Computer Speech and Language 96 (2026) 101873

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Deep feature representations and fusion strategies for speech emotion recognition from acoustic and linguistic modalities: A systematic review

Andrea Chaves-Villota [a,*], Ana Jimenez-Martín [a], Mario Jojoa-Acosta [b], Alfonso Bahillo [b], Juan Jesús García-Domínguez [a]

[a] Department of Electronics, University of Alcalá, E.P.S. Campus universitario s/n, Alcalá de Henares (Madrid), 28805, Spain
[b] Department of Signal Theory and Communications, University of Valladolid, School of Telecommunications Engineering, Valladolid, 47011, Spain

ARTICLE INFO

Keywords:
Emotion recognition
Speech
Linguistic
Acoustic
Fusion
Deep learning
Machine learning
Low and high-level features

ABSTRACT

Emotion Recognition (ER) has gained significant attention due to its importance in advanced human-machine interaction and its widespread real-world applications. In recent years, research on ER systems has focused on multiple key aspects, including the development of high-quality emotional databases, the selection of robust feature representations, and the implementation of advanced classifiers leveraging AI-based techniques. Despite this progress in research, ER still faces significant challenges and gaps that must be addressed to develop accurate and reliable systems. To systematically assess these critical aspects, particularly those centered on AI-based techniques, we employed the PRISMA methodology. Thus, we include journal and conference papers that provide essential insights into key parameters required for dataset development, involving emotion modeling (categorical or dimensional), the type of speech data (natural, acted, or elicited), the most common modalities integrated with acoustic and linguistic data from speech and the technologies used. Similarly, following this methodology, we identified the key representative features that serve as critical emotional information sources in both modalities. For acoustic, this included those extracted from the time and frequency domains, while for linguistic, earlier embeddings and the most common transformer models were considered. In addition, Deep Learning (DL) and attention-based methods were analyzed for both. Given the importance of effectively combining these diverse features for improving ER, we then explore fusion techniques based on the level of abstraction. Specifically, we focus on traditional approaches, including feature-, decision-, DL-, and attention-based fusion methods. Next, we provide a comparative analysis to assess the performance of the approaches included in our study. Our findings indicate that for the most commonly used datasets in the literature: IEMOCAP and MELD, the integration of acoustic and linguistic features reached a weighted accuracy (WA) of 85.71% and 63.80%, respectively. Finally, we discuss the main challenges and propose future guidelines that could enhance the performance of ER systems using acoustic and linguistic features from speech.

Contents

* Corresponding author.
E-mail address: andrea.chaves@uah.es (A. Chaves-Villota).

https://doi.org/10.1016/j.csl.2025.101873
Received 2 April 2025; Received in revised form 21 June 2025; Accepted 15 August 2025
Available online 1 September 2025

1. Introduction

Emotion recognition (ER) plays a fundamental role in the understanding of cognitive processes, human behaviors, and social dynamics faced by human beings in different facets of their lives. Nowadays, with the rise of artificial intelligence (AI), the study of ER has received great attention from the scientific community, since these systems allow a better understanding of how emotions are expressed, experienced, and regulated across different cultures and contexts (Cambria et al., 2017). This is also due to their wide range of practical applications, which include developing therapies for mental health disorders, improving user experiences in areas like virtual assistants and educational tools, enhancing human–computer interaction through more intuitive and sympathetic technologies, and improving marketing strategies by understanding consumer emotions (Kołakowska et al., 2014). In addition, ER systems can reveal patterns of collective behavior, offering potential benefits at the group or community level. For instance, in education, such systems can enhance learning by adapting content and providing targeted support based on students' emotional states. Wang et al. (2020) and Yang et al. (2018). Moreover, from the field of mental health, the identification of how people identify and express their emotions contributes to the diagnosis of emotional disorders such as depression (Bourke

et al., 2010), anxiety (Amstadter, 2008), and disorders of individuals with autism (Rump et al., 2009), and therefore allows for the development of effective techniques and therapies that improve emotional regulation. Hence, research in ER systems allows a better understanding of human nature and improves the design of technologies and interventions aimed at promoting mental health and social well-being (Koolagudi and Rao, 2012).

ER methods based on a single modality or technique, such as self-assessment questionnaires, are often subjective and susceptible to inaccuracies due to behavioral biases (Imani and Montazer, 2019). Furthermore, emotion is a transient state that requires continuous monitoring for accurate recognition, a disadvantage for self-assessment questionnaires. Additionally, some methods may assume that participants can accurately recognize and report their own emotions, which may not always be the case (Lopatovska and Arapakis, 2011). Besides, emotions are inherently complex and multifaceted, often expressed through a combination of modalities like speech, facial expressions, gestures, and linguistic content (Deng and Ren, 2021). Relying solely on one modality risks missing complementary emotional signals present in other ones. For instance, acoustic features may capture tone and prosody but overlook semantic meaning, while linguistic data may fail to reflect paralinguistic cues such as pitch or intensity (Ahmed et al., 2023). Moreover, single-modality systems are typically more sensitive to noise or missing data; for example, background noise can degrade the performance of audio-based systems, while poor lighting or occlusion can affect visual inputs (Salazar et al., 2021). This lack of redundancy and context limits the robustness of unimodal approaches, making them less effective in real-world scenarios where emotional expressions are often multimodal in nature.

To develop more robust models, the latest research focuses on training multimodal models that consider different aspects of human behavior, including but not limited to facial expressions, body language, speech patterns, and physiological responses such as heart rate, electromyograms, and galvanic skin responses (Ragot et al., 2018; Sebe et al., 2005; Ingale and Chaudhari, 2012; Tarnowski et al., 2017). This approach, known as multimodal emotion recognition (MER), leverages the strengths of multiple data modalities to achieve a more accurate and holistic understanding of emotional states (Flanagan et al., 2020). By integrating various types of data, MER systems effectively capture the complexity and nuances of human emotions. Nowadays, these systems employ advanced technologies based on Machine Learning (ML), DL, and attention mechanisms to provide a more comprehensive representation of emotions (Ahmed et al., 2023).

Speech is naturally available in many real-world applications, such as virtual assistants, call center analytics, and interactive systems, which makes it practical and widely applicable for ER. Moreover, a key advantage of speech emotion recognition (SER) is its non-intrusive nature, unlike MER systems, which often rely on physiological sensors, cameras, or wearables, leading to higher costs, increased storage demands, and greater computational requirements due to additional hardware. In addition, some modalities, such as video, present higher privacy risks (Shoumy et al., 2020), making it more complex to integrate into ER models than speech. Thus, the fusion of their acoustic and linguistic modalities could overcome these limitations.

In previous decades, the development of SER systems has progressively emphasized the integration of acoustic and linguistic modalities (Schuller, 2018). SER systems have attracted increasing research interest due to their capacity to capture two distinct and complementary modalities: acoustic and linguistic. These modalities represent information through different channels, providing explicit and implicit emotional cues (Ramanarayanan et al., 2022). Linguistic information captures semantic content, while acoustic information conveys paralinguistic features such as tone, pitch, rhythm, and intensity. Research in the neuroscience field has also demonstrated the correlation among acoustics, linguistics, and emotions in human communication; their experimental findings suggest that individuals use emotional words to amplify the emotional tone of their communication (Liebenthal et al., 2016). In this review, the term *speech* is understood as a spoken language, from which two distinct modalities are derived: the *acoustic* modality, denoting vocal characteristics (*how* something is said), and the *linguistic* modality, including the content of speech (*what* is said).

The rest of the paper is organized as follows: Section 2 explains the process of study selection, using the PRISMA methodology. Section 3 discusses all the findings corresponding to the most commonly used datasets, highlighting main features such as emotion modeling, dataset type, modalities, and technologies adopted. Section 4 contains information on the approaches implemented for extracting low- and high-level features, including the fusion techniques applied. The benchmarking model performances are detailed in Section 5. Section 6 presents discussion and the future directions of research. Finally, Section 7 summarizes the outcome and the conclusion of this research.

1.1. Main contributions

Despite progress, a cohesive synthesis of recent findings, methodologies, datasets, and challenges in this interdisciplinary field has been lacking. Our systematic review aims to bridge this gap, offering:

1. Exploration of emotion modeling theories: we present a detailed evaluation of the most used emotion modeling from speech, covering dimensional and categorical frameworks.
2. Analysis of key aspects of SER datasets: this study examines essential factors of current speech databases, such as emotion modeling, speech types (natural, elicited, or acted), the latest technologies used, and other relevant considerations. These factors provide valuable information for developing SER datasets.
3. Detailed assessment of both low- and high-level features obtained from the acoustic and linguistic modalities of speech: we provide a detailed analysis of the selection of robust features commonly employed in state-of-the-art ER techniques, focusing on DL- and attention-based approaches.
4. Thorough evaluation of fusion techniques for SER: we explore an in-depth analysis of the fusion stage in SER, with particular emphasis on linguistic and acoustic, especially leveraging DL- and attention-based methods.

5. Detailed benchmark analysis of SER approaches: we summarize and compare results of the leading SER methods using linguistic and acoustic features from the most widely used databases, according to common metrics (WA or UA).
6. Detailed selection of research articles under a systematic review approach: We selected current literature published between 2015 and 2024, sourced exclusively from peerreviewed journals, and the most recent studies presented in conferences between 2022 and 2025, following the PRISMA methodology (Page et al., 2021).
7. Critical discussion of the current state and emerging research opportunities: we discuss the existing gaps and outline the directions and challenges for future research aimed at advancing the development of more robust and generalizable SER systems from acoustic and linguistic features. This includes insights related to data scarcity, integration of approaches and modalities, among others.

1.2. Related work

Currently, diverse research has been conducted to provide significant relevant insights into developing ER systems. Most of them are mainly applied to multimodal approaches (Kalateh et al., 2024; Lian et al., 2023; Geetha et al., 2024), which as previously established, exhibit weaknesses as they are more intrusive technologies, costly in terms of computation, hardware, and storage, with higher privacy risks that must be addressed in the development stages. Hence, developing ER systems leveraging acoustic and linguistic information provides benefits not only in the development stages but also in having higher usability for real-world applications due to their advantages in implementation.

The overall purpose of this review is to determine the current challenges and limitations in developing ER systems to optimize fusion strategies by integrating linguistic and acoustic modalities. Thus, this review seeks to answer the following main question:

What is the current state-of-the-art of SER based on its acoustic and linguistic modalities?

To answer this question, it is essential to formulate subquestions that will serve as the review's objectives. This study specifically examines the common SER datasets, focusing on those incorporating their acoustic and linguistic modalities. Additionally, it explores the extraction of low and high-level features and their respective fusion techniques. Thus, the research questions will be divided into three main groups. The first centers on dataset reviews *RQ1-RQ4*, the second addresses the approaches to extract representative features and fusion mechanisms based on AI techniques *RQ5-RQ6*, and finally, the third covers a benchmark analysis for proposed SER models with the most used emotions and datasets *RQ7-RQ8*. Below, we describe the research questions:

Datasets

RQ1: What is the most commonly used emotion modeling in the SER field?

RQ2: What is the predominant type of speech in datasets used for ER?

RQ3: What further data modalities have been integrated with linguistic and acoustic modalities of speech for ER?

RQ4: What technology and devices have been used to construct datasets for SER?

Approaches

RQ5: What are the most representative low- and high-level features of linguistic and acoustic modalities for SER? And what is the most common feature vector dimension?

RQ6: Which fusion techniques have been applied to develop SER systems from their acoustic and linguistic modalities?

Benchmark analysis

RQ7: What are the most common emotions evaluated in the literature of SER?

RQ8: What are the best performances in literature for SER from acoustic and linguistic modalities?

To highlight the main contributions of the present review in comparison to related surveys, we summarize the key distinctions in Table 1. In particular, we contrast the eight research questions addressed in this study against those explored in previous works.

From a broad perspective, a key contribution of this paper is the identification of current limitations in the field of ER using speech, particularly using acoustic and linguistic information. As highlighted in Table 1 and supported by existing studies, numerous surveys focus on MER (Ahmed et al., 2023; Yang et al., 2023; Wang et al., 2022; Pepa et al., 2023; Shoumy et al., 2020; Jiang et al., 2020). However, to the best of the authors' knowledge, only one prior study, published in 2022, specifically addresses this area.

A common limitation of existing overviews in MER is their broad analytical approach, which does not always provide a detailed examination of crucial aspects specific to SER from linguistic and acoustic information. The inclusion and analysis of different data modalities vary across studies, leading to inconsistencies in the depth of evaluation for speech-based ER, as Kalateh et al. (2024), Lian et al. (2023) and Geetha et al. (2024). Some may exclude entirely, focusing instead on physiological signals, static images, or text-only datasets. Even in some cases, using text may imply that it originates from written sources, such as descriptions or opinions (Du et al., 2022) rather than specific transcribed speech (linguistic), a key difference this review seeks to address. Another limitation arises in analyzing relevant features extracted from each data modality and the corresponding fusion methods employed. Since these reviews do not focus specifically on including studies based on acoustic and linguistic modalities from speech, they lack in-depth insights into the challenges and optimizations required for their acquisition, feature extraction, and fusion mechanisms. Whether or not to include different modalities directly influences the performance of the ER models. Similarly, in the benchmark analysis of SER approaches, it is essential to establish parameter equivalence, such as the number and set of emotions, metrics, and datasets considered. However, none of the previous studies have addressed this aspect. Consequently, the questions addressed in this review are still not thoroughly analyzed or concluded.

Table 1
Comparative review of prior surveys.

Paper	Modalities	RQ1	RQ2	RQ3	RQ4	RQ5	RQ6	RQ7	RQ8	Details
Kalateh et al. (2024)	Verbal, physiological signals, facial, body gesture, and speech	✓	✗	✗	✗	✗	✗	✗	✗	This paper reviews studies from journal publications or articles following a PRISMA methodology. The type of speech, technology, and specific high- and low-level features from linguistic and acoustic information, and the comparative analysis of performance benchmarking are not covered.
Lian et al. (2023)	Speech, text, and face	✗	✗	✗	✗	✗	✓	✗	✗	This paper reviews the studies under a non-systematic review. It focuses on multimodal approach where further modalities to the speech focus specially on face images. Emotion modeling, predominant speech type, dimension of extracted feature vectors, and performance benchmarking are not covered.
Geetha et al. (2024)	Text, audio, visual, and physiological	✓	✗	✗	✗	✗	✓	✗	✓	This paper presents a systematic review following PRISMA methodology focusing on DL-based MER. The predominant type of speech data, the technology, and the dimension of extracted feature vectors are not covered.
Atmaja et al. (2022)	Speech and text	✓	✗	✓	✗	✓	✓	✗	✓	This paper reviews the studies under a non-systematic review in the area of bimodal SER for studies reported in the literature until 2021, mainly those from INTERSPEECH 2020 conference.
This review	Linguistic and acoustic from speech	✓	✓	✓	✓	✓	✓	✓	✓	This review presents the current key findings for SER from peer-reviewed journal articles, and studies presented at high-impact conferences related to the field. Then, they were systematically selected and analyzed following the PRISMA methodology, with a specialized focus on acoustic and linguistic modalities extracted from the speech

In contrast to Atmaja et al. (2022), which provides an overview primarily evaluating studies presented at the INTERSPEECH 2021 conference, this paper offers an updated review of the state of the art published at INTERSPEECH and other related conferences, as well as focusing on results from journal-published articles. Additionally, we classify the proposed methods into DL- and attention-based approaches, enabling a clearer understanding of how these technologies are evolving within the field. While the previous work primarily examines data fusion technologies rather than acquisition methods or database construction, this review provides valuable insights that can benefit both academia and industry by keeping them informed about the latest advancements in this rapidly developing domain.

2. Method

2.1. Study selection

In this section, the process of study selection is described. To identify relevant studies concerning the research questions, this review was completed following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021). The systematic review flowchart is depicted in Fig. 1. It includes four main steps: identification, screening, eligibility, and inclusion. The development of each step will be explained in detail in the following subsections.

2.1.1. Identification

The relevant studies were obtained from four scientific digital libraries: WoS, IEEE, ACM, and Scopus. To identify potential studies and evaluate the evolution of SER from acoustic and linguistic modalities in recent years, we limited the time window from January 1st, 2015, to April 2024. SER can be done using many different synonyms, thus, the search strategy was conducted by selecting relevant keywords for the review scope and their related words. The keywords were categorized by recognition task, AI-based technique, application field, data modality, and modal approach (detailed in Table A.11). The search filtered studies whose

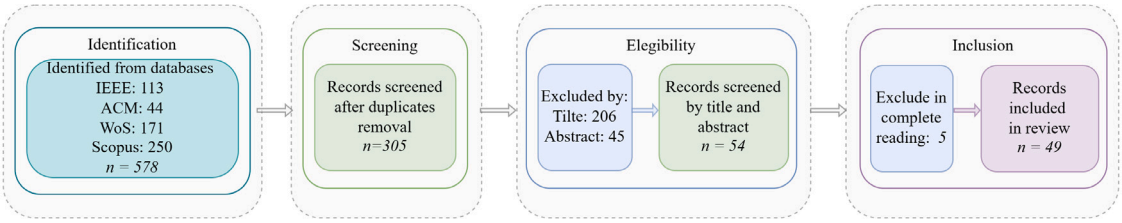

Fig. 1. Flowchart of systematic review based on PRISMA.

Table 2
Inclusion and exclusion Criteria.

Criteria	
Inclusion	Exclusion
Complete records	Unavailable records
Studies published in journals and selected conferences	Studies not written in English, reviews, editorials or opinion papers
Focus on recognition/monitoring of a mood/emotion	Does not focus on recognition/monitoring of a mood/emotion
It is based on bimodal/multimodal approach and includes acoustic and linguistic modalities from speech	It is not based on bimodal/multimodal approach and does not include acoustic and linguistic modalities
Define a method of dataset acquisition or reference the dataset	Does not include dataset acquisition or reference it.
Studies that describe the evaluation metric(s) and AI-based method(s) used	Does not use an AI-based approach

keywords of categories, such as application field, data modality, and approach, were found in the title. All other technical terms were expanded to abstract or title. Hence, the search strategy employed is as follows: *Title-Abstract* (Task related keywords) AND *Title-Abstract* (AI-based techniques related keywords) AND *Title* (Application field related keywords) AND *Title* (Data modality related keywords) AND *Title* (Approach related keywords) AND (1 January 2015: April 2024)

Following the same methodology, to evaluate the most recent advances in the field, we conducted an additional search for studies presented at conferences related to speech and AI, from January 1st, 2022, to June 2025. These conferences include: Interspeech; the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); the ACM International Conference on Multimedia (MM); the Multimodal Interfaces and Machine Learning for Multimodal Interaction (ICMI-MLMI) conference; and the IEEE Spoken Language Technology (SLT).

