Artificial Intelligence in Speech Emotion Detection: Trends, Challenges, and Future Directions

Noor Alwan Malk, Sinan Adnan Diwan

Computer science and information technology, Wasit University, Al-Kut, Iraq

Abstract

Speech Emotion Detection (SED) has become a pivotal component in the development of emotionally aware artificial intelligence systems. This paper presents a comprehensive review of recent advancements in the field, focusing on signal processing techniques, machine learning and deep learning approaches, real-time implementation, and multimodal integration. The study highlights the critical role of feature extraction and classification methods in improving emotion recognition accuracy and robustness. Additionally, it discusses emerging trends such as personalization, explainable AI (XAI), and adaptation to noisy and culturally diverse environments. Ethical considerations and legal implications surrounding emotion-aware systems are examined, along with practical applications in healthcare, education, customer support, and entertainment. The review concludes by outlining the unresolved challenges and proposing future research directions to bridge existing gaps and enable more human-centric and trustworthy emotion recognition technologies.

Keywords

Speech Emotion Detection (SED), Deep Learning; Feature Extraction, Explainable AI, Multimodal Emotion Recognition, Human–Computer Interaction, Real-time Emotion Detection, Ethical AI

1. Introduction

Speech emotion detection is necessary to help machines develop an empathetic relationship with users. Humans incorporate emotional expressions into daily, verbal communication and it plays a crucial role in their interaction. Natural communication is only possible if a machine recognises the emotions of the user and hence the machine also needs to understand the emotional cues of the user and responds accordingly [1]. To recognize emotions, the machine is capable of using visual data or speech data as an input of which speech is the most direct mode of communication and is a natural choice for the task. Presently, only a limited number of speech databases exist in the public domain and this increases the difficulty in designing recognition frameworks that are capable of working satisfactorily across all types of speech data relating to different emotions. In the process of designing frameworks, various approaches to feature extraction can be adopted and experiments can be conducted to select the best approach for the task. A crucial part of the emotion recognition framework is the selection of a classifier and based on the properties of some commonly used machine learning classifiers, the strengths and weaknesses of each algorithm can be analysed for emotion classification. Research shows that building an integrated feature space as opposed to employing selective single features can result in much better recognition rates. Efforts toward improving the efficiency, accuracy, and ultimate usability of the system for practical applications constitute a future direction of research. The scope of emotion recognition systems can also be extended to other emotions such as depression and mood changes for example, and help therapists monitor mood swings in patients over time. An additional task that can be addressed alongside a general emotion recognition system is sarcasm detection. Sarcasm cannot be identified just by depending solely on the words or tone of the speaker, but can be addressed when combined with sentiment analysis. Several applications of speech-based emotion recognition can be expected in the near future.

2. Overview of Speech Emotion Detection

Continuous recognition of emotion from speech is essential for numerous applications, such as call centre analysis, intelligent virtual assistants, and e-learning. Speech is the natural medium for human communication; hence, speech-based emotion recognition plays an important role in human–computer interaction. The main objective of speech emotion recognition (SER) is to identify discriminative voice features that reflect emotional states. However, the task remains challenging because a reliable system should also be robust to speaking style and rate. Various other applications of SER include pain and lie detection, tutorial systems, and content-based recommendation systems. Consequently, a lot of research has been conducted in this area [2]. In SER, discrimination of emotions is based on two main approaches: classification models and regression models [3]. Classification models categorise emotions into different classes, such as neutral, happy, sad, and angry, whereas regression models produce continuous-valued outputs, such as valence–arousal vectors [4] or attractiveness values.

3. Historical Context

A variety of computational approaches have been developed to detect and interpret human emotions. In the past decade, emotion recognition has emerged as one of the most challenging research directions for the artificial-intelligence community [2]. Emotion detection approaches rely on different kinds of data, such as facial expressions, speech, or text. In particular, speech-emotion detection is crucial for improving the quality of interactions between humans and computers [5]. Humans have an extraordinary ability to recognize emotional cues embedded in the human voice, but machines are not yet able to perform a similar task. When asked, a machine produces answers that tend to generate mistrust and disappointment. AI applications using speech channels for interacting with humans, such as intelligent personal assistants, diagnostic applications, and scientific research, inevitably demand the development of fast and reliable speech-emotion recognition systems. Inspired by human-group decision making, a novel semi-supervised prediction model, based on consensus, was developed to tackle speech-emotion recognition. A series of tests demonstrated its effectiveness, and labeling performances were evaluated on a public database of spontaneous speeches. The extraction of an emotional state from speech is an important tool for enhancing human-computer interaction. If a machine can interpret speech to understand the speaker's emotional state, it can be used in complementary applications, such as pain detection, lie detection, and personalized recommendations.

4. Technological Foundations

The development of Artificial Intelligence (AI) in Speech Emotion Detection (SED) has been profoundly influenced by advances in speech enhancement techniques and breakthroughs in deep neural networks. The technology employs a learning-based approach that extracts descriptive information from the source speech signal. Standard speech signal features are combined with additional features used in speech classification tasks. Deep learning models are usually trained using raw signal frames and various acoustic, spectral, voice-quality, and prosodic features. Typically, the classification network has a single-stage structure that performs feature extraction and emotional categorization simultaneously [2].

In a complete SED system, an architecture for both feature extraction and classification is often integrated to improve framework performance, as the features are essential for inference and inference aids in feature learning. Most frameworks include four modules, which may be trained end-to-end or independently; these modules are speech production, pre-processing, feature extraction, and classification [3]. Analogous to the discrete Fourier transform of a signal, the phases and magnitudes of poles characterize a linear system. Based on the relationship between formants and emotion, system-theoretic approaches can thus be suitable for SED. The speech signal is modelled by an autoregressive moving average process. Poles are extracted to determine the emotion using a Support Vector Machine (SVM) with a Gaussian kernel.

