, 3,475 are research articles, 67 are encyclopedias, 433 are book chapters, 72 are conference abstracts, 4 are book reviews, and the rest are other types of articles. On the other hand, MDPI shows 62 search results on this topic. Springer shows 19,614 search results. Out of this, 6,499 are chapters, 3,956 are articles, 3,788 are books, 5,040 are conference papers and proceedings, the rest are other documents. Finally, IEEEXplore shows 1,964 search results. This has 1,689 conferences, and 246 journals, the rest are other documents. Figure 3 shows the year-wise development works from 1999 to 2022. The graph shows that in the past three years, there is a huge development in the research work. This suggests that this topic is gaining huge attention but there is still too much research work that needs to be done in this field. According to the literature, there are significant differences across the databases in terms of the number of performers, the number of emotions recognized, and the methodology. Speech-emotional databases are used in both psychological investigations to understand the patient's behavior and in circumstances when it is desirable to automate emotion recognition. When real-time data is used, the system gets

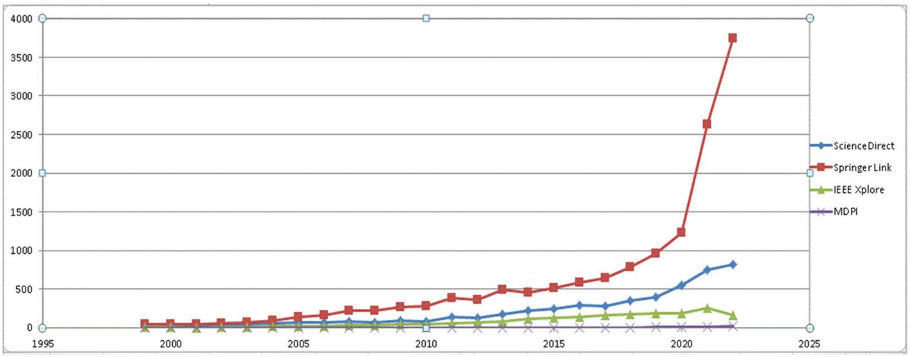

Fig. 3 Advancement of the SER research work in recent years

complicated and emotion recognition is challenging. Figure 3 hows the advancement of the SER papers in the recent years.

Extraction of features and selection is also a major focus of current research, which aims to increase performance accuracy by selecting the best characteristics. To improve system performance and recognize the appropriate emotions, classifier selection is a difficult process, according to data analysis [24]. Although many classifiers have been selected for the speech emotion identification system, no clear victor has emerged. SER is a complex problem. Emotions and facial expressions are two possible ways to recognize emotions when conducting sentiment analysis in the future [25]. If a future study is to be done on emotion recognition in general or on hazardous speech detection in particular, certain difficulties have been noted.

In the past ten years, contributions to the SER system focused on the novel approach based on a statistical method name extracted statistical feature works on a unimodal approach, gender, speaker-independent, and real-time [26]. Feature learning approaches from speech data are used to extract the feature statistically in terms of the degree of standard derivation [27]. The real-time issue with human–computer interaction is to catch the human voice and reply with accuracy like a human. The design to resolve this problem used an automatic speech recognition system that can figure out different emotional classes coming from the human mind and select the major feature extraction from the speech signals [28]. The improvement of SER used a novel feature-learning method based on an adaptive time–frequency coefficient to improve the accuracy of SER using the simulation performance based on the Persian Drama Radio Emotional Corpus, the Surrey Audio-Visual Expressed Emotion Database, and Berlin Emotional Speech Database [29]. The experimental result set shows that methodology adaptive time–frequency based on FFT with Cepstral features works effectively resulting in SAVEE (80% accuracy), EMO-DB(97.57% accuracy), and PDREC (91.46% accuracy) data sets [30]. The SER is a complicated problem because it extracts the natural feature emotion from the real audio data set, and it is playing too. SER plays a significant role in human–computer interaction [31]. The study's main goal is to improve classification accuracy and extract eight emotions from human speech. Emotion prediction from the speech used MFF-Aug research by white noise injection, pitch tuning, and noise removal [32]. On pre-processed speech signals, the feature extraction techniques MFCC, zero crossing rate, and root mean square [33]. To analyze the voice emotional classification and speech representation used the CNN approach. The next step is to compare the LSTM and CNN method. The TESS, CREMA, RAVDESS, and SAVEE datasets were used for analyzing the experimental methodology and accuracy of 92.6%, 89.94%, 84.9%, and 99.6% [34]. SER has a broad range in the smart application field of medical science, human–robot interaction, and online gaming apps [35]. Smart SER system applications are two main major problems computational cost and time to figure out this issue used preprocessing steps on six databases i.e. EmoDB, RAVDESS, IEMOCAP, ShEMO, DEMoS, and MSP-Improv which speech segments with similar formate characteristics [36].

3 Traditional emotional extraction in speech

Signal pre-processing, feature extraction, and classification are the three core elements of emotion identification systems based on digitized speech [37]. To establish meaningful units of the signal, acoustic pre-processing techniques like denoising and segmentation are used. To find the pertinent features present in the signal, feature extraction is used. Speech signal processing, feature extraction, and classification are all covered in-depth in this section [38]. Due to their importance to the subject, the distinctions between spontaneous and performed speech are also examined. Speech enhancement is carried out in the initial step of speech-based signal processing, where the noisy components are eliminated [39]. Feature extraction and feature selection make up the second stage. The pre-processed speech signal is used to extract the necessary features, and the extracted features are then used to make the selection. The study of speech signals in the temporal and frequency domains is typically the foundation for such feature extraction and selection [40]. In the third stage, different classifiers, including Gaussian Mixture Model and Hidden Markov Model, are used to categorize these features. Last but not least, several emotions are identified based on feature classification [41].

3.1 Improving speech input data for speech emotion recognition

During the data collection phase, noise frequently taints the input data used for emotion recognition. These flaws make the feature extraction and classification less precise. This means that for emotion detection and recognition algorithms to function properly, the input data must be improved. The speaker and recording variance is removed during this preprocessing stage while the emotional discrimination is retained [42].

3.2 Extraction and selection of features in speech emotion recognition

Following augmentation, segments are used to characterize the speech stream. Based on the information gathered, pertinent traits are extracted and divided into several groups. Short-term categorization, which is based on properties that last just a short while, like energy, formants, and pitch, is one type of classification. The other is known as long-term categorization, and two often employed long-term features are mean and standard deviation [43]. The intensity, pitch, pace, and variance of uttered words are among the prosodic qualities that are typically significant in identifying different types of emotions from the input speech signal [44].

3.3 Acoustical measures in speech emotion recognition

Each emotion's information availability is encrypted. Among the most studied subjects in this area are vocal parameters and how they relate to emotion identification. Many factors are typically taken into account, including spoken word characteristics like intensity, pitch, pace, and quality of voice [45]. The assumption that emotions are separate categories with independent existence is a common one in the simple view of emotion. In many cases, the relationship between intensity and pitch, and activation is such that the intensity value rises with a high pitch and falls with a low pitch [45]. If the speaker is acting, whether there are many different speakers, and the person's mood or personality all have an impact on how acoustic factors transfer to emotion. Emotions in HCI are typically not the conventional discrete emotions; rather, they are frequently weakly expressed, jumbled, and difficult to identify from one another [46]. Based on the feelings exhibited by a person, emotional remarks in literature are classified as either good or negative. Some research suggests that actors exaggerate their emotional expressions because listener-based performed emotions are significantly stronger and more accurate than real emotions. Areas within the space, according to the study, can describe basic emotions. While valence depicts the impact of positivity and negativity on emotions, arousal shows the intensity of serenity or excitement [47].

3.4 Classification of speech emotion recognition features

Different classifiers have been researched in the literature to create systems like SER, speech recognition, and speaker verification [48]. On the other hand, the reasons for selecting a specific classifier for a given speech task are frequently left out of most applications. Typically, classifiers are chosen based on an empirical evaluation of some signs or a rule of thumb, as was previously discussed. Ordinarily, the two primary types of pattern recognition classifiers used for SER can be broadly divided into linear classifiers and non-linear classifiers [49]. With a linear arrangement of numerous objects, linear classifiers typically conduct classification based on object attributes. Most of the time, these objects are assessed as an array known as a feature vector [50]. On the other hand, non-linear classifiers are used to characterize things before creating a non-linear weighted combination of those objects.

3.5 Databases for recognition of speech emotion

Many academics use speech emotional databases for several research projects. The most crucial aspects of evaluation for emotion recognition are the caliber of the databases used and the performance attained. Depending on the reason for developing speech systems, different techniques and goals are used to collect voice databases [51]. The basic categories of speech databases are used to construct emotional speech systems. The speech data in these databases was captured by skilled and seasoned actors. This database is regarded as the one that makes it easiest to collect the speech-based dataset of different emotions out of all the others [52]. It is estimated that this method is used to compile over 60% of speech datasets. This is a different kind of database where the emotional set is gathered by fabricating a fake emotional circumstance. This is done without the artist or speaker's awareness. This database is more lifelike than actor-based databases [53].