2.1.2. Screening

The screening process was carried out by applying an initial filter designed to eliminate duplicate records. A total of 274 studies were found to be repeated across four databases, leaving 305 unique records remaining.

2.1.3. Eligibility

The review aimed to identify original research focused on developing or evaluating methods that could fuse the features extracted mainly from linguistic and acoustic modalities of speech, specifically within the AI field. Consequently, we outline the inclusion and exclusion criteria for screening records in Table 2. Thus, in this phase, the studies were excluded according to the criteria based on title or abstract. This step resulted in the elimination of 206 papers because of the title, leaving 99 for further review. Then, the abstracts were screened to ensure alignment with the eligibility criteria; in total, 45 were extracted, resulting in 54 remaining articles to be read completely.

2.1.4. Inclusion

In this stage, we performed a complete reading of the 54 remaining documents for a full-text assessment. After reviewing, 49 studies were selected for inclusion in this review. Fig. 2 shows the number of articles included per publication year.

3. Datasets

With the study selection process completed, we will discuss the main characteristics of the most commonly used datasets for ER in literature, especially but not limited to linguistic and acoustic modalities from speech.

There are different MER datasets, however, Table 3 lists only the datasets used by the studies that met the inclusion criteria. They are arranged according to their frequency of use in the selected studies (highest to lowest). It can be noted that the most

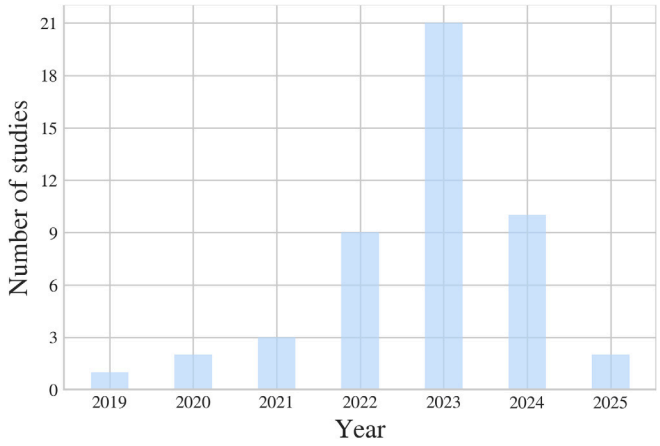

Fig. 2. Number of studies included per publication year.

Table 3
Review of ER datasets.

Dataset	Language	Year	Size	Dataset usage
IEMOCAP (Busso et al., 2008)	English	2007	12 h	46.99%
MELD (Poria et al., 2018)	English	2018	30 h	26.51%
CMU-MOSEI (Zadeh et al., 2018)	English	2018	65 h	8.43%
CMU-MOSI (Zadeh et al., 2016)	English	2016	2–3 h	2.41%
RAVDESS (Livingstone and Russo, 2018)	English	2018	3–4 h	2.41%
EmoInt-MD (Singh et al., 2023)	English	2022	534 h	1.20%
CREMA-D (Cao et al., 2014)	English	2014	5 h	1.20%
SAVEE (Haq and Jackson, 2010)	English	2011	<1 h	1.20%
MSP-Podcast (Lotfian and Busso, 2019)	English	2019	27 h	1.20%
EmoReact (Nojavanasghari et al., 2016)	English	2016	2 h	1.20%
ElderReact (Ma et al., 2019)	English	2019	2–3 h	1.20%
LSSED (Fan et al., 2021)	English	2021	200 h	1.20%
Spanish MEACorpus (Pan et al., 2024)	Spanish	2023	13 h	1.20%
EmoDB (Burkhardt et al., 2005)	German	2005	<1 h	1.20%
MTED (Zhao et al., 2024b)	Chinese	2024	9 h	1.20%
Korean Emotional Speech Dataset (Byun et al., 2021)	Korean	2021	8–12 h	1.20%

popular datasets for SER are: the Interactive Emotional Dyadic Motion Capture (IEMOCAP) (Busso et al., 2008) and Multimodal EmotionLines Dataset (MELD) (Poria et al., 2018). These datasets are used in 39 (46.99%) and 22 (26.51%) of the studies in this review, respectively (some studies use more than one dataset to test their recognition models).

All datasets were published between 2005 and 2024, suggesting that ER is a relatively new field and has received research attention in recent years. Furthermore, considering there are around 6000 speaking languages worldwide, it is evident that corpora incorporating both acoustic and linguistic data modalities encompass only a limited number of languages, with English being the most widely represented. In this review, 75% of the corpora use English speech data, as opposed to Spanish (Pan et al., 2024), German (Burkhardt et al., 2005), Chinese (Zhao et al., 2024b) and Korean (Byun et al., 2021), which are used only once for a specific dataset. Hence, it has been and continues to be one of the biggest challenges that industry and academia face when developing ER systems. This concern is one of the main reasons why efforts are being made to develop systems using relatively small-size datasets (Latif et al., 2019).

It is also evident from the data reported in Table 3, that 11 out of 16 corpora (68%) have a speech duration ranging from 1 h to 13 h. Longer datasets such as MELD (Poria et al., 2018) or CMU-MOSEI (Zadeh et al., 2018) have speech lengths of 30 h and 65 h, respectively. The datasets with the longest speech duration are EmoInt-MD (Singh et al., 2023) and LSSED (Fan et al., 2021) which include 534 h and 200 h of data, respectively. Notably, creating large-scale datasets for ER is generally limited since this process is time-consuming and requires expert human knowledge for proper data labeling (Latif et al., 2021).

3.1. Emotion modeling

Datasets are fundamental components to ensure the suitable performance of ER systems. A common factor among the datasets used in the literature is that labels are annotated based on specific emotional models. Hence, focusing on the first subquestion of this review *RQ1: What is the most commonly used emotion modeling in the SER field?* We will discuss emotion modeling commonly used in literature, more specifically, two primary theories remain in use, leading to the classification of emotional databases based on either categorical or dimensional models. They are explained in the following subsections.

3.1.1. Categorical modeling

This modeling has been commonly adopted as a basic model in ER tasks for its simplicity and intuitive nature (Cambria et al., 2017). The categorical model refers to dividing emotions into discrete categories, thus, ER systems in this modeling must solve classification tasks to assign one or more discrete emotions to input information. Even though this modeling does not present a consensus on categorizing basic emotions, Paul Ekman's six basic emotions model is most commonly accepted. In this modeling, Ekman proposes categorizing six basic universal emotions: Anger, Disgust, Fear, Happiness, Sadness, and Surprise (Ekman and Friesen, 1971).

Table 4 describes the emotion modeling used by datasets introduced in Table 3 for ER, the most predominant labeling model for representing emotions from speech data is categorical annotation, which is supported by the fact that ten datasets (62.5%) use only this modeling. Five works (31.25%) use both categorical and dimensional modeling. The most common emotion categories include Happiness, Sadness, and Anger; Fear and surprise labels are the least represented. MELD (Poria et al., 2018), The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) (Livingstone and Russo, 2018), Surrey Audio-Visual Expressed Emotion (SAVEE) (Haq and Jackson, 2010), and MSP-Podcast (Lotfian and Busso, 2019) contain Ekman's categorization, and a neutral class. The work presented in Singh et al. (2023) adopts the categorization of 32 emotions. This dataset was constructed considering nine different movie genres: drama, action, fantasy, comedy, horror, crime, romance, thriller, and science fiction. They categorized the labels into a set of 32 distinct emotions, aiming to capture a wide range of emotions across various movie genres that reflect real-life scenarios. In addition to the basic emotions typically included in the datasets, there are corpora with categorical annotations including Calm (Livingstone and Russo, 2018), Boredom (Burkhardt et al., 2005), Excited and Confused (Zhao et al., 2024b) as well.

3.1.2. Dimensional modeling

Dimensional modeling theories consider emotions according to a continuous structure; thus, each emotion state is represented as a multi-dimensional vector, where for each dimension, there is a continuous value within a given range, the extremes of the range indicating two polarities (Deng and Ren, 2021). One of the most widely recognized dimensional models is the Pleasure-Arousal-Dominance (PAD) Emotional State Model, which establishes that three nearly orthogonal dimensions provide a sufficiently comprehensive description of emotional states (Mehrabian, 1996). The IEMOCAP and MSP-Podcast corpus, which use both types of modeling, includes continuous emotional content as variants of this model, hence, an utterance can be represented in a three-dimensional space in terms of Valence, Activation (or Arousal), and Dominance (Busso et al., 2008). Another widely used dimensional approach is Russell's circumplex model, where emotion states are represented by a continuous value on an Arousal vs. Valence dimensional plane. The Valence extends from negative to positive range with a neutral state between the polarities. Arousal is set from low to high values with a neutral state in between (Russell, 1980). The EmoReact (Nojavanasghari et al., 2016) and ElderReact (Ma et al., 2019) corpus are based on this modeling, representing emotional information in a one-dimensional space in terms of Valence. Other works such as CMU-MOSEI (Zadeh et al., 2018) and CMU-MOSI (Zadeh et al., 2016) consider emotion states in a single dimension as either strongly positive (labeled as +3), positive (+2), weakly positive (+1), neutral (0), weakly negative (−1), negative (−2), or strongly negative (−3). As shown, constructing high-quality dimensional datasets is more challenging than categorical annotation, resulting in fewer available corpora. In this review, only six studies (37.5%) use dimensional labeling and five of them also include discrete annotations. Having annotations based on the two approaches allows a mapping between the two annotation strategies, and the use of additional information could improve the performance of the ER systems.

3.2. Emotional speech data

As stated above, an adequate corpus design is crucial for developing ER systems with optimal performance. One of the main criteria for preparing a corpus is the scope, which refers to the different types of variations of a dataset, such as the number of speakers, speaker gender, language, kind of emotions, among others (Swain et al., 2018). Some of these variations covered by the datasets in this review have been included in Table 5. In this subsection particularly, we want to focus on answering the second subquestion *RQ2: What is the predominant type of speech in datasets used for ER?* Consequently, the type of database for SER can be categorized into three groups: Natural, Acted (Simulated), or Elicited (Induced) (Ververidis and Kotropoulos, 2006), they are described below.

3.2.1. Natural

A natural corpus includes spontaneous speech in real-world situations, such as: call center conversations, discussions between patients and doctors, educational environments, family, or couple conversations, among others. These emotions, also known as *underlying emotions*, are mildly expressed and difficult to identify. The labeling process is highly subjective and typically relies on the judgment of experts. In addition, datasets that include this type of speech cover a limited emotional range (El Ayadi et al., 2011).

As shown in Table 5, seven (43.75%) of corpora encompass natural speech. Four are taken from YouTube opinion videos, this can be observed in CMU-MOSEI (Zadeh et al., 2018), which contains 23453 monologue videos from YouTube segments with 1000 distinct speakers and 250 topics. The most frequent topics include reviews (16.2%), debate (2.9%), and consulting (1.8%). They also limit their videos where the speaker's attention is directed at the camera, excluding videos that involve moving, such as those mounted on bicycles or videos recorded while walking. A similar case is presented in CMU-MOSI (Zadeh et al., 2016), where the authors also include YouTube monologue videos collected from 89 different speakers. The authors remark that to achieve a fine-grained

Table 4
Emotion modeling used in the literature.

Dataset	Model		Emotions								
	C	D	No.	Ekman's basic emotions						N	Other
				Ang	Disg	Fear	Hap	Sad	Sur		
IEMOCAP	✓	✓	9	✓	✓	✓	✓	✓	✓	✓	Frustration, excitement **D:** valence, activation, and dominance.
CMU-MOSEI	✓	✓	6	✓	✓	✓	✓	✓	✓	–	**D:** Sentiment on a Likert scale from strongly negative to strongly positive [−3,3]
MSP-Podcast	✓	✓	9	✓	✓	✓	✓	✓	✓	✓	**C:** Contempt and other **D:** Valence, arousal and dominance on a seven-point Likert scale
ElderReact	✓	✓	6	✓	✓	✓	✓	✓	✓	–	**D:** Valence
EmoReact	✓	✓	16	✓	✓	✓	✓	✓	✓	✓	**C:** Curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment and frustration **D:** Valence
MELD	✓	–	7	✓	✓	✓	✓	✓	✓	✓	–
RAVDESS	✓	–	8	✓	✓	✓	✓	✓	✓	✓	Calm
EmoInt-MD	✓	–	32	✓	✓	–	✓	✓	✓	✓	Grateful, afraid, lonely, impressed, hopeful, furious, confident, disappointed, jealous, nostalgic, apprehensive, trusting, worried, anticipating, excited, caring, proud, anxious, terrified, ashamed, content, faithful, embarrassed, guilty, sentimental
CREMA-D	✓	–	6	✓	✓	✓	✓	✓	–	✓	–
SAVEE	✓	–	7	✓	✓	✓	✓	✓	✓	✓	–
LSSED	✓	–	11	✓	✓	✓	✓	✓	✓	✓	Disappointment, boredom, excitement and other
Spanish MEACorpus	✓	–	6	✓	✓	✓	✓	✓	–	✓	–
EmoDB	✓	–	7	✓	✓	✓	✓	✓	–	✓	Boredom
MTED	✓	–	6	–	–	–	✓	✓	✓	✓	Excited, confused
Korean Emotional Speech Dataset	✓	–	4	✓	–	–	✓	✓	–	✓	–
CMU-MOSI	–	✓	–	–	–	–	–	–	–	–	**D:** Sentiment on a Likert scale from strongly negative to strongly positive [−3,3]

C: Categorical, D: Dimensional, Ang: Anger, Disg: Disgust, Hap: Happiness, Sad: Sadness, Sur: Surprise, N: Neutral.

sentiment analysis, it is necessary to conduct a subjectivity segmentation at the opinion level. In Pan et al. (2024), the authors selected videos from diverse Spanish-language YouTube channels, where the users express opinions in different contexts that could evoke basic emotions, including political channels to express disgust, sports channels to identify anger, and entertainment channels to capture joy. In this case, some of them were recorded in outdoor environments subjected to unwanted noise conditions, as well as recordings that have been previously processed and edited in recording studios. In Nojavanasghari et al. (2016), authors present a multimodal emotion dataset that contains 1102 videos of children between the ages of four and fourteen years old; these videos were downloaded from the YouTube React channel and contain children reacting to subjects that include food, technology, YouTube videos, and gaming devices. Authors in Lotfian and Busso (2019) introduce a corpus of 2317 utterances of existing spontaneous recordings obtained from audio-sharing websites (podcasts). The data contains natural conversations between many different people on various topics, such as political debates, movie reviews, or sports discussions.

3.2.2. Elicited

Elicited or induced datasets gather data simulating artificial emotional situations; it is generally expected that participants involved are unaware of the problem so that the process is as natural as possible. Usually, different contextual situations are presented through emotional conversations to elicit different emotions. Unlike acted data corpus, elicited datasets are closer to

Table 5
Modalities and technology employed by datasets.

Dataset	Source	Type			Data modalities				Technology				Speakers	
		N	E	A	S	V	SL	O	Mic	Vid	Unk	O	No.	Description
IEMOCAP	Recordings of improvisations or scripted scenarios	–	✓	–	✓	✓	✓	✓	✓	✓	–	✓	10	Actors (5 F, 5 M)
MELD	Dialogues from TV series	–	–	✓	✓	✓	✓	–	–	–	✓	–	–	Actors
CMU-MOSEI	Opinion videos from YouTube	✓	–	–	✓	✓	✓	–	–	–	✓	–	1000	YouTube users (430 F, 570 M)
CMU-MOSI	Opinion videos from YouTube	✓	–	–	✓	✓	✓	–	–	–	✓	–	89	YouTube users (41 F, 48 M)
RAVDESS	Recordings of predefined statements	–	–	✓	✓	✓	–	–	✓	✓	–	–	24	Actors (12 F, 12 M)
EmoInt-MD	Movies of different genres	–	–	✓	✓	✓	✓	–	–	–	✓	–	–	Actors
CREMA-D	Recordings of predefined statements	–	–	✓	✓	✓	–	–	✓	✓	–	–	91	Actors (43 F, 48 M)
SAVEE	Recordings of predefined statements	–	–	✓	✓	✓	–	–	✓	✓	–	–	4	Actors (4 M)
MSP-Podcast	Podcast recordings	✓	–	–	✓	–	–	–	–	–	✓	–	83	Speakers from podcasts
EmoReact	Videos from YouTube REACT channel	✓	–	–	–	✓	–	–	–	–	✓	–	63	Children (32 F, 31 M)
ElderReact	Videos from YouTube REACT channel	–	✓	–	–	✓	–	–	–	–	✓	–	46	Elderly people (26F, 20M)
LSSED	Video sessions in an indoor lab environment	–	✓	–	✓	–	–	–	–	–	✓	–	820	Youtube users (485 F, 335 M)
Spanish MEACorpus	Opinion videos from YouTube	✓	–	–	✓	–	✓	–	–	–	✓	–	–	YouTube users
EmoDB	Recordings of predefined statements	–	–	✓	✓	–	–	–	✓	–	–	–	10	Professional speakers (5 F, 5 M)
MTED	Classroom videos	✓	–	–	✓	✓	✓	–	–	–	✓	–	235	Teachers
Korean Emotional Speech Dataset	Recordings of Korean predefined statements	–	–	✓	✓	–	–	–	–	–	✓	–	4	Actors (2 F, 2 M)

N: Natural, E: Elicited, A: Acted; S: Speech, V: Video, SL: Speech linguistic, O: Other; Mic: Microphone, Vid: Video Camera, Unk: Unknown; F: Female, M: Male.

natural ones. As can be seen in Table 5, three datasets (18.75%) include this type of emotional speech data. In Busso et al. (2008), the authors employed two approaches for eliciting emotions from the actors. First, subjects memorized and practiced scripts selected and supervised by a theater professional to ensure the portrayal of target emotions. In the second approach, participants improvised emotional expressions based on hypothetical scenarios involving common experiences such as loss or separation. Unlike the structured nature of scripted sessions, this method captures more authentic responses, offering a broader understanding of emotional expression in both controlled and spontaneous contexts.