4.1 Signal Processing Techniques

To enable automatic detection from voice recordings, raw speech signals must be preprocessed to remove irrelevant components, such as silence, background noise, and accents. Denoising algorithms are typically applied [2]. The speech signal is then segmented into fixed-length blocks, usually 16,000 samples for a 1-second duration at a 16 kHz sampling frequency [6]. Each block is subsequently analysed by a signal processing algorithm to extract a set of features that provide an informative representation of the speech segment. The application of frequency-domain analysis to these segments generally enhances information quality. The Mel-Frequency Cepstral Coefficients (MFCC) technique derives coefficients for each segment, specifying the distribution of the signals within the Mel frequency bands. The Mel scale is a perceptual scale of pitches perceived by human ears, and therefore MFCC features closely mimic the hearing characteristics of the human auditory system. Other effective features may also be used as complementary inputs in speech emotion recognition (SER) such as Local Binary Patterns (LBP), Chroma, Mel Spectrogram, and Contrast [3].

4.2 Machine Learning Algorithms

Machine learning algorithms are either generative or discriminative classification techniques. Generative models use the joint probability, P(x, C), to compute the posterior probability, P(C j x) using Bayes' theorem and then classify the test instance into the category with the maximum posterior distribution. Conversely, discriminative algorithms model the boundary between two classes by directly estimating the posterior probability, P(C j x) or the conditional class density function. The machine learning algorithms investigated in this work represent a variety of regression approaches showing promising approaches in the context of SER. The description of each approach is given below: Support Vector Machines (SVM) finds a separating hyperplane that maximizes the margin between the two classes [7]. Artificial Neural Networks (ANN) are inspired by the studying of the functioning of the neurons in the human brain combined in a network. ANN utilizes a bottom-up approach to generalize from training data and can model very complex functions [8]. Decision Trees (DT) learn decision rules directly from the training data to construct a tree-like model of decisions [5]. K-Nearest Neighbour (KNN) assigns a class of a test instance by finding the K-nearest neighbours in the feature space and using a majority voting mechanism to assign a class.

Table 1. Comparative Summary of Machine Learning and Deep Learning Models in Speech Emotion Detection

Model / Algorithm	Features Used	Common Datasets	Strengths	Limitations
Support Vector Machine (SVM)	MFCC, Pitch, Energy	RAVDESS, EMO-DB	Fast training, effective for binary emotion classification	Limited scalability, less effective for multi-class problems
Random Forest (RF)	MFCC, Chroma, Zero-Crossing Rate (ZCR)	TESS, SAVEE	Resistant to overfitting, interpretable	Lower performance compared to deep learning methods
Artificial Neural Network (ANN)	MFCC, LPC, Spectral Centroid	IEMOCAP, EMOVO	Learns complex feature relationships	Sensitive to parameter tuning, prone to overfitting
Convolutional Neural Network (CNN)	Spectrogram, Log-Mel, MFCC	RAVDESS, IEMOCAP	Captures spatial patterns in spectrograms	Requires large data, lacks temporal modeling
Recurrent Neural Network (RNN) / LSTM	Sequential MFCC, Time-based Features	IEMOCAP, CREMA-D	Excellent for modeling temporal dependencies	Slower training, may suffer from vanishing gradient issues
3D Convolutional Neural Network (3D CNN)	Spectrogram tensors (from k-means keyframes)	Custom datasets ([2])	Integrates spatial and temporal features simultaneously	High computational cost, complex to implement
Hybrid (CNN + LSTM)	Spectrograms + MFCC + Text Embeddings	EMO-DB, IEMOCAP, RAVDESS	Combines spatial and temporal dynamics effectively	High model complexity, difficult to interpret

Table 1 presents a comparative summary of widely used machine learning and deep learning models in the field of speech emotion detection. It highlights the typical features these models utilize, datasets commonly applied for training and evaluation, and outlines both their practical advantages and technical challenges. This overview helps to clarify the rationale behind model selection in various research scenarios and practical applications.

4.3 Deep Learning Approaches

Deep-learning approaches have dominated the field of speech emotion detection. The development of artificial intelligence (AI) has motivated the investigation of the relationship between intelligent algorithms and emotions [9]. Ability to understand and manage emotions, called emotional intelligence, plays an important role in decision-making. Some researchers suggest that emotional intelligence can be learned and strengthened. For machine intelligence, it also is important to understand and generate such emotional intelligence. The Deep-EMO framework, therefore, provides a simple yet effective speech emotional-recognition deep-learning technique. It has two main pipelines: extracting strong speech features and exploiting deep-transfer learning for emotion recognition. The Deep-EMO framework is primarily applied to English emotional speech but can easily be adapted to other languages. Owing to the significant technological demands for speech-emotion recognition, systematic experiments on real-world data demonstrated the effectiveness of the proposed approach.

Motion plays a fundamental role in daily life for effective communication and underlies human abilities to interact, collaborate, and empathize [10]. Researchers have investigated human emotion and behavior from psychological and computational perspectives and emphasized the importance of emotion recognition in human-machine interaction. Speech contains the linguistic content of a message together with information about the speaker and environment. This complexity makes emotion recognition challenging, owing to the entanglement of multiple signals. Traditional features include pitch, log-Mel filter-bank energies, mel-frequency cepstral coefficients (MFCCs), and perceptual linear prediction in the acoustic modality; and Haar, local binary pattern, histogram of oriented gradients, and scale-invariant feature transform in visual modalities. Various classifiers using these features have shown good performance. A deep-learning alternative makes it possible to automatically learn features from data, and convolutional neural networks (CNNs) have demonstrated outstanding results in image recognition, object detection, and speech-acoustic modeling. Applying convolution in speech-emotion recognition helps capture high-level features from large datasets, which facilitates the automatic identification of affective information.

