CSE3000 Research Project

Survey of Affect Representation Schemes used in Automatic Affect Prediction for Speech Emotion Recognition:

A Systematic Review

Aditi Rawat[1]

Supervisor(s): Chirag Raman[1]**, Bernd Dudzik**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Aditi Rawat
Final project course: CSE3000 Research Project
Thesis committee: Chirag Raman, Bernd Dudzik, Cynthia Liem

Abstract

Automatic affect prediction systems usually assume its underlying affect representation scheme (ARS). This systematic review aims to explore how different ARS are used for in affect prediction systems based on spoken input. The focus is only on the audio input from speakers. Various datasets for speech emotion recognition were also involved in the study to understand the motivation for certain (categorical or dimensional) schemes used for emotions. The basis, popularity, advantages and target affective states were investigated. We used Scopus and Web of Science to extract the papers, focusing on the systems in the field of Computer Science in English language. In summary, our exploration of affect representation schemes in Speech Emotion Recognition (SER) reveals a predominant focus on categorical representations of affect, particularly variations of Ekman's six basic emotions. Behavior and attitude, although rare, are also represented sometimes. Emotions like anger, happiness, and sadness receive the most attention, while the recognition of the neutral state as an emotional state remains controversial. Dimensional affect representation schemes are less common, possibly due to the difficulty in estimating valence solely from audio input. Researchers often combine multiple categorical schemes to accommodate different datasets used in SER systems, aligning the popularity of the schemes with the corresponding datasets. However, issues such as a lack of explanation for chosen categories, interchangeable use of terminology, and a weak psychological foundation for category selection pose challenges in achieving a comprehensive understanding of affect representation in SER research.

1 Introduction

Affect refers to a feeling experienced by an individual [1], while affect prediction involves detecting affective states such as emotions, moods, attitudes and so on based on various forms of input. In this paper, we consider speech as the input to automatically predict the affective state represented, emphasizing on speech emotion recognition (SER) as a growing field where the ability to detect emotional undertones of speech allows for several useful developments in the field of machine learning and artificial intelligence. For example, it allows emotionally intelligent agents that can interact with humans, potentially also creating emotionally-supported statements in speech [2]. For example, Alexa and other conversational agents could improve by taking into account the user's affective state and responding appropriately.

In order to be able to predict affective states, it is important to have a scheme to represent these states. These will be referred to as Affect Representation Schemes (ARS) throughout this paper. Several schemes exist currently as either a categorical scheme or a dimensional one. A widely accepted categorical is Ekman's 6 basic emotions, which are fear, anger, joy, sadness, disgust, and surprise [3]. Here, we consider emotions to either belong to one category or the other. This is in contrast to dimensional schemes such as Russell's Valence and Arousal [4] or Warriner et al. Valence, Arousal and Dominance [5]. Valence can be seen as how positive or negative a person may be feeling. Activation or arousal can be explained in terms of the dynamic or lethargic a person may feel. While these 2 dimensions are common, often dominance is also utilised as the third dimension to represent how much a person feels in control of their current situation. The representation of emotions is quite subjective, considering that even humans themselves may disagree on this in several situations [2]. Thus, it becomes more important to be critical in comparing the different representation models available to us, how they work (together or separately), as well as their advantages and popularity in SER.

The main research question addressed by this paper is ***"How are different affect representation schemes utilised in the context of speech emotion recognition?"*** More concretely, the sub-questions shown in table 1 have been devised for the purpose of answering the main research question. The paper begins with highlighting the methodology employed to answer the aforementioned research question, explaining the steps taken in conducting a systematic review of existing literature in the field of affective representation for automatic speech emotion recognition in section 2. Afterwards, the paper explains the results of abiding by the methodology described in section 3 as well as the ethical considerations involved in the research process in section 4. Finally, section 5 discusses and evaluates the findings of the paper, followed by concluding remarks and future recommendation in section 6.

2 Methodology

The methodology of carrying out a systematic review is followed for the purpose of answering the research question. The main idea is to ensure generalizability by following a systematic procedure that improves the reproducibility of this findings of this paper [6]. Additionally, the PRISMA guidelines [7] were followed since it provides a comprehensive framework for conducting systematic reviews. It includes detailed checklists to guide the research procedure in a reliable manner as well as highlights the relevant information to be described in each step of the research. This section first provides an explanation of the eligibility criteria established to decide what records will be included for answering the research questions in section 2.1. The search strategy and selection process are highlighted in sections 2.2 and 2.3 respectively. Finally, the set of papers obtained as a result of implementing the aforementioned stages for extraction are provided in section 2.5.

2.1 Eligibility criteria

The inclusion and exclusion criteria is developed to lay out the scope of the systematic review. In order to have a clear idea of the type of literature that is expected from the search and selection procedure, the criteria was established and explained below.

Table 1: Sub-questions to be considered as part of answering the main research question, as well as their corresponding motivation.

Sub-Question	Explanation
SQ1. *What types of affective states have been targeted by prediction systems?*	Before diving into the representation of different affective states, it is important to know which type is being targeted. For example, the representation of mood may differ from the representation of emotions.
SQ2. *What different affect representation schemes have been used and their motivation?*	After observing the types of affective states targeted, we want to analyze the affective states being represented themselves. It is important to understand the rationale behind a specific way of representin gthem since it is possible that certain schemes are useful in particular contexts or for particular affective state types. This is relevant to investigate since the subjectivity of affect suggests that there is no one universal way to represent emotions that works for every scenario.
SQ3. *Are systems using more than one emotion representation scheme simultaneously? If so, what is their motivation for doing so?*	Considering the subjectivity of the matter, it is possible that certain systems are unable to choose exactly one system suitable for their goal and thus, resort to multiple schemes. This is basically an extension of the previous sub-question.
SQ4. *Are there differences in the popularity of schemes used for modeling different affective states?*	We want to look into the acceptability of different ARS and the reasons for the same. By highlighting the motivations of various ARS, we aim to understand which of these motivations are most relevant.
SQ5. *Has the popularity of specific schemes changed over time?*	Similar to how emotions may change over time, it is important to consider how popularity of schemes have changed as affect prediction systems receive more attention.
SQ6. *Is the majority of representation schemes used based on psychological theory?*	While the motivation for ARS is already previously explored, it is also important to evaluate whether these models arise from psychology-based foundations or does the task of speech emotion recognition indicate the need for adaptations or unexpected representation schemes.

Tables 2 describe and motivate the type of papers to be included and those to be excluded, in that order. This establishes the criteria for a paper to be considered eligible for analysis as the means to answer the research question.

Table 2: Inclusion and exclusion criteria along with their corresponding motivation.

Inclusion Criteria	**Motivation**
System introduced in the record is an affect prediction system	This is the main scope of this research.
The affect prediction systems takes speech as input for the system in order to make the prediction.	This review focuses on speech emotion recognition (SER) systems specifically.
The papers are in English.	All parties involved only have English language in common.
Exclusion Criteria	**Motivation**
Speaker Recognition in an emotional context	Several papers in the query results focus on identifying the speaker(s) based on the spoken input, or they aim to identify whether a speaker is emotional or not in a binary manner. This results in lack of information in terms of the underlying ARS since the purpose is not to distinguish between the various affective states.
Multimodal Affect Prediction Models	Out of the scope for this project.
The paper is a review.	This paper relies on primary sources for affect prediction systems, instead of indirect information obtained from other reviews. This helps eliminate biases in findings.
The paper is not in English.	All parties involved only have English language in common.
The paper is not in the field of Computer Science.	The paper only focuses on affect prediction in the field of computer science.