Alternatively, authors in Fan et al. (2021) introduce a Large-Scale Speech English dataset (LSSED) for ER, which contains data collected from 820 subjects induced by random questions as their utterances are associated with an emotional label. Speech data was processed from videos recorded in sessions in an indoor lab environment. On the other hand, in Ma et al. (2019), authors propose a multimodal dataset of 1323 video clips with human annotations of six discrete emotions. Videos were collected from the YouTube REACT channel in which elders react to different emotion elicitation stimuli, covering a wide range of topics including video games, social events, and online challenges.

3.2.3. Acted

Simulated or acted (also known as *full blown* emotions) refers to data collected with professional and experienced artists. When collecting these datasets, experts are usually asked to express neutral sentences on different emotions (El Ayadi et al., 2011). Specifically, this method is used in the data collection process of RAVDESS (Livingstone and Russo, 2018), CREMA-D (Cao et al., 2014), SAVEE (Haq and Jackson, 2010), EmoDB (Burkhardt et al., 2005), and Korean Emotional Speech Dataset (Byun et al., 2021), where the source corresponds to the recordings of predefined statements. Other studies including this type of speech use already existing audiovisual material such as TV series, MELD (Poria et al., 2018), or movies from various genres, EmoInt-MD (Singh et al., 2023). This is a convenient method since the efforts in dataset collection no longer involve the definition and implementation of the recording stage. Therefore, most datasets involve this type of speech since they are an easier and more reliable method for collecting expressive speech. This tendency is noticeable in the datasets included in this review since 43.75% of the corpora cover this type of speech. This type of data collected involves covering a wider range of emotions, incorporating greater intensity and expressiveness, and involving most aspects relevant to emotions' expression (Koolagudi and Rao, 2012).

3.3. Data modalities

In this section, we will address the question *RQ3: What further data modalities have been integrated with linguistic and acoustic modalities of speech for ER?* As previously stated, this review essentially focuses on exploring SER systems involving their extracted

linguistic and acoustic modalities. The modalities included by the datasets are listed in Table 5. In the following sections, we will discuss the main factors encompassing the datasets depending on modalities.

3.3.1. Speech

Speech refers to spoken human language, considered the primary mode of communication among people. It includes not just the spoken words (*what* is said) but also the nuances in tone, pitch, speed, and rhythm (*how* it is said). Therefore, emotional expressions from speech can be conveyed through explicit (linguistic) and implicit (acoustic/paralinguistic) messages. Usually, acoustic information of speech provides a wide range of indicators that reliably reflect the speaker's emotional state in contrast to linguistic content, where language-dependent factors such as variability in word choice for expressing emotions make their generalization across languages a more challenging task (Latif et al., 2021). There is widely available speech data from numerous sources, including webcasts, conversations, music, meetings, voice messages, lectures, television, and radio.

As outlined in Table 5, four datasets covered in this review include speech as unimodal data source, known as EmoDB (Burkhardt et al., 2005; Lotfian and Busso, 2019; Fan et al., 2021), and the Korean Emotional Speech Dataset (Byun et al., 2021). The rest of the corpora (75%) provide other modalities, the most common being video, besides speech. EmoDB is a German database of acted emotional speech that contains 535 utterances with an average of 3 s. The utterances were generated with ten sentences interpreted in target emotions (neutral, anger, fear, joy, sadness, disgust, and boredom) by ten actors (5 female and 5 male), the sentences were previously defined from everyday sentences to prioritize the naturalness of the speech. Hence, this database has been adopted to analyze prosodic and articulatory features in SER. The Korean Emotional Speech Dataset includes utterances using Korean scripts from dramas and movies recorded by four professional actors (2 female and 2 male). The dataset contains 4000–5000 audio files with around 3–10 s each categorized into four emotions: anger, happiness, neutrality, and sadness.

Speech acoustic modality

Due to the scope of this review, studies whose methods for ER involve only the acoustic modality of speech (*how* it is said) have not been considered. However, it is important to note that the acoustic modality can be extracted from any dataset containing speech or video, which applies to all the datasets included in this review. Notably, existing literature includes studies for MER in which the acoustic modality is shared through its representation in low- and high-level features, aiming to comply with ethical and privacy regulations (Miranda Calero et al., 2024).

Speech linguistic modality

Linguistic modality from speech refers to the written or spoken language that conveys explicit information through words and sentences (*what* is said). It includes not only the content of the communication but also contextual elements like syntax, semantics, and word choice, which contribute to the meaning (Deng and Ren, 2021). It should be noted that, unlike text emotion recognition (TER), which is based only on written text, the linguistic modality from speech is often paired with its respective acoustic modality for SER (Atmaja et al., 2022). Studies have also established that linguistic SER provides an efficient approach to interpreting emotional dialogue, enhancing the accuracy and intelligence of human–computer interactions (Suero Montero and Suhonen, 2014). Furthermore, studies have found that exploration in data integration can contribute significantly to high-quality emotional datasets and thus to better accuracy for ER (Abdullah et al., 2021). From the datasets included in this review, all corpora that contain *Speech* do not necessarily include its *linguistic* modality or transcriptions. Some studies incorporate manual or automatic techniques, such as Automatic Speech Recognition (ASR), to extract this information from speech. These corpora will be evaluated in the following section.

3.3.2. Acoustic and linguistic from speech

Training SER systems using linguistic and acoustic modalities requires preprocessing phases to extract the linguistic modality when datasets do not directly provide it. Generally, this phase is completed manually to ensure greater accuracy, although ASR methods, such as speech-to-text, are also commonly employed due to the time-consuming nature of manual transcription. As depicted in Table 5, seven out of sixteen datasets, including speech data, provide the corresponding linguistic modality. The authors of CMU-MOSEI and CMU-MOSI provide this data obtained through manual transcriptions made by experts. In EmoInt-MD, the authors offer subtitles and transcripts from movies. In IEMOCAP, the authors used the Ubiqus service. Modern systems such as ASR methods are also used in the conformation of the different datasets; in MTED where the teachers' speeches were transcribed with Baidu's auto speech transcription API and Spanish MEACorpus whose transcriptions are developed using Whisper (Radford et al., 2023), together with some manual ones to confirm the quality of the transcriptions. It is important to highlight that, to develop SER systems that are accurate, fast, and easily applicable to real-life scenarios based on their two primary modalities (as will be discussed in the following section), recent studies have proposed methods aimed at enhancing performance by reducing the word error rate (WER) of ASR systems. Despite these advancements, while the use of automatic tools to obtain the transcriptions of speech data is fast and cost-effective, manual transcription remains the gold standard for applications where precision and clarity are essential; they can handle language changes, speaker identification, and errors in real-time, improving the overall quality and reliability of the transcription. In addition, manual transcription of speech can, in some cases, include some specifications on pause fillers (such as "um", "uh", etc.), stress indicators, and speech pauses (Zadeh et al., 2016). These phonetic parameters provide relevant information for the monitoring of emotional changes in people (Lin et al., 2020).

3.3.3. Speech and other modalities

The video modality is the most frequent in ER datasets that include speech. In this review, eleven of the corpora (68.75%) include this data source. This data refers to visual information captured through images or video sequences, including facial expressions, body language, gestures, and environmental context. This modality provides rich, non-verbal cues essential for interpreting human behavior, emotions, and interactions, allowing a more holistic understanding of communication in MER systems.

As indicated in Table 5, usually in natural and some acted datasets, videos include monologues that focus on gestures and facial expressions. This is a critical design consideration. While a single-speaker emotional corpus may suffice for emotional speech synthesis, ER demands a more diverse dataset with multiple speakers and expressive styles to accurately capture and interpret emotional variability (Koolagudi and Rao, 2012). A significant disadvantage of using this modality lies in the complexity and resources required for accurate analysis, which can be costly and slow down real-time applications (Shoumy et al., 2020). Additionally, video data introduces higher privacy concerns, as visual information can reveal sensitive personal details. These challenges make video data harder to integrate compared to more streamlined modalities, such as speech, which require less processing and often yield consistent emotional insights.

According to the results, another data modality that provides valuable content when determining emotions corresponds to motion, which refers to the dynamic aspects of human movement, including gestures, body language, and facial expressions that occur over time (Kleinsmith and Bianchi-Berthouze, 2012). It is captured by sensors other than video cameras, such as inertial measurement units. This modality captures the physical actions and changes in posture that convey important non-verbal information about a person's emotional state, intent, or focus. From the datasets in this review, only IEMOCAP (Busso et al., 2008) includes motion data, where actors used markers on the face, head, and hands, providing detailed information about their facial expressions and hand movements during recording scenarios.

3.4. Technology

The fourth research question is addressed in this section *RQ4: What technology and devices have been used to construct datasets for SER?* The aim is to provide an analysis of the tools and technologies employed for data collection to support SER. Recent advancements in data collection have increasingly involved the use of pre-existing sources such as YouTube videos, television series, movies, and podcasts. Thus, in these cases, the information on the technological tools used is unknown. In IEMOCAP (Busso et al., 2008), the audio was recorded using high-quality shotgun microphones (Schoeps CMIT 5U) directed at each participant in the dialogue, and the sample rate was set to 48 kHz. For video data, a semi-frontal view of participants was recorded with high-resolution digital cameras (Sony DCR-TRV340). For motion, to capture facial expression information, fifty-three markers were attached to the face and a headband with two markers on it to identify the head rotation. The hands' movements are estimated with the information provided by wristbands with two markers and an extra marker on each hand. The trajectories of the markers were recorded using a VICON motion capture system with eight cameras placed one meter from the subject with markers, with a sample rate of 120 frames per second.

In RAVDESS (Livingstone and Russo, 2018), a multimodal database that includes video, speech, and video-and-speech formats where 24 professional actors vocalize lexically matched statements. The registers were done in a professional recording studio using a Sony Handycam HDR-SR11 with a resolution of 1920 × 1080 pixels at 30 fps. Speech data was captured by a Rode NTK vacuum tube condenser microphone, fitted with a Stedman prosc-reen XL pop filter, placed 20 cm from the actor at a sampling rate of 48 kHz, 16 bit. CREMA-D (Cao et al., 2014) includes audio-only, visual-only, or audio-visual data from 91 actors and actresses of various ages and ethnicities. Video data was recorded using a Panasonic AG-HPX170 at a resolution of 960 × 720. The speech was recorded with a far-field directional microphone at 48 kHz.

SAVEE (Haq and Jackson, 2010) includes 480 utterances of audio, visual, and audio-visual modalities. 2D frontal color video and Beyerdynamic microphone signals were collected in a 3dMD dynamic face capture system with a sample rate of 44.1 kHz for audio and 60 fps for video. In EmoDB (Burkhardt et al., 2005), a Sennheiser MKH 40P 48 microphone and a Tascam DA-P1 portable DAT recorder were used to achieve high-quality recordings with a sampling frequency of 48 kHz. In addition, electro-glotto-grams were recorded using the portable laryngograph. In the acquisition of audio data, microphones with sampling rates of at least 44.1 kHz are generally used, with 48 kHz and 16 bits being the most commonly used. The distance between the speaker and the microphone is in the range of 20 to 30 cm, provided that individuals are free to use body language (Busso et al., 2008; Burkhardt et al., 2005).

4. Approaches

In this section, we will discuss the methods and techniques used by the included studies to recognize emotions. This review will emphasize acoustic and linguistic modalities from speech with others (when they are used). An ER system primarily involves extracting features from each modality and integrating information from multiple modalities using fusion methods, as shown in Fig. 3 (Deng and Ren, 2021; Imani and Montazer, 2019).

4.1. Feature representation

The feature extraction in SER systems aims to identify patterns that effectively distinguish between various emotions. Focusing on the fifth research question *RQ5:What are the most representative low- and high-level features of linguistic and acoustic modalities for SER? And what is the most common feature vector dimension?* In SER, feature engineering and developing DL models for classification are

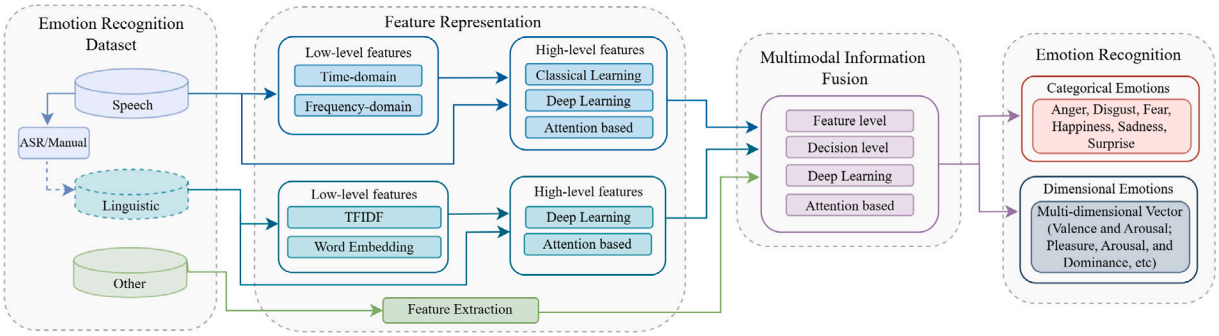

Fig. 3. Emotion recognition system based on acoustic and linguistic modalities from speech.

generally handled as independent tasks (Atmaja et al., 2022). On the one hand, the process of manually transforming speech signals into meaningful information and a manageable set of attributes based on domain expertise is known as feature engineering, this method is often considered labor-intensive and time-consuming (Latif et al., 2021). The features obtained by this method are also known as hand-crafted features. On the other hand, representation learning involves generating representations by automatically transforming input data, typically within DL approaches, to produce abstract and functional representations for DL tasks. In this review, we categorized these feature extraction processes into two groups: low- and high-level for feature engineering and representation learning tasks, respectively. Depending on the feature type and approach, some methods may or may not include the two processes. In the following sections, we will discuss the features most commonly used by studies for SER using acoustic and linguistic information.

4.1.1. Acoustic feature representation

The acoustic features from speech typically fall into one of the following categories: frequency-domain, time-domain, hybrid, or deep features (Hashem et al., 2023). As mentioned in this review, deep ones will be considered high-level features. To determine which are the most frequent, the low-level features were divided into time-domain and frequency-domain.

Low-level features

Typically, features derived from speech signals are referred to as acoustic features (in this review, we distinguish this term from the acoustic modality, i.e. *how* it is said). Several studies have explored the relationship between acoustic features and emotional states (Scherer, 2005).

The frame period for speech analysis typically ranges from 30 to 100 ms, during which suprasegmental features, such as rhythm, melody, and expressiveness are captured. These elements are characterized by attributes including duration, intensity, intonation, and sound units. Prosodic features, which encompass pitch, energy, duration, and their respective derivatives (Rao and Yegnanarayana, 2006), play a crucial role in conveying speech dynamics. Therefore, in this study, prosodic features also will be incorporated into the time-domain features. These features perform well by differentiating between emotions associated with differing levels of arousal, such as happiness and sadness. However, they are less effective in differentiating emotions that share similar arousal levels, such as anger and happiness, which are more strongly correlated to valence (Fahad et al., 2021). As shown in Table 6, 20 (40.81%) of the studies include time-domain features, among the most common are zero crossing rate, energy, and pitch frequency.

Spectral features are related to frequency domain analysis, which usually reflects vocal tract characteristics, where emotion-specific information is present (Koolagudi and Rao, 2012). Due to the frame duration (about 10–30 ms), spectral features are sometimes called segmental features. In literature, Mel-frequency cepstral coefficients (MFCCs) have been widely known as the principal set of features for ER (Furui, 1986). The process of extracting MFCCs involves four key steps. First, perform a Fourier transform to analyze the signal frequency content. Second, the power spectrum is mapped onto the Mel scale. Third, applying a logarithmic transformation to Mel-scaled frequencies, and finally, utilizing a discrete cosine transform (DCT) or an alternative suitable transformation to obtain a compact representation of the features (sometimes, this step is omitted because of loss of information and destruction of spatial relations) (Latif et al., 2021). As shown in Table 6, from the 19 studies using frequency-domain features, 14 use MFCCs where the speech signal represents the short-term power spectrum.

Often, in some studies (20.49%), both time- and frequency-domain features are used together as hybrid features. Several studies use this approach to extract representative information. Among the most commonly used are the well-known Minimalist feature sets like GeMAPs and eGeMAPs, which contain 62-D and 88-D time and frequency-domain features, respectively (Eyben et al., 2015). As reported in several studies, spectral and temporal features are also extracted using the Munich open-source Media Interpretation by Large feature-space Extraction (openSMILE), a toolkit that allows feature extraction for signal processing and ML applications.

Concerning the feature vector dimension, where mentioned, most of them range from less than 100-D, with 34-D being the most common. In two specific studies, despite the approach being the same, the dimension is different between datasets. As in Huddar et al. (2021), a 73-D vector is obtained for MOSI and IEMOCAP and 384-D for MOSEI. In Ma et al. (2024), the vector dimensions are longer; the IEMOCAP speech data are represented by a 1582-D feature vector, while a 300-D representation is used for MELD.

Table 6
Low and high-level features for SER from acoustic and linguistic modalities.