5. Current Trends in AI for Emotion Detection

Based on the literature and recent research, there are several current trends in AI for emotion detection. Emotion modeling frameworks include Ekman's theory of basic emotions, Plutchik's wheel of emotion, Russel's circumplex model, and others like Shaver, OCC, VAD, and Lovheim [11]. Emotions such as joy, sadness, anger, fear, and love can be recognized from various modalities including facial expressions, textual contents, spoken expressions, and physiology measured from wearable devices. Emotion recognition and sentiment analysis are sometimes used interchangeably but differ in their definitions; emotion refers to strong feelings from circumstances or relationships, while sentiment reflects opinions or views. Several open-source and commercial emotion detection platforms are available, such as IBM Watson NLU™, which analyzes text, voice transcripts, and other data to identify emotions like anger, disgust, fear, joy, and sadness, providing likelihood values. In addition, an emotion recognition system based on

the analysis of speech signals has been proposed. It extracts audio features such as MFCC, pitch, and intensity from speech frames, then selects keyframes using k-means clustering. The spectrograms of these keyframes are encapsulated in 3D tensors used to train a 3D CNN with two convolutional layers. Experiments on multiple databases show the proposed method outperforms current state-of-the-art techniques [2]. Speech emotion recognition (SER) remains an important research direction for improving human–machine interactions and applications like pain detection, lie detection, and emotion-based recommendations. The main goal of SER is to extract discriminative features from voice signals to predict emotional states accurately.

5.1 Integration of Multimodal Data

Multimodal sentiment analysis still faces many challenges, including the lack of large, diverse datasets that incorporate multiple languages and accurate annotations. Current datasets primarily include visual, speech, and text modalities and lack physiological signals. Analyzing hidden emotions such as sarcasm, complex emotions, and context-dependent feelings remains difficult, highlighting the gap between human and artificial intelligence. Video data analysis is challenging due to noise, low-resolution videos, and the need to capture micro-expressions and micro-gestures. Text data often involves cross-lingual comments and mixed emotions, making sentiment recognition complex. Analyzing emotional cues in transcribed speech, especially with multiple speakers and cultural differences, is particularly difficult. The future of multimodal sentiment analysis is promising, with applications in mental health assessment, deception detection, offensive language identification, and emotion-aware robotics. Enhancing models with more parameters and multimodal data will improve accuracy, potentially reaching human-like sentiment understanding [12]. Multimodal sentiment analysis is potentially urban with applications in mental health assessment, deception detection, offensive language identification, and emotion-aware robotics. Models endowed with greater numbers of parameters and modalities will demonstrate heightened accuracy, aspiring toward human-level sentiment understanding. Emotion recognition plays a key role in affective computing and human-computer interaction. Integration of verbal and non-verbal information, such as speech and images, is often used in current emotion recognition approaches. A proposed multimodal fusion technique combines audio-visual modalities from a temporal window with different temporal offsets for each modality. Experiments conducted on the open-access multimodal dataset CREMA-D show that this method outperforms other approaches and human accuracy ratings [13].

5.2 Real-time Emotion Recognition

Speech emotions are effectively captured and conveyed using audio or video data. Conventional methods combine shallow classifiers, such as Support Vector Machines (SVMs), and feature extraction algorithms like Mel-Frequency Cepstral Coefficients (MFCCs) [2]. Achieving robust speech emotion recognition (SER) is challenging, especially with non-stationary, unpredictable data, and general models often require adaptation to specific speakers or corpora. Despite substantial research, a universally robust SER system remains elusive. In speech analysis, the temporal characteristics of emotion can be well represented by individual frames, while spectral information is more effectively modeled by full spectrograms, which can be processed using Convolutional Neural Networks (CNNs). A multi-stream system integrating a 3D CNN framework has been proposed, offering promising results with limited parameter requirements.

According to Martinelli et al. [5], SER has numerous potential applications in clinical diagnostics and psychological research, but current solutions continue to evolve. Human listeners can easily recognize non-verbal emotional cues in speech, whereas Artificial Listeners often find this task difficult, leading to user frustration during interactions with virtual assistants. An innovative SER system based on a semi-supervised consensus approach has been developed and compared with state-of-the-art methodologies. Fast online annotation of extended speech sequences is emphasized as vital for real-time emotion detection.

Pulatov et al. [3] introduce a dual-channel feature extraction and encoding model for SER that exploits complementary information from speech articulations and semantic cues. Speech spectrograms are transcribed using a CNN, while MFCC features undergo semantic encoding through Speech2Vec. These encoded representations are subsequently combined and processed via a Long Short-Term Memory (LSTM) network to enhance emotion detection. Experimental evaluation on the RAVDESS and EMO-DB datasets yields high accuracy figures, demonstrating the approach's efficacy in capturing speech-based emotional nuances.

5.3 Personalization and Adaptation

Personalization and adaptability constitute important requirements for the deployment of intelligent systems designed to operate in dynamic settings or dealing with individual data, and are therefore relevant in the field of speech emotion detection. Adopting paradigms of adaptation may allow the refinement of a model, retrained starting from current parameters with new data that become available at run time. Certain approaches to personalization afford coping with situations in which a limited amount of individual data is available by mapping model parameters to a lower-dimensional space that can be quickly traversed, potentially also enabling simultaneous fine-tuning to multiple users, or by identifying speech segments that support training data selection for each user, producing personalized models for speaker-independent systems [14].

Personalization may span multiple levels of detail on the part of the speaker, and several works in the literature explore individual subjects or subject cohorts. Emotional agent generators can leverage the profiles of individuals being

modeled; these profiles can be synthesized, allowing the modification of the emotional output depending on predefined characteristics, such as personality traits, gender, age, and accent. Where data include particular information or circumstances, models can be created by extracting sets of features with the relative emotion labels; subsequently, training can be performed for each of these sets, producing frameworks that include a prediction for each individual characteristic. This facilitates conditioning the output of predictive models according to various demographic or cultural features [15].

6. Challenges in Speech Emotion Detection

Challenges in speech emotion detection are manifold, arising primarily from the intrinsic nature of emotional characterization and variability in speech. Firstly, emotions are inherently difficult to characterize due to their wide range of ambiguity. Studies illustrate that emotions atop the same category, such as the "anger" class distributed between "anger" and "frustration," can occasionally confuse classification algorithms [2]. It is necessary to select carefully various spectral properties and suprasegmental properties due to the intrinsic nature of emotions. The second challenge concerns the importance of speech in terms of text content, which can vary substantially and affect the evaluation of the speaker's emotional state [4]. More experiments are required to investigate how the semantic content of speech influences the expressed emotion. Thirdly, each person demonstrates different emotional expressions, caused by differences in gender, group, age, and cultural background; preprocessing, feature extraction, and classification techniques must account for these factors [3].