The speaker should be aware that they have been videotaped for research purposes, therefore an ethical question might arise. These databases are tough to collect owing to the difficulties in recognition even though they are the most realistic. Conversations from contact centers, the general public, and other situations are typically recorded for natural emotional speech databases [54]. When research on speech-based emotion identification began to take off in the early 1990s, researchers frequently started using acted databases before switching to realistic databases [55]. The most often utilized performed databases are the 10 actors' recorded voices included in the Berlin emotional speech database and the Danish emotional speech database [56]. Four test subjects were asked to speak a variety of words in five distinct emotional states. German-Aibo emotion and Smart-Kom data, where the actors' voices are captured in a lab, are included in the data. Additionally, real-world call center interactions captured during live recordings have been utilized [57]. According to the literature, there are significant differences across the databases in terms of the number of performers, the number of emotions recognized, and the methodology [58]. Speech-emotional databases are used in both psychological investigations to understand the patient's behavior and in circumstances when it is desirable to automate emotion recognition [59].

4 Methodology

This literature review uses the PRISMA methodology. We screened 46,131 articles from the WoS and Scopus websites. To find relevant articles, we use an advanced filtering system on different scientific websites. We use logical operators to find the most relevant

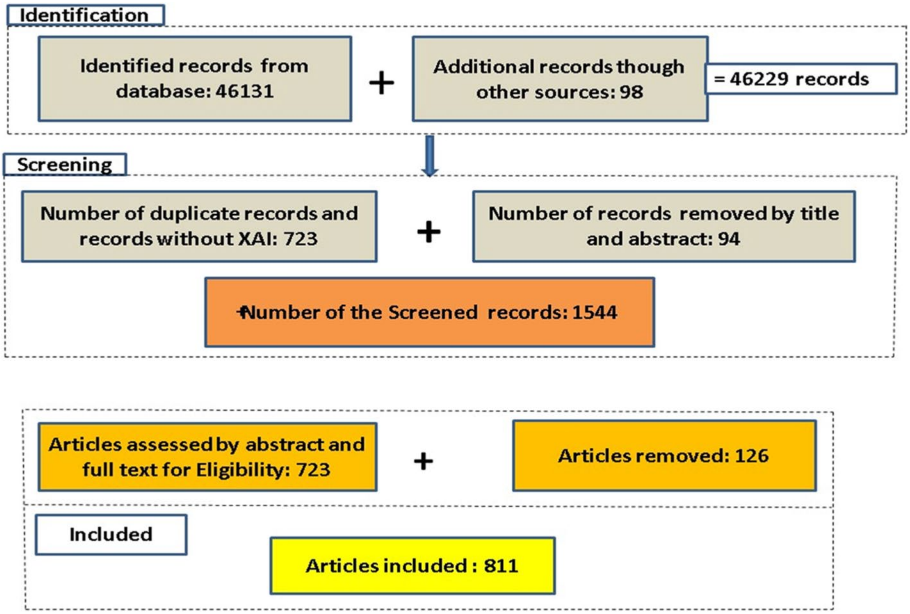
Identification
Identified records from database: 46131
+
Additional records though other sources: 98
= 46229 records
Screening
Number of duplicate records and records without XAI: 723
+
Number of records removed by title and abstract: 94
Number of the Screened records: 1544
Articles assessed by abstract and full text for Eligibility: 723
+
Articles removed: 126
Included
Articles included : 811

Fig. 4 Flowchart of the used PRISMA methodology

documents. Figure 4 shows the flowchart of the PRIMSA methodology used to select the papers. Table 1 shows the methodology used for shorting the relevant papers for this review.

Table 1 Keyword filtering system used for shorting the papers

Springer: 23,350 documents	
Keywords used for Filters	Number of documents found
Emotion speech ML OR semantic DL	4,059
Emotion, NLP “OR” audio, DL	13,303
Emotion, NLP “OR” speech, extraction, CNN	739
Speech, emotion, LSTM “OR” extraction, CNN	607
Speech, emotion, MFCC “OR” extraction, Perceptual Linear Predictive	105
Speech processing, fast Fourier transform “OR” human emotion, extraction, Discrete Cosine Transform	157
Speech, CNN “OR” extraction, SVM	3,310
Speech, Linguistic Features, neural network “OR” emotion, acoustic features	1,270
IEEE Xplore: 13827	
Emotion, speech, ML “OR” semantic, DL	2,288
Emotion, NLP “OR” audio, DL	2,374
Emotion, NLP “OR” speech, extraction, CNN	304
Speech, emotion, LSTM “OR” extraction, CNN	3,602
Speech, emotion, MFCC “OR” extraction, Perceptual Linear Predictive	43
Speech, fast Fourier transform “OR” human emotion extraction, Discrete Cosine Transform	383
Speech, CNN “OR” extraction, SVM	4,699
Speech, Linguistic Features, neural network “OR” emotion, acoustic features	134
ScienceDirect: 8,599	
Emotion, speech, ML, semantic DL	1,250
Emotion, NLP, audio, DL	1,508
Emotion, NLP, speech, extraction, CNN	623
Speech emotion, LSTM, extraction, CNN	562
Speech emotion, MFCC, extraction, Perceptual Linear Predictive	132
Speech processing, fast Fourier transform, human emotion extraction, Discrete Cosine Transform	84
Speech, Linguistic Features, neural network, emotion, acoustic features	478
MDPI: 355	
Emotion speech, ML “OR” speech, DL	21
Emotion, NLP “OR” audio, DL	18
Emotion, NLP “OR” speech extraction, CNN	2
Speech, emotion, LSTM “OR” extraction, CNN	207
Speech, emotion, MFCC “OR” extraction, Perceptual Linear Predictive	1
Speech, fast Fourier transfosrm “OR” Speech, Discrete cosine transform	1
Speech, CNN “OR” extraction, SVM: 103	103
Speech, Linguistic Features, neural network “OR” emotion, acoustic features	02

5 Literature review

As a developing field of ML research, DL has attracted more attention in recent years [60]. DL techniques for SER have several advantages over conventional techniques, including the ability to recognize composite systems and structures without the need for tuning and physical feature extraction, the capacity to work with unlabeled data [61], and a propensity to extract low-level features from raw data. For a machine, identifying human emotions is a challenging task, but for us, it is straightforward [62]. To improve communication between humans and robots, an emotion recognition system makes use of knowledge about emotions [63]. The basic frequencies, Linear Prediction Cepstrum Coefficient, and Mel Frequency Cepstrum Coefficient are a few of the speech aspects that have been researched [64]. There is a possibility of speaker-dependent or speaker-independent emotional information recognition. Other classifiers include K-nearest Neighbors (KNN), SVM, CNN, and others [65]. Various strategies for identifying emotional states in speech using selected papers from the period 2005 to 2018 help to create a model that can recognize and classify six different emotions using a Deep Neural Network (DNN) for emotion identification [66]. The research concludes by averaging the accuracy of the two databases, which were used in the study. Both sets of data are utilized to extract features using MFCC, SVM, and Gaussian mixture models using the retrieved feature to categorize the speaker's age [67]. The emotional approach utilized in the study, which focuses on transformation, is then used to forecast the training data. Real-time CNN models were suggested as a way to recognize emotions [68]. This paradigm includes subcategories for being angry, joyful, and depressed. The model's accuracy is 66.1% on average. A few different emotions were identified. It is therefore difficult to predict any other feeling but their own [69]. Text categorization was employed to examine the content of speech. The main method for translating emotions from audio to text is text mining. Huang and his coworkers created a novel method for identifying emotions. Using a nonlinear SVM classifier, four kinds of emotions were identified [70]. The Deep Belief Network model took too long to extract features compared to other feature extraction techniques, which was the flaw in this unique approach [71].

2010 through 2022 saw the exploration of this work. All ML techniques are combined with "hate speech identification" in the inquiry. It had been done to categorize the procedure using an ML approach. The state-of-the-art review methodology had been modified from [77–89]. Tables 2 and 3 provided a quick summary of some current research in this area.

A 1D CNN performs better in classification tasks than traditional machine learning algorithms. By learning low-level or spectral information, SER technology is used to categorize emotions. A CNN-based method for identifying emotions employs feature space for low-level data such as pitch and energy as well as spectral information such as a log-Mel spectrogram, STFT. Calculating the spectral flux, which assesses the spectral change between two frames, involves squaring the difference between the normalized magnitudes of the spectra of two successive short-term windows [90].

$$F1_{(i,i-1)} = \sum_{K=1}^{Wf_L} \left(EN_i(K) - WN_{i-1}(K)\right)^2$$

where $EN_i(K)$ is the K^{th} normalized discrete Fourier transform, i is the frame, Wf is the weighted frequency, N is the bin number, and F(n) is the center frequency of the bin (Table 4).