Article	Low-level features					High-level features							Fusion technique				
	Acoustic features			Linguistic features			Acoustic features				Linguistic features			F	D	DL	A	Technique
	TD	FD	Features	TF	WE	Features	ML	DL	A	Approach	DL	A	Approach					
Byun et al. (2021)	x	x	43-D (13 MFCCs, 11 spectral-domain, 12 chroma, 7 harmonic features)		x	256-D Tacotron's encoder		x		LSTM and softmax function.	x		LSTM and softmax function.		x			Average from probability
Cai et al. (2019)	x	x	34-D (zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, MFCCs, 12-D chroma vector, and standard deviation of chroma vector)		x	300-D GloVe embedding		x	x	CNN-BiLSTM-Attention (CBLA) and dense layer.	x		Bi-LSTM.			x		DNN and softmax layer
Chauhan et al. (2024)		x	MFCC			–		x	x	CNNs and head pooling attention.		x	BERT			x		FC and softmax layer.
Singh et al. (2023)	x		Tonal low-level descriptors		x	200-D GloVe embedding				–	x		Bi-LSTM.			x	x	Fusion of modalities, contextual transformer, and softmax layer
Ma et al. (2024)	x		1582-D for IEMOCAP and 300-D for MELD			–		x	x	Intra- and inter-modal transformers	x	x	1024-D RoBERTa large model, intra- and inter-modal transformers			x	x	Hierarchical gated fusion (unimodal- and multimodal-level gated fusions, FC and softmax layer)
Ho et al. (2020)		x	12 MFCC, 13-delta and 13-acceleration coefficients			–		x		Batch normalization and a gated recurrent unit module (GRU)	x	x	768-D BERT model, batch normalization layer and GRU			x	x	Multi-Level Multi-Head fusion attention, global average pooling, FC, dropout, and softmax
Hosseini et al. (2024)	x	x	34-D (energy, zero crossing rate, energy entropy, spectral center, spectral expansion, spectral entropy, spectral flux, spectral start, and MFCC)		x	300-D Word2Vec embedding (sequence length of 500)		x	x	CNN and Bi-LSTM-Attention	x		Bi-LSTM	x		x		Concatenation, DNN, and softmax layer
Huddar et al. (2021)	x		73-D for MOSI, 73-D for IEMOCAP, and 384-D for MOSEI		x	Word2Vec for MOSI and IEMOCAP, and GloVe embedding for MOSEI		x	x	RNN and weighted-pooling attention	x	x	RNN and weighted-pooling attention	x		x		Biomdal and trimodal fusion, then softmax layer.
Liu et al. (2022a)	x		Not listed		x	Word embedding		x		CNN and BiLSTM	x		BiLSTM				x	Cross-attention with gate-control mechanism, and transformer encoder.
Liu et al. (2022b)	x	x	Zero-crossing-rate, root mean square frame energy, pitch frequency, harmonics to noise ratio, MFCC		x	300-D Word2Vec embedding		x	x	Self-attention MCNN (two embedding layers, an attention layer, three convolutional layers, and three pooling layers)	x		Self-attention bc-LSTM (BLS-TM layer, an attention layer and a dense layer)	x		x		Multi-scale fusion: FF by concatenation DLF by Dempster-Shafer

(*continued on next page*)

Table 6 (*continued*).

Article	Low-level features						High-level features							Fusion technique				
	Acoustic features			Linguistic features			Acoustic features				Linguistic features			F	D	DL	A	Technique
	TD	FD	Features	TF	WE	Features	ML	DL	A	Approach	DL	A	Approach					
Braunschweiler et al. (2022)		x	log-Mel filterbank features			–		x	x	CNNRNNATT encoder (CNN, RNN, FC, attention)		x	BERT		x			Concatenation and softmax layer
Pan et al. (2024)		x	MFCC			–	x	x	x	SVM, CNN, LSTM and Wav2Vec 2.0 (F-W2V, J-W2V, W2VB)		x	BETO, ALBETO, dBETO, BERTIN, MarIA, XLM		x		x	Late Fusion (LF concat, LF mean), Multi-head cross-attention (Fusion+attn.) and Ensemble Learning (EL mean, EL max)
Ai et al. (2024)			–			–		x		Bidirectional gated recurrent unit (Bi-GRU)	x	x	RoBERTa and Bi-GRU				x	Cross-attention

TD: Time-domain, FD: Frequency-domain, TF: TFIDF, WE: Word Embedding, ML: Machine learning, DL: Deep learning, A: Attention, F: Feature-level, D: Decision-level, n-D: n-Dimensional feature vector.

Article	Low-level features						High-level features							Fusion Technique					
	Acoustic features			Linguistic features			Acoustic features				Linguistic features			F	D	DL	A	O	Technique
	TD	FD	Features	TF	WE	Features	ML	DL	A	Approach	DL	A	Approach						
Wang et al. (2023a)	x	x	199-D (MFCC, chroma, pitch, zero-crossing rate, spectral, and their mean, standard deviation, minimum, and maximum.)	x		1890-D TFIDF		x	x	Linear layer and Self-Transformer encoder	x	x	Linear layer and Self-Transformer encoder	x			x		MF: Cross-Transformer encoder FF: Self-Transformer encoder
Wen et al. (2023)			–			–		x		VGGish		x	BERT						–
Xie et al. (2021)	x	x	Spectogram and waverform			–		x		512-D WaveRNN		x	GPT				x		EmbraceNet and Crossmodality Transformer Fusion
Zhang et al. (2024)	x		Not listed			–				–		x	RoBERTa	x					Three modality-specific graphs and Dual-Stream Propagation (intra- and inter-modal)
Zhang et al. (2022)		x	34-D (13 MFCC, 8 spectrum features such as zero-crossing rate, and 13 spectral features)			–			x	Two-layer Transformer's encoder		x	768-D BERT and two-layer Transformer's encoder			x	x		Multi-head attention and BiLSTM
Zhang et al. (2023c)		x	40-D MFCCs			–		x	x	Bi-LSTM and local intra-attention		x	768-D BERT and local intra-attention network			x	x		Global Inter-modal attention, two cross-modal feature and FC
Zhang et al. (2023a)	x		–			–			x	Variable-length feature extraction (VQ-Wav2Vec and Wav-RoBERTa)		x	GPT-2 tokenizer 1024-D features using RoBERTa			x	x		Adaptative interactive attention and softmax layer
Zhang et al. (2023b)	x		Not listed			–				–		x	RoBERTa	x					Intra and inter-modal features, followed by dynamic fusion

(*continued on next page*)

Table 6 (*continued*).

Article	Low-level features						High-level features							Fusion Technique					
	Acoustic features			Linguistic features			Acoustic features				Linguistic features			F	D	DL	A	O	Technique
	TD	FD	Features	TF	WE	Features	ML	DL	A	Approach	DL	A	Approach						
Zhao et al. (2024b)	x	x	Time-domain, frequency-domain, energy-domain, and perception-domain			–		x	x	Prosody encoder (1D-convolution block, attentive pooling, and fully connected)		x	768-D BERT	x		x			Concatenation and Bi-GRU layer
Lian et al. (2021)	x	x	88-D (extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), energy, MFCC, and pitch)		x	300-D word-level lexical features			x	Multi-Head Attention, Single-Modal and Cross-Modal transformers		x	Single-modal and cross-modal transformers			x	x		ATS-Fusion, FC and Softmax
Kim and Kang (2022)			–			–			x	Wav2vec 2.0		x	BERT				x		Cross-attention and self-attention
Lin and Wang (2023)			–			–			x	Wav2vec 2.0		x	Automatic ASR error adaptation and BERT	x		x			Concatenation, FC and Softmax
Yao and Shi (2024)	x		Opensmile			–				–		x	ASR and RoBERTa	x		x		x	BiGRU-based Intramodal and Speaker-centric Cross-modal Fusions
Li et al. (2023)	x		OpenSmile			–				–		x	Robert-large					x	Contribution-aware Fusion Mechanism (CFM) and Context Refusion Mechanism (CRM)
Van et al. (2025)			–			–		x		FC network		x	Transformer encoder					x	ConxGNN:Inception Graph Module (IGM) and the Hypergraph Module (HM).
Cai et al. (2024)			–			–		x	x	CNN Encoder, WavLM, and HuBERT using LoRA fine-tuning		x	ASR and RoBERTa	x				x	Audio-Text Fusion, FC, and Mutual Information Neural Estimation
Gao et al. (2024)			–			–			x	HuBERT		x	ASR and BERT				x		Self-attention for intra-modal and cross-attention for inter-modal interactions
Ghosh et al. (2022)			–			–			x	Wav2vec 2.0		x	RoBERTa				x		Crossmodal Encoder: head cross-modal attention, residual connections, and feedforward layers
Zhao et al. (2025)			–			–		x		Multi-Layer Perceptrons (MLPs)	x		Bi-GRUs					x	Graph-based multi-frequency propagation and context filtering
Huang et al. (2024)			–			–		x	x	Wav2Vec2.0 encoder and Bi-GRU	x	x	RoBERTa encoder and Bi-GRU				x		Multiple Transformer encoders and Softmax
Luo et al. (2023)	x	x	MFCCs, pitch, and others by OpenSmile			–			x	Pre-trained OpenL3		x	RoBERTa				x		Self- and cross-attention to capture inter- and intra-modal interactions

(*continued on next page*)

Table 6 (*continued*).

Article	Low-level features						High-level features							Fusion Technique					
	Acoustic features			Linguistic features			Acoustic features				Linguistic features			F	D	DL	A	O	Technique
	TD	FD	Features	TF	WE	Features	ML	DL	A	Approach	DL	A	Approach						
Wei et al. (2023)			–			–			x	Wav2vec 2.0		x	RoBERTa				x		Joint attention weight learning.
Kim and Cho (2023)		x	Spectrogram and MFCCs			–		x	x	CNN, BiLSTM encoder and Wav2Vec2.0		x	GPT-2 and RoBERTa				x		Crossmodal Transformer
Kyung et al. (2024)			–			–		x		Latent embedding	x	x	ASR and latent embedding from pre-trained LLM				x		Crossmodal transformer
Priyasad et al. (2023)			–			–			x	Wav2vec 2.0		x	BERT					x	Memory fusion
Wang et al. (2023b)			–			–			x	HuBERT		x	MPNet				x		Modality interaction transformer (MIT)
Rasendrasoa et al. (2022)	x	x	COVAREP and OpenSmile			–				–		x	RoBERTa				x		Multi-head attention mechanism
Zou et al. (2023)	x		OpenSmile			–				–		x	RoBERTa				x	x	Prompt transformer with hybrid contrastive learning
Anand et al. (2023)			–			–		x		Tokenization and position Embedding	x		Tokenization and position Embedding					x	Fusion network by minimizing the Kullback– Leibler divergence
Chen et al. (2022b)			–			–			x	Wav2vec 2.0		x	RoBERTa				x		Key-sparse attention
Li et al. (2022)		x	MFCCs			–		x	x	Bi-LSTM, Wav2vec	x	x	Wav2vec, CTC decoder and Bi-LSTM	x	x		x		Concatenation, concatenation with co-attention, and hierarchical co-attention
Zhang and Li (2023)		x	Spectograms			–		x		Dilated CNN, Leaky -ReLU and skip connection residual block		x	RoBERTa				x		Multi-head co-attention
Zhao et al. (2023)			–			–			x	WavLM		x	BERT				x		Sliding window attention (SliWa)
Zhao et al. (2024a)			–			–			x	WavLM		x	BERT				x		Sliding Adaptative Window attention
Li et al. (2024)			–			–			x	Wav2vec 2.0, HuBERT WavLM, Whisper, and their variations		x	ASR error-robust framework and RoBERTa	x	x		x	x	Early, late, cross-attention, tensor, non-local gate-based (NL-gate), modality-invariant and -specific fusion (MISA)

TD: Time-domain, FD: Frequency-domain, TF: TFIDF, WE: Word Embedding, ML: Machine learning, DL: Deep learning, A: Attention, F: Feature-level, D: Decision-level, n-D: n-Dimensional feature vector, O: Other.

High-level features

Regarding high-level features, representation learning is a more efficient automated process that requires minimal human expertise in the domain while achieving superior performance compared to hand-engineered features. Furthermore, this process demonstrates greater generalization capability across various tasks (Latif et al., 2021). In addition, research in SER has demonstrated that no single discriminative feature universally provides an accurate representation of emotions across diverse datasets (Jahangir et al., 2021). Hence, in recent years, there has been growing interest in using DL techniques to automatically extract speech features from large datasets, enabling models to learn intrinsic patterns and enhance classification performance.

In this review, these methods have been categorized into three main groups. The first corresponds to classical ML methods. From the included studies, only one (2%) uses a method based on SVM. The authors report that the SVM-based method performs worse than other DL- and attention-based methods (Pan et al., 2024).

The second category corresponds to DL-based techniques. For ER, these features are typically extracted using convolutional neural networks (CNNs) to capture spatial patterns and recurrent neural networks (RNNs) to model temporal dependencies within sequential data. Studies have also evaluated the performance of neural network architectures, such as fully connected (FC) networks (Van et al., 2025) and multilayer perceptrons (MLPs) (Zhao et al., 2025). Another widely used approach is the long-short-term memory network (LSTM) whose units are based on a cell and three gates (input, output, forget) that learn temporal information, while the gates regulate the flow of information. A simpler and faster version of the LSTM architecture, gated recurrent unit (GRU), has also been evaluated. Moreover, in recent years, the use of the improved version of LSTM, Bi-LSTM, has also acquired great relevance, where sequence processing is performed in both directions, forward and backward. Some combinations of DL methods such as CNN-BiLSTM are also used to integrate both spatial and temporal features. From Table 6 it can be seen that there are ten studies that are purely DL-based techniques, such as WaveRNN (Xie et al., 2021) LSTM (Byun et al., 2021), GRU (Ho et al., 2020), CNN-BiLSTM (Liu et al., 2022a), and Bi-GRU (Ai et al., 2024). Moreover, in Wen et al. (2023), the performance of the VGGish architecture has been evaluated. This architecture was developed by Google based on the VGGNet architecture for image recognition. In this case, it was specifically adapted for feature extraction of speech using convolutional layers (Hershey et al., 2017). Among these feature representations, only (Xie et al., 2021) reports the feature vector dimension corresponding to 512-D. This dimension corresponds to the default output of the WaveRNN model used Kalchbrenner et al. (2018).

It is noticed that in recent years attention-based methods (also known as Transformers) have been used in different applications, where the main idea is to pay attention to the crucial data features, assigning them a higher weight than the rest. In this way, these approaches learn temporal correlations from sequential data and enable them to capture broader temporal contexts with reduced computational complexity (Vaswani et al., 2017). Thus, the third category corresponds to attention-based approaches. In this category, fifteen studies utilizing exclusively attention mechanisms were included. Of these 15 studies, 11 employ the Wav2Vec architecture (or its 2.0 version), indicating the high relevance and effectiveness of high-level feature representations for SER. This architecture, designed by Facebook, combines CNN and self-attention to learn the contextual representation of speech (Baevski et al., 2020). Additionally, architectures such as HuBERT (Hsu et al., 2021), WavLM (Chen et al., 2022a), and Whisper (Radford et al., 2023) have also been evaluated to capture contextual and semantic information at this stage of SER. HuBERT and WavLM employ self-supervised learning to extract rich speech representations, while Whisper, developed by OpenAI, is a multitask model capable of transcription and translation, offering robust feature representations even in noisy conditions. Their ability to model long-range dependencies and context-sensitive cues makes them particularly well-suited for capturing the dynamic nature of emotional expression in speech.

Besides, in this review, fifteen studies employ hybrid techniques, where more than one approach is combined to allow the models to learn more complex patterns of essential features. These include the use of DL architectures such as RNN, CNN, LSTM, and Bi-LSTM with different attention-based techniques. The most widely used corresponds to the Bi-LSTM with attention mechanism (Cai et al., 2019; Hosseini et al., 2024).

4.1.2. Linguistic feature representation

With the aim of machines understanding the text provided at the input, it must be transformed or mapped into a numerical representation. In text processing, these procedures are considered methods for feature extraction. The most common techniques involve the statistical method Term frequency-inverse document frequency (TFIDF), in which words are represented based on their frequency in a document relative to a collection of documents. It quantifies the information these terms carry in a given document, obtaining a high-dimensional representation where each word is independent without capturing contextual or semantic relationships (Nandwani and Verma, 2021). The increasing availability of large-scale data has facilitated the development of word embeddings generated by DL networks based on distributional semantics (Deng and Ren, 2021). Through this method, word embeddings capture semantic relationships by positioning similar words closer together in vector space, often learning latent, low-dimensional representations from the language structure (Souma et al., 2019). Examples of word embeddings in the literature include Word2Vec (Mikolov, 2013) developed by Google, GloVe introduced by researchers at Stanford University (Pennington et al., 2014), and FastText developed by Facebook (Joulin et al., 2016).

Low-level features

In this review, low-level features extracted from the text will be defined as those obtained by TFIDF or earlier word embedding, (i.e. models that represent a word by a single vector). As shown in Table 6, only 16% of the studies include low-level features, most obtained from GloVe (Cai et al., 2019; Singh et al., 2023; Huddar et al., 2021) and Word2Vec embeddings (Hosseini et al., 2024; Liu et al., 2022b; Huddar et al., 2021). In most cases, feature vector dimensions range from 200-D to 300-D, with 300-D being the most prevalent. A single approach employs the features extracted using the TFIDF method (Wang et al., 2023a).

In current literature, emotional word embeddings have been developed, showing significant contributions in tasks such as emotion classification and emotion intensity prediction. In Xu et al. (2018), Emo2Vec is proposed to encode emotional semantics into vectors. A sentiment-specific word embedding (SSWE) and domain-sensitive and sentiment-aware embedding models are introduced in Tang et al. (2014) and Shi et al. (2018), respectively, the emotional information integration into word embeddings resulting in improved performance for ER. Although these guidelines could have greater relevance to the ER task, possibly leading to better performance. They have not yet been explored in ER, this opens a gap to be studied in the future.

High-level features

The main limitation of earlier word embedding models is known as meaning conflation, where each word, regardless of whether it has one meaning (monosemous) or multiple meanings (polysemous), is represented by a single vector ignoring the role of contextual information (Deng and Ren, 2021). With the successful application of transfer learning, pre-trained language models produce contextualized word embeddings with general knowledge; they could be adapted for a wide range of downstream tasks (Deng and Ren, 2021). These models have achieved state-of-the-art performance on several NLP tasks, which can be easily transferred to ER tasks. Thus, the high-level features were divided into two categories: extracted by DL or by attention-based models.