6.1 Data Quality and Availability

Speech datasets have been extensively employed in a wide range of modalities and domains, with the expectation of offering content-rich representations. Nevertheless, these datasets are frequently collected from the Internet, and their recording environments and conditions are not explicitly documented. Consequently, many such datasets may contain low-quality data that can degrade the performance of SDR systems [16]. It is therefore critical to carefully consider the quality of datasets when addressing the challenges of SDR systems.

SDR system quality is highly dependent on good acoustic conditions. No matter how sophisticated a deep learning architecture is, poor acoustic conditions will degrade the quality of information extracted from speech and consequently worsen the effectiveness of the system [3]. In speech research, especially for speech recognition, quality and particularly intelligibility are key aspects in predicting the performance of a given architecture or algorithm. To enhance the quality of speech for SDR systems, research has increasingly focused on reducing noise in audio while preserving the dose of information contained in the speech. According to ISO 532-1, objective quality is understood to quantify psychoacoustic cues related to loudness, sharpness, tonality, and roughness. While objective quality is reasonably linked with emotional content for categorical modes, the moderate correlations observed here may be caused by the scattered distribution of prosodic cues among different emotions.

6.2 Cultural and Linguistic Variability

Culture plays an essential role in theorizing on human auditory perception and emotional bonding. Explorations of the pertinence of culture in the context of speakers from different cultures point out that cultural factors are indispensable to understanding human behavior in general and human auditory cognition in particular. There have been numerous exploratory studies on the particular role that audio cues perform in communication and the influence of culture on the ability of humans to interpret vocal behavior and express emotions. Therefore, it is of utmost importance to make emotion recognition accurately sensitive also to the cultural and/or linguistic variations.

In many real-life situations, emotion recognition may require the systems to have the ability to recognize emotions at different accents within a particular language or among different cultures. That is, in order to make automatic emotion recognition effective to real-life emotions, it must be adept at recognizing any range of language. Existing research in this direction has identified the influence of culture-specific factors and have proposed accent-sensitive speech emotion recognition systems. They have even analyzed a methodology toward identifying the most representative acoustic features for each corpus and the evaluation of human performance in similar recognition tasks. However, it remains an open question whether the models are flexible enough to distinguish between the linguistic and emotional status of the input speech.

6.3 Ethical Considerations

Automated emotion recognition (AER) technology has the ability to detect humans' emotional states in real-time by analyzing facial expressions, voice attributes, text, body movements, and neurological signals [17]. AER appears to be a fertile ground for Artificial Intelligence (AI) and is undergoing rapid progress through the advances in the modelling of human affect. AI-based emotion recognition has many potential applications ranging from healthcare to education, from customer experience to safety, and security. Emotion recognition technology offers great promise for enhancing many fields of human endeavour yet, at the same time, it also presents significant dangers if applied irresponsibly and without thoughtful consideration of the wider ethical and social context of automated emotion-recognition systems. The human emotions that AER instruments measure and interpret are deeply personal and highly intimate features of individuals. Users frequently express grave concerns about the potential invasion of privacy, emotional manipulation, and bias that could result from AER technology. The same algorithmic, privacy, and social problems that now engage the public

imagination when it comes to facial- and speech-recognition software apply in an even more concentrated form to AI-driven AER services. The ethical problems of automated emotion and sentiment recognition cannot be resolved merely by appealing to the beneficence of manufacturers, governments, and companies that develop and deploy such technologies. The thorny questions raised by these systems require passionate and resolute public engagement, and urgency in designing the social environment in which these technologies operate. The recent rapid advancements in artificial intelligence research and deployment have sparked discussion about the potential ramifications of socially- and emotionally-aware AI. Such affectively-aware AI have a potential impact on society since the technology can effectively read people's minds and emotions [18]. A multi-stakeholder analysis framework clarifies the ethical responsibilities of AI Developers and those who deploy such AI, Operators, by establishing two main pillars: effectiveness of the AI and responsible collection, use, and storage of data and the decisions made from such data. The framework facilitates the evaluation of the ethical consequences of affectively-aware AI and guides researchers, industry professionals, and policymakers.

6.4 Technological Limitations

Speech emotion recognition systems still encounter numerous technical limitations. Speech contains a variety of acoustical and linguistic features depicting the speaker's emotional state, such as utterance duration, voice quality, and pitch sequence. It is therefore imperative to extract features that preserve emotions while removing non-related information, such as emotion-irrelevant intensity variations and vocal tract length alterations [3].

Convolutional neural networks (CNNs) have shown considerable potential for the task [2]. Instead of analyzing raw audio, employing spectrograms as inputs effectively captures spectro-temporal variations by converting speech into a two-dimensional format. A 3D CNN model can then extract emotion-salient features from the spectrograms. To reduce computational burden, k-means clustering selects keyframes that best summarize the speech signal, and these spectrograms form a 3D tensor fed into the 3D CNN. Incorporating an optimal number of clusters and keyframes allows rich feature extraction from limited data.

Further refinement employs dual feature extraction encoders. A fully convolutional neural network transcribes the speech spectrograms, while a Mel-frequency cepstral coefficient (MFCC) abstraction approach integrates with Speech2Vec to encode semantic features. Evaluations on RAVDESS and EMO-DB databases demonstrate notable improvements over existing methods in speech emotion recognition accuracy, underscoring the effectiveness of combining both acoustical and semantic representations.