2.2 Search Strategy

A search strategy was developed to obtain an initial, broad set of records with the potential to be analysed for the review. The databases used for this purpose were Scopus[1] and Web of Science[2] due to a large range of records being available there. In order to develop the initial query, certain concepts integral to the research question were established. This included three aspects:

1. *Speech or spoken input*
2. *Affect representation or emotion models*
3. *Prediction or recognition of emotion*

The intersection of these concepts formed the foundation of the query executed on Scopus and Web of Science. These also gave rise to the relevant key terms and their respective synonyms to ensure that the papers falling within the scope laid out are

[1]Scopus: https://www.scopus.com

[2]Web of Science: https://www.webofscience.com

included in the search results. However, considering that this is an active field of research, there was an overwhelming amount of records, which were not feasible to cover over a period of only ten weeks. Additionally, the initial query required additional filtering to ensure that the eligibility critera highlighted in section 2.1 was enforced. Thus, a rigorous selection strategy had to be followed to ensure the feasibility of the project and the implementation of the inclusion and exclusion criteria.

Table 3: Keywords established for each relevant concept.

Affect	**Speech**	**Recognition**
(emotion represent)* *(emotion model*)* *(affect represent*)* *(mood*)* *(affective computing)*	*(speech)* *(speaker)*	*(recogni*)* *(detect*)* *(predict*)*

2.3 Selection Strategy

The strategy for filtering the results arising from Scopus and the results arising from Web of Science were separated due to differences in the features of these two databases and the significant difference between the number of records obtained from each. More specifically, Scopus initially resulted in 16,356 papers while Web of Science resulted in 4,998 results. While the latter was considered to be feasible for manual filtering through title and abstracts, the former needed additional filtering. This was achieved in the following steps:

1. *Base Query*: This is the same as the query highlighted in section 2.2.
2. *Filtering by Title*: The first step of screening involved evaluating if the record abides by the inclusion and exclusion criteria based on its title. Ambiguous or vague titles were not eliminated so that they could be further looked into in the following steps.
3. *Filtering by Abstract*: The next step was to read the abstract of the remaining records to evaluate their eligibility.
4. *Filtering by Full-Text*: Finally, the remaining records were reviewed by their full texts to decide their eligibility.

A brief overview of the results of these steps followed is visualized in figure 1.

2.4 Feasibility Filtering

Considering the high amount of records available on speech emotion recognition, it was considered infeasible to analyse all of the papers that were considered according to eligibility criteria. To tackle this issue, certain filtering criteria and prioritization procedures were followed to make the review's methodology achievable within the given timeframe of 10 weeks. These are as follows:

1. *Filtering by Micro-topics*: Web of Science allows advanced features such as excluding or limiting the search results based on the topics. The topics related to Computer Science were chosen and the full list of topics can be found in the appendix A.1.
2. *Filtering by Keywords*: Scopus allows advanced features such as excluding or limiting the search results based on the keywords highlighted. It provides a list of keywords found in the existing search results. This list was used to implement the inclusion and exclusion criteria. For example, excluding multimodal emotion recognition could be achieved by excluding this keyword. The full list of keywords can be found in the appendix A.2.
3. *Filtering by Database Names*: The search results were further limited by mentioning existing databases that are commonly employed by speech emotion recognition systems. This list of database names was extracted from an existing systematic review conducted in the field of speech emotion recognition[8]. The search results were thus limited to only those records that mention at least one of the mentioned databases. The full list of datasets can be found in the appendix A.3.
4. *Prioritization by Year*: While the previously mentioned filters helped reduce the number of search results by a significant amount, this was still deemed not enough. Thus, the decision was made to divide the papers extracted so far into batches. In order to ensure the most recent developments for SER systems are included in the study, this procedure started with taking 10 records from each year starting with the current year, i.e. 2023, up to and including 2021 so far. This was achieved by sorting the all papers from *Relevance* on the respective databases and interleaving between the results from Scopus and Web of Science. This has led to 30 papers (2023, 2022, and 2021) being analyzed so far, with the potential to include 20 more records at least (from 2020 and 2019). It was important to sample from multiple years in order to address the research question with respect to how the popularity of different ARS has changed over time.

By following this procedure, we could establish a clear direction and priority to allow for flexibility in a systematic and feasible manner.

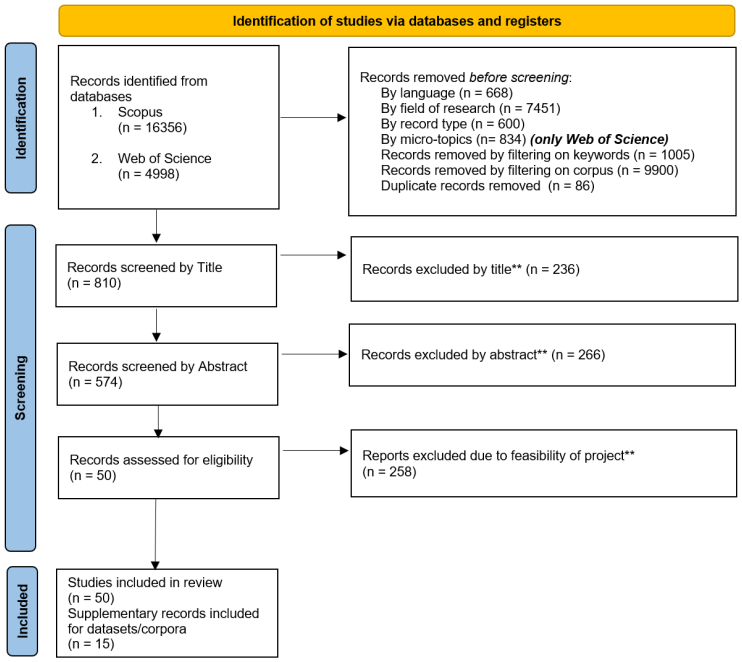


Figure 1: PRISMA Flow Diagram

2.5 Search Results

The aforementioned steps taken by establishing an eligibility criteria as well as pursuing a search and selection strategy resulted in the final set of extracted papers shown in table 4, grouped by year. In addition to these,

Table 4: Results of the search and selection strategy, i.e. final set of papers to be extracted as well as their relevant features.

Record(s)	Year
[9; 10; 11; 12; 13; 14; 15; 16; 17; 18]	2023
[19; 20; 21; 22; 23; 24; 25; 26; 27; 28]	2022
[29; 30; 31; 32; 33; 34; 35; 36; 37; 38]	2021
[39; 40; 41; 42; 43; 44; 45; 46; 47; 48]	2020
[49; 50; 51; 52; 53; 54; 55; 56; 57; 58]	2019

In addition to the results extracted from the search and selection strategy, additional records were required to understand the choice of emotion models for each dataset that was used for training, testing or validating the prediction systems of the extracted results. The papers to be analyzed for this purpose are shown in table 5. Table 8 in appendix B shows the questions asked for each paper in the search results to extract relevant information in order to answer the sub-questions of the research.

During the data extraction, we realized that there is often a lack of motivation provided for the ARS within the papers extracted. However, it was observed that the papers often employed the ARS of the dataset that they were using. Thus, these explanations had to be extracted from supplementary dataset papers, which are highlighted in table 5 showing all the datasets used across the final search results. Certain dataset papers were not analyzed (shown as N/A in the table) for two main reasons. Firstly, the documentation of the dataset may not be an official record. For example, TESS and SAVEE were websites. Second reason was in the case when the actual ARS of the dataset was not used. For example, certain papers used CHEAVD and AVEC 2016 dataset but their ARS was not employed or was already explained by other datasets. Thus, we finally have 50 extracted papers from our search and selection strategy as well as 15 dataset papers to supplement them.

Table 5: Records obtained for each dataset used in the extracted papers. *Related Papers* refers to the papers that utilize the dataset.