Table 2 Review of DL in SER

Reference	Recognized emotions	Database	Deep learning approach	Application
[72]	Disgust, boredom, fear, sad, happy	DEAP Database	LSTM and Bi-LSTM	Biofeedback and Psychological wellbeing
[73]	Happiness, Sadness, Neutral	eNTERFACE and EMOD database	CNN	Smartphone application (TFL Auth app) and Tensorflow lite
[74]	Sad, anger, fear, happy, neutral	BAVED, EMO-DB, SAVEE, and EMOVO	Lightweight CNN and White noise, Time stretch	Arabic vocal emotions and Human–computer interaction
[75]	Anger, stress, fear, and happy, netural	IEMOCAP, Ryerson, and RAVDESS	CNN EfficientNetB0	Emotion change detection and Emotional transformation
[76]	Happy, sad, neutral, joy. Anger emotions	Emotion Challenge, FAUAEC	Qualitative method	Offensive language and Malicious comment
[77]	Neutral, Fear, Disgust, Anger, Sadness, Joy	IEMOCAP and TIMIT for emotion and phenomenon recognition	Societal implications	Attacks against reputation, hate speech and fake news
[78]	Attribute detection	Arabic MGB	CNNs and DBN-DNN	Arabic speech recognitions and Phonologically distinctive articulatory features

Table 3 Recent ML methods and their applications in hate speech detection

Reference	Year	Source	Application	Algorithms/Methods
[79]	2022	Information Processing & Management	Harmful news identification	Sentimental Analysis, BERT Module
[80]	2022	Speech Communication, Elsevier	Emotional speech recognition	Hidden Markov models, Artificial Neural Networks
[81]	2022	Computation and Language	Tackling Cyber-Bullying	SVM
[82]	2022	ACM Transactions on Internet Technology	Negative Information Measurement	k-nearest neighbors, principal component analysis (PCA)
[83]	2022		Radicalization and Hate Speech Detection	Similarity-based sentiment projection, TF-IDF method
[84]	2022	IEEE Access	Hate Speech Detection	Convolutional, BiGRU, and Capsule network-based deep learning model
[85]	2022	Computer Speech & Language	Hate speech and offensive language detection in Dravidian languages	Deep Learning, Machine Learning algorithms
[86]	2022	International Journal of Advanced Computer Science and Applications	Detecting Hate Speech on Twitter Network	Linear Regression, Decision Tree, Support Vector Machine
[87]	2022	Expert Systems with Applications	Sentiment detection	CNN and LSTM
[88]	2022	arXiv preprint	Large-Scale Hate Speech Detection	Transformer-based language models
[89]	2022	Applied Science	Sentiment Analysis of COVID-19	LSTM-RNN
[90]	2022	2nd International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications	Emotion Detection in Tweets	LSTM
[91]	2022	IEEE Access	Racism Detection	GCR-NN Model
[92]	2022	Springer Nature Computer Science 3	Multimodal Hate Speech Detection	BERT model, SMOTE oversampling technique and random under sampling
[83]	2022	Computational Intelligence and Neuroscience	Detection of Speaker Emotions	1D CNN

Table 4 Notable preliminary research using ML

References	Years	Sources	Application	Module/Algorithms
[93]	2010	International Conference on Intelligent Tutoring Systems	AutoTutor	SVM
[94]	2010	International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction	user's eye gaze behavior	SVM and NB
[95]	2011	Speech Communication	Parent-infant interaction analysis	Supervised Learning
[96]	2011	IEEE Transactions on Audio, Speech, and Language Processing	Convex Combination of Multiple Statistical Models	Statistical distributions
[97]	2012	Springer	Forensic Speaker Recognition	Gaussian Mixture Model, Speaker Verification
[98]	2013	Journal on Multimodal User Interfaces	Depression diagnosis and monitoring	LBP-TOP, SVM
[99]	2014	Intelligent Decision Technologies	Speech activity detection	Hidden Markov-Models
[100]	2014	Journal of Intelligent Systems	Speaker Verification	Adaptive VAD Algorithm
[101]	2015	International Journal of Multimedia and Ubiquitous Engineering 10	Hate Speech Detection	SVM and maximum entropy
[102]	2015	Speech Communication	Acoustic space variability	Gaussian Mixture Models, Probabilistic Acoustic Volume
[103]	2016	IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies	Smartphone-Based Acoustic	BigEAR
[104]	2016	Lecture Notes in Computer Science	Sentiment analysis	RNN and LSTM

Subjectivity analysis can separate objective sentences from subjective statements and remove the latter from the corpus. A rule-based classifier for identifying hate speech was constructed using a vocabulary built on semantic, anti-, and theme-based components. Two machine learning methods that can be used directly to raise precision and recall scores are SVM and maximal entropy [99]. MFCC and Gaussian Mixture Models are widely combined to identify or forecast the presence or severity of depression in speakers [100]. The context of the discussions, which may include the speakers' emotions or mood, is inferred by the BigEAR architecture using a psychological audio processing chain (PAPC). The advantage of the BigEAR framework is that psychologists are no longer required to evaluate the expanding body of acoustic big data, which calls for them to carefully listen to each audio recording and classify emotions [101]. Bi-grams, SentiWordNet, and stop word removal have all been demonstrated to improve accuracy when it comes to Twitter feature selection [102]. The most popular machine learning algorithms for sentiment analysis, emotion analysis, and hate speech identification on social media platforms are shown in a block diagram in Fig. 5. This demonstrates that LSTM and SVM algorithms are frequently used to produce the most accurate outcomes (Table 5).

6 Results and discussion

6.1 Methods for identifying emotions in speech

The fundamental ER system is made up of the processes listed below, according to speech as depicted in Fig. 6.

The speech samples are utilized as input in the initial stage. If a standard database is not used, samples are preprocessed to eliminate noise using a variety of trade-offs, including Audacity, WavePad, Sony Creative Noise, Once Audio, and PRATT Reduction [110]. To obtain the final result, a classification is then applied to the samples. The main reason for doing this is to get signals with high-frequency characteristics.

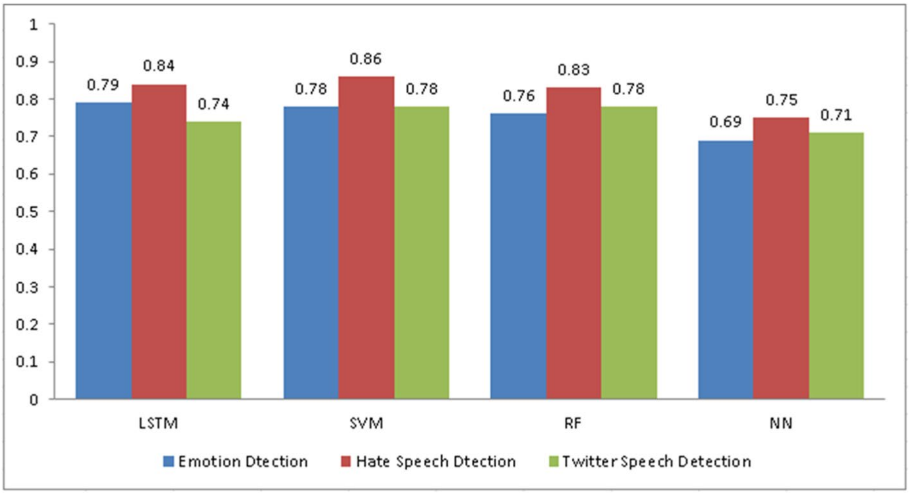

Fig. 5 Accuracy of most suitable ML algorithms application in hate speech detection

Table 5 Advanced ML application for hate speech detection

References	Year	Source	Application	Module/Algorithm
[105]	2017	Proceedings of the First Italian Conference on Cybersecurity	Hate speech detection	SVM and LSTM
[106]	2017	Computational Intelligence and Neuroscience	Recognizing Emotions from Speech Signals	Random deep belief networks
[107]	2018	LREC	Improving Hate Speech Detection	NN
[109]	2018	Computation and Language	Detecting Hate Speech and Offensive Language	N-gram and TFIDF
[12]	2019	ICAEM	Hate Speech Detection	CNN
[108]	2019	Computation and Language	Hate Speech Detection	CBOW and RoBERTa
[109]	2020	Computing-Springer	Automatic hate speech detection	KNLPEDNN
[110]	2020	Advances in Intelligent Systems and Computing	Arabic Hate Speech Detection in OSNs	RNN
[111]	2020	12th ACM Conference on Web Science	Hate Speech Detection via Multi-Faceted Text Representations	DeepHate
[112]	2020	ACM Transactions on Internet Technology	Multilingual evaluation	LSTM, GRU
[113]	2021	SAGE Journal	Intelligent detection of hate speech	SCM, DT, TFIDF
[114]	2021	Journal of Intelligent Systems	Bangla hate speech detection	Attention-based encoder–decoder model
[115]	2021	5th ICCMC	Toxic Speech Detection	Bert and fastText
[116]	2021	SEPLN 2021	Sexism Identification	RF

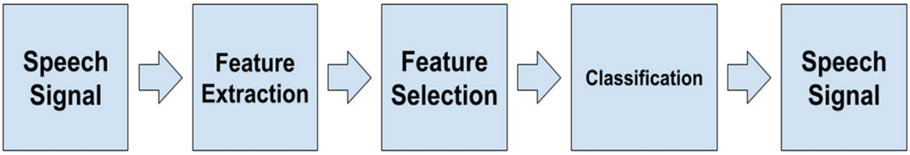

Fig. 6 Emotion recognition system [157]

6.2 Standard SER systems

Traditional SER Systems follow the steps of speech normalization, feature extraction, feature selection, and classification, as indicated in Fig. 7 which illustrates the fundamental process for identifying emotions in incoming speech. Following the separation of the noise components, feature extraction and selection are carried out in the process of normalizing speech [111]. The first step in the analysis of speech signals for emotion detection is the extraction and counting of speech features. The majority of the time, a time- and frequency-domain analysis of the spoken data produces the speech features [112]. The creation of a database of speech features generated from input voice signals follows. The classifiers can identify emotions in the final stage. In order to recognize emotions, classifiers use a variety of pattern-matching algorithms [113].