Despite these advancements, machines still struggle to detect emotional cues embedded in speech signals. Current approaches neglect the non-linguistic content of speech, which includes commands, interpersonal attitudes, and other emotional cues. As a result, virtual receptionists lack empathy and fail to predict a user's mood, which can lead to dissatisfaction. Nonetheless, speech is a rich source of information, allowing humans to easily recognize non-verbal vocal signals like murmuring, sighing, or choking. Fast and reliable systems for online labeling of long speech sequences remain a priority for advancing practical speech emotion recognition [5].

7. Applications of Speech Emotion Detection

Speech emotion detection facilitates human-machine communication based on human emotional information, such as speech, language, and facial expressions. The process involves four key steps: preprocessing raw signals to reduce noise and assist feature extraction; feature extraction to identify distinctive audio features representing emotional vocal patterns; feature selection to eliminate irrelevant features; and classification algorithms for final emotion recognition and quantification [1]. The automatic detection and recognition of speech emotions can be employed in various applications, including smart home devices, call centers, eLearning, online tutoring, and personal assistants.

7.1 Healthcare and Therapy

Low-cost, noninvasive, and easily accessible disease detection from speech has attracted a great deal of interest due to the affordability of high-end microphones. Recent developments in speech-based AI have demonstrated the potential of such approaches for many diseases and disorders, but crucial challenges remain. The acquisition of sufficient quantities of high-quality and well-controlled data, with reliable annotations and a large variety of healthy and diseased subjects, poses a major limitation to further progress, especially for rare diseases and languages other than English. Only a small subset of health disorders has been considered up to now, leaving many research questions open. Applications may be envisaged in various scenarios, including hospitals and clinics, where such devices may be used to screen patients for different diseases as a first step before further examinations. In test centers, they may serve as rapid, nonintrusive, and inexpensive preliminary diagnosis during pandemics. In remote locations or developing countries, where many diseases remain underdiagnosed, such devices could be used regardless of the availability of healthcare professionals, and for the targeted monitoring of patients. In such contexts, the self-administration of tests opens the possibility to obtain longitudinal recordings more easily, which in turn facilitates the discrimination between short-term pathologies and long-lasting effects and sequelae that usually require longer recording periods. Within a developmental context, children's speech recorded in the environment of child health-care centers could enable the earlier diagnosis of developmental disorders as soon as they manifest in the voice. Other potentials lie in automatic monitoring—at home or in the workplace. As a result, a number of use cases can be expected, as outlined below: • Pre-screening: Healthcare

practitioners, especially outside the hospital environment, have limited time for each patient and often need to decide whether a laboratory-based test is justified. A tool based on speech analysis could help reduce the number of unnecessary tests. • Remote monitoring: Availabile telemonitoring systems addressing cardiopulmonary and neurological diseases involve the analysis of vital signs such as blood pressure, heart rate, oxygen saturation, or temperature. Buffers limit the time range over which patients can be monitored remotely, and indeed, vital signs sensors are often removed after a few days. The presence of dedicated devices that monitor diseases from speech could offer a cheap and noninvasive alternative for long-term remote monitoring. • Screening at the point of care or during triage: At present, speech assessment cannot replace laboratory tests. Nevertheless, it could be used in combination with other point-of-care instruments to provide feedback before laboratory results are available. • Telemedicine support: Long-term remote monitoring could enable the building of statistical models that characterize the subject's health status. If the speech is analysed continuously or on a regular basis through a smartphone application, the subject could be alerted whenever such a model shows a significant increase of anomalies either for the subject itself or with respect to the general population. This could be triggered directly or by an advice to a healthcare provider. In addition to monitor the health status, an application that continuously analyses speech could detect the presence of particular respiratory markers in a similar way as a smoke or carbon monoxide detector in the home. Monitoring by inference is not limited to patient status. Public health authorities could analyse data submitted by compulsory platforms on which the population is screened during an epidemic or pandemic. The acquired data could be analysed to identify the early appearance of new markers and may lead to the faster detection of a potential outcome of the disease or a new disease altogether. Data may be gathered in two different contexts, and will yield two different types of data: in-clinic or in-the-wild. Data captured in clinics benefit from highquality recording conditions and reliable ground-truth information, but the situation is unnatural and therefore speech production is less representative. In-the-wild data corresponds to more natural and representative recordings with respect to the acoustic environment, linguistic content, and health status, but the quality is uncontrolled and annotation strategies can only be approximated. Ethically complex problems remain, including data protection, digital access, and storage of recordings. The probabilistic nature of AI outputs also requires that detection results are carefully interpreted by professionals, even though tools that provide better insight into the decision-making process are emerging [19].

7.2 Customer Service and Support

Discerning acoustics and tone is crucial for virtual personal assistants and chatbots engaged in call centers, customer support for electronic devices, and automated help desks addressing common problems and frequently asked questions. Machines that interact with humans via voice and gestures are expected to recognize emotions, humor, and sarcasm, comprehend the intention behind messages, and comprehensibly respond to queries [20]. Unlike conventional information retrieval systems, human-machine systems incorporate an empathy module that extracts emotion from speech, dialogue, and behavior and determines an appropriate emotive response. A prevailing approach employs signal processing, sentiment analysis, and machine learning algorithms to create empathetic robots. The human-robot interaction modality includes facial expression and gesture recognition, as well as robot movement that conveys emotional signals and intent. The empathetic agent "Zara the Supergirl" exemplifies a virtual android that autonomously learns and enhances its abilities via machine learning algorithmic feedback and a rule-based system. The long-term objective involves devising machines that understand and appraise emotions and various high-level signals so as to make people's lives easier.