Record	Dataset	Categorical, Dimensional or Both	Count of Related Papers	Related Papers
[59]	IEMOCAP	Both	25	[9; 16; 10; 14; 25; 27; 20; 22; 24; 28; 32; 34; 36] [38; 40; 44; 46; 48; 50; 52; 56; 58; 43; 51; 55]
[60]	EMO-DB	Categorical	24	[9; 13; 15; 16; 19; 21; 23; 27; 22; 24; 26; 29; 31] [37; 30; 32; 46; 52; 56; 58; 41; 43; 51; 55]
[61]	RAVDESS	Both	13	[9; 13; 15; 18; 12; 14; 25; 26; 28; 29; 32; 39; 47]
N/A	SAVEE	Categorical	13	[13; 12; 21; 25; 26; 29; 31; 37; 30; 32; 39; 47; 57]
[62]	CASIA	Both	8	[23; 20; 26; 31; 56; 49; 53; 55]
N/A	TESS	Categorical	5	[11; 13; 25; 35; 47]
[63]	CREMA-D	Categorical	4	[13; 25; 37; 47]
[64]	RECOLA	Both	4	[33; 42; 54; 45]
[65]	EMOVO	Categorical	3	[17; 18; 12]
N/A	CHEAVD	Categorical	2	[56; 55]
[66]	FAU Aibo	Both	1	[52]
[67]	ABC	Categorical	1	[20]
[68]	eNTERFACE'05	Categorical	1	[30]
[69]	URDU	Categorical	1	[12]
[70]	MSP-IMPROV	Categorical	1	[10]
[71]	AVEC 2014	Dimensional	1	[33]
[72]	SEMAINE	Both	1	[42]
[73]	USC CreativeIT	Dimensional	1	[42]
N/A	AVEC 2016	Dimensional	1	[54]

3 Results

After extracting the records, we synthesized the data to understand how different ARS are used. We recorded the ARS utilized in each paper and the list of all distinct representations of affect are shown in table 6, along with the categories and/or dimensions provided, the count of these categories or dimensions and the dataset pertaining to the ARS. In most cases, there is a one-to-one relationship between the ARS and the dataset used, whereas in certain cases, there may be multiple datasets associated with a specific ARS. Also note that there is often a one-to-many relationship between the paper and the ARS or dataset used.

In this section, each of the sub-research questions is answered with respect to the scope of the 50 records analyzed from 2019 up to 2023 using the aforementioned list of ARS. We start with describing the types of affective states that are usually targeted by the records overall to answer *SQ1*. Afterward, we answer *SQ2* to understand the advantages and disadvantages of each ARS in terms of the motivation provided either by the papers extracted themselves or the paper corresponding to the dataset(s) employing the scheme. As mentioned previously, automatic affect recognition systems tend to use multiple datasets and consequently, multiple ARS. Thus, we dive into the common combinations used as well as the reasoning behind the same in order to answer *SQ3*. Following this, we answer the overall popularity and the change in popularity of the schemes over time (corresponding to *SQ4* and *SQ5*). Finally, the theoretical basis for these ARS are highlighted to answer *SQ6*.

Table 6: Affect Representation Schemes present in extracted records of the systematic review.

ARS	Categories or Dimensions Count	Related Datasets	ARS ID
Anger, Anxiety/Fear, Boredom, Disgust, Happiness, Sadness	6	EMO-DB	C.6ed
Anger, Calm, Disgust, Fear, Happiness, Sadness, Surprised + Intensity	7	RAVDESS	C.7r
Anger, Happiness, Sadness	3	IEMOCAP, MSP-IMPROV, URDU, Custom	C.3i
Anger, Disgust, Fear, Happiness/Joy, Sadness, Surprised	6	TESS, SAVEE, eNTERFACE'05, EMOVO (joy instead of happy)	C.6e
Anger, Disgust, Fear, Happiness, Sadness + Intensity	5	CREMA-D	C.5cr
Disappointment, Fear, Sadness, Surprised	4	Customized EMOVO	C.4e
Anger, Fear, Happiness, Sadness, Surprised	5	CASIA	C.5cs
Aggressive, Cheerful, Intoxicate, Nervous, Tired	5	ABC	C.5a
Anger, Frustrated, Happiness, Sadness	4	Modified IEMOCAP	C.4i
Anger, Happiness, Sadness, Scared, Surprised	5	Modified IEMOCAP	C.5i.1
Anger, Fear, Frustrated, Happiness, Sadness, Surprised	6	Modified IEMOCAP	C.6i
Anger, Excited, Frustrated, Happiness, Sadness	5	Modified IEMOCAP	C.5i.2
Anger, Happiness	2	Modified all of IEMOCAP, EMO-DB, CASIA, CHEAVD	C.2i
Anger, Emphatic, Positive, Rest	4	FAU Aibo	C.4f
Valence, Arousal and Dominance	2	AVEC 2014	D.va
Valence and Arousal	2	RECOLA, SEMAINE, USC Creative IT	D.va
Arousal	1	Customized RECOLA, customized AVEC 2016	D.a

3.1 Target Affective States

Automatic speech emotion recognition systems predict the affective state of the speaker based on the audio-input from the datasets. In this section, we discuss the type of affective states that are targeted as part of the automatic affect prediction

system, i.e the different emotions that the system distinguishes between. Table 7 shows the types of affective states targeted by the SER systems.

Table 7: Types of affective states targeted

Type of Affective State Targeted	Corresponding ARS	Related Papers
Emotion	All except ***D.5f, D.5a***	All
Emotion-related state / Affect / Behaviour	***D.5f***	[52]
Behaviour	***D.5a***	[20]
Attitude	***D.7r, D.6t***	[39]

Overall, most affect representation schemes tend to be focus mainly on emotions. However, a few papers do target other types of affective states as well. The SER system developed in [52] employs the FAU Aibo corpus (and thus, ARS ***D.5f***). As explained in [66], the corpus distinguishes between emotions in a more specific sense and common user states (for example, helplessness) and behavioral patterns (such as reprimanding), acknowledging that the latter two are not strictly classified as emotions, but rather states associated with emotions. It is claimed that these categories do not encompass all emotions in a broad sense, but are considered suitable for representing human behavior. We also have paper [20], which obtains its dataset from ABC, i.e. Airplane Behaviour Corpus. Thus, the ARS of ABC (***D.5a***) is considered to target behaviour instead of the usual emotions. Finally, paper [39] itself outlines its aim as that of identifying the attitudes of the speakers by identifying emotional states. Nevertheless, the datasets employed by the system described relies on those that target emotions, not attitude.

After highlighting the various types of affective states targeted, we will go more in depth into targeting emotions since all SER systems achieve this in one way or another. We can also see from table 6, that categorical ARS are more common, while the actual categories may differ quite a bit. To understand which of these affective states categories are targeted most often, we observed the percentage of systems that target *anger, boredom, calm, disgust, fear, frustrated, happiness, joy, neutral, sadness, surprised and miscellaneous*. Emotion categories that were only targeted once, namely *aggressive, cheerful, disappointment, emphatic, excited, intoxicate, nervous, positive, rest* and *tired* were moved to a *Miscellaneous* category. Only the discrete type of ARS were considered for this analysis. Here, a system is said to target an emotion if it exists as a category in at least one ARS followed by the system. For systems with multiple ARS, duplicate emotion categories across the multiple ARS are counted only once.

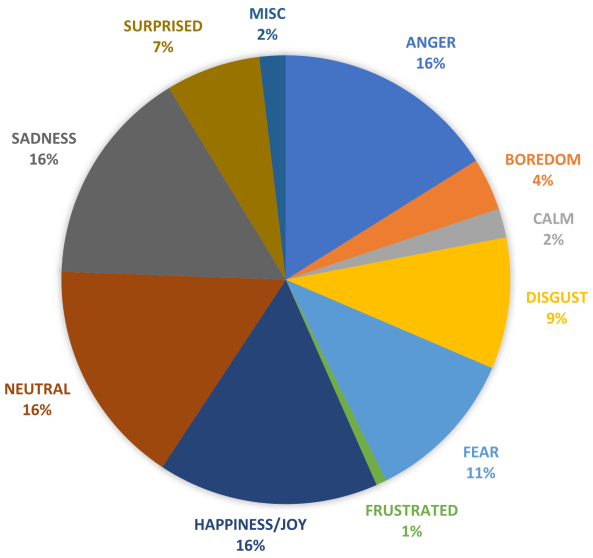


Figure 2: Target emotions by categorical ARS

The results of the analysis are shown in figure 2. We can see that *anger*, *happiness* and *sadness* are most often targeted. *Disgust* and *fear* are also relatively frequent, while *frustrated*, *calm* and *boredom* are least often the target of automatic affect prediction systems. Note that *neutral* is used as a baseline label in the dataset but is not a target affective state in most cases. It has been included in this analysis since most SER systems in the study predict this state. While most papers themselves do not reflect on the role of neutral themselves, the dataset papers often specify this. For example, [66],[65] and [68] describe

neutral as a default state or a baseline, while [67] considers neutral a behavior and [62] reports neutral as an emotion that is most common in our day-to-day lives. It is also worth noting that datasets employing the same ARS may also disagree on the role of neutral. MSP-IMPROV [70] and URDU [69] both use ***C.4i*** with the latter recognizing neutral as an emotion while the former does not.