As shown in Table 6, approaches relying on DL-based features represent a smaller proportion, used for only 16% of the studies reviewed. The most frequent architectures are LSTM and Bi-LSTM networks. On the other hand, it is evident that current approaches emphasize the extraction of high-level features from the linguistic modality through contextualized embeddings generated by transformer-based models. These include the best-known Generative Pre-trained Transformer (GPT) proposed by OpenAI (Radford, 2018), the Bidirectional Encoder Representations for Transformers (BERT) (Devlin, 2018), and its Robustly Optimized version Approach (RoBERTa) (Liu et al., 2019). Among the approaches analyzed, 65.3% incorporate this feature extraction process, with RoBERTa as the most frequently used, appearing in 18 of the approaches.

In the same way, studies have explored the integration of linguistic modality obtained from ASR systems with attention-based methods. This combination leverages the textual modality derived from transcribed speech to capture representative emotional features (Yao and Shi, 2024; Cai et al., 2024; Gao et al., 2024; Kyung et al., 2024). Furthermore, some approaches emphasize the importance of reducing speech recognition errors prior to linguistic feature extraction, as improvements in ASR accuracy directly enhance the quality of the input to attention-based models, thereby leading to more reliable and robust ER performance (Lin and Wang, 2023; Li et al., 2024).

Hybrid methods have also been developed to improve performance of ER. The most common involve techniques such as RNN, LSTM, and BiLSTM with attention-based methods or well-known transformers. Furthermore, it is also shown that the representation vectors of high-level features range in values of 768-D and 1024-D. This is one of the main differences compared to acoustic modality, where it is common to find smaller dimension representation vectors. This difference in dimensionality potentially presents a challenge that must be addressed when modalities are fused to construct robust ER systems.

4.2. Fusion techniques

Once the relevant features of data modalities are extracted, fusion methods to build ER systems are needed, either in a discrete or continuous approach. In this way, in this section, we aim to answer our sixth research question *RQ6: Which fusion techniques have been applied to develop SER systems from their acoustic and linguistic modalities?*.

As the literature states, how to fuse different data modalities is a challenging task; depending on the approach taken, it can be determined at which level of abstraction to combine the modalities, as well as which procedure must be used to integrate them (Yang et al., 2021). There are different fusion techniques; particularly in this review, they have been mainly categorized based on the level or method into the traditional techniques: decision or late-fusion, feature-fusion, DL, or attention-based. As shown in Table 6, it is worth highlighting that several studies evaluate various types of fusion strategies, including hybrid approaches that combine early or late fusion levels using DL or attention-based mechanisms. Additionally, some methods rely exclusively on a single type of fusion. It will be discussed in the following sections.

4.2.1. Decision-level fusion

In decision-level fusion or late fusion, each decision corresponds to the results obtained by the methods that process each modality independently. Subsequently, a fused vector is formed with these decisions, which will be further analyzed to obtain a final decision on the respective recognition task (Kaya et al., 2017; Zhalehpour et al., 2016). Thus, this approach aims to leverage the strengths of each modality's model to produce a more accurate and robust overall decision or prediction. In ML, this term is also known as ensemble learning. This method has certain advantages over other techniques, as it allows each modality to find the optimal classifier for the task (Liu et al., 2021). Additionally, at this level, the modality representations do not present dimensional differences. Also, scalability is considered in terms of the number of modalities used. Nevertheless, it has the limitation of not leveraging the correlations between modalities. In this review, at this level of abstraction, this technique has been the least utilized compared to the others, with only five studies (10.20%) employing it. Among these, the most commonly applied methods include: concatenation (Braunschweiler et al., 2022) and averaging (Byun et al., 2021).

4.2.2. Feature-level fusion

At this level of abstraction, also known as early fusion, the representative features are previously extracted by different identification techniques for each unimodal source, then they are combined into a high-dimensional vector (latent representation)

to be processed by a given model (Shoumy et al., 2020). Thus, this technique combines features from acoustic modality (e.g., pitch, volume), and linguistic modality (e.g., word embeddings) with attributes either visual or drawn from other modalities, to create a comprehensive feature set that captures the complementary information from each modality.

This method presents an advantage in retrieving information on the correlations between the different modalities, and additionally, it is possible to obtain a better performance when merging additional data than unimodal ones. However, when shaping the feature representation into a high-dimensional vector, it is necessary in some cases to use selection or transformation techniques to minimize the extracted features. In the same way, another relevant problem with this technique is to determine the appropriate dimensions of the extracted feature vectors (Ahmed et al., 2023). This problem is fairly notorious in the studies of SER from acoustic and linguistic modalities, as previously stated.

Table 6 shows that recent approaches have applied techniques for feature-level fusion such as intra- and inter-modal, where the aim is to leverage the quality of single-modality representations and to capture richer context by modeling relationships across modalities, respectively (Zhang et al., 2024, 2023b), in some cases bimodal or trimodal (Huddar et al., 2021). Some studies also employ the concatenation of high-level representations from each modality as a feature-level fusion technique (Hosseini et al., 2024; Liu et al., 2022b; Zhao et al., 2024b). According to results in this review, this fusion technique is commonly employed in a hybrid manner, with DL- or attention-based methods.

4.2.3. Deep learning fusion

As the interest in employing DL- and attention-based techniques in the extraction of meaningful features has increased, it is worth noting the use of these techniques in modality fusion as well, with 67.34% of studies using these approaches. This may be attributed to DL-based fusion offers significant advantages by enabling automatic learning of complex, non-linear relationships between modalities, reducing the need for manual feature engineering. These models can adaptively weight and combine multimodal information, leading to more robust and generalizable emotion representations (Ahmed et al., 2023).

In ER (especially from acoustic and linguistic modalities), a common factor in this type of technique is to use fully connected (FC), dense or deep layers, with a softmax output layer that generates the probabilities of belonging to the emotion classes evaluated (Cai et al., 2019; Chauhan et al., 2024; Ma et al., 2024; Hosseini et al., 2024; Lian et al., 2021). Multimodal approaches in this category have also evaluated Dempster Shafer's theory (DST) for the determination of emotions (Liu et al., 2022b). Similarly, advanced RNN algorithms such as Bi-LSTM (Zhang et al., 2022) and Bi-GRU (Zhao et al., 2024b) have been employed.

Nevertheless, models based on DL fusion present notable limitations for real-time or resource-constrained applications due to their high computational demands. Furthermore, their black-box nature poses challenges to interpretability, making it difficult to understand the specific contribution of each modality to the final predictions.

4.2.4. Attention-based fusion

Attention-based fusion further enhances the fusion process by dynamically focusing on the most relevant features or modalities. By assigning learned importance weights, attention mechanisms help the model selectively integrate information, making them particularly effective in handling modality specific noise and varying levels of informativeness across inputs (Vaswani et al., 2017; Ahmed et al., 2023). This method is one of the most evaluated for MER, as can be seen in Table 6, where 29 approaches include this mechanism.

Cross-attention which plays a crucial role in capturing relevant relationships between modalities (inter-modal), represents the predominant approach; in this review, 15 out of 33 studies classified in this category used this method (Liu et al., 2022a; Pan et al., 2024; Ai et al., 2024; Kim and Kang, 2022). Similarly, studies also introduce its combination with the use of self-attention as a mechanism for extracting intra-modal features (Ma et al., 2024; Wang et al., 2023a; Luo et al., 2023; Gao et al., 2024; Zhang et al., 2023c).

Furthermore, some studies evaluated techniques based on multi-head attention, where the model captures relationships by processing multiple attention distributions in parallel, focusing on emotional cues within or across modalities (Zhang et al., 2022; Lian et al., 2021; Rasendrasoa et al., 2022). On the other hand, diverse approaches used multi-level attention, which operates across various layers of feature abstraction facilitating a hierarchical understanding of emotional content (Ho et al., 2020).

However, one of the main limitations of using this type of fusion lies in the need for careful extraction and preparation of features for each data modality. This is particularly important because such approaches typically assume that all modalities are temporally aligned and equally informative, an assumption that may not hold in real-world scenarios, especially when additional modalities such as motion or video are involved. Additionally, the dimensionality of feature representations for each modality plays a crucial role when employing mechanisms such as cross-attention. These techniques typically require feature vectors to have matching dimensions, which introduces an additional constraint. This requirement can further complicate efforts to address modality imbalance, as it may require artificial adjustments that do not reflect the true informativeness of each modality.

Hybrid approaches are also widely used, where combinations of fusion methods are studied depending on the technique employed (usually at feature level). The most common is to employ DL techniques combined with attention techniques to improve the models' performance. Typically, as mentioned above, it is also common to combine fully connected or dense networks and softmax layers with diverse attention techniques (Singh et al., 2023; Ma et al., 2024; Ho et al., 2020).

5. Benchmark analysis

This stage presents a challenge, as the results depend on variables such as the proposed approach, the data modalities, the datasets, the emotion modeling, and the number of emotions employed, as well as the evaluation metrics selected to assess the performance of the models. Thus, to make the results comparable, the metrics obtained have been organized by the datasets used. The results of the best performances included in this review are shown in Tables 7–10 for the three most used datasets IEMOCAP, MELD, MSU-MOSEI, and datasets that are evaluated only once. Each table shows the results grouped by evaluation metrics for each dataset, number and emotion classes, and data modalities used. Note that if a method uses several datasets to evaluate performance, it will be presented in the respective tables. Furthermore, depending on the metrics and approaches reported by the studies, in Tables 7–10, the metrics achieved for the independent modalities of speech acoustic or linguistic are consigned in columns (A) and (L), respectively. As well as the fusion of both in (A+L) and for the cases where more modalities are used, we detailed in (A+L+O), and the modality is described in the 'Other' column. Additionally, the Tables also report, whether the linguistic modality was derived from ASR systems, alongside the highest reported performance metrics (when this information is provided by the authors). These studies have been organized from the highest to the lowest scores achieved (where comparable) and the scores correspond to results in the test phases.

Considering that one of the main problems presented by the datasets is the class imbalance among the reported emotions. The most reported metrics in ER correspond to weighted accuracy (WA) and unweighted accuracy (UA) (Braunschweiler et al., 2022; Chauhan et al., 2024). WA calculates a weighted contribution for each class, where the weight is proportional to the number of samples belonging to that class in the dataset. In contrast, UA assigns equal importance to each class, regardless of the number of samples representing that class in the dataset.

5.1. Evaluated emotions

In this section, we will discuss our seventh research question *RQ7: What are the most common emotions evaluated in the literature of SER?*

As seen in Tables 7–10, most studies evaluate their approaches for four frequent emotions: Neutral, Happy, Sad, and Angry. This makes it somewhat easier to compare performances. However, many factors need to be considered for a fair comparative analysis.

For the IEMOCAP dataset (Table 7), most studies assess model performance using the four most frequent emotions, with the emotion 'Happy' often considered as 'Excited'. However, in certain studies, these classes are evaluated independently. In contrast, studies that use the MELD dataset (Table 8) typically report results for the recognition of seven emotions: Neutral, Sad, Surprise, Fear, Angry, Joy, and Disgust. Regarding the CMU-MOSEI dataset (Table 9), reported results are generally based on the four most frequent emotions, with three studies employing a dimensional approach to ER (Liu et al., 2022a). Similarly, recent research trends have proposed evaluating the presence of multiple co-occurring emotions in speech over the analyzed utterances (Anand et al., 2023). This method was tested using multi-label emotions from the CMU-MOSEI, EmoReact, and ElderReact datasets.

On the other hand, as shown in Tables 7–10, it can be observed that current literature employs the categorical model for ER. One of the main reasons for this tendency (previously discussed in the manuscript) is the limited availability of datasets that support dimensional modeling. The dimensional approach presents greater challenges in terms of annotation, as it requires human labelers to make continuous or scaled judgments, which are more cognitively demanding. In contrast, the categorical model offers greater simplicity and interpretability, providing discrete and intuitively understandable labels (e.g., anger or sadness), which are often more straightforward and precise than interpreting complex combinations of e.g. valence and arousal values. According to results, the evaluation of dimensional modeling, typically involves reporting metrics such as binary accuracy for positive or negative classes, seven-class accuracy for the sentiment integer score from −3 to 3, or Mean Absolute Error (MAE) since it involves a regression problem (Zhang et al., 2023a).

Finally, for studies using unique datasets, the emotions most commonly considered correspond to the four most frequent ones, with the exception of Singh et al. (2023), which evaluates a broader range of 32 emotions included in the proposed dataset.

5.2. Performances

In this section, based on the results presented in Tables 7–10, we focus on addressing the final research question *RQ8: What are the best performances in literature for SER from acoustic and linguistic modalities?* The following subsections provide a detailed description of the results for each case.

5.2.1. IEMOCAP

Table 7 reports the results achieved by the different articles using the IEMOCAP dataset. The number indicated in the first column shows the studies that are directly comparable in a block. We can see that among comparable studies, the maximum scores of 85.50%, 85.71%, and 82.67% are reported for the metrics UA, WA, and F1, respectively. These approaches mainly employ the bimodal fusion of acoustic and linguistic modalities of speech in classifying the four typical emotions (Neutral, Angry, Sad, Happy/Excited). A common feature, when making use of this dataset, is that authors consider the 'Excited' class as 'Happy'.

In particular, the one achieving the best performance proposes an architecture named CNNRNNATTBERT. In this model, the acoustic modality is processed using log-Mel filterbank features, which are fed through CNN layers to capture spatial patterns, followed by Bi-LSTM layers for temporal modeling. This is further refined by fully connected layers with batch normalization and

Table 7
Results and performance of studies with IEMOCAP dataset.

	Article	Emotions	Metric	Scores (%)				Other
				A	L	A+L	A+L+O	
1	Braunschweiler et al. (2022)	4	UA	76.40	83.80	85.50	–	–
2	Cai et al. (2024)	4	UA	72.39	60.94	80.18	–	–
3	Zhao et al. (2024a)	4	UA	–	–	79.20	–	–
4	Zhao et al. (2023)	4	UA	–	–	78.50	–	–
5	Zhang and Li (2023)	4	UA	60.30	69.10	77.80	–	–
6	Gao et al. (2024)	4	UA	77.25	67.13	77.64	–	–
7	Priyasad et al. (2023)	4	UA	–	–	77.30	–	–
8	Lin and Wang (2023)	4	UA	72.00	69.5	77.20	–	–
9	Kyung et al. (2024)	4	UA	68.21	66.88 (ASR)	77.16 (ASR)	–	–
10	Ho et al. (2020)	4	UA	66.19	56.5	76.98	–	–
11	Lin and Wang (2023)	4	UA	72.00	65.50 (ASR)	76.90 (ASR)	–	–
12	Li et al. (2024)	4	UA	–	–	76.66 (ASR)	–	–
13	Wang et al. (2023b)	4	UA	–	–	76.40	–	–
14	Kim and Cho (2023)	4	UA	77.70	74.90	76.30	–	–
15	Chen et al. (2022b)	4	UA	–	–	75.30	–	–
16	Liu et al. (2022b)	4	UA	–	–	75.05	–	–
17	Cai et al. (2019)	4	UA	–	–	71.25	–	–
18	Hosseini et al. (2024)	4	UA	–	–	69.8	–	–
1	Chauhan et al. (2024)	4	WA	76.41	–	85.71	–	–
2	Ghosh et al. (2022)	4	WA	73.90	69.20	81.20	–	–
3	Cai et al. (2019)	4	WA	–	–	70.4	–	–
4	Li et al. (2022)	4	WA	–	–	63.40 (ASR)	–	–
1	Kim and Kang (2022)	4	F1	–	–	82.67	–	–
2	Zhang et al. (2023a)	4	F1	–	–	79.2	–	–
3	Zhao et al. (2024b)	4	F1	–	–	78.8	–	–
4	Zhang et al. (2023c)	4	F1	56.06	68.96	73.06	–	–
1	Huang et al. (2024)	6	WA	58.64	71.73	72.82	74.20	Video
2	Wei et al. (2023)	6	WA	66.39	63.28	71.60	–	–
3	Yao and Shi (2024)	6	WA	–	–	71.21	–	–
4	Lian et al. (2021)	6	WA	–	–	67.5	–	–
5	Li et al. (2023)	6	WA	38.62	65.46	65.89	71.75	Video
6	Ai et al. (2024)	6	WA	61.4	63.2	65.8	–	–
1	Zhao et al. (2025)	6	WA	–	–	–	72.10	Video
2	Zhang et al. (2024)	6	WA	–	–	–	71.11	Video
3	Van et al. (2025)	6	WA	–	–	–	68.64	Video
4	Rasendrasoa et al. (2022)	6	WA	–	–	–	65.70	Video
1	Zhang et al. (2022)	4	WA	63.6	66.0	–	75.6	Motion
2	Wen et al. (2023)	4	WA	–	–	–	71.84	Image
1	Hosseini et al. (2024)	4	UA	–	–	–	82.9	Video
1	Ma et al. (2024)	6	UA	–	–	–	73.95	Video
2	Zou et al. (2023)	6	UA	46.58	68.92	70.09	72.83	Video
1	Wang et al. (2023a)	7	WA	–	–	72.31	–	–

4: Neutral, Angry, Sad, Happy/Excited.
6: Neutral, Angry, Sad, Happy, Excited, Frustrated.
7: Neutral, Angry, Sad, Happy, Excited, Frustrated, Surprise.

an attention mechanism. The linguistic modality is encoded using a pretrained BERT model, where the output representation is extracted using the *CLS pooled* method. The outputs from both encoders are concatenated and passed through a softmax layer for final emotion classification. Additionally, the study compares model performance using manual transcriptions versus those generated by the Google ASR system, showing that the use of hand transcriptions yields superior results.

Moreover, since this dataset additionally incorporates data modalities such as video and motion (video more frequent), the scores achieved in the classification tasks for four emotions do not show higher scores than those obtained with acoustic and linguistic features, reaching 82.9% and 75.6% for UA and WA metrics, respectively.

5.2.2. MELD

Table 8 reports the evaluation metrics achieved by the approaches evaluating the MELD dataset. In this case, researchers generally use the 7 emotions provided (Neutral, Sad, Surprise, Fear, Angry, Joy, Disgust). This could be one reason why the metrics evaluated achieve lower scores compared to IEMOCAP. According to results for the two modalities, the maximum ones reach 66.28%, 65.09%, and 64.71% for WA, UA, and F1, respectively. The authors who achieved state-of-the-art performance on this dataset proposed a novel architecture called Cross-Modal RoBERTa (CM-RoBERTa), designed to enhance SER by integrating audio information into the fine-tuning process of a pretrained RoBERTa model. The architecture incorporates cross- and self-attention layers to model inter-

Table 8
Results and performance of studies with MELD dataset.