7.3 Entertainment and Gaming

Emotion recognition is important for improving human–machine interaction by understanding emotions behind spoken words. Applications include pain and lie detection, tutorial systems, and emotion-based recommendations. The goal is to extract discriminative voice features in various emotional states, making the system robust to different speaking styles and rates. Enhancing human–machine interaction with machines capable of recognizing emotions has become an important issue in recent years [2]. Human–machine communication that exhibits human-like characteristics will be enhanced if an intelligent agent can recognise the emotional state of human beings during interaction and respond appropriately. Being able to communicate natural emotions, as shown by Williams and Stevens (1972), is necessary for an artificial system to achieve natural interaction. Voice, in face-to-face conversation, is one of the most powerful mediums for recognising a person's emotional state. Society is increasingly spending more hours in front of a television or playing computer games. Machines, therefore, should be designed to communicate with humans in a way that is more natural and enjoyable for the user. This may be achieved by equipping an interactive system with the capability of recognising the emotional state of the user and responding appropriately. Automating this process gives computers and interactive machines the ability to detect emotional states such as anger, frustration, disappointment, or satisfaction; where the knowledge of these states can be fed back to the system to make its interaction more natural and human-like [16]. Increasingly, the analysis of emotion from speech is being applied to enhance entertainment applications. The goal is to determine the emotional state of a user from a speech utterance and to use this information to augment the application so that it responds in a more appropriate and entertaining manner. Previous example applications include call centre monitoring, analysis of students' affect in computer-aided tuition, personal agents, and interactive games.

7.4 Education and Training

Artificial intelligence applications in speech emotion detection find ready use in training and education. Commercial systems provide software to detect student confusion through speech patterns in online courses. Empirical data from the domain of speech emotion may be incorporated into the development and evaluation of relevant projects. Training tools exposing common communication errors through interactions with embodied conversational agents can help students build confidence and develop next-generation interview skills. Ubiquitous computing, natural language processing, and speech-based emotion detection augmented with conversational agents have also provided a medium for education in recent years.

8. Future Directions

Speech emotion recognition (SER) systems still encounter serious difficulties when applied to real-world, long-term acoustic recordings [3]. Data sparsity, the availability of large, high-quality labelled datasets, the multitude of disparate data sources, and the requirement of low-latency SER in artificial intelligence applications are among the obstacles encountered. The next step involves integrating various recognition paradigms encompassing audio and non-verbal messages, linguistic cues, and domain knowledge in order to develop economically viable blended architectures [2]. Machine emotional intelligence continues to constitute a major research issue primarily requiring cooperative contributions from varied disciplines.

The reliability required to setup an efficient interaction in diagnosing, distance rehabilitation, malady ensnarement, and research activities can still be guaranteed by limited speech emotion recognition methods [5]. Semi-supervised prediction models based on consensus have been devised to redress the prediction performance exhibited by a single architecture. Tests in an established emotion recognition framework indicate that the solution improves performance of single methods significantly, both in terms of prediction accuracy and the number of sequences that could be assigned with a quantified degree of confidence.

8.1 Advancements in AI Techniques

Speech emotion recognition (SER) has received significant attention in recent years due to the growth of human–computer interaction applications. Deep learning models have been widely adopted to enhance the performance of these systems, as they can learn generalizable representations directly from raw data. An innovative SER framework was proposed that employs dual feature extraction encoders to improve accuracy. One of the encoders transcribes spectrograms of speech signals using a convolutional neural network (CNN), while the other derives semantic features from Mel-frequency cepstral coefficients combined with Speech2Vec. These modality-specific features are integrated through a weighting fusion mechanism and processed by a long short-term memory (LSTM) network that incorporates an attention mechanism to capture long-range dependencies across speech samples [3]. In a separate approach inspired by group decision-making, a semi-supervised consensus-based prediction model, termed artificial listeners, was developed to assess emotion strength rather than discrete emotional states. Applied to spontaneous speech, this framework achieved superior labeling performance compared to traditional methods [5]. A third technique converts speech signals into spectrograms, from which keyframes are selected using k-means clustering. The spectrograms of these keyframes form a 3D tensor that serves as input to a 3D CNN architecture comprising two convolutional layers and one fully connected layer. Evaluated on multiple emotion databases, this method demonstrated a substantial performance improvement over state-of-the-art solutions [2]. These advances underscore the progress in deep-learning-based SER and its importance in facilitating effective human–machine interaction and understanding spoken language.

8.2 Interdisciplinary Collaborations

The increasing use of spoken language interfaces is drawing attention to the automatic recognition of emotional states in speech. Several systems and datasets for speech emotion recognition have been proposed, but there is no conclusive methodology or consistent results, since emotions are complex and their definitions vary widely. Data sources often consist of actors' voices or broadcasts, and procedures such as Mood Induction Procedures may also be employed. Translating these fluid cognitive states into discrete categories involves a choice between dimensional and categorical approaches: dimensional models characterize continuous emotion spectra, and categorical models classify emotions into fixed classes. A typical machine-learning pipeline for speech emotion recognition includes speech input, feature extraction, feature selection, classification, and emotion output modules. It follows a categorical paradigm and rests on the assumption that measurable voice parameters reflect affective states; physiological reactions such as changes in respiration or muscle tension impact vocal characteristics and undergird the discernibility of emotions [14]. Interdisciplinary cooperation, combining the contributions of the humanities and the sciences, has produced some promising ideas. A novel system based on consensus, inspired by group decision-making dynamics, incorporates these interdisciplinary results to improve speech emotion recognition. Comparisons against traditional approaches demonstrate the effectiveness of this strategy. Despite recent progress, virtual receptionists typically fail to detect callers' emotional states. This limitation diminishes users' perception of empathy, reducing the overall quality of the interaction. That shortcoming persists because machines still lack the ability to recognize the nonverbal cues carried by voice and that humans interpret effortlessly [5].

8.3 Regulatory Frameworks

Emotion detection technologies, which infer emotional states based on various data inputs, are not inherently harmful except when used for harmful purposes. The assessment of when they serve such purposes remains a challenge [21]. The recently adopted Regulation on the transparency and accountability of content moderation (DSA) subjects the largest online platforms and search engines in the EU to obligations that affect emotion detection: risk assessments, independent audits, and public reporting. These requirements may illuminate the manner in which such systems—and the data generated—are utilized, although important limitations must be considered. The DSA also prohibits the profiling of minors and the use of "special categories" of personal data for targeted advertising. While compliance with existing law is necessary, it is not sufficient to address serious ethical concerns and ongoing gaps that create loopholes exploitable for unjustified and harmful uses of emotion data.