3.2 Affect Representation Schemes and their Motivation

While the speech emotion recognition systems themselves were examined for mentioning the motivation behind their chosen ARS. However, they do not mention the motivation for using certain ARS, the motivation of these models can be extracted from the papers of the databases themselves. These are as follows:

Categorical ARS

Most systems tend to prefer categorical ARS by viewing SER as a classification problem. All of the categorical ARS mentioned in table 6 are based on the wiely accepted Ekman's 6 Basic Emotions [3]. In this section, we explain the motivation behind any deviation and additional modifications made to this model for each of the affect representations schemes. The datasets themselves may be chosen due to practical reasons such as audio quality, duration of recordings, number of recordings, balance of gender, language et cetera. For SER systems using categorical ARS, the datasets are not chosen based on the ARS, rather the other way around. Thus, most of explanations below were extracted from the supplementary dataset papers, and not the papers themselves due to a lack of discussion in terms of affect representation decisions.

- *C.6ed*
  This ARS arises from the EMO-DB dataset [60], featuring a range of emotions captured for analysis. It includes seven distinct emotions: *sadness, boredom, neutral, disgust, happiness, anxiety/fear,* and *anger*. The ctageorical representation is considered easily understandable when constructing the dataset by both the performer eliciting the required emotion and the listener annotating the audio recordings.

  One notable aspect is the consistent use of the same emotions as in their previous studies. This decision enables researchers to compare and contrast results across different studies, enhancing the understanding of emotional expression. However, beyond this rationale, the motivation for using these specific emotions in the EMO-DB dataset is not explicitly stated in the available information. Further information or justification for the inclusion of these emotions is not provided in the referenced source.
- *C.7r*
  This ARS correponds to the RAVDESS dataset [61], encompassing the largest number of emotional states. It includes eight distinct categories: *neutral, calm, happy, sad, angry, fearful, surprise,* and *disgust*. The inclusion of both neutral and calm emotions as baseline states is noteworthy. While neutral emotions can sometimes convey a negative valence due to performer uncertainty. Additionally, more positive states such as pleasure, joy, pride, and amusement could have been added. Nevertheless, the decision was made to not include them in order to prioritize high discriminability since these additional positive states typically achieve recognition at or below a recognition threshold of 40. Instead, calm emotions were introduced as a compensatory measure for neutral's potential negative perception. They are described as perceptually similar to neutral emotions but may be perceived as having a mild positive valence. Finally, the inclusion of surprise in the RAVDESS dataset has been a subject of controversy. However, given its presence in many existing datasets, it was included to ensure compatibility and consistency across different emotional databases.
- *C.3i, C.4i, C.5i.1, C.5i.2, C.6i, C.2i*

  The IEMOCAP corpus [59] offers valuable insights into the expression of emotions by encompassing a range of emotional states. It has 9 labels in total (*anger, disgust, excitement, happiness, neutral, sadness, scared and surprised*). It was recognized that having too many categories would result in low agreement between evaluators, while having too few categories would lead to poor emotional description and less accurate characterization. It is worth noting that using four emotion categories is common practice for the IEMOCAP corpus of this nature, namely, *anger*, *happiness*, *neutral* and *sadness* due to imbalance between the emotion labels in the dataset. The most common modification is to merge the emotions of happiness and excitement due to their close proximity in the activation and valence domains. A lot of modifications to the ARS of IEMOCAP can be seen in table 6. However, the specific rationale behind choosing the specific labels is seldom clarified but mostly relate to dataset balance and amount of representative samples. Other datasets using this scheme is MSP-IMPROV [70] and [69]. The former chose this scheme since they are the most commonly occurring classes while the latter did not consider it feasible due to logistical issues.
- *C.6e, C.4e*
  This ARS is pertains to EMOVO [65], eNTERFACE'05 [68], TESS and SAVEE that use the basic 6 emotions by Ekman with a neutral state. In paper [15], two modifications were made to ***C.6e*** to obtain ***C.4e***. *Anger* and *disgust* were considered to be similar and merged into one class of *disappointment*. *Happy* or *joy* were removed keeping in mind the aim of the system, which was to investigate emotions for COVID contact tracing application. Therefore, *happiness* was not an expected outcome in patients testing positive for COVID.

- *C.5cr*
  This corresponds to CREMA-D dataset [63] using the aforementioned basic 6 emotions except *surprise* due to unknown reasons.
- *C.5cs*
  This ARS arises from CASIA dataset [62], which distinguishes between "prototypical" and "non-prototypical" emotional states in daily life, with the former being what is predominantly used by most datasets arising from the basic 6 emotions. According to [62], it is desirable in human-computer interaction to recognize the non-prototypical emotions that are more subtle, for example shy, worried et cetera. Therefore, in its annotation process, no specific emotional categories were provided for the process and categories were added in a dynamic manner based on the feelings of the annotators. This lead to 24 emotions being annotated, namely: *happy, neutral, sad, nervous, disgust, surprise, worried, angry, anxious, proud, embarrassing, frustrated, fearful, anticipated, blamed, helpless, sarcastic, suspicious, confused, curious, aggrieved, shy, hesitant* and *contemptuous*. These were grouped to form main emotion labels of *anger, fear, happiness, sadness* and *surprised*. Along with these labels, there were also accompanying emotions to describe the subtler emotions highlighted earlier. However, none of the SER systems analyzed in this systematic review use or mention the accompanying labels.
- *C.5a*
  This ARS is with respect to the Airplane Behavior Corpus [67]. The choice of labels for the behaviors as *aggressive, cheerful, intoxicate, nervous, neutral* and *tired* are not explained but a mood induction procedure was utilized to reflect realism in the behavior that was being enacted.
- *C.4f*
  This ARS is introduced by the FAU Aibo corpus [66] to represent the spontaneous emotion-related states by children when they play with the robot **Aibo**. Therefore, the labels are specific to this context. For example, *bored* refers to when a child is uninterested in interacting with **Aibo** currently, *motherese* was an initially proposed label to descibe if the child is addressing Aibo in the way a mother would or *emphatic*, where a the child talks with strong articulation but does not show emotion. 10 initial categories were used. The first 5, i.e., *angry, touchy* or *irritated, joyful, surprised, bored* were considered to be emotions in a much limited sense. The rest, i.e. *helpless, motherese, reprimanding, emphatic* and *other* were used to describe the state of the user or patterns in behavior. These labels were grouped by majority vote to provide the final labels of *anger, emphatic, neutral, positive* and *rest*.

Dimensional ARS (*D.vad, D.va, D.a*)

There are several advantages that dimensional ARS have to offer over the limitations of categorical ARS. Firstly, the motivation behind the RECOLA dataset [64] design stems from the limitations of categorizing mixed emotions into discrete categories. Such restrictive categorization may not adequately capture the complex and nuanced nature of mixed emotional states. Secondly, it is difficult to capture the rise and fall of emotion with time [72; 33]. Thirdly, and the agreement between annotators of dataset in terms of which emotion is shown can be relatively low.[72; 73]. Dimensional affect representation schemes are reflected in the RECOLA [64] and SEMAINE [72] datasets with *valence* and *arousal* as dimensions, i.e. ARS ***D.va*** based on Russell's Valence and Arousal [4]. Additionally, AVEC 2014 dataset [71] and USC-CreativeIT [73] also have the dimension of *dominance*, i.e. ARS ***D.vad*** motivated by Warriner et al. Valence, Arousal and Dominance scheme [5].