	Article	Emotions	Metric	Scores (%)				Other
				A	L	A+L	A+L+O	
1	Luo et al. (2023)	7	WA	43.10	63.83	66.28	–	–
2	Yao and Shi (2024)	7	WA	–	–	66.25	–	–
3	Huang et al. (2024)	7	WA	42.51	63.49	64.37	66.09	Video
4	Wei et al. (2023)	7	WA	62.97	48.14	63.92	–	–
5	Ai et al. (2024)	7	WA	62.10	62.80	63.80	66.80	Video
6	Wang et al. (2023a)	7	WA	–	–	48.12	–	–
1	Zhang et al. (2023a)	7	UA	–	–	65.09	–	–
2	Ho et al. (2020)	7	UA	48.84	61.66	63.26	–	–
3	Zou et al. (2023)	7	UA	44.12	61.89	62.84	65.86	Video
1	Zhao et al. (2024b)	7	F1	–	–	64.7	–	–
2	Zhang and Li (2023)	7	F1	36.7	59.80	62.2	–	–
3	Zhang et al. (2023c)	7	F1	40.5	54.88	60.22	–	–
1	Ma et al. (2024)	7	UA	–	–	–	67.55	Video
1	Zhao et al. (2025)	7	WA	–	–	–	66.68	Video
2	Rasendrasoa et al. (2022)	7	WA	–	–	–	66.00	Video
3	Van et al. (2025)	7	WA	–	–	–	65.69	Video
1	Xie et al. (2021)	7	F1	43.80	61.80	–	64.00	Image
1	Li et al. (2023)	5	WA	41.72	64.54	64.95	67.03	Video
1	Kim and Kang (2022)	4	F1		–	83.41	–	–

4: Neutral, Happy, Sad, Angry.
5: Neutral, Happy, Sad, Angry, Surprise.
7: Neutral, Sad, Surprise, Fear, Angry, Joy, Disgust.

Table 9
Results and performance of studies with CMU-MOSEI dataset.

	Article	Emotions	Metric	Scores (%)				Other
				A	L	A+L	A+L+O	
1	Braunschweiler et al. (2022)	4	UA	77.40	78.40	79.00	–	–
1	Kim and Kang (2022)	4	F1	–	–	80.18	–	–
1	Huddar et al. (2021)	6	UA	–	–	–	81.29	Video
1	Liu et al. (2022a)	Dim.	UA	–	–	81.1	–	–
1	Zhao et al. (2024a)	Dim.	F1	–	–	84.1	–	–
2	Zhao et al. (2023)	Dim.	F1	–	–	83.8	–	–
1	Anand et al. (2023)	Mult.	F1	73.17	84.75	85.04	89.04	Video

4: Neutral, Happy, Sad, Angry.
6: Sad, Surprise, Fear, Angry, Happy, Disgust.
Dim: Dimensional from strongly negative to strongly positive.
Mult.: Multiple co-occurring emotion states.

and intra-modality interactions, enabling adaptive learning of correlations between linguistic and acoustic features. To further refine bimodal representation, the authors include a temporal calibration module, which preserves both local information within each modality and global contextual information.

In this case, when using another type of data modality such as video or image, the performances show better results, concerning the unimodal or bimodal models.

5.2.3. CMU-MOSEI

Table 9 shows the metrics achieved by the studies using the CMU-MOSEI dataset. In this case, as noted in Table 4, this dataset has been used less frequently compared to IEMOCAP and MELD. Consequently, there are fewer studies evaluating approaches on this dataset, which limits the possibility of conducting performance comparisons. Nevertheless, it is worth mentioning that, using acoustic and linguistic features for the recognition of four emotions, existing works have reported scores of 79% and 80.10% for UA and F1, respectively. Notably, the authors who proposed the CNNRNNATTBERT architecture (Braunschweiler et al., 2022) (which achieved state-of-the-art performance on the IEMOCAP) also obtained strong performance in ER on CMU-MOSEI.Furthermore, being a database that provides dimensional modeling, studies have also evaluated their approaches for the prediction from strongly negative to strongly positive scale, achieving highest scores of 81.1% and 84.1% for the UA and F1 metrics, respectively. The overall metrics achieved do not exceed those with the IEMOCAP dataset, nor are they lower than those with the MELD dataset. Similarly, the authors in Anand et al. (2023) propose the use of the CMU-MOSEI dataset in their approach to multi-label ER, based on existing evidence that humans are capable of experiencing multiple emotions simultaneously, as outlined in their work. They argue that methods limited to estimating a single emotion may reduce applicability in real-world conversational contexts. Their

Table 10
Results and performance of studies with unique datasets.

	Article	Dataset	Emotions	Metric	Scores (%)				Other
					A	L	A+L	A+L+O	
1	Chauhan et al. (2024)	RAVDESS	4	WA	93.35	–	76.71	–	–
1	Braunschweiler et al. (2022)	RAVDESS	4	UA	85.9	–	75.6	–	–
1	Chauhan et al. (2024)	EmoDB	4	WA	93.47	–	79.52	–	–
1	Chauhan et al. (2024)	CREMA-D	4	WA	61.3	–	73.96	–	–
1	Chen et al. (2022b)	LSSED	4	WA	–	–	65.10	–	–
1	Byun et al. (2021)	Korean Emotional Speech Dataset	4	UA	94.86	68.11	95.97	–	–
1	Pan et al. (2024)	Spanish MEACorpus	6	F1	84.48	69.38	87.74	–	–
1	Zhao et al. (2024b)	MTED	6*	F1	–	–	81.7	–	–
1	Singh et al. (2023)	EmoInt-MD	32	UA	–	–	–	69.38	Video
1	Zhang et al. (2023a)	CMU-MOSI	Dim.	UA	–	–	89.33	–	–
1	Li et al. (2024)	CMU-MOSI	Dim.	UA	–	–	83.23 (ASR)	–	–
1	Huddar et al. (2021)	CMU-MOSI	Dim	UA	–	–	79.71	–	–
1	Anand et al. (2023)	EmoReact	Mult.	F1	86.03	–	–	–	–
1	Anand et al. (2023)	ElderReact	Mult.	F1	85.29	–	–	–	–
1	Li et al. (2024)	MSP-Podcast	–	CCC	–	–	61.80 (ASR)	–	–

4: Neutral, Happy, Sad, Angry.
6: Neutral, Happy, Sad, Angry, Fear, Disgust.
6*: Neutral, Happy, Sad, Excited, Confused, Surprise.
32: described in Table 4.
Dim: Dimensional Positive and negative.
Mult: Multiple co-occurring emotion states.

approach achieved F1-scores of 85.04% and 89.04% for ER using acoustic and linguistic modalities, and with the inclusion of video data, respectively.

5.2.4. Datsets used once

The last Table 10 lists all datasets that are used less frequently (mostly only once) by the studies included in this review. Within this category, only the results obtained using the CMU-MOSEI dataset are directly comparable. In this context, authors typically report performance metrics related to the prediction of positive or negative sentiment states. The methodology proposed in Zhang et al. (2023a) achieves the highest UA, reaching 89.33%. Most studies report metrics obtained in the evaluation of model performance for bimodal approaches involving acoustic and linguistic modalities from speech. Likewise, as in the previous cases, the most commonly evaluated emotions correspond to the four typical categories. Additionally, it is worth noting that this category includes datasets consisting of data in the languages: Spanish, German, Chinese, and Korean. Hence, it is important to explore ER methods that can classify emotions in languages other than English.

6. Discussion

This section discusses existing challenges and potential opportunities of the main topics covered in this review regarding datasets, approaches, and benchmark analysis used in SER.

6.1. Datasets

As shown in previous sections, high-quality datasets are necessary, as the quality of the trained models will depend on them. One of the main challenges faced when creating ER datasets is to determine a uniform labeling scheme. This means that the use of different emotion labels across datasets often results in cross-corpus emotion resources being incompatible. For this reason, the latest research reports the results generally evaluated for the four most frequent emotions. Nevertheless, these emotion labels reduce the data, which could provide valuable information for ER models. Hence, it is important to define benchmarks in the labeling schemes, as well as to explore possible relationships between the data provided by the existing corpora.

On the other hand, another challenge faced by methods in ER corresponds to unbalanced databases. In ER, it is typically assumed that the cost of misclassifying each emotion label is equal. However, in scenarios involving data imbalance, it is often desirable for the classifier to focus more on accurately identifying minority emotion labels. The recognition accuracy for certain emotions tends to be lower due to the limited amount of training data available. To address this, some approaches aim to adapt the classification algorithm by, for example, altering the loss function to prioritize minority classes during the training process, this is a common technique known as class weighting. Hence, future research should prioritize techniques aimed at addressing class imbalance among emotional categories. Common approaches such as under-sampling and over-sampling may contribute to improved performance in emotion classification tasks by ensuring more balanced model training.

Annotating an emotional corpus is a complex and challenging task, even for humans. While most existing emotional datasets are manually labeled, large-scale emotional corpora remain limited. Although manual annotation can provide a certain level

of accuracy, the process is highly time- and labor-intensive. Additionally, the subjective nature of annotators' perceptions often leads to inconsistencies in inter-annotator agreement. The scarcity of high-quality emotional datasets poses a significant challenge. Addressing this low-resource issue and developing efficient methods for creating large, high-quality emotional corpora are crucial for enhancing the performance of ER models.

As shown in Table 4, most of the available corpora are developed in English, and, to a lesser extent for other languages such as Spanish, German, Chinese, and Korean. Hence, some research has addressed this challenge by taking advantage of emotional information from corpora with greater resources such as English (Deng and Ren, 2021). Thus, future work should focus on transferring emotional knowledge from English corpora to languages that are more limited in this information. Using techniques such as cross-lingual embeddings would facilitate the capture of this relevant information across languages (Deng and Ren, 2021).

According to Anand et al. (2023), future studies should also consider labeling processes and evaluating the co-occurrence of multiple emotions in speech simultaneously. This recommendation is supported by existing evidence indicating that humans are capable of experiencing more than one emotion at the same time.

Future research should also explore approaches that incorporate both categorical and dimensional emotion modeling. Such efforts could have the potential to enhance the robustness of ER systems by leveraging the complementary strengths of both frameworks. Notably, among the datasets reviewed, only five provide annotations for both types of emotional representation, highlighting the need for further investigation in this area.

6.2. Approaches

From Table 6, different findings can be observed, both in low-level and high-level feature extraction in acoustic and linguistic speech, as well as modality fusion methods. In the two primary stages of ER, most studies focus on employing hybrid approaches that are characterized by using DL- and attention-based methods. Bi-LSTM and Bi-GRU-based RNNs, when integrated with CNNs, are widely utilized and well-suited for extracting emotional attributes in supervised learning scenarios. Nevertheless, in ER, semi-supervised learning applications have not been explored so far. Further research is required to explore this application, enabling effective representation learning from both labeled and unlabeled data, this also helps to reduce the time spent on the emotion labeling process.

Considering that emotions depend on context-based information, unlike linguistic, where techniques such as BERT, RoBERTa, and GPT, among others, have been widely used. In speech, future research to extract context-dependent features should be explored, in this review some studies evaluate these techniques using Wav2Vec. Nevertheless, proposed methods to increase the knowledge of audio patterns to leverage speech representation learning such as AudioBERT (Ok et al., 2024) and Audio DistilBERT (Yu et al., 2021) have not yet been explored.

Similarly, as mentioned in Section 4.1.2, some embeddings specifically represent emotional information in their respective latent space. Although, so far, these models have not been explored in depth for the representation of linguistic features. Hence, from the perspective of a multimodal approach, studies should examine the construction of embeddings that represent emotional information from different data modalities. Taking advantage of the correlations that may exist between these multiple information sources.

The techniques discussed in this review do not provide a definitive approach for determining the most effective fusion method for identifying emotional states from SER modalities. Most techniques focus on employing attention-based methods. However, there is no clear line to build a robust model with appropriate performance in ER.

As mentioned in Section 4.1.2, there is also a marked discrepancy in the dimensionality of the feature vector representation for both acoustic and linguistic modalities. This is an issue that needs to be addressed when defining the fusion technique since if it is based on feature level, this difficulty needs to be solved. Particularly, in low-dimensional acoustic features versus high-dimensional linguistic embeddings (e.g. 40-D of MFCCs, to 1024-D by RoBERTa model (Pan et al., 2024; Kim and Cho, 2023)) the fusion technique may disproportionately prioritize the modality with higher dimensionality, not necessarily because it is more informative, but due to its greater numerical weight. This imbalance can lead to suboptimal performance, as the contributions of lower-dimensional but potentially valuable modalities may be underrepresented.

Additionally, it is important to emphasize that the application of emerging feature fusion techniques, particularly those based on cross-attention mechanisms, often requires equality in the dimension of input feature representations. This poses a significant challenge when attempting to fuse low-level and high-level features across modalities. On the other hand, based on the findings of this review, the combination of high-level features extracted from widely used pre-trained models, such as Wav2Vec, Wav2Vec2.0, WavLM, and HuBERT for the acoustic modality, and contextual embeddings like BERT and RoBERTa for the linguistic modality, reveals certain difficulties when the modalities differ in dimensionality (Pan et al., 2024; Zhang et al., 2023a), since their dimensionality is highly dependent on the specific version of the model. However, it was found that studies are emerging to employ models that allow equality in dimensionalities. A commonly adopted is 768-D, which is supported by models such as Wav2Vec BERT, Wav-RoBERTa, Wav2Vec 2.0, HuBERT, and WavLM for acoustic features, and by RoBERTa and BERT-base for linguistic features (Kim and Kang, 2022; Lin and Wang, 2023; Priyasad et al., 2023; Cai et al., 2024; Zhao et al., 2023). This standardization offers a significant advantage when employing high-level representations across both modalities, as they facilitate a more straightforward adoption of fusion techniques.

In particular, to develop more effective automatic systems that enhance the performance of SER from both modalities, while leveraging existing ASR technologies, promising future directions should focus on evaluating transcription errors when linguistic modality is generated. As suggested by recent studies, understanding and mitigating the impact of ASR errors is essential for improving the reliability and robustness of ER systems in real-world applications.

In current trends, exploring retrieval-augmented generation (RAG) models in ER also presents a promising avenue for advancing the field (Fan et al., 2024; Zhai, 2024). RAG models combine generative capabilities with retrieval mechanisms, allowing them to integrate external knowledge and context into the ER process (Guu et al., 2020). In MER, which involves analyzing multiple data modalities, the ability to retrieve relevant contextual information could significantly enhance model performance. For instance, RAG models can access domain-specific knowledge or situational cues to better interpret ambiguous emotional expressions. Moreover, their generative nature allows them to synthesize insights across modalities, fostering a more comprehensive understanding of complex emotional states.

Similarly, the rise of generative AI has opened new avenues for the development of approaches applied to ER, although such applications remain limited at present. Among the studies included in this review, some authors propose the use of prompt-based techniques during the fusion stage (Zou et al., 2023), as well as to guide a Large Language Model (LLM) to generate the most likely transcript generated by ASR systems (Li et al., 2024). These early explorations underscore the potential of leveraging emerging generative technologies to enhance the performance and adaptability of SER systems.

Regarding fusion techniques, it is important to note that one of the main challenges in ER lies in the imbalance of available data. DL-based approaches can be particularly vulnerable to this issue if not properly addressed, potentially leading to biased models and degraded performance. Therefore, it is essential that fusion strategies incorporate mechanisms to mitigate the effects of data imbalance, ensuring fair and robust learning across all emotion classes. Future work should also focus on reducing the computational cost associated with both DL- and attention-based fusion techniques. Although these methods have demonstrated strong performance in SER, as evidenced by this review, they typically require substantial computational resources for both training and inference. This high demand may limit their feasibility for real-time or resource-constrained applications, underscoring the need for more efficient and lightweight fusion strategies.

6.3. Benchmark analysis

The most common datasets used for SER using acoustic and linguistic modalities correspond to IEMOCAP, MELD, and CMU-MOSEI. Usually, studies model emotions using a discrete approach, such as Ekman's six basic emotions, which fails to capture the complexity and diversity of human emotions. In reality, emotions often exhibit interconnections and lack clear boundaries, making it challenging to assign precise labels to emotional expressions. Additionally, emotions are inherently subjective, different emotional responses may be evoked in individuals based on their unique experiences. Also, guidelines should be defined to determine which emotions can be treated as similar. For example, in some cases, it is common to consider the 'Happy' label as 'Excited'. In other datasets, however, joy is considered different from happiness. As some research has already been exploring, one gap in dealing with this issue focuses on assigning multiple emotion labels simultaneously (Deng and Ren, 2020). These factors, including ambiguous emotional boundaries and the subjective nature of human feelings, contribute to the difficulty of ER. It is noteworthy that these cases have not yet been considered in ER datasets.

As previously established, most studies evaluate their approaches based on the recognition of the four most frequent emotions (Neutral, Happy, Sad, Angry). However, by relying exclusively on data from these classes, valuable information contained in the less frequent emotions is disregarded. Achieving high accuracy on limited, controlled datasets may not translate to real-world scenarios where emotions are more fluid and context-dependent. Therefore, it is essential to assess performance not only for the most common emotions but also for the remaining ones, depending on the dataset used. This comprehensive evaluation may enable a more thorough assessment of the generalization capacity of the proposed approaches in ER, improving its applicability to better align with real-world scenarios.