6.3 Speech normalization

Speech normalization is the process of the emotional data that is recorded and is usually diminished by outside noise (like the "hiss" of the recording device). This change will lead to inaccurate feature extraction and categorization. Therefore, normalization is an important step in the identification of emotions [114]. With the preservation of emotional distinction, this pre-processing stage gets rid of speaker and recording fluctuation. The two most popular methods of normalization are energy normalization and pitch normalization [115].

6.4 Selection and extraction of emotional speech features

After being normalized, the emotional speech signal is divided into segments and then decomposed into meaningful units. These components often express the speaker's emotional state through speech signals [116]. The following step is the feature set extraction

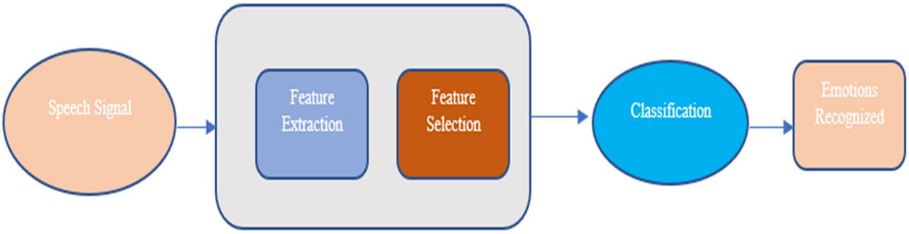

Fig. 7 Standard SER

as shown in Table 6. These emotional speech characteristics can be categorized in a variety of ways.

Two categories can be used to separate long-term and short-term traits. Examples of short-term qualities are formants, pitch, and energy because they only last for a small period. Long-term characteristics are a statistical tool for examining a digital audio stream. The Mean and Standard Deviation are two of the most often applied long-term measures. If more features are employed, the categorization process will be more accurate [117].

6.5 Training data

Numerous databases have been created by the voice-processing community [118]. The databases contain training and test data sets. There is an English version of the emotional prosody speech and documents database from 2002. Three different types of databases are used by SER, and the review in Table 6 shows examples of some researche works which works on the training datasets in order to extraction emotions from the speech. The Table 6 also discuss the type of databases they used and what kind of emotions they were abled to figures out form the the available sources of database (Table 7).

Type 1 emotional speech includes personal labels. Acting out or simulating speech is done professionally. To obtain these, actors are requested to speak with a specific emotion, such as DES or EMO-DB [122]. Realistic, human-like expressive speech is type 2. Natural speech is simply unplanned speech that conveys an individual's actual feelings. These databases are based on actual applications from the real world, like contact centers. Instead of labeling, the speaker employs self-reporting to elicit feelings in Type 3 to manage to label. Emotional speech is prompted by type 3. Expressed long short-term memory speech is not fictional or neutral [119].

6.6 Classifiers for emotion recognition

Only a few systems have had classifiers explored in the literature: SER, voice recognition, and speaker recognition [120]. In contrast, most implementations hardly ever explain why a particular classifier was chosen for a specific speech task. Selecting classifiers often involves either a general rule or empirical analyses of some indicators.

Table 6 Characteristics of emotions

Emotion	Anger	Happy	Surprise	Fear	Inquiry
Pitch range	High	Very high	High	High	High
Mean	High	Incline	Medium	Very High	Very high
Variance	Moderate	High	More	Medium	Incline
Rate of speaking	More	High	Medium	High	Moderate
Contour	Very high	High	Medium	High	Very High

Table 7 Types of databases with emotions

Types of databases	Example	Emotions	Reference
Simulated Database (Berlin)	LDC speech corpus	Happiness, sadness, disgust	[12], Emo-DB (Burkhardt et al. 2005), IITKGP-SESC (Koolagudi et al. 2009)
Induced Database (Danish)	Wizard of OZ databases	Anger, sadness, surprise	(McMahon et al. 2003)
Natural Database	Call center conversations	Happiness, anger, sadness	[142]

6.7 MFCC

The MFCCs represent some aspects of human speech perception and production. For instance, MFCC displays the logarithmic volume and pitch perception of the human auditory system [121]. The MFCC cepstral coefficients are produced using a twist frequency scale centered on human auditory perception. By using windowing, the voice signal is first divided into frames before being subjected to MFCC computation [122]. Since their amplitude is smaller than that of low-frequency formants, high-frequency formants are highlighted. This guarantees that the amplitude of each formant is the same. After windowing, Fast Fourier Transform is used to get the power spectrum from each frame [123]. After that, filter banks are processed on the power spectrum using mel-scale. After the power spectrum has been transformed into the logarithmic domain, the speech signal is subjected to the function to derive the MFCC coefficients [124].

6.8 Feature extraction and feature set classification

A crucial step in emotion identification is selecting and extracting relevant characteristics. The overall performance of the system. They can be classified into two primary groups: Spectral Features and Prosodic Features. The features are selected using a variety of techniques in order to be processed. Using LPA[125], PLPCS, PLP, FT, RASTA, MFCC[126], and FFT to extract emotions such as burden, anxiety, surprise, natural, grief, and happiness, the CASIA and EMODB dataset has an average recognition rate of 87.5% [127].

6.9 The ML and DL methods for researching emotions

Speech Emotion Recognition is a field of study that seeks to infer the speaker's emotional state from speech data. Progress in emotion identification, according to various surveys, would simplify many systems. Consequently, raising the standard of living [123]. We'll go over SER's applications in more detail in the section that follows. For instance, it is not possible to reliably deduce an emotion from the surroundings, culture, a person's facial expression, or speech corpus. One of the last significant challenges that an operating system in the actual world must overcome is the knowledge of dealing with bilingual inputs [128].

Due to the ubiquity of mixed-language speech in everyday situations, cross-language recognition demands more performance experience. A survey was carried out to better understand speech emotion identification. The method of feature extraction is used to identify the most crucial components of a signal. The extracted feature vectors are mapped to the appropriate emotions in the final stage using classifiers. In-depth discussions of feature extraction, classification, and speech signal processing can be found in [129, 133]. The differences between spontaneous and performed speech are also looked at because they are important to the topic [130]. A noisy component is removed in the first stage of speech-based signal processing. The second stage is divided into two components: feature extraction and feature selection. The desired features are extracted from the preprocessed voice input and used to make a selection [94]. Speech is recorded via microphone sounds and utilized by the system. The sound card of a computer is then used to build a digital representation of the received sounds. Feature extraction and selection Out of the 300 different emotional states, the returned speech features are selected based on emotional relevance [131]. Figure 8 shows the increasing development of the SER topics in the scientific areas that makes it one of the most vulnerable topic areas.

The main objective of speech emotion recognition systems is classification. It is challenging to classify emotions because the average set of emotions includes more than 300 different emotional states. Since some of the most frequent human emotions are fear, surprise, fury, joy, contempt, and sadness, the naturalness of a speech emotion recognition system is what is evaluated [132]. ML algorithms can be used to recognize emotions in speech. It has been done using a variety of methods, including RF, SVM [133], GMMs, HMMs, CNNs, KNN, and MLP. In the past, these algorithms have been routinely used to identify emotions.

6.10 Emotions and database type

The two basic methods for categorizing emotions are the dimensional approach and the use of categories. The category method breaks down emotions into more manageable categories. The six main emotions are anger, joy, happiness, sadness, fear, surprise, and disgust [134]. There are two categories for emotions, the second of which is axis-based and has several dimensions [135]. Tao found 89.6% of the emotions in the CASIA Chinese emotion corpus using a decision tree diagram. In the work of [136], a GMM was employed as a classifier to categorize emotion-founded MFCCs. Several Berlin emotional datasets were identified using a three-stage classical SVM. In order to categorize emotions in the Marathi voice dataset, MFCCs extracted the features from the Berlin EmoDB database [137]. To determine the emotional content of a person's speech, the KNN algorithm was applied. The Berlin emotive speech database operated flawlessly 90 to 99.5 percent of the time. Hossain and Shamim presented cooperative media systems in 2014 that make use of MFCCs and standard characteristics like emotions from voice signals. To identify emotions in speech, Alonso et al. exploited paralinguistic and prosodic characteristics. They used SVM, a radial basis function neural network, and an auto-associative neural network after integrating two characteristics from a music library, the residual phase and MFCCs [138]. Researchers used a database of scholarly publications from China to investigate SVMs and