Most of the SER models using ***D.va*** within the scope of this systematic review use RECOLA dataset [64]. It offers continuous values for valence and arousal dimensions. It also incorporates five social behaviors, however, none of the systems in this study utilize this aspect of the dataset. Therefore, only valence and arousal predictions are included in this analysis. The creators of the RECOLA dataset acknowledge that reducing the emotional dimensions to only two or three may result in a significant loss of information. However, it is still considered more practical. Additionally, in order to enhance the differentiation of emotional valence among participants within a team, the RECOLA dataset [64] employs a mood induction procedure. This procedure aims to increase the difference in emotional valence while slightly raising arousal levels. Participants receive either a positive or negative mood induction based on their self-reported valence, with the goal of balancing the emotional states within the dataset. This is similar the the mood induction procedure followed for ABC [67]. For ARS ***D.vad*** was considered is suitable in representing the ambiguous expressions of the CreativeIT database [73]. The AVEC 2014 [71] dataset was a part of a challenge between SER systems. It also uses this ARS and chose this based on its relevance to the main task of the challenge which was to estimate depression, explaining *dominance* to be "an individual's sense of how much they feel to be in control of their current situation" [71]. This dataset also provides continuous values for depression. Finally, ARS ***D.a*** arose from a more practical perspective. It was considered that valence prediction performs better only when supported with with visual input, instead of only relying on spoken input. This is why paper [54] utilizes this specific scheme.

ARS Combinations

22 out of the 50 systems analyzed use a single ARS, while the others use multiple. Primarily, it has been observed that multiple categorical schemes are commonly utilized together. In the case of most speech emotion recognition (SER) systems, which are often based on neural networks, variations of the model are created to match the specific database used for testing. This flexibility allows for easy expansion of the systems to accommodate different emotion classes by adjusting the number of

neurons in the outermost layer [31], consequently enabling the utilization of diverse data from various datasets. As a result, the majority of the SER systems discussed in the extracted papers are capable of leveraging multiple datasets.

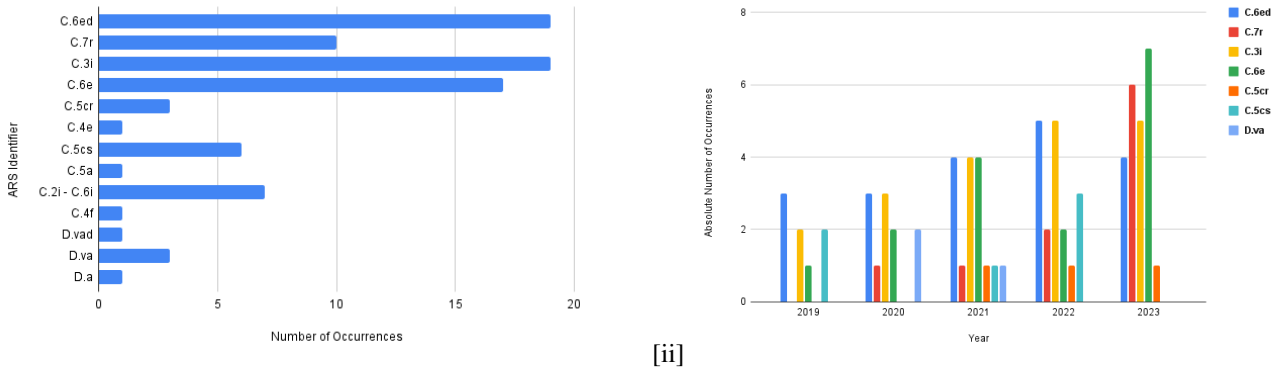
Certain corpora, such as IEMOCAP [59] and CASIA [62], incorporate a dimensional representation scheme alongside their categorical representation schemes. This inclusion of dimensional schemes allows for a more comprehensive analysis of the diversity in emotions, offering valuable insights into how emotions manifest across different contexts within the dataset. By combining dimensional and categorical schemes, we can obtain complimentary information on how individuals express emotions and how these cues can be effectively recognized or synthesized to enhance human-machine interfaces. Although these considerations are taken into account in the datasets themselves, the systems themselves do not simultaneously employ both dimensional and categorical ARS.

An interesting approach of combining two categorical representation schemes was taken by the system explained in [54], employing two affect representation schemes (***C.2i*** and ***C.4i***), with one being considered as a "coarse" classification and the other defined as a "fine" classification of emotion. Firstly, a two-dimensional coordinate system utilizing valence and arousal as the horizontal and vertical axes was constructed, respectively. This system creates four quadrants, which were consider as four coarse types. Traditional discrete emotion types can be mapped to one of the coarse types based on their corresponding annotation values. For example, happy and excited were assigned to the first quadrant, while another quadrant accommodated angry and frustrated as emotion categories. The coarse space as the overarching emotion type, with the discrete types within the coarse space serving as the finer classifications. By employing the coarse type to aid in fine type classification, the aim was to achieve more accurate recognition results, resembling the concept of top-down classification. Although the end result is the classification of emotion into the categories of the corresponding dataset, i.e. IEMOCAP in this case, each step in the two-step procedure of coarse and subsequently, fine classification used its own ARS.

In this way, we have explored how and why different ARS can be combined in SER systems, with categorical ARS being the most convenient to employ together. Additionally, dimensional and categorical schemes are never used simultaneously despite datasets allowing the possibility to do so in their design. Finally, we saw a hierarchical approach taken in [54] in terms of coarse-to-fine classification, when combining two distinct ARS.

3.3 Popularity of Affect Representation Schemes

In this section, we discuss the popularity of the ARS described in table 6. We show the overall popularity across the 50 papers included in this systematic review (see figure 3i), followed by showing the change in popularity of these ARS throughout the years 2019 to 2023 (see figure 3ii).



[i] [ii]

Figure 3: Popularity of Affect Representation Schemes where (i) Overall Dataset Popularity (ii) ARS Popularity Over Time

From figure 3(i), we can see that the 3 most popular ARS are ***C.6ed, C.3i*** and ***C.6e***. It is also worth noting that some of the datasets corresponding to these (i.e. EMO-DB, IEMOCAP, SAVEE) are the most popular datasets as well, which can be seen in table 5. ***C.7r*** is also relatively average in terms of popularity, corresponding to the RAVDESS dataset, another relatively popular dataset. Note that ***C.2i-C.6i*** refer to the modifications of IEMOCAP as well, namely ARS ***C.2i, C.4i, C.5i.1, C.5i.2*** and ***C.6i***. It is important to note that all of the aforementioned ARS are categorical ones. Dimensional models overall are not as commonly found as shown by the plots of ***D.vad, D.va*** and ***D.a***.

In figure 3(ii), we can observe the trend of 7 distinct ARS over the years 2019-2023. ARS that only occurred once or twice were not included in this analysis as they are not considered representative enough for extrapolation of potential trends in popularity, leading to the final 7 ARS. Both ***C.6ed*** and ***C.3i*** follow a similar trend of steady growth over the years with a slight decrease in 2023. On the other hand, ***C.7r*** remains relatively unpopular upto and including 2021, subsequently becoming significantly more popular by 2023. ***C.6e*** also experiences an increasing pattern in general (with the exception of 2022), becoming the most popular ARS in 2023. Out of the dimensional ARS, ***D.va*** is the most popular but still does not compete with the popularity of its categorical counterparts. Its presence is observed only in 2020 and 2021. ***C.5cr*** remains constant in

its comparatively low popularity from 2021-2023 and ***C.5cs*** was only sporadically popular above average, in years 2019 and 2022.