9. Case Studies

Speech emotion detection (SED) has become a thriving field with applicability to various domains [22]. Several successful implementations illustrate the benefits of SED and offer insights for future developments [11].

Multimodal SED can strengthen its accuracy by integrating information from multiple modalities, such as speech and vision, but few frameworks employ both [14]. The HUMAINE database, containing audiovisual data of human-agent interactions, stimulated MPQM-HUMAINE, a novel SED approach for Emotion Production Questions (EPQs) that combines speech and visual cues through fusion and score averaging. The database was split into training, validation, and test sets comprising 540, 60, and 200 samples, respectively. Both the fusion and score-averaging models outperformed unimodal configurations, with score averaging requiring less training and achieving better results overall.

9.1 Successful Implementations

The utilization of automatic systems to discern emotional states within speech signals has received considerable attention over the past decades [5]. Modern applications range from enhancing human–machine dialogue to potentially augmenting therapeutic and diagnostic procedures [3]. Concerted efforts have focused on devising mechanisms that derive models from annotated collections to deduce emotional characteristics embedded in speech. Pioneering systems have predominantly applied classifiers to feature vectors encompassing assorted attributes deduced from extracted acoustic cues. While such acoustic cues exhibit elevated signal processing fidelity, complications arise owing to the similarity of cues across varied emotions. Consequently, approaches aimed at the direct identification of raw utterances have been proposed. The study described herein investigates the enhancement of this paradigm through the adoption of spectral representations of audio data, thereby confronting the challenges associated with speech emotion recognition with the intelligence born of collective judgment.

9.2 Lessons Learned

Speech emotion recognition (SER) is a complex research field that involves capturing the speaker's voices—naturally containing emotions—and directly classifying them. This type of system must be able to classify the emotion present in speech, regardless of cultural and linguistic differences, the available amount of emotional data, the context, or the specific domain. In terms of operation, an SER method must classify emotions—the signals that indicate different feelings experienced by a person—via computational methods, reducing the natural complexities of the detection process, such as different voices, signal quality, sentiment dynamics, noise, different channels, different cultural contexts, or emotion diversity.

To build an effective SER technique, basic signal processing, machine learning, and deep learning procedures are required in artificial intelligence. The process must begin with signal acquisition, followed by preprocessing for data quality management. These stages prepare the raw audio signal for the next stage, feature extraction, which implements voice-frequency analysis to reveal the emotional characteristics in the speech signal. The resulting vectors, including features such as voice properties, timbre or tone, vocal tract qualities, Mel-frequency cepstral coefficients, and pitch, serve as a compact representation of the whole speech signal. After transformation, these vectors can be trained with supervised learning algorithms. Training involves implementing and developing a speech emotion classification model using labeled emotional speech data, with the aim of generating an emotion model that can accurately classify emotions in unlabeled speech signals. Finally, the models are tested with unseen audio signals to detect the emotion of the input data and reveal how different techniques perform in classification.

9.3 Open Research Issues and Gaps

Despite notable advancements in the field of speech emotion detection (SED), several unresolved research issues and methodological gaps continue to hinder its practical deployment and generalizability. One of the primary concerns is the **lack of a standardized evaluation framework**. Studies often rely on different datasets, emotional taxonomies, and performance metrics, making it difficult to compare results across models or replicate findings. The absence of universally accepted benchmarks prevents objective assessment of progress and model effectiveness in real-world conditions.

Another critical limitation is the **scarcity of large-scale, multilingual, and culturally diverse speech emotion datasets**. Most available corpora are limited to a few major languages, predominantly English, and often involve acted

rather than spontaneous emotional expressions. This restricts the ability of current models to generalize across languages, dialects, and culturally embedded emotional cues, thus reducing their applicability in global or multicultural contexts.

Moreover, current deep learning architectures used in SED are often perceived as "black-box" systems. There is a growing demand for **explainable AI (XAI)** in this domain—models that can not only provide accurate emotion predictions but also offer interpretable justifications for their decisions. Explainability is crucial in sensitive applications such as healthcare, education, and law enforcement, where understanding why a certain emotional label was assigned can influence downstream decisions and ethical implications.

Furthermore, the **integration of paralinguistic and contextual factors** such as background noise, environmental setting, and speaker-specific variables (e.g., age, gender, mood, health status) remains underexplored. Many existing systems are fragile in the face of variability and lack robustness to real-time, noisy, or adversarial inputs.

Lastly, while multimodal approaches show great promise, there is **limited research on fusion strategies** that effectively combine audio, textual, visual, and physiological data without significantly increasing model complexity or latency.

Addressing these open issues is essential for transitioning speech emotion recognition systems from controlled laboratory settings to reliable tools in real-world environments.

Table 2. Summary of Key Research Gaps and Future Directions in Speech Emotion Detection

Research Gap	Description	Proposed Direction
Lack of Standard Evaluation Framework	Inconsistent use of metrics, datasets, and emotion taxonomies	Establish unified benchmarks and standardized protocols for cross-study comparison
Limited Multilingual and Multicultural Datasets	Most datasets are monolingual (mainly English) and rely on acted emotions	Develop large-scale, spontaneous, multilingual corpora with diverse cultural representation
Lack of Explainable AI (XAI)	Deep models provide accurate results but lack interpretability	Design interpretable models and integrate attention mechanisms or post-hoc explanation methods
Low Robustness to Real-world Conditions	Systems fail in noisy environments or under speaker variability	Incorporate noise augmentation, domain adaptation, and real-world deployment testing
Underutilization of Paralinguistic and Contextual Features	Emotion cues from prosody, environment, and user profile are often ignored	Combine acoustic features with contextual metadata (e.g., speaker identity, environment descriptors)
Challenges in Multimodal Fusion	Difficulty integrating audio with visual/textual/physiological data efficiently	Develop lightweight and dynamic multimodal fusion frameworks with cross-modal attention mechanisms
Ethical and Privacy Concerns	Emotion detection from speech raises issues of consent, surveillance, and data misuse	Create ethical frameworks and regulatory guidelines; ensure transparency, fairness, and data protection
Generalization Across Domains and Tasks	Models trained on specific datasets often fail to generalize to new domains or applications	Use domain generalization, transfer learning, and meta-learning approaches

Table 2 highlights the key research gaps currently limiting the advancement of speech emotion detection systems and outlines possible research avenues to address these challenges. Bridging these gaps is essential for building robust, ethical, and widely deployable AI-based emotional intelligence systems.