Likewise, according to the information presented in Tables 7 and 8,it can be observed that adding an additional modality (Other) to the proposed multimodal approaches generally leads to improved performance metrics compared to those obtained using each individual modality alone (where such comparisons are available and reported). However, it is important to emphasize that, in the case of the IEMOCAP dataset, the inclusion of an additional modality does not surpass the best results achieved by the bimodal approaches that combine acoustic and linguistic features. For instance, in the bimodal recognition of four emotions, the proposed approaches have achieved UA and WA scores of 85.50% (Braunschweiler et al., 2022) and 85.71% (Chauhan et al., 2024), respectively. These results are higher than the UA score of 82.9% (Hosseini et al., 2024) (when incorporating video) and the WA score of 75.6% (Zhang et al., 2022) (when including motion)

In contrast, for the MELD dataset, the inclusion of the video modality (Zhao et al., 2025) resulted in a higher WA of 66.68%, compared to the 66.28% achieved by the corresponding bimodal approach (Luo et al., 2023). This highlights the importance of carefully evaluating whether a marginal improvement (such as the 0.4% gain observed with the inclusion of an additional modality) justifies the increased hardware requirements needed to support it. Moreover, as previously discussed, the use of modalities such as video and motion poses greater technological challenges and limitations in terms of adaptability to real-world applications.

7. Conclusion

This review addresses the research that has been developed in recent years in SER, focusing on acoustic and linguistic modalities. In this way, the present work will be of interest to those working in these fields, as well as a basis for those interested in starting.

The main research question that this review addresses is *what is the current state-of-the-art of SER based on its acoustic and linguistic modalities?* This question was answered through 8 subquestions discussed in three main sections: datasets, approaches, and benchmark analysis. The main insights for each subquestion are provided below.

RQ1: What is the most commonly used emotion modeling in the SER field?
As established, the two most widely used theories in the literature are the discrete categories (Paul Ekman's six basic emotions most widely used) and the continuous one, where emotions are represented as a multidimensional vector (Valence and Arousal). Most studies in SER use discrete categorization.
RQ2: What is the predominant type of speech in datasets used for ER?
The datasets most used by the studies included in ER correspond to the IEMOCAP, MELD, and CMU-MOSEI (all of them in English), which were used by 46.99%, 26.51% and 8.43% of the studies in this review, respectively.
RQ3: What further data modalities have been integrated with linguistic and acoustic modalities of speech for ER?
Most datasets integrate video data with acoustic and linguistic modalities. However, other modalities such as motion are also evaluated. Thus, all four modalities are provided by the IEMOCAP dataset, making it a fairly complete model for evaluating ER models.
RQ4: What technology and devices have been used to construct datasets for SER?
In recent years, most of the datasets that have been created leverage audio-visual material from YouTube videos, series, movies, or podcasts. Thus, the technological devices used are unknown. Nevertheless, for older datasets, devices are used that were considered to be high-resolution technologies at the time of recording. In audio data acquisition, microphones use a common sampling rate of 48 kHz and 16 bits and are usually positioned at a distance of 20 to 30 cm from the speaker.
RQ5: What are the most representative low- and high-level features of linguistic and acoustic modalities for SER? And what is the most common feature vector dimension?
In terms of low-level features, the most commonly used in acoustic modality are those coming from features such as zero crossing rate, energy, and pitch frequency; among the spectral/ cepstral features, the most frequent one corresponds to MFCCs. Additionally, some studies use sets of features in both the time and frequency domain, known as GeMAPs and eGeMAPs. Where the most common dimension is 34-D. As for high-level features in acoustic modality, most focus on integrating both spatial and temporal features, hence DL-based techniques such as BiLSTM, Bi-GRU, and CNN-BiLSTM are widely used. Attention-based methods and their combination with DL techniques such as the Bi-LSTM-Attention architecture are also being explored. Contextual embeddings, such as the well-known Wav2Vec, stand out among the proposed approaches.
For the linguistic modality, low-level features are extracted by earlier embeddings using the well-known GloVe and Word2Vec models with a common feature vector of 300-D. Nevertheless, it is important to highlight that recent trends indicate a shift away from non-contextual (low-level) embedding techniques toward models that provide rich contextual information. Among these, RoBERTa has emerged as the most widely adopted, reflecting its effectiveness in capturing linguistic features critical for ER tasks. Models such as BERT and GPT are also evaluated. Finally, some studies present common DL techniques such as RNN, LSTM, and BiLSTM combined with attention-based methods as well.
RQ6: Which fusion techniques have been applied to develop SER systems from their acoustic and linguistic modalities?
In terms of the methodologies employed by the studies, there is no uniform structure in which a guideline in the fusion of modalities is determined. However, the latest advances focus on using DL-based architectures, among the most frequent, fully connected, or dense layers, Bi-GRU and Bi-LSTM. Regarding attention-based techniques, recent studies increasingly leverage cross-attention and self-attention mechanisms to facilitate both intra- and inter-modal fusion across modalities. Additionally, multi-head and multi-level attention strategies are widely employed, as they enable the modeling of complex relationships within and between modalities.
RQ7: What are the most common emotions evaluated in the literature of SER?
For the IEMOCAP dataset, most studies evaluate their approaches based on the recognition of the four most frequent emotions (Neutral, Angry, Sad, Happy/Excited.). In contrast, for the MELD dataset, researchers typically consider the seven emotions provided (Neutral, Sad, Surprise, Fear, Angry, Joy, Disgust).
RQ8: What are the best performances in literature for SER from acoustic and linguistic modalities?
Among comparable studies using the bimodal fusion from acoustic and linguistic modalities of speech, for the classification of the four typical emotions provided by IEMOCAP (Neutral, Angry, Sad, Happy/Excited), the metrics reported maximum scores of 85.50%, 85.71%, and 82.67% for the metrics UA, WA, and F1, respectively. For ER provided by MELD (Neutral, Sad, Surprise, Fear, Angry, Joy, Disgust), the studies reach 66.28%, 65.09%, and 64.71% for WA, UA, and F1, respectively. For both, when using another type of data modality such as video or image, the performances show better results.

CRediT authorship contribution statement

Andrea Chaves-Villota: Writing – original draft, Investigation, Writing – review & editing, Methodology, Conceptualization. **Ana Jimenez-Martín:** Supervision, Project administration, Funding acquisition, Validation, Resources, Methodology. **Mario Jojoa-Acosta:** Writing – review & editing, Formal analysis, Supervision. **Alfonso Bahillo:** Writing – review & editing, Resources, Supervision, Project administration. **Juan Jesús García-Domínguez:** Funding acquisition, Project administration, Resources.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author used DeepL and ChatGPT to improve the grammar and readability of some sentences. After using these tools/services, the author reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Andrea Chaves-Villota reports financial support was provided by University of Alcala. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the following projects: FrailAlert (SBPLY/21/180501/000216, co-financing from both the Junta de Comunidades de Castilla-La Mancha, Spain and the European Union through the European Regional Development Fund), ActiTracker (TED2021-130867B-I00 funded by MCIN/AEI, Spain/10.13039/501100011033 and by European Union NextGenerationEU/PRTR), INDRI (PID2021-122642OB-C41 /AEI/10.13039/501100011033/ FEDER, Spain, UE), and by the Spanish Ministry of Science and Innovation, Spain under the Aginplace project (ref. PID2023-146254OB-C41)

Appendix. Search strategy

See Table A.11.

Table A.11
Search strategy keywords.

Topic	Related keywords
Recognition task	Identification OR recognition OR classification OR regression OR clustering OR categorization OR grouping OR measurement OR Quantification OR Evaluation OR discrimination
AI-based techniques	‘‘Deep learning’’ OR ‘‘Machine learning’’ OR ‘‘Neural Networks’’ OR ‘‘Deep Neural Networks’’ OR ‘‘Convolutional Neural Networks’’ OR ‘‘Recurrent Neural Networks’’ OR ‘‘Supervised learning’’ OR ‘‘Unsupervised learning’’ OR ‘‘Predictive model’’ OR Reinforcement Learning OR Feature extraction OR Natural Language Processing OR Image Recognition OR ‘‘Pattern recognition’’ OR ‘‘Computer vision’’ OR Pytorch OR Tensorflow OR ScikitLearn OR HuggingFace.
Application field	Emotions OR feelings OR sentiments OR mood OR sensation OR temperament OR temper OR ‘‘emotional state’’ OR ‘‘mental state’’ OR ‘‘emotional condition’’
Data modality	Speech OR ‘‘speech patterns’’ OR ‘‘speech features’’ OR conversation OR speaking OR communication OR audio OR ‘‘verbal expression’’ OR dialogue OR expression OR linguistic OR sound OR ‘‘voice features’’ OR ‘‘vocal cues’’.
Approach	Bimodal OR Multimodal OR ‘‘Multi-modal’’ OR Multimodel

Data availability

No data was used for the research described in the article.

References

Abdullah, S.M.S.A., Ameen, S.Y.A., Sadeeq, M.A., Zeebaree, S., 2021. Multimodal emotion recognition using deep learning. J. Appl. Sci. Technol. Trends 2 (01), 73–79.

Ahmed, N., Aghbari, Z.A., Girija, S., 2023. A systematic survey on multimodal emotion recognition using learning algorithms. Intell. Syst. Appl. 17, 200171. http://dx.doi.org/10.1016/j.iswa.2022.200171, URL https://www.sciencedirect.com/science/article/pii/S2667305322001089.

Ai, W., Shou, Y., Meng, T., Li, K., 2024. DER-GCN: Dialog and event relation-aware graph convolutional neural network for multimodal dialog emotion recognition. IEEE Trans. Neural Netw. Learn. Syst. 1–14. http://dx.doi.org/10.1109/TNNLS.2024.3367940.

Amstadter, A., 2008. Emotion regulation and anxiety disorders. J. Anxiety Disord. 22 (2), 211–221. http://dx.doi.org/10.1016/j.janxdis.2007.02.004, URL https://www.sciencedirect.com/science/article/pii/S0887618507000679.

Anand, S., Devulapally, N.K., Bhattacharjee, S.D., Yuan, J., 2023. Multi-label emotion analysis in conversation via multimodal knowledge distillation. In: Proceedings of the 31st ACM International Conference on Multimedia. MM ’23, Association for Computing Machinery, New York, NY, USA, pp. 6090–6100. http://dx.doi.org/10.1145/3581783.3612517.

Atmaja, B.T., Sasou, A., Akagi, M., 2022. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. Speech Commun. 140, 11–28. http://dx.doi.org/10.1016/j.specom.2022.03.002, URL https://www.sciencedirect.com/science/article/pii/S0167639322000413.

Baevski, A., Zhou, Y., Mohamed, A., Auli, M., 2020. Wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. 33, 12449–12460.

Bourke, C., Douglas, K., Porter, R., 2010. Processing of facial emotion expression in major depression: A review. Aust. N. Z. J. Psychiatry 44, 681–696. http://dx.doi.org/10.3109/00048674.2010.496359.

Braunschweiler, N., Doddipatla, R., Keizer, S., Stoyanchev, S., 2022. Factors in emotion recognition with deep learning models using speech and text on multiple corpora. IEEE Signal Process. Lett. 29, 722–726. http://dx.doi.org/10.1109/LSP.2022.3151551.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., et al., 2005. A database of German emotional speech. In: Interspeech. Vol. 5, pp. 1517–1520.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S., 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. 42, 335–359.

Byun, S.-W., Kim, J.-H., Lee, S.-P., 2021. Multi-modal emotion recognition using speech features and text-embedding. Appl. Sci. 11 (17), 7967.

Cai, L., Hu, Y., Dong, J., Zhou, S., 2019. Audio-textual emotion recognition based on improved neural networks. Math. Probl. Eng. 2019 (1), 2593036.
Cai, Y., Wu, Z., Jia, J., Meng, H., 2024. LoRA-MER: Low-rank adaptation of pre-trained speech models for multimodal emotion recognition using mutual information. In: Proc. Interspeech 2024. pp. 4658–4662.
Cambria, E., Das, D., Bandyopadhyay, S., Feraco, A., 2017. Affective computing and sentiment analysis. In: A Practical Guide to Sentiment Analysis. Springer, pp. 1–10.
Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R., 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Trans. Affect. Comput. 5 (4), 377–390.
Chauhan, K., Sharma, K.K., Varma, T., 2024. Multimodal emotion recognition using contextualized audio information and ground transcripts on multiple datasets. Arab. J. Sci. Eng. 49 (9), 11871–11881.
Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F., 2022a. WavLM: Large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process. 16 (6), 1505–1518. http://dx.doi.org/10.1109/JSTSP.2022.3188113.
Chen, W., Xing, X., Xu, X., Yang, J., Pang, J., 2022b. Key-sparse transformer for multimodal speech emotion recognition. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 6897–6901. http://dx.doi.org/10.1109/ICASSP43922.2022.9746598.
Deng, J., Ren, F., 2020. Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning. IEEE Trans. Affect. Comput. 14 (1), 475–486.
Deng, J., Ren, F., 2021. A survey of textual emotion recognition and its challenges. IEEE Trans. Affect. Comput. 14 (1), 49–67.
Devlin, J., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
Du, G., Zeng, Y., Su, K., Li, C., Wang, X., Teng, S., Li, D., Liu, P.X., 2022. A novel emotion-aware method based on the fusion of textual description of speech, body movements, and facial expressions. IEEE Trans. Instrum. Meas. 71, 1–16. http://dx.doi.org/10.1109/TIM.2022.3204940.
Ekman, P., Friesen, W.V., 1971. Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. 17 (2), 124.
El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognit. 44 (3), 572–587.
Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., et al., 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. 7 (2), 190–202.
Fahad, M.S., Ranjan, A., Yadav, J., Deepak, A., 2021. A survey of speech emotion recognition in natural environment. Digit. Signal Process. 110, 102951.
Fan, W., Xu, X., Xing, X., Chen, W., Huang, D., 2021. LSSED: A large-scale dataset and benchmark for speech emotion recognition. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 641–645. http://dx.doi.org/10.1109/ICASSP39728.2021.9414542.
Fan, Q., Yuan, H., Zuo, H., Liu, R., Gao, G., 2024. Leveraging retrieval augment approach for multimodal emotion recognition under missing modalities. arXiv:2410.02804. URL https://arxiv.org/abs/2410.02804.
Flanagan, O., Chan, A., Roop, P., Sundram, F., 2020. Using acoustic speech patterns from smartphones to investigate mood disorders: A scoping review (Preprint). JMIR mHealth uHealth 9, http://dx.doi.org/10.2196/24352.
Furui, S., 1986. Speaker-independent isolated word recognition based on emphasized spectral dynamics. In: ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 11, IEEE, pp. 1991–1994.
Gao, Y., Shi, H., Chu, C., Kawahara, T., 2024. Speech emotion recognition with multi-level acoustic and semantic information extraction and interaction. In: Proc. Interspeech 2024. pp. 1060–1064.
Geetha, A.V., Mala, T., Priyanka, D., Uma, E., 2024. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. Inf. Fusion 105, 102218. http://dx.doi.org/10.1016/j.inffus.2023.102218, URL https://www.sciencedirect.com/science/article/pii/S1566253523005341.
Ghosh, S., Tyagi, U., Ramaneswaran, S., Srivastava, H., Manocha, D., 2022. Mmer: Multimodal multi-task learning for speech emotion recognition. arXiv preprint arXiv:2203.16794.
Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.-W., 2020. REALM: retrieval-augmented language model pre-training. In: Proceedings of the 37th International Conference on Machine Learning. ICML '20, JMLR.org.
Haq, S., Jackson, P., 2010. In: Wang, W. (Ed.), Machine Audition: Principles, Algorithms and Systems. IGI Global, Hershey PA, pp. 398–423, Ch. Multimodal Emotion Recognition.
Hashem, A., Arif, M., Alghamdi, M., 2023. Speech emotion recognition approaches: A systematic review. Speech Commun. 102974.
Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al., 2017. CNN architectures for large-scale audio classification. In: 2017 Ieee International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 131–135.
Ho, N.-H., Yang, H.-J., Kim, S.-H., Lee, G., 2020. Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. IEEE Access 8, 61672–61686.
Hosseini, S.S., Yamaghani, M.R., Poorzaker Arabani, S., 2024. Multimodal modelling of human emotion using sound, image and text fusion. Signal Image Video Process. 18 (1), 71–79.
Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Trans. Audio Speech Lang. Proc. 29, 3451–3460. http://dx.doi.org/10.1109/TASLP.2021.3122291.
Huang, Z., Mak, M.-W., Lee, K.A., 2024. MM-NodeFormer: Node transformer multimodal fusion for emotion recognition in conversation. In: Proc. Interspeech 2024. pp. 4069–4073.
Huddar, M.G., Sannakki, S.S., Rajpurohit, V.S., 2021. Attention-based multi-modal sentiment analysis and emotion detection in conversation using RNN.
Imani, M., Montazer, G.A., 2019. A survey of emotion recognition methods with emphasis on E-Learning environments. J. Netw. Comput. Appl. 147, 102423.
Ingale, A.B., Chaudhari, D., 2012. Speech emotion recognition. Int. J. Soft Comput. Eng. (IJSCE) 2 (1), 235–238.
Jahangir, R., Teh, Y.W., Hanif, F., Mujtaba, G., 2021. Deep learning approaches for speech emotion recognition: State of the art and research challenges. Multimedia Tools Appl. 80 (16), 23745–23812.
Jiang, Y., Li, W., Hossain, M.S., Chen, M., Alelaiwi, A., Al-Hammadi, M., 2020. A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. Inf. Fusion 53, 209–221. http://dx.doi.org/10.1016/j.inffus.2019.06.019, URL https://www.sciencedirect.com/science/article/pii/S1566253519301381.
Joulin, A., Grave, E., Bojanowski, P., Mikolov, T., 2016. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.
Kalateh, S., Estrada-Jimenez, L.A., Nikghadam-Hojjati, S., Barata, J., 2024. A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges. IEEE Access 12, 103976–104019. http://dx.doi.org/10.1109/ACCESS.2024.3430850.
Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., van den Oord, A., Dieleman, S., Kavukcuoglu, K., 2018. Efficient neural audio synthesis. In: Dy, J., Krause, A. (Eds.), Proceedings of the 35th International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 80, PMLR, pp. 2410–2419, URL https://proceedings.mlr.press/v80/kalchbrenner18a.html.
Kaya, H., Gürpınar, F., Salah, A.A., 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image Vis. Comput. 65, 66–75.
Kim, K., Cho, N., 2023. Focus-attention-enhanced crossmodal transformer with metric learning for multimodal speech emotion recognition. In: Proc. Interspeech 2023. pp. 2673–2677.
Kim, D., Kang, P., 2022. Cross-modal distillation with audio–text fusion for fine-grained emotion classification using BERT and Wav2vec 2.0. Neurocomputing 506, 168–183.