3.4 Basis for Affect Representation Schemes

We have seen that most (if not all) of Ekman's 6 Basic Emotions [3] are reflected in the emotions present in an overwhelming majority of categorical ARS. The deviations from these 6 emotions are highlighted and explained in section 3.2, but it still remains the basis of most ARS. The model is considered widely accepted in the field of psychology as per most supplementary dataset papers. Moreover, all of the dimensional models are motivated by Russell's Valence and Arousal [4] or Warriner et al. Valence, Arousal and Dominance [5] in psychological theory. There are two exceptional cases of ARS that are not based on psychological theory. Firstly, ARS ***C.5a*** as explained in section 3.2 represents labels for different behaviors. The ARS ***C.5f*** as motivated in section 3.2 represents labels for emotion-related states. However, for both of these schemes, a theoretical basis was not provided.

4 Responsible Research

A systematic review was chosen for this project to ensure reproducibility of results [6]. Ensuring that a systematic review can be replicated and verified by other researchers is crucial for its reproducibility. This is achieved by providing transparency in the research process. Thus, in section 2, we provide details with respect to searching, selecting and analyzing the papers with the goal that if the steps laid out there are followed by other independent researchers, it should lead to the same results. Although, it is still possible that there are some differences in the findings in the end, the procedure of a systematic review still mitigates the likelihood of this discrepancy. Additionally, the PRISMA guidelines followed also offer a comprehensive framework and checklist that was utilized to maintain transparency and consistency throughout the process of this research. The guidelines advise essential elements of the method, such as the search strategy, study selection criteria, data extraction methods, and quality assessment procedures. By adhering to PRISMA guidelines, we aim to achieve reproducibility in results to a large extent.

Other ethical issues could arise from different sources of biases. For example, the presence of graphs could instigate statistical bias. As explained in section 5, the number of datapoints relative to the number of distinct ARS might show certain ARS being much more popular than others. However, the scale is much smaller, potentially causing volatility in the results. It is possible to be mislead in a manner that the difference in popularity is overestimated this way. To prevent this, this has been clarified in the paper and in the future, a more representative sample can be taken into consideration.

Another source of risk could be the lack of discussion with respect to affect in the extracted papers. This might mean that certain rationale was not taken into consideration because the data could not be extracted from the papers itself. Additionally, since the motivation for various ARS were often retrieved from supplementary papers instead of the primary extracted ones, it is possible that the considerations made in the former were not actually made in the primary ones.

Therefore, we have made several ethical considerations in terms of reproducibility of results, potential statistical bias and risk due to lack of information. The procedure of a systematic review allows the opportunity to be transparent as well as provide results in a consistent and predictable manner.

5 Discussion

There are several limitations to be considered when analyzing the use of ARS in speech emotion recognition systems. Firstly, the utilization of different datasets introduces variations in contexts, purposes, gender balance, and types such natural, acted, or elicited data. This leads to a variety of ARS applications with the difficulty of deducing a universally accepted ARS for all scenarios. While the extracted papers do not explicitly justify the use of affect representation schemes, they primarily focus on the justification for selecting specific datasets. However, it is worth noting that the choice of datasets indirectly motivates the adoption of emotion models, thus indirectly motivating the use of affect representation schemes. Nevertheless, this does not reflect the existence of considerations made for a specific SER model for affect representation, rather a specific dataset.

Additionally, categorical representation schemes appear to be more prevalent than dimensional models, potentially due to their convenience in implementation. It is noteworthy that the motivation behind utilizing affect representation schemes often stems from practical or convenience-related considerations, rather than a deliberate reflection on how affect should be represented. For instance, the merging of happiness and excitement into a single category exemplifies this approach. Motivation for using specific ARS also sometimes relies on logistic matters with respect to creation and annotation of the dataset.

The ABC and FAU Aibo corpora, designed for different affective types, are still used alongside datasets that primarily target specific emotions. This suggests that the focus of the models seems to revolve around discriminating between different categories without thoroughly reflecting on the nature or significance of these categories. Additionally, the IEMOCAP dataset offers annotations for valence, arousal, and dominance, which provide additional dimensions for characterizing emotions. However, these attributes are seldom utilized by systems or models within the field, highlighting a missed opportunity for a more comprehensive understanding and representation of emotions.

Overall, there is also a tendency to mix up terminology related to attitude, mood, behavior, and emotion, further adding to the complexity and lack of consistency in the field. For example, paper [52] establishes its aim of predicting attitude

but employing datasets targeting emotions. Moreover, categories of emotions are often arbitrarily merged together for better performance on discriminability or because the emotions are considered similar enough (for example, anger and frustration or disgust and anger). Often papers Due to the lack of explicit motivation provided in the papers, the reliance on existing literature and establishing clear terminology becomes necessary. Sometimes, emotion label names are used interchangeably, such as happiness and joy, potentially leading to confusion and inconsistency.

Other limitations include perhaps a low number of papers analyzed relative to the variety of distinct ARS found. For both graphs provided to compare popularity of ARS (see figure 3), the values in terms of number of occurrences are not significantly high values, which could lead to instability in the results. A larger and more representative sample would lead to more robust results.

6 Conclusions and Future Work

Overall, we have successfully explored how different affect representation schemes are used in the field of SER. Emotion is the most targeted affective state and within this, categorical representations of emotions are commonly found. While miscellaneous modifications are often made to emotion categories, most of these are based on Ekman's basic 6 emotions. From these emotions, anger, happiness and sadness are most often targeted. Neutral state is also often targeted, although its recognition as an emotional one is controversial. Dimensional ARS are particularly less prevalent in the context of SER. This could potentially be explained by some papers claiming that valence, which is an integral aspect of dimensional models, is considered to be difficult to estimate via audio-only input. Most common combinations of ARS involve several categorical schemes applied to variations of the SER system due to multiple datasets being utilized. We also assessed the popularity of different ARS. The popularity of the ARS seems to have high correlation with the popularity of the dataset overall, with ***C.6ed*** and ***C.3i*** being particularly popular owing to the extensive use of EMO-DB and IEMOCAP corpora. ***C.7r*** respective to RAVDESS dataset is also more popular currently. Some ARS tend to aim to target behavior and/or attitude. However, there are several issues with respect to lack of explanation of categories chosen, interchangeable use of terminology and lack of a psychological foundation.

For future work, it would be beneficial to explore the role of neutral and potentially other baseline states such as the use of *calm* for ARS ***C.7r*** and how it is approached in a greater scope. Additionally, it would be useful to attempt to understand the reason for the current relative deficiency of dimensional models for speech emotion recognition. Moreover, it would be interesting to investigate multimodal approach to contrast with audio-only approach. Finally, a comparative analysis of different datasets could provide more useful insight into affect representation schemes than the papers corresponding to the automatic affect prediction models itself. Finally, the affect of other factors on affect prediction can also be explored such as age, gender, linguistic barriers and so on. In this way, we could gain more in-depth understanding of how affect can be represented for prediction through speech.

7 Acknowledgments

We would like to thank our supervisor for enthusiastically introducing us to this field of research and guiding us in the process of conducting the systematic review as well. We are also grateful to our responsible professor for the guidance and valuable feedback provided. It was a pleasure working with them as well as the whole team and we had some great discussions on the topic.