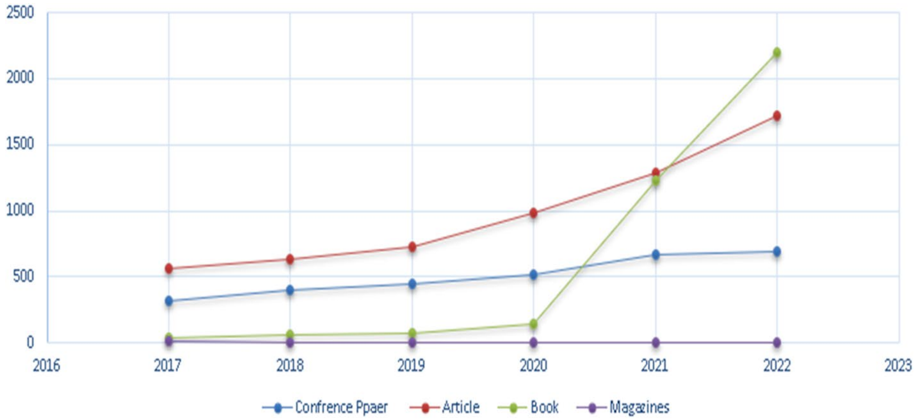

Fig. 8 Recent development of SER using ML and DL techniques

Table 8 ML comparison of speech features

Feature/Characteristics	Purpose of approach
MFCC	Insight between music and spoken words. It's more common for lower-order MFCCs to include more speech particular data than it is for higher-order MFCCs to contain music-specific data
MFCCs, coefficients, power coefficients, log frequency	The four-emotion classification system in Mandarin Chinese. This study takes into account feelings of anger, happiness, neutrality, and sadness
An amalgamation of MFCCs and minimally useful characteristics	Emotional speech databases in Swedish and English were used to classify emotions
Vowels stressed and unstressed are included in the MFCC (class-level MFCCs)	Emo-DB and LDC datasets are used to classify emotions in English
FFT and Chirp modifications were used to extract spectral information from the data	The emotional states of people under pressure are being modeled

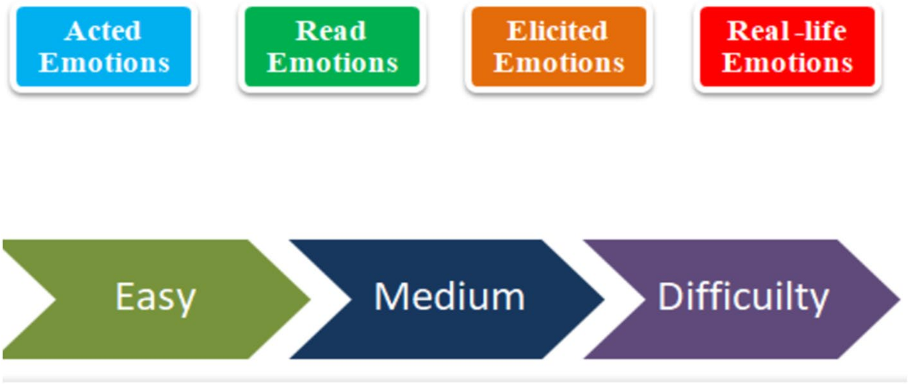

Fig. 9 Emotions and their complexities [159]

DBNs. DBNs had an accuracy of 94.5% whereas SVMs was about 85% accurate. High-order statistical traits and characteristics based on particle swarm optimization were used in this work. Following the extraction of spectral information from voice recordings, [139] categorized speech emotions using an HMM and SVM. Performance analysis for different languages uses a variety of ML techniques. The comparison shows that different ML methods have been used to identify speech emotions for several languages. In light of this, the best-case accuracy for the best-case scenario has been determined. Although emotions vary, the research selects the most accurate example utilizing a variety of feature extraction techniques and ML techniques [140].

In a range of research projects, several academics use emotional speech databases. The performance and quality of the databases employed are the most important factors in assessing emotion recognition systems [141]. Depending on why speech systems are being developed, different data collection techniques and objectives may be used. Table 8 provides a summary of several publicly accessible datasets of emotional speech [142]. The creation of emotional speech systems uses three different types of speech databases. A continuum that can be used to illustrate database classification is shown in Fig. 9. The intricacy of several emotion recognition databases is represented in the image.

Actors with a high level of training and experience recorded the voice information in these databases [143]. From any of the previous databases, this is the simplest way to get a speech-based dataset of various moods. It is estimated that this method is used to collect around 60% of speech datasets.

Due to the fact that they collect emotional data by creating an artificial emotional state, these databases are also known as induced databases [144]. The speaker or performer is unaware that this is taking place. Compared to actor-based databases, this database is more naturalistic. There can be an ethical issue because the speaker should be aware that they are being filmed for research [145]. Natural databases are the most realistic because they are the hardest to recognize, but they are also the hardest to obtain. Typically, emotional speech databases are compiled from conversations in contact centers, the general public, and other sources [146]. Emotion Recognition is used by contact centers to classify incoming calls according to their emotional content. Emotion Recognition as a performance criterion for conversational analysis can be used to determine satisfied and dissatisfied clients [147]. The SER in-car board system can intervene to keep the driver safe and prevent accidents when it recognizes the driver's mental state.

The performance evaluation provided in [148] investigates a variety of speeches' acoustic properties and classifier algorithms, which helps explore modern ways of emotion recognition. The design of DL makes it possible to utilize it for modalities other than NLP, like SER and voice recognition. It is possible to use the RNN for natural language phrase classification and natural image processing [149]. As a final point, DL is replacing traditional SER techniques as the favored approach. Unsupervised and multimodal SER, as well as NLP and speech recognition, are all on the rise [150]. It is effective to identify emotions while simultaneously employing both aural and visual information. This trait to incorporate is a crucial decision in the development of any vocal system [151]. The features chosen should represent the information being delivered through them. The representation of speech information by various speech components, speaker, emotion, speech, etc. substantially overlaps [152]. The speech characteristics comparison is shown in the table below. As a result, many features in speech research are chosen experimentally, while others are picked with the use of Principal Component Analysis [153]. The ML comparison of the speech features for various techniques is shown in Table 8.

The impact of emotional expression is also influenced by the speech's linguistic substance [154]. To increase the precision of emotion recognition, emotional speech can be utilized to identify prominent words and traits that can be recovered from these words, in addition to more conventional aspects [155]. A real-time application where it is crucial to authenticate requests is call monitoring in the ambulance and fire brigade. Under the umbrella of emotion verification, pertinent aspects and models may be researched in this regard [156]. Figure 10 shows the most used keywords for the emotion extraction. In this research, LSTM, MFCC, RNN, CNN are found to be the most useful keywords.

7 Conclusion and future work

A new taxonomy was introduced, and the main ML techniques for hate speech identification were illustrated. According to the study, among the various DL techniques, RF, CNN, SVM, and LSTM had the greatest practical uses. These algorithms work well for

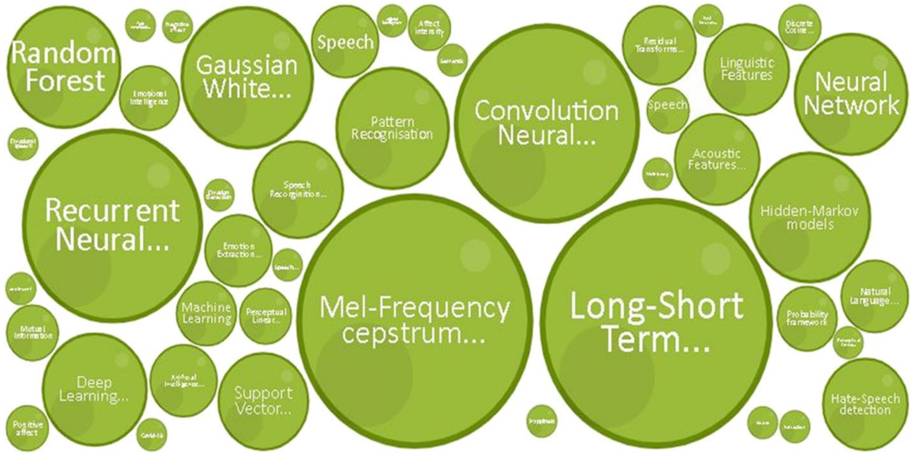

Fig. 10 Most used keywords for the emotion extraction

sentimental and emotional analysis as well as the detection of hate speech. Incontrovertible analytical data from a range of sources, such as common documents, business reports, social media monitoring, and customer support tickets, are provided by emotional analysis. On the other hand, DL enables the employment of more potent tools and algorithms for data analysis. The classifier and database to use to accurately assess emotions can be chosen using the data presented in the earlier articles. The most often searched-for emotions are neutral, disgust, happiness, and sadness, along with other characteristics like a burden, joy, surprise, and fear. The classifier used has an impact on the extraction rate. Due to the drawbacks of using subpar sample recordings in the databases, the accuracy in DBN networks is between 56 and 57 percent, and the recognition ratio has decreased. In this paper, both deep learning and machine learning for SER have been carefully analyzed.The paper includes a block diagram of the voice emotion detection system and a brief introduction to SER. To classify a speech recognition system, they must be able to distinguish between isolated, connected, spontaneous, and continuous words. There are several different techniques to research and assess approaches for recognizing emotions, including Emotion Recognition, DL, and ML. Researchers have recently paid a lot of attention to the topic of speech-based emotion recognition. This study examined a huge number of research publications using databases, feature extraction, and classifiers. The research on emotion recognition systems conducted between 1994 and 2022 is summarized in this document. In order to improve performance accuracy, current research has placed a lot of emphasis on the extraction of features and feature selection. Data analysis shows that classifier selection is a challenging task to enhance system performance and recognize the proper emotions. No obvious winner has emerged despite the selection of several classifiers for the speech emotion identification system.