10. Conclusion

Speech emotion detection (SED) continues to be a critical area of research in the intersection of artificial intelligence and human-computer interaction. As machines increasingly participate in emotionally nuanced communication, the ability to perceive and interpret human affect becomes essential for building trust, empathy, and natural interaction. This review has explored the technological, methodological, and practical dimensions of SED systems, including preprocessing, feature extraction, classification models, and emerging trends such as deep learning and multimodal integration.

Recent developments demonstrate that combining **deep learning techniques with multimodal data sources**—such as speech, facial expressions, textual content, and physiological signals—offers substantial promise in improving recognition accuracy and robustness. However, these systems must contend with several unresolved challenges, including the lack of standard evaluation frameworks, limited multilingual datasets, cultural biases, and model interpretability. The need for **explainable AI (XAI)** in speech emotion recognition is particularly pressing, especially in

critical fields such as education, healthcare, and law, where understanding the rationale behind an emotion classification is as important as the classification itself.

In addition to technical limitations, the advancement and deployment of SED technologies raise significant **ethical, legal, and social concerns**. The extraction and analysis of emotional data from speech introduces risks related to user privacy, emotional manipulation, surveillance, and potential misuse. Therefore, future research must be accompanied by the development of clear **regulatory frameworks and ethical guidelines** to ensure transparency, consent, fairness, and accountability in the design and application of emotion-aware systems.

Looking forward, SED research must focus on addressing **open research questions** such as generalization across diverse populations and domains, personalization of emotion models, robust real-time inference under noisy conditions, and scalable architectures that remain interpretable and adaptable. Interdisciplinary collaboration between computer scientists, psychologists, linguists, and legal experts will be essential in developing the next generation of human-centric, emotionally intelligent AI systems.

References

[1] N. Sundarprasad, "SPEECH EMOTION DETECTION USING MACHINE LEARNING TECHNIQUES," 2018.
[2] N. Hajarolasvadi and H. Demirel, "3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms," 2019.
[3] I. Pulatov, R. Oteniyazov, F. Makhmudov, and Y. I. Cho, "Enhancing Speech Emotion Recognition Using Dual Feature Extraction Encoders," 2023.
[4] H. Binali, C. Wu, and V. Potdar, "Computational Approaches for Emotion Detection in Text," 2010.
[5] E. Martinelli, A. Mencattini, E. Daprati, and C. Di Natale, "Strength Is in Numbers: Can Concordant Artificial Listeners Improve Prediction of Emotion from Speech?," 2016.
[6] M. Jain, S. Narayan, P. Balaji, B. K P et al., "Speech Emotion Recognition using Support Vector Machine," 2020.
[7] G. Sahu, "Multimodal Speech Emotion Recognition and Ambiguity Resolution," 2019.
[8] M. Kamruzzaman Sarker, K. Md. Rokibul Alam, and M. Arifuzzaman, "Emotion Recognition from Speech based on Relevant Feature and Majority Voting," 2018.
[9] E. Togootogtokh and C. Klasen, "DeepEMO: Deep Learning for Speech Emotion Recognition," 2021.
[10] C. W. Huang and S. S. Narayanan, "Characterizing Types of Convolution in Deep Convolutional Recurrent Neural Networks for Robust Speech Emotion Recognition," 2017.
[11] B. Abu-Salih, M. Alhabashneh, D. Zhu, A. Awajan et al., "Emotion detection of social data: APIs comparative study," 2022.
[12] S. Lai, X. Hu, H. Xu, Z. Ren et al., "Multimodal Sentiment Analysis: A Survey," 2023.
[13] A. Birhala, C. Nicolae Ristea, A. Radoi, and L. Cristian Dutu, "Temporal aggregation of audio-visual modalities for emotion recognition," 2020.
[14] G. Costantini, E. Parada-Cabaleiro, D. Casali, and V. Cesarini, "The Emotion Probe: On the Universality of Cross-Linguistic and Cross-Gender Speech Emotion Recognition via Machine Learning," 2022.
[15] D. Cevher, S. Zepf, and R. Klinger, "Towards Multimodal Emotion Recognition in German Speech Events in Cars using Transfer Learning," 2019.
[16] S. Binti Lebai Lutfi, F. Fernández Martínez, J. Manuel Lucas Cuesta, L. López Lebón et al., "A Satisfaction-based Model for Affect Recognition from Conversational Features in Spoken Dialog Systems," 2013.
[17] S. Latif, H. Shehbaz Ali, M. Usama, R. Rana et al., "AI-Based Emotion Recognition: Promise, Peril, and Prescriptions for Prosocial Path," 2022.
[18] D. C. Ong, "An Ethical Framework for Guiding the Development of Affectively-Aware Artificial Intelligence," 2021.
[19] M. Milling, F. B. Pokorny, K. D. Bartl-Pokorny, and B. W. Schuller, "Is Speech the New Blood? Recent Progress in AI-Based Disease Detection From Audio in a Nutshell," 2022.
[20] P. Fung, D. Bertero, Y. Wan, A. Dey et al., "Towards Empathetic Human-Robot Interactions," 2016.
[21] A. Hauselmann, A. M. Sears, L. Zard, and E. Fosch-Villaronga, "EU law and emotion data," 2023.
[22] A. G. Sabea, M. J. Kadhim, A. F. Neamah, and M. I. Mahdi, “Enhancing medical image analysis with CNN and MobileNet: A particle swarm optimization approach,” *Journal of Information Systems Engineering and Management*, vol. 10, no. 13s, pp. 28–40, Feb. 2025.