A Search Query Details

A.1 Micro-topics for Search Query (Web of Science Only)

- Neuroscanning
- Digital Signal Processing
- Artificial Intelligence and Machine Learning
- Human Computer Interaction
- Models of Computation

A.2 Keywords for Search Query

1. "Speech Recognition"
2. "Emotion Recognition"
3. "Speech Emotion Recognition"
4. "Speech"
5. "Speech Communication"
6. "Emotions"
7. "Speech Analysis"

8. ”Speech Processing”
9. ”Emotional Speech”
10. ”Emotion Detection”
11. ”Speech Emotions”
12. ”Human Emotion”
13. ”Human Emotion Recognition”
14. ”Continuous Speech Recognition”
15. ”Recognizing Emotions”
16. ”Speaker Recognition”
17. ”Recognition Systems”
18. ”Emotion Recognition From Speech”
19. ”Speech Perception”
20. ”Speech Emotion Recognition Systems”
21. ”Automatic Speech Recognition”

A.3 Datasets for Search Query

1. Berlin Emotional Database (EMO-DB)
2. Surrey Audio-Visual Expressed Emotion (SAVEE)
3. RECOLA Speech Database
4. SAMAINE Database
5. eNTERFACE’05 Audio-Visual Emotion Database
6. Interactive Emotional Motion Capture (USC-IEMOCAP)
7. FAU Aibo Emotion Corpus
8. BAUM-1 Speech Database
9. Situation Analysis in a Fictional and dEmotional corpus (SAFE)
10. Chinese Emotional Speech Corpus (CASIA)
11. Toronto Emotional Speech Database (TESS)
12. Beihang University Database of Emotional Speech (BHUDES)
13. Chinese Annotated Spontaneous Speech Corpus
14. Chinese Natural Emotional Audio-Visual Database (CHEAVD)
15. Danish Emotional Speech Database (DES)
16. Chinese Elderly Emotional Speech Database (EESDB)
17. Electromagnetic Articulography Database (EMA)
18. Italian Emotional Speech Database (EMOVO)
19. Keio University Japanese Emotional Speech Database (Keio-ESD)
20. LDC Emotional Speech Database
21. Speech Under Simulated and Actual Stress Database (SUSAS)
22. Vera Am Mittag Database (VAM)
23. TUM AVIC Database
24. AFEW Database
25. Turkish Emotional Speech Database (TURES)

B Data Extraction Questionnaire

Table 8: Guiding questions to extract data from the papers

Extracted Paper Questions	Possible Answers
What type of affective state is being targeted?	Emotion, mood or as claimed by the paper
What type of ARS is used?	Categorical or Dimensional
If categorical, then what are the categories targeted? OR If dimensional then what are the dimensions?	Happy, sad, angry et cetera for categorical OR Valence, arousal, dominance et cetera for dimensional
What is the source of ARS?	The system uses a custom-made ARS OR The system uses the ARS of the dataset(s) used
If the dataset's ARS is used then: 1. What is the motivation for choice of ARS and/or dataset? 2. What modifications, if any, were made to the ARS of the dataset?	Open question
What is the motivation for the ARS chosen, if provided?	Open question
How were the categories and/or dimensions decided?	Open question
Is the ARS based on psychological theory?	- Ekman's 6 basic emotions, - Russell's Valence and Arousal \cite{russell} - Warriner et al. Valence, Arousal and Dominance \cite{warriner} et cetera
Are multiple ARS used? OR Are multiple datasets with different ARS used?	Yes or No
If multiple ARS or dataset(s) used then which ones are used?	Open question

References

[1] Murray Alpert and Anna Rosen. A semantic analysis of the various ways that the terms "affect," "emotion," and "mood" are used. *Journal of Communication Disorders*, 23(4):237–246, 1990.

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

[3] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[4] James Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 12 1980.

[5] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45:1191–1207, 2013.

[6] Angela Boland, Rumona Dickson, and Gemma Cherry. Doing a systematic review: A student's guide. *Doing a Systematic Review*, pages 1–304, 2017.

[7] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, 88:105906, 2021.

[8] Renan Vinicius Aranha, Cleber Gimenez Correa, and Fatima L. S. Nunes. Adapting software with affective computing: A systematic review. *IEEE Transactions on Affective Computing*, 12(4):883 – 899, 2021.

[9] K.L. Ong, C.P. Lee, H.S. Lim, and K.M. Lim. Speech emotion recognition with light gradient boosting decision trees machine. *International Journal of Electrical and Computer Engineering*, 13(4):4020–4028, 2023.

[10] Lu-Yao Liu, Wen-Zhe Liu, and Lin Feng. Sdtf-net: Static and dynamic time-frequency network for speech emotion recognition. *SPEECH COMMUNICATION*, 148:1–8, MAR 2023.

[11] K.A. Kumar and J.L.M. Iqbal. Machine learning technique-based emotion classification using speech signals. *Soft Computing*, 27(12):8331–8343, 2023.

[12] Kummari Ramyasree and Chennupati Sumanth Kumar. Multi-attribute feature extraction and selection for emotion recognition from speech through machine learning. *TRAITEMENT DU SIGNAL*, 40(1):265–275, FEB 2023.

[13] M. Rayhan Ahmed, S. Islam, A.K.M. Muzahidul Islam, and S. Shatabda. An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition. *Expert Systems with Applications*, 218, 2023.

[14] Yalamanchili Bhanusree, Samayamantula Srinivas Kumar, and Anne Koteswara Rao. Time-distributed attention-layered convolution neural network with ensemble learning using random forest classifier for speech emotion recognition. *JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY-MALAYSIA*, 22(1):49–76, JAN 2023.

[15] A.S. Alluhaidan, O. Saidani, R. Jahangir, M.A. Nauman, and O.S. Neffati. Speech emotion recognition through hybrid features and convolutional neural network. *Applied Sciences (Switzerland)*, 13(8), 2023.

[16] Z. Chen, J. Li, H. Liu, X. Wang, H. Wang, and Q. Zheng. Learning multi-scale features for speech emotion recognition with connection attention mechanism. *Expert Systems with Applications*, 214, 2023.

[17] F. Pucci, P. Fedele, and G.M. Dimitri. Speech emotion recognition with artificial intelligence for contact tracing in the covid-19 pandemic. *Cognitive Computation and Systems*, 5(1):71–85, 2023.

[18] S. Sekkate, M. Khalil, and A. Adib. A statistical feature extraction for deep speech emotion recognition in a bilingual scenario. *Multimedia Tools and Applications*, 82(8):11443–11460, 2023.

[19] Y. Qi, H. Huang, and H. Zhang. Research on speech emotion recognition method based a-capsnet. *Applied Sciences (Switzerland)*, 12(24), 2022.

[20] Cheng Lu, Wenming Zheng, Hailun Lian, Yuan Zong, Chuangao Tang, Sunan Li, and Yan Zhao. Speech emotion recognition via an attentive time-frequency neural network. *IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS*, 2022 DEC 7 2022.

[21] A. Thakur and S.K. Dhull. Language-independent hyperparameter optimization based speech emotion recognition system. *International Journal of Information Technology (Singapore)*, 14(7):3691–3699, 2022.

[22] Fazliddin Makhmudov, Alpamis Kutlimuratov, Farkhod Akhmedov, Mohamed S. Abdallah, and Young-Im Cho. Modeling speech emotion recognition via attention-oriented parallel cnn encoders. *ELECTRONICS*, 11(23), DEC 2022.

[23] Z. Yang and Y. Huang. Algorithm for speech emotion recognition classification based on mel-frequency cepstral coefficients and broad learning system. *Evolutionary Intelligence*, 15(4):2485–2494, 2022.

[24] Xinlei Xu, Dongdong Li, Yijun Zhou, and Zhe Wanga. Multi-type features separating fusion learning for speech emotion recognition. *APPLIED SOFT COMPUTING*, 130, NOV 2022.

[25] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor. Recognition of emotions in speech using convolutional neural networks on different datasets. *Electronics (Switzerland)*, 11(22), 2022.

[26] Jia-Xin Ye, Xin-Cheng Wen, Xuan-Ze Wang, Yong Xu, Yan Luo, Chang-Li Wu, Li-Yan Chen, and Kun-Hong Liu. Gm-tcnet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition*. *SPEECH COMMUNICATION*, 145:21–35, NOV 2022.

[27] M.S. Fahad, A. Ranjan, A. Deepak, and G. Pradhan. Speaker adversarial neural network (sann) for speaker-independent speech emotion recognition. *Circuits, Systems, and Signal Processing*, 41(11):6113–6135, 2022.