This work gives an in-depth analysis of all the properties, databases, classifiers, and methods utilized to address the complicated challenge of SER Inferring that SVM performs better than the other model across all studies is possible. In the future, sentiment analysis may be used to identify emotions through facial expressions and emotions. We hope to be able to recognize offensive speech in the future from a variety of monitoring data. We want to consider visual information in addition to comment text to distinguish between dangerous emotions. Online texts can also be handled using the adaptive bagging approach, which enriches the processing at the level of dynamic processing by processing the texts as streams. Future research on this subject might examine how to improve the performance of our model by using BERT to build the embedding layer.

Funding Open access funding provided by Óbuda University.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interests The authors declare that they have no conflict of interest.

References

1. Garcia-Garcia JM, Penichet VM, Lozano MD (2017) Emotion detection: a technology review. In: Proceedings of the XVIII international conference on human computer interaction, pp 1–8
2. Todd B, Tucker C, Hopkinson K, Bilén SG (2014) Increasing the veracity of event detection on social media networks through user trust modeling. In: IEEE international conference on big data, pp 636–643
3. Sun S, Luo C, Chen J (2017) A review of natural language processing techniques for opinion mining systems. Inf Fusion 36:10–25
4. Rajput K, Kapoor R, Mathur P, Kumaraguru P, Shah RR (2020) Transfer learning for detecting hateful sentiments in code switched language. Deep learning-based approaches for sentiment analysis, pp 159–192
5. Zhu X, Lou Y, Deng H, Ji D (2022) Leveraging bilingual-view parallel translation for code-switched emotion detection with adversarial dual-channel encoder. Knowl-Based Syst 235:107436
6. Zhang L, Wang S, Liu B (2018) Deep learning for sentiment analysis: a survey. WIREs Data Mining Knowl Discov 8(4):e1253
7. Mohanty AK (2018) The multilingual reality: living with languages. Multiling Matters
8. Cummins J (1979) Linguistic interdependence and the educational development of bilingual children. Rev Educ Res 49(2):222–251
9. Li Y (2021) Dual-attention generative adversarial network and flame and smoke analysis. PhD diss. Université d'Ottawa/University of Ottawa
10. Zhang S, Zheng D, Hu X, Yang M (2015) Bidirectional long short-term memory networks for relation classification. In: Proceedings of the 29th Pacific Asia conference on language, information and computation, pp 73–78
11. Lin SY, Kung YC, Leu FY (2022) Predictive intelligence in harmful news identification by BERT-based ensemble learning model with text sentiment analysis. Inf Process Manage 59(2):102872
12. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Commun 48(9):1162–1181
13. Zeighami M, Mangolian Shahrbabaki P, Dehghan M (2023) Iranian nurses' experiences with sexual harassment in workplace a qualitative study. Sex Res Social Policy 20(2):575–588

14. Chen M et al (2022) Negative information measurement at AI edge: a new perspective for mental health monitoring. ACM Trans Internet Technol (TOIT) 22(3):1–16
15. Araque O, Iglesias CA (2022) An ensemble method for radicalization and hate speech detection online empowered by sentic computing. Cogn Comput 14:48–61
16. Khan S et al (2022) HCovBi-caps: hate speech detection using convolutional and bi-directional gated recurrent unit with capsule network. IEEE Access 10:7881–7894
17. Roy PK, Bhawal S, Subalalitha CN (2022) Hate speech and offensive language detection in Dravidian languages using deep ensemble framework. Comput Speech Lang 75:101386
18. Mutanga RT, Naicker N, Olugbara OO (2022) Detecting hate speech on twitter network using ensemble machine learning. Int J Adv Comput Sci Appl 13(3)
19. Rani S, Bashir AK, Alhudhaif A, Koundal D, Gündüz ES (2022) An efficient CNN-LSTM model for sentiment detection in #BlackLivesMatter. Expert Syst Appl:116256
20. Toraman C, Şahinuç F, Yilmaz EH (2022) Large-scale hate speech detection with cross-domain transfer. arXiv preprint:220301111
21. Singh C, Imam T, Wibowo S, Grandhi S (2022) A deep learning approach for sentiment analysis of COVID-19 reviews. Appl Sci 12(8):3709
22. Javed N, Muralidhara BL (2021) Emotions during Covid-19: LSTM models for emotion detection in tweets. In: Proceedings of the 2nd international conference on recent trends in machine learning, IoT, smart cities and applications: ICMISC, vol 2022, pp 133–148
23. Lee E, Rustam F, Washington PB, El Barakaz F, Aljedaani W, Ashraf I (2022) Racism detection by analyzing differential opinions through sentiment analysis of tweets using stacked ensemble GCR-NN model. IEEE Access 10:9717–9728
24. Sai S, Srivastava ND, Sharma Y (2022) Explorative application of fusion techniques for multimodal hate speech detection. SN Comput Sci 3(2):1–13
25. Alnuaim AA, Zakariah M, Alhadlaq A, Shashidhar C, Hatamleh WA, Tarazi H, Shukla PK, Ratna R (2022) Human-computer interaction with detection of speaker emotions using convolution neural networks. Comput Intell Neurosci
26. Dybala MPP, Masui TMF, Rzepka R, Araki K (2010) Machine learning and affect analysis against cyber-bullying. In: Proceedings of the linguistic and cognitive approaches to dialog agents symposium
27. Bee N, Wagner J, André E, Vogt T, Charles F, Pizzi D, Cavazza M (2010) Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application. In: International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction, pp 1–8
28. Mahdhaoui A, Chetouani M (2011) Supervised and semi-supervised infant-directed speech classification for parent-infant interaction analysis. Speech Commun 53(9–10):1149–1161
29. Petsatodis T, Boukis C, Talantzis F, Tan ZH, Prasad R (2011) Convex combination of multiple statistical models with application to VAD. IEEE Trans Audio Speech Lang Process 19(8):2314–2327
30. Hansen JH, Sangwan A, Kim W (2012) Speech under stress and Lombard effect: impact and solutions for forensic speaker recognition. In: Forensic speaker recognition: law enforcement and counter-terrorism, pp 103–123
31. Joshi J, Goecke R, Alghowinem S, Dhall A, Wagner M, Epps J, Parker G, Breakspear M (2013) Multimodal assistive technologies for depression diagnosis and monitoring. J Multimodal User Interfaces 7(3):217–228
32. Sztahó D, Vicsi K (2014) Speech activity detection and automatic prosodic processing unit segmentation for emotion recognition. Intell Decis Technol 8(4):315–324
33. Rudramurthy MS, Kamakshi Prasad V, Kumaraswamy R (2014) Speaker verification under degraded conditions using empirical mode decomposition based voice activity detection algorithm. J Intell Syst 23(4):359–378
34. Gitari ND, Zhang Z, Damien H, Long J (2015) A lexicon-based approach for hate speech detection. Int J Multimed Ubiquitous Eng 10(4):215–230
35. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J (2015) Analysis of acoustic space variability in speech affected by depression. Speech Commun 75:27–49
36. Dubey H, Mehl MR, Mankodiya K (2016) Bigear: inferring the ambient and emotional correlates from smartphone-based acoustic big data. In: 2016 IEEE first international conference on connected health: applications, systems and engineering technologies (CHASE), pp 78–83
37. Yan L, Tao H (2016) An empirical study and comparison for tweet sentiment analysis. In: International conference on cloud computing and security, pp 623–632