Kleinsmith, A., Bianchi-Berthouze, N., 2012. Affective body expression perception and recognition: A survey. IEEE Trans. Affect. Comput. 4 (1), 15–33.

Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., Wróbel, M.R., 2014. Emotion recognition and its applications. In: Hippe, Z.S., Kulikowski, J.L., Mroczek, T., Wtorek, J. (Eds.), Human-Computer Systems Interaction: Backgrounds and Applications 3. Springer International Publishing, Cham, pp. 51–62. http://dx.doi.org/10.1007/978-3-319-08491-6_5.

Koolagudi, S.G., Rao, K.S., 2012. Emotion recognition from speech: a review. Int. J. Speech Technol. 15, 99–117.

Kyung, J., Heo, S., Chang, J.-H., 2024. Enhancing multimodal emotion recognition through ASR error compensation and LLM fine-tuning. In: Proc. Interspeech 2024. pp. 4683–4687.

Latif, S., Qadir, J., Bilal, M., 2019. Unsupervised adversarial domain adaptation for cross-lingual speech emotion recognition. In: 2019 8th International Conference on Affective Computing and Intelligent Interaction. ACII, IEEE, pp. 732–737.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., Schuller, B., 2021. Survey of deep representation learning for speech emotion recognition. IEEE Trans. Affect. Comput. 14 (2), 1634–1654.

Li, Y., Bell, P., Lai, C., 2022. Fusing ASR outputs in joint training for speech emotion recognition. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 7362–7366. http://dx.doi.org/10.1109/ICASSP43922.2022.9746289.

Li, Y., Bell, P., Lai, C., 2024. Speech emotion recognition with ASR transcripts: a comprehensive study on word error rate and fusion techniques. In: 2024 IEEE Spoken Language Technology Workshop. SLT, pp. 518–525. http://dx.doi.org/10.1109/SLT61566.2024.10832143.

Li, B., Fei, H., Liao, L., Zhao, Y., Teng, C., Chua, T.-S., Ji, D., Li, F., 2023. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. MM '23, Association for Computing Machinery, New York, NY, USA, pp. 5923–5934. http://dx.doi.org/10.1145/3581783.3612053.

Lian, Z., Liu, B., Tao, J., 2021. CTNet: Conversational transformer network for emotion recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 985–1000. http://dx.doi.org/10.1109/TASLP.2021.3049898.

Lian, H., Lu, C., Li, S., Zhao, Y., Tang, C., Zong, Y., 2023. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. Entropy 25 (10), http://dx.doi.org/10.3390/e25101440, URL https://www.mdpi.com/1099-4300/25/10/1440.

Liebenthal, E., Silbersweig, D.A., Stern, E., 2016. The language, tone and prosody of emotions: Neural substrates and dynamics of spoken-word emotion perception. Front. Neurosci. 10, http://dx.doi.org/10.3389/fnins.2016.00506, URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2016.00506.

Lin, H., Karjadi, C., Ang, T.F., Prajakta, J., McManus, C., Alhanai, T.W., Glass, J., Au, R., 2020. Identification of digital voice biomarkers for cognitive health. Explor. Med. 1, 406.

Lin, B., Wang, L., 2023. Robust multi-modal speech emotion recognition with ASR error adaptation. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49357.2023.10094839.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, W., Qiu, J.-L., Zheng, W.-L., Lu, B.-L., 2021. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. IEEE Trans. Cogn. Dev. Syst. 14 (2), 715–729.

Liu, F., Shen, S.-Y., Fu, Z.-W., Wang, H.-Y., Zhou, A.-M., Qi, J.-Y., 2022a. Lgcct: A light gated and crossed complementation transformer for multimodal speech emotion recognition. Entropy 24 (7), 1010.

Liu, Y., Sun, H., Guan, W., Xia, Y., Zhao, Z., 2022b. Multi-modal speech emotion recognition using self-attention mechanism and multi-scale fusion framework. Speech Commun. 139, 1–9.

Livingstone, S.R., Russo, F.A., 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLOS ONE 13 (5), 1–35. http://dx.doi.org/10.1371/journal.pone.0196391.

Lopatovska, I., Arapakis, I., 2011. Theories, methods and current research on emotions in library and information science, information retrieval and human-computer interaction. Inf. Process. Manag. 47 (4), 575–592. http://dx.doi.org/10.1016/j.ipm.2010.09.001, Cited by: 200. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-79957582811&doi=10.1016%2fj.ipm.2010.09.001&partnerID=40&md5=b66a54e49ece1df6db68362ff3c7e5ec.

Lotfian, R., Busso, C., 2019. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. IEEE Trans. Affect. Comput. 10 (4), 471–483. http://dx.doi.org/10.1109/TAFFC.2017.2736999.

Luo, J., Phan, H., Reiss, J., 2023. Cross-modal fusion techniques for utterance-level emotion recognition from text and speech. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49357.2023.10096885.

Ma, H., Wang, J., Lin, H., Zhang, B., Zhang, Y., Xu, B., 2024. A transformer-based model with self-distillation for multimodal emotion recognition in conversations. IEEE Trans. Multimed. 26, 776–788. http://dx.doi.org/10.1109/TMM.2023.3271019.

Ma, K., Wang, X., Yang, X., Zhang, M., Girard, J.M., Morency, L.-P., 2019. ElderReact: A multimodal dataset for recognizing emotional response in aging adults. In: 2019 International Conference on Multimodal Interaction. ICMI '19, Association for Computing Machinery, New York, NY, USA, pp. 349–357. http://dx.doi.org/10.1145/3340555.3353747.

Mehrabian, A., 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Curr. Psychol. 14, 261–292.

Mikolov, T., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 3781.

Miranda Calero, J.A., Gutiérrez-Martín, L., Rituerto-González, E., Romero-Perales, E., Lanza-Gutiérrez, J.M., Peláez-Moreno, C., López-Ongil, C., 2024. Wemac: Women and emotion multi-modal affective computing dataset. Sci. Data 11 (1), 1182.

Nandwani, P., Verma, R., 2021. A review on sentiment analysis and emotion detection from text. Soc. Netw. Anal. Min. 11 (1), 81.

Nojavanasghari, B., Baltrušaitis, T., Hughes, C.E., Morency, L.-P., 2016. EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction. ICMI '16, Association for Computing Machinery, New York, NY, USA, pp. 137–144. http://dx.doi.org/10.1145/2993148.2993168.

Ok, H., Yoo, S., Lee, J., 2024. AudioBERT: Audio knowledge augmented language model. arXiv:2409.08199. URL https://arxiv.org/abs/2409.08199.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D., 2021. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, http://dx.doi.org/10.1136/bmj.n71, arXiv:https://www.bmj.com/content/372/bmj.n71.full.pdf. URL https://www.bmj.com/content/372/bmj.n71.

Pan, R., García-Díaz, J.A., Rodríguez-García, M.Á., Valencia-García, R., 2024. Spanish MEACorpus 2023: A multimodal speech–text corpus for emotion analysis in Spanish from natural environments. Comput. Stand. Interfaces 90, 103856. http://dx.doi.org/10.1016/j.csi.2024.103856, URL https://www.sciencedirect.com/science/article/pii/S0920548924000254.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. EMNLP, pp. 1532–1543.

Pepa, L., Spalazzi, L., Capecci, M., Ceravolo, M.G., 2023. Automatic emotion recognition in clinical scenario: A systematic review of methods. IEEE Trans. Affect. Comput. 14 (2), 1675–1695. http://dx.doi.org/10.1109/TAFFC.2021.3128787.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R., 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508.

Priyasad, D., Fernando, T., Sridharan, S., Denman, S., Fookes, C., 2023. Dual memory fusion for multimodal speech emotion recognition. In: Proc. INTERSPEECH. Vol. 2023, pp. 4543–4547.
Radford, A., 2018. Improving language understanding by generative pre-training.
Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2023. Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. PMLR, pp. 28492–28518.
Ragot, M., Martin, N., Em, S., Pallamin, N., Diverrez, J.-M., 2018. Emotion recognition using physiological signals: laboratory vs. wearable sensors. In: Advances in Human Factors in Wearable Technologies and Game Design: Proceedings of the AHFE 2017 International Conference on Advances in Human Factors and Wearable Technologies, July 17-21, 2017, the Westin Bonaventure Hotel, Los Angeles, California, USA 8. Springer, pp. 15–22.
Ramanarayanan, V., Lammert, A.C., Rowe, H.P., Quatieri, T.F., Green, J.R., 2022. Speech as a biomarker: Opportunities, interpretability, and challenges. Perspect. ASHA Spec. Interes. Groups 7 (1), 276–283.
Rao, K.S., Yegnanarayana, B., 2006. Prosody modification using instants of significant excitation. IEEE Trans. Audio Speech Lang. Process. 14 (3), 972–980.
Rasendrasoa, S., Pauchet, A., Saunier, J., Adam, S., 2022. Real-time multimodal emotion recognition in conversation for multi-party interactions. In: Proceedings of the 2022 International Conference on Multimodal Interaction. ICMI '22, Association for Computing Machinery, New York, NY, USA, pp. 395–403. http://dx.doi.org/10.1145/3536221.3556601.
Rump, K., Giovannelli, J., Minshew, N., Strauss, M., 2009. The development of emotion recognition in individuals with autism. Child Dev. 80, 1434–1447. http://dx.doi.org/10.1111/j.1467-8624.2009.01343.x.
Russell, J.A., 1980. A circumplex model of affect. J. Pers. Soc. Psychol. 39 (6), 1161.
Salazar, C., Montoya-Múnera, E., Aguilar, J., 2021. Analysis of different affective state multimodal recognition approaches with missing data-oriented to virtual learning environments. Heliyon 7 (6).
Scherer, K.R., 2005. What are emotions? And how can they be measured? Soc. Sci. Inf. 44 (4), 695–729.
Schuller, B.W., 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM 61 (5), 90–99.
Sebe, N., Cohen, I., Huang, T.S., 2005. Multimodal emotion recognition. In: Handbook of Pattern Recognition and Computer Vision. World Scientific, pp. 387–409.
Shi, B., Fu, Z., Bing, L., Lam, W., 2018. Learning domain-sensitive and sentiment-aware word embeddings. arXiv preprint arXiv:1805.03801.
Shoumy, N.J., Ang, L.-M., Seng, K.P., Rahaman, D.M., Zia, T., 2020. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. J. Netw. Comput. Appl. 149, 102447.
Singh, G.V., Firdaus, M., Ekbal, A., Bhattacharyya, P., 2023. EmoInt-trans: A multimodal transformer for identifying emotions and intents in social conversations. IEEE/ACM Trans. Audio Speech Lang. Process. 31, 290–300. http://dx.doi.org/10.1109/TASLP.2022.3224287.
Souma, W., Vodenska, I., Aoyama, H., 2019. Enhanced news sentiment analysis using deep learning methods. J. Comput. Soc. Sci. 2 (1), 33–46.
Suero Montero, C., Suhonen, J., 2014. Emotion analysis meets learning analytics: online learner profiling beyond numerical data. In: Proceedings of the 14th Koli Calling International Conference on Computing Education Research. pp. 165–169.
Swain, M., Routray, A., Kabisatpathy, P., 2018. Databases, features and classifiers for speech emotion recognition: a review. Int. J. Speech Technol. 21, 93–120.
Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B., 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1555–1565.
Tarnowski, P., Kołodziej, M., Majkowski, A., Rak, R.J., 2017. Emotion recognition using facial expressions. Procedia Comput. Sci. 108, 1175–1184.
Van, C.T., Tran, T.V.T., Nguyen, V., Hy, T.S., 2025. Effective context modeling framework for emotion recognition in conversations. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49660.2025.10888112.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.
Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: Resources, features, and methods. Speech Commun. 48 (9), 1162–1181.
Wang, Y., Gu, Y., Yin, Y., Han, Y., Zhang, H., Wang, S., Li, C., Quan, D., 2023a. Multimodal transformer augmented fusion for speech emotion recognition. Front. Neurorobotics 17, 1181598.
Wang, S., Ma, Y., Ding, Y., 2023b. Exploring complementary features in multi-modal speech emotion recognition. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49357.2023.10096709.
Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., Zhang, W., 2022. A systematic review on affective computing: emotion models, databases, and recent advances. Inf. Fusion 83–84, 19–52. http://dx.doi.org/10.1016/j.inffus.2022.03.009, URL https://www.sciencedirect.com/science/article/pii/S1566253522000367.
Wang, W., Xu, K., Niu, H., Miao, X., 2020. Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation. Complexity 2020, 1–9.
Wei, J., Hu, G., Tuan, L.A., Yang, X., Zhu, W., 2023. Multi-scale receptive field graph model for emotion recognition in conversations. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49357.2023.10094596.
Wen, G., Ye, S., Li, H., Wen, P., Zhang, Y., 2023. Multimodal and multitask learning with additive angular penalty focus loss for speech emotion recognition. Int. J. Intell. Syst. 2023 (1), 3662839.
Xie, B., Sidulova, M., Park, C.H., 2021. Robust multimodal emotion recognition from conversation with transformer-based crossmodality fusion. Sensors 21 (14), 4913.
Xu, P., Madotto, A., Wu, C.-S., Park, J.H., Fung, P., 2018. Emo2vec: Learning generalized emotion representation by multi-task training. arXiv preprint arXiv:1809.04505.
Yang, D., Alsadoon, A., Prasad, P.C., Singh, A.K., Elchouemi, A., 2018. An emotion recognition model based on facial recognition in virtual learning environment. Procedia Comput. Sci. 125, 2–10.
Yang, K., Tag, B., Wang, C., Gu, Y., Sarsenbayeva, Z., Dingler, T., Wadley, G., Goncalves, J., 2023. Survey on emotion sensing using mobile devices. IEEE Trans. Affect. Comput. 14 (4), 2678–2696. http://dx.doi.org/10.1109/TAFFC.2022.3220484.
Yang, K., Wang, C., Gu, Y., Sarsenbayeva, Z., Tag, B., Dingler, T., Wadley, G., Goncalves, J., 2021. Behavioral and physiological signals-based deep multimodal approach for mobile emotion recognition. IEEE Trans. Affect. Comput. 14 (2), 1082–1097.
Yao, B., Shi, W., 2024. Speaker-centric multimodal fusion networks for emotion recognition in conversations. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 8441–8445. http://dx.doi.org/10.1109/ICASSP48485.2024.10447720.
Yu, F., Guo, J., Xi, W., Yang, Z., Jiang, R., Zhang, C., 2021. Audio DistilBERT: A distilled audio BERT for speech representation learning. In: 2021 International Joint Conference on Neural Networks. IJCNN, pp. 1–8. http://dx.doi.org/10.1109/IJCNN52387.2021.9533328.
Zadeh, A.B., Liang, P.P., Poria, S., Cambria, E., Morency, L.-P., 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2236–2246.
Zadeh, A., Zellers, R., Pincus, E., Morency, L.-P., 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259.
Zhai, W., 2024. Self-adaptive multimodal retrieval-augmented generation. arXiv:2410.11321. URL https://arxiv.org/abs/2410.11321.
Zhalehpour, S., Onder, O., Akhtar, Z., Erdem, C.E., 2016. BAUM-1: A spontaneous audio-visual face database of affective and mental states. IEEE Trans. Affect. Comput. 8 (3), 300–313.
Zhang, D., Chen, F., Chang, J., Chen, X., Tian, Q., 2024. Structure aware multi-graph network for multi-modal emotion recognition in conversations. IEEE Trans. Multimed. 26, 3987–3997. http://dx.doi.org/10.1109/TMM.2023.3238314.

Zhang, X., Li, Y., 2023. A dual attention-based modality-collaborative fusion network for emotion recognition. In: Proc. Interspeech. Vol. 2023, pp. 1468–1472.
Zhang, T., Li, S., Chen, B., Yuan, H., Philip Chen, C.L., 2023a. AIA-Net: Adaptive interactive attention network for text–audio emotion recognition. IEEE Trans. Cybern. 53 (12), 7659–7671. http://dx.doi.org/10.1109/TCYB.2022.3195739.
Zhang, T., Tan, Z., Wu, X., 2023b. HAAN-ERC: hierarchical adaptive attention network for multimodal emotion recognition in conversation. Neural Comput. Appl. 35 (24), 17619–17632.
Zhang, J., Xing, L., Tan, Z., Wang, H., Wang, K., 2022. Multi-head attention fusion networks for multi-modal speech emotion recognition. Comput. Ind. Eng. 168, 108078.
Zhang, S., Yang, Y., Chen, C., Liu, R., Tao, X., Guo, W., Xu, Y., Zhao, X., 2023c. Multimodal emotion recognition based on audio and text by using hybrid attention networks. Biomed. Signal Process. Control. 85, 105052.
Zhao, H., Gao, Y., Chen, H., Li, B., Ye, G., Zhang, Z., 2025. Enhanced multimodal emotion recognition in conversations via contextual filtering and multi-frequency graph propagation. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 1–5. http://dx.doi.org/10.1109/ICASSP49660.2025.10888592.
Zhao, Z., Gao, T., Wang, H., Schuller, B.W., 2023. SWRR: feature map classifier based on sliding window attention and high-response feature reuse for multimodal emotion recognition. In: Proc. Interspeech. Vol. 2023, pp. 2433–2437.
Zhao, Z., Gao, T., Wang, H., Schuller, B., 2024a. MFDR: Multiple-stage fusion and dynamically refined network for multimodal emotion recognition. In: Proc. Interspeech 2024. pp. 3719–3723.
Zhao, G., Zhang, Y., Chu, J., 2024b. A multimodal teacher speech emotion recognition method in the smart classroom. Internet Things 25, 101069.
Zou, S., Huang, X., Shen, X., 2023. Multimodal prompt transformer with hybrid contrastive learning for emotion recognition in conversation. In: Proceedings of the 31st ACM International Conference on Multimedia. MM '23, Association for Computing Machinery, New York, NY, USA, pp. 5994–6003. http://dx.doi.org/10.1145/3581783.3611805.