38. Del Vigna F et al (2017) Hate me, hate me not: hate speech detection on Facebook. In: Proceedings of the first Italian conference on cybersecurity, pp 86–95
39. 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
40. Zimmerman S, Kruschwitz U, Fox C (2018) Improving hate speech detection with deep learning ensembles. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)
41. Gaydhani A, Doma V, Kendre S, Bhagwat L (2018) Detecting hate speech and offensive language on twitter using machine learning: an n-gram and tfidf based approach. arXiv preprint:1809.08651
42. Sajjad M, Zulifqar F, Khan MUG, Azeem M (2019) Hate speech detection using fusion approach. In: International conference on applied and engineering mathematics (ICAEM):251–255
43. Nguyen TB, Nguyen QM, Nguyen TH, Pham NP, Nguyen TC, Do QT (2019) VAIS hate speech detection system: a deep learning based approach for system combination. arXiv preprint:1910.05608
44. Al-Makhadmeh Z, Tolba A (2020) Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. Computing 102(2):501–522
45. Cao R, Ka-Wei Lee R, Hoang T-A (2020) DeepHate: hate speech detection via multi-faceted text representations. In: 12th ACM conference on web science, pp 11–20
46. Corazza M, Menini S, Cabrio E, Tonelli S, Villata S (2020) A multilingual evaluation for online hate speech detection. ACM Trans Internet Technol 20(2):1–22
47. Aljarah I, Habib M, Hijazi N, Faris H, Qaddoura R, Hammo B, Abushariah M, Alfawareh M (2021) Intelligent detection of hate speech in Arabic social network: a machine learning approach. J Inf Sci 47(4):483–501
48. Das AK, Asif AA, Paul A, Hossain MN (2021) Bangla hate speech detection on social media using attention-based recurrent neural network. J Intell Syst 30(1):578–591
49. Malik P, Aggrawal A, Vishwakarma DK (2021) Toxic speech detection using traditional machine learning models and BERT and fastText embedding with deep neural networks. In: 5th international conference on computing methodologies and communication (ICCMC), pp 1254–1259
50. Butt S, Ashraf N, Sidorov G, Gelbukh A (2021) Sexism identification using BERT and data augmentation-EXIST2021. In: International conference of the Spanish Society for Natural Language Processing SEPLN, pp 381–389
51. Chen M, Zhou P, Fortino G (2016) Emotion communication system. IEEE. Access 5:326–337
52. Tolksdorf NF, Siebert S, Zorn I, Horwath I, Rohlfing KJ (2021) Ethical considerations of applying robots in kindergarten settings: towards an approach from a macroperspective. Int J Soc Robot 13(2):129–140
53. Dong B, Shi Q, Yang Y, Wen F, Zhang Z, Lee C (2021) Technology evolution from self-powered sensors to AIoT enabled smart homes. Nano Energy 79:105414
54. Gnanamanickam J, Natarajan Y, Sri Preethaa KR (2021) A hybrid speech enhancement algorithm for voice assistance application. Sensors 21(21):7025
55. Riya KS, PoornaPushkala K (2022) A healthcare system for detecting stress from ECG signals and improving the human emotional. In: International conference on advanced computing technologies and applications (ICACTA), pp 10–18
56. Stoll S, Camgoz NC, Hadfield S, Bowden R (2020) Text2Sign: towards sign language production using neural machine translation and generative adversarial networks. Int J Comput Vis 128(4):891–908
57. Boyd RL, Schwartz HA (2021) Natural language analysis and the psychology of verbal behavior: the past, present, and future states of the field. J Lang Soc Psychol 40(1):21–41
58. Zhang Y, Yan G, Chang W, Huang W, Yuan Y (2023) EEG-based multi-frequency band functional connectivity analysis and the application of spatio-temporal features in emotion recognition. Biomed Signal Process Control 79:104157
59. Anagnostou M et al (2022) Characteristics and challenges in the industries towards responsible AI: a systematic literature review. Ethics Inf Technol 24(3):1–18
60. Zhu Q, Wang Z, Dou Y, Zhou J (2022) Whispered speech conversion based on the inversion of mel frequency cepstral coefficient features. Algorithms 15(2):68
61. Patel N, Patel S, Mankad SH (2022) Impact of autoencoder based compact representation on emotion detection from audio. J Ambient Intell Humaniz Comput 13(2):867–885
62. Basel D, Aviram T, Lazarov A (2022) Lack of an attention bias away from relatively negative faces in dysphoria is not related to biased emotion identification. Behav Ther 53(2):182–195

63. Nassif AB, Shahin I, Lataifeh M, Elnagar A, Nemmour N (2022) Empirical comparison between deep and classical classifiers for speaker verification in emotional talking environments. Information 13(10):456
64. Ozdamli F, Aljarrah A, Karagozlu D, Ababneh M (2022) Facial recognition system to detect student emotions and cheating in distance learning. Sustainability 14(20):13230
65. Mariz JLV, Soofastaei A (2022) Advanced analytics for rock blasting and explosives engineering in mining. Adv Anal Min Eng:363–477
66. Jahangir R, Teh YW, Mujtaba G, Alroobaea R, Shaikh ZH, Ali I (2022) Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion. Mach Vis Appl 33(3):1–16
67. Kaur AP, Singh A, Sachdeva R, Kukreja V (2023) Automatic speech recognition systems: a survey of discriminative techniques. Multimed Tools Appl 82(9):13307–13339
68. Wang Y et al (2022) A systematic review on affective computing: emotion models, databases, and recent advances. Inform Fusion 83:19–52
69. Selvan AK, Nimmi K, Janet B, Sivakumaran N (2022) Emotion detection on phone calls during emergency using ensemble model with hyper parameter tuning. Int J Inform Technol 15(2):1–13
70. Matveev Y, Matveev A, Frolova O, Lyakso E, Ruban N (2022) Automatic speech emotion recognition of younger school age children. Mathematics 10(14):2373
71. Kumar Shashi GS, Arun A, Sampathila N, Vinoth R (2022) Machine learning models for classification of human emotions using multivariate brain signals. Computers 11(10):152
72. Al-onazi BB, Nauman MA, Jahangir R, Malik MM, Alkhammash EH, Elshewey AM (2022) Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion. Appl Sci 12(18):9188
73. Kapoor S, Kumar T (2022) A novel approach to detect instant emotion change through spectral variation in single frequency filtering spectrogram of each pitch cycle. Multimed Tools Appl 82(6):9413–9429
74. Azman NF, Zamri NAK (2022) Conscious or unconscious: the intention of hate speech in cyberworld—a conceptual paper. Proceedings 82(1):29
75. Jougleux P (2022) Hate speech, fake news, and the moderation problem. In: Facebook and the (EU) law. Springer, pp 183–212
76. Al Dabel M (2022) Speech attribute detection to recognize arabic broadcast speech in industrial networks. Mobile Inform Syst 2022
77. Kumar K, Pande BP (2022) Applications of machine learning techniques in the realm of cybersecurity. Cyber Secur Digit Forensics:295–315
78. Rodrigues AP, Fernandes R, Shetty A, Lakshmanna K, Shafi RM (2022, 2022) Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. Comput Intell Neurosci
79. Chiriacescu (2009) Distributed model predictive control of irrigation canals. Netw Heterog Media 4(2):359–380
80. Emerich S, Lupu E, Rusu C (2009) A new set of features for a bimodal system based on on-line signature and speech. Digit Signal Process 23(3):928–940
81. Deng L (2014) Deep learning methods and applications. Found Trends® Signal Process 7(3–4):197–387
82. Patel P, Chaudhari A, Pund M, Deshmukh D (2017) Speech emotion recognition system using gaussian mixture model and improvement proposed via boosted GMM. IRA-International J Technol Eng 7(2):56
83. Fayek H, Lech M, Cavedon L (2017) Evaluating deep learning architectures for speech emotion recognition. Neural Netw 92:60–68
84. Huang C, Gong W, Fu W, Feng D (2014) A research of speech emotion recognition based on deep belief network and SVM. Math Probl Eng 2014:1–7
85. Chavhan Y, Dhore M, Yesaware P (2010) Speech emotion recognition using support vector machine. Int J Comput Appl 1(20):8–11
86. Jin Q, Wu H, Li C, Chen S (2007) 2008 IEEE international conference on acoustics, speech, and signal processing (ICASSP). IEEE Transact Audio, Speech, Lang Process 15(5):1737–1737
87. Song T, Zheng W, Song P, Cui Z (2020) EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Transact Affect Comput 11(3):532–541
88. Koolagudi SG, Krothapalli SR (2012) Emotion recognition from speech using sub-syllabic and pitch synchronous spectral features. Int J Speech Technol 15(4):495–511

89. Rong J, Li G, Chen YPP (2009) Acoustic feature selection for automatic emotion recognition from speech. Inf Process Manag 45(3):315–328
90. Noroozi F, Akrami N, Anbarjafari G (2017) Speech-based emotion recognition and next reaction prediction. In: 25th signal processing and communications applications conference (SIU), pp 1–4
91. Graves A, Mohamed A, Hinton G (2013) Speech recognition with deep recurrent neural networks. IEEE International Conference on Acoustics, Speech and Signal Processing, pp 6645–6649
92. Huang CW, Narayanan SS (2017) Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition. arXiv preprint:1706.0290
93. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
94. Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117
95. Ngiam J, Khosla A, Kim M, Nam J, Lee H, Ng AY (2011) Multimodal deep learning, proceedings of the 28th international conference on machine learning, pp 689–696
96. Yadav J, Kumari A, Rao KS (2015) Emotion recognition using LP residual at sub-segmental, segmental and supra-segmental levels. In: International conference on communication, Information & Computing Technology, pp 1–6
97. Latif S, Rana R, Qadir J (2018) Adversarial machine learning and speech emotion recognition: utilizing generative adversarial networks for robustness. arXiv preprint:1811.11402
98. Bernard M, Poli M, Karadayi J, Dupoux E (2021) Shennong: a Python toolbox for audio speech features extraction. arXiv preprint:2112.05555
99. Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116:56–76
100. Yanase J, Triantaphyllou E (2019) A systematic survey of computer-aided diagnosis in medicine: past and present developments. Expert Syst Appl 138:112821
101. Vázquez-Romero A, Gallardo-Antolín A (2020) Automatic detection of depression in speech using ensemble convolutional neural networks. Entropy 22(6):688
102. Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E (2021) A comprehensive review of speech emotion recognition systems. IEEE Access 9:47795–47814
103. Higgins JP, Li T, Deeks JJ (2019) Choosing effect measures and computing estimates of effect. Cochrane handbook for systematic reviews of interventions:143–176
104. Mohammed MA et al (2020) Voice pathology detection and classification using convolutional neural network model. Appl Sci 10(11):3723
105. Alnuaim AA et al (2022) Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier. J Healthc Eng 2022
106. Fu L, Mao X, Chen L (2008) Speaker independent emotion recognition based on SVM/HMMs fusion system. IntConf Audio, Lang Image Process Proc, pp 61–65
107. Wang K, An N, Li BN, Zhang Y, Li L (2015) Speech emotion recognition using fourier parameters. IEEE Trans Affect Comput 6(1):69–75
108. Batliner A, Schuller B, Seppi D, Steidl S, Devillers L, Vidrascu L, Vogt T, Aharonson V, Amir N (2011) The automatic recognition of emotions in speech. Emotion-Oriented Systems:71–99
109. Mower E, Mataric MJ, Narayanan S (2011) A framework for automatic human emotion classification using emotion profiles. IEEE Trans Audio Speech Lang Process 19(5):1057–1070
110. Han J, Zhang Z, Ringeval F, Schuller B (2017) Prediction-based learning for continuous emotion recognition in speech. IEEE international conference on acoustics, speech and signal processing (ICASSP):5005–5009
111. Jokinen E, Takanen M, Vainio M, Alku P (2014) An adaptive post-filtering method producing an artificial Lombard-like effect for intelligibility enhancement of narrowband telephone speech. Comput Speech Lang 28(2):619–628
112. Sezgin M, Gunsel B (2012) Kurt G (2012) Perceptual audio features for emotion detection. EURASIP J Audio, Speech, Music Process 1:1–21
113. Ekman P (1992) An argument for basic emotions. Cognit Emot 6(3–4):169–200
114. Costanzi M et al (2019) The effect of emotional valence and arousal on visuo-spatial working memory: incidental emotional learning and memory for object-location. Front Psychol 10:2587
115. Kandali A, Routray A, Basu T (2009) Vocal emotion recognition in five native languages of Assam using new wavelet features. Int J Speech Technol 12(1):1–13
116. Demircan S, Kahramanli H (2014) Feature extraction from speech data for emotion recognition. J Adv Comput Netw 2(1):28–30

117. Nalini N, Palanivel S (2016) Music emotion recognition: the combined evidence of MFCC and residual phase. Egypt Inform J 17(1):1–10
118. Chourasia M, Haral S, Bhatkar S, Kulkarni S (2021) Emotion recognition from speech signal using deep learning. Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020:471–481
119. Martin O, Kotsia I, Macq B, Pitas I (2006) The enterface'05 audio-visual emotion database. In: 22nd international conference on data engineering workshops (ICDEW'06):8-8
120. Jackson P, Haq S (2014) Surrey audio-visual expressed emotion (savee) database. University of Surrey, Guildford
121. Nevendra M, Singh P (2021) Software defect prediction using deep learning. Acta Polytech Hung 18(10):173–189
122. Lutsiv N, Maksymyuk T, Beshley M, Lavriv O, Vokorokos L, Gazda J (2022) Deep semisupervised learning-based network anomaly detection in heterogeneous information systems. CMC-Comput Mater Continua 70(1):413–431
123. Koolagudi S, Rao K (2012) Emotion recognition from speech: a review. Int J Speech Technol 15(2):99–117
124. Salovey P, Mayer D, Kokkonen JM, Lopes PN (2007) Feelings and emotions: the Amsterdam symposium. Cambridge University Press, pp 321–340
125. Cowie R, Douglas-Cowie E, Tsapatsoulis N, Votsis G, Kollias S, Fellenz W, Taylor JG (2001) Emotion recognition in human-computer interaction. IEEE Signal Process Mag 18(1):32–80
126. Kwon OW, Chan K, Hao J, Lee TW (2003) Emotion recognition by speech signals. Eighth European Conference on Speech Communication and Technology
127. Picard RW (1995) Affective computing. In: Perceptual computing section, media laboratory. Massachusetts Institute of Technology
128. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recogn 44(3):572–587
129. Dhawan M, Sharma D (2020) Hate speech detection and sentiment analysis. Int Res J Eng Technol 7(5):3905–3908
130. Deng L, Yu D et al (2014) Deep learning: methods and applications. Foundations and trends®. Signal Process 7(3–4):197–387
131. Hemanth Kumar H, Gowramma Y, Manjula S, Anil D, Smitha N (2021) Comparison of various ML and DL models for emotion recognition using twitter. In: Third international conference on intelligent communication technologies and virtual Mobile networks (ICICV), pp 1332–1337
132. Fu L, Mao X, Chen L (2008) Relative speech emotion recognition based artificial neural network. In: PACIIA'08 Pacific-Asia workshop on computational intelligence and industrial application, vol 2, pp 140–144
133. Stuhlsatz A, Meyer C, Eyben F, Zielke T, Meier G, Schuller B (2011) Deep neural networks for acoustic emotion recognition: raising the benchmarks, acoustics, speech and signal processing (ICASSP), pp 5688–5691
134. Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans Multimedia 16(8):2203–2213
135. Barros P, Weber C, Wermter S (2016) Learning auditory neural representations for emotion recognition. In: International joint conference on neural networks (IJCNN), pp 921–928
136. Mao Q, Xue W, Rao Q, Zhang F, Zhan Y (2016) Domain adaptation for speech emotion recognition by sharing priors between related source and target classes. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2608–2612
137. Zhang Y, Liu Y, Weninger F, Schuller B (2017) Multi-task deep neural network with shared hidden layers: breaking down the wall between emotion representations. In: Acoustics, speech and signal processing (ICASSP), pp 4990–4994
138. Zhang S, Zhang S, Huang T, Gao W (2017) Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Trans Multimedia 20(6):1576–1590
139. Ayadi ME, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recog 44:572–587
140. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: resources, features, and methods. Speech Commun 48(9):1162–1181
141. Koolagudi S, Rao K (2012) Emotion recognition from speech: a review. Int J Speech Technol 15(2):99–117
142. Lee CM, Narayanan S (2005) Toward detecting emotions in spoken dialogs. IEEE Trans Speech Audio Process 13(2):293–303

143. Pierre-Yves O (2003) The production and recognition of emotions in speech: features and algorithms. Int J Hum Comput Stud 59(1–2):157–183
144. Ajibola Alim S, KhairAlang Rashid N (2018) Some commonly used speech feature extraction algorithms, from natural to artificial intelligence - algorithms and applications. IntechOpen
145. Koolagudi SG, Krothapalli SR (2012) Em ti recognition from speech using sub-syllabic and pitch synchronous spectral features. Int J Speech Technol 15(4):495–511
146. Rong J, Li G, Chen YPP (2009) Acoustic feature selection for automatic emotion recognition from speech. Inf Process Manag 45(3):315–328
147. Gharavian D, Sheikhan M, Nazerieh A, Garoucy S (2011) Peech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Comput Appl 21(8):2115–2126
148. Chew LW, Seng KP, Ang L-M, Ramakonar V, Gnanasegaran A (2011) Audio-emotion recognition system using parallel classifiers and audio feature analyzer. In: Third international conference on computational intelligence, modelling & simulation, pp 210–215
149. Krishna Kishore KV, Krishna Satish P (2013) Emotion recognition in speech using MFCC and wavelet features. In: 3rd IEEE international advance computing conference (IACC), pp 842–847
150. Yixiong P, Shen P, Shen L (2012) Speech emotion recognition using support vector machine. Int J Smart Home 6(2):101–108
151. Ingale AB, Chaudhari DS (2012) Speech emotion recognition using hidden markov model and support vector machine. Int J Adv Eng Res Stud 1(3):316–318
152. Le D, Aldeneh Z, Provost EM (2017) Discretized continuous speech emotion recognition with multi-task deep recurrent neural network. Interspeech:1108–1112
153. Morgan N (2012) Deep and wide: multiple layers in automatic speech recognition. IEEE Trans Audio Speech Lang Process 20(1):7–13
154. Steidl S, Levit M, Batliner A, Noth E, Niemann H (2005) “Of all things the measure is man” automatic classification of emotions and interlabeler consistency. In: IEEE international conference on acoustics, speech, and signal processing (ICASSP05), pp 1–317
155. Vogt T (2010) Real-time automatic emotion recognition from speech, Dissertation
156. Lugovic S, Dunder I, Horvat M (2016) Techniques and applications of emotion recognition in speech. In: 39th international convention on information and communication technology, electronics and microelectronics (MIPRO), vol 2016. MIPRO, pp 1278–1283
157. Khalil RA, Jones E, Babar MI, Jan T, Zafar MH, Alhussain T (2019) Speech emotion recognition using deep learning techniques: a review. IEEE Access 7:117327–117345
158. Kamble VV, Deshmukh RR, Karwankar AR, Ratnaparkhe VR, Annadate SA (2015) Emotion recognition for instantaneous Marathi spoken words. In: Proceedings of the 3rd international conference on Frontiers of intelligent computing: theory and applications (FICTA), pp 335–346
159. Ganapathy A (2016) Speech emotion recognition using deep learning techniques. ABC J Adv Res 5(2):113–122