[28] Shalini Kapoor and Tarun Kumar. A novel approach to detect instant emotion change through spectral variation in single frequency filtering spectrogram of each pitch cycle. *MULTIMEDIA TOOLS AND APPLICATIONS*, 82(6):9413–9429, MAR 2023.

[29] Mustaqeem and Soonil Kwon. Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network. *International Journal of Intelligent Systems*, 36(9):5116–5135, 2021.

[30] Shihan Huang, Hua Dang, Rongkun Jiang, Yue Hao, Chengbo Xue, and Wei Gu. Multi-layer hybrid fuzzy classification based on svm and improved pso for speech emotion recognition. *ELECTRONICS*, 10(23), DEC 2021.

[31] H. Zhang, H. Huang, and H. Han. A novel heterogeneous parallel convolution bi-lstm for speech emotion recognition. *Applied Sciences (Switzerland)*, 11(21), 2021.

[32] Ammar Amjad, Lal Khan, and Hsien-Tsung Chang. Effect on speech emotion classification of a feature selection approach using a convolutional neural network. *PEERJ COMPUTER SCIENCE*, 7, NOV 3 2021.

[33] Y. Dong and X. Yang. Affect-salient event sequence modelling for continuous speech emotion recognition. *Neurocomputing*, 458:246–258, 2021.

[34] Ning Jia and Chunjun Zheng. Two-level discriminative speech emotion recognition model with wave field dynamics: A personalized speech emotion recognition method. *COMPUTER COMMUNICATIONS*, 180:161–170, DEC 1 2021.

[35] S. Toliupa, I. Tereikovskyi, L. Tereikovska, S. Mussiraliyeva, and K. Bagitova. Deep neural network model for recognition of speaker's emotion. pages 172–176, 2021.

[36] Yangwei Ying, Yuanwu Tu, and Hong Zhou. Unsupervised feature learning for speech emotion recognition based on autoencoder. *ELECTRONICS*, 10(17), SEP 2021.

[37] R. Dhiman, G.S. Kang, and V. Gupta. Modified dense convolutional networks based emotion detection from speech using its paralinguistic features. *Multimedia Tools and Applications*, 80(21-23):32041–32069, 2021.

[38] Qiupu Chen and Guimin Huang. A novel dual attention-based blstm with hybrid features in speech emotion recognition. *ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE*, 102, JUN 2021.

[39] L. Matsane, A. Jadhav, and R. Ajoodha. The use of automatic speech recognition in education for identifying attitudes of the speakers. 2020.

[40] Zijiang Zhu, Weihuang Dai, Yi Hu, and Junshan Li. Speech emotion recognition model based on bi-gru and focal loss. *PATTERN RECOGNITION LETTERS*, 140:358–365, DEC 2020.

[41] A. Bakhshi, S. Chalup, A. Harimi, and S.M. Mirhassani. Recognition of emotion from speech using evolutionary cepstral coefficients. *Multimedia Tools and Applications*, 79(47-48):35739–35759, 2020.

[42] Zhaocheng Huang and Julien Epps. An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, 11(4):653–668, OCT 1 2020.

[43] M. Hou, J. Li, and G. Lu. A supervised non-negative matrix factorization model for speech emotion recognition. *Speech Communication*, 124:13–20, 2020.

[44] Huijuan Zhao, Ning Ye, and Ruchuan Wang. Coarse-to-fine speech emotion recognition based on multi-task learning. *JOURNAL OF SIGNAL PROCESSING SYSTEMS FOR SIGNAL IMAGE AND VIDEO TECHNOLOGY*, 93(2-3, SI):299–308, MAR 2021.

[45] Y. Dong and X. Yang. Affect-salient event sequences modelling for continuous speech emotion recognition using connectionist temporal classification. pages 773–778, 2020.

[46] Xusheng Ai, Victor S. Sheng, Wei Fang, and Charles X. Ling. An optimal model with a lower bound of recall for imbalanced speech emotion recognition. *MULTIMEDIA TOOLS AND APPLICATIONS*, 79(33-34):24281–24301, SEP 2020.

[47] P. Mishra and R. Sharma. Gender differentiated convolutional neural networks for speech emotion recognition. volume 2020-October, pages 142–148, 2020.

[48] Shruti Gupta, Md. Shah Fahad, and Akshay Deepak. Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition. *MULTIMEDIA TOOLS AND APPLICATIONS*, 79(31-32):23347–23365, AUG 2020.

[49] H. Chen, Z. Liu, X. Kang, S. Nishide, and F. Ren. Investigating voice features for speech emotion recognition based on four kinds of machine learning methods. pages 195–199, 2019.

[50] Jian-Hua Tao, Jian Huang, Ya Li, Zheng Lian, and Ming-Yue Niu. Semi-supervised ladder networks for speech emotion recognition. *INTERNATIONAL JOURNAL OF AUTOMATION AND COMPUTING*, 16(4):437–448, AUG 2019.

[51] S. Sekkate, M. Khalil, A. Adib, and S.B. Jebara. An investigation of a feature-level fusion for noisy speech emotion recognition. *Computers*, 8(4), 2019.

[52] Suman Deb and Samarendra Dandapat. Emotion classification using segmentation of vowel-like and non-vowel-like regions. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, 10(3):360–373, JUL-SEP 2019.

[53] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang. Speech emotion recognition based on dnn-decision tree svm model. *Speech Communication*, 115:29–37, 2019.

[54] Mia Atcheson, Vidhyasaharan Sethu, and Julien Epps. Using gaussian processes with lstm neural networks to predict continuous-time, dimensional emotion in ambiguous speech. In *2019 8TH INTERNATIONAL CONFERENCE ON AFFECTIVE COMPUTING AND INTELLIGENT INTERACTION (ACII)*, International Conference on Affective Computing and Intelligent Interaction, 2019. 8th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, ENGLAND, SEP 03-06, 2019.

[55] M. Gao, J. Dong, D. Zhou, X. Wei, and Q. Zhang. Speech emotion recognition based on convolutional neural network and feature fusion. pages 1145–1150, 2019.

[56] Mengna Gao, Jing Dong, Dongsheng Zhou, Qiang Zhang, and Deyun Yang. End-to-end speech emotion recognition based on one-dimensional convolutional neural network. In *3RD INTERNATIONAL CONFERENCE ON INNOVATION IN ARTIFICIAL INTELLIGENCE (ICIAI 2019)*, pages 78–82, 2019. 3rd International Conference on Innovation in Artificial Intelligence (ICIAI), Suzhou, PEOPLES R CHINA, MAR 15-18, 2019.

[57] A.B. Abdul Qayyum, A. Arefeen, and C. Shahnaz. Convolutional neural network (cnn) based speech-emotion recognition. pages 122–125, 2019.

[58] Lili Guo, Longbiao Wang, Jianwu Dang, Zhilei Liu, and Haotian Guan. Exploration of complementary features for speech emotion recognition based on kernel extreme learning machine. *IEEE ACCESS*, 7:75798–75809, 2019.

[59] Carlos Busso, Murtaza Bulut, Chi Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12 2008.

[60] F Burkhardt, A Paeschke, M Rolfes, W Sendlmeier, and B Weiss. A database of german emotional speech.

[61] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. 2018.

[62] Wei Bao, Ya Li, Mingliang Gu, Minghao Yang, Hao Li, Linlin Chao, and Jianhua Tao. Building a chinese natural emotional audio-visual database. volume 2015-January, pages 583–587. Institute of Electrical and Electronics Engineers Inc., 2014.

[63] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5:377–390, 10 2014.

[64] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the recola multimodal corpus of remote collaborative and affective interactions.

[65] Giovanni Costantini, Iacopo Iadarola, Andrea Paoloni, and Massimiliano Todisco. Emovo corpus: an italian emotional speech database.

[66] Der Technischen and Stefan Steidl. Automatic classification of emotion-related user states in spontaneous children's speech, 2009.

[67] Bjorn Schuller, Dejan Arsic, and Gerhard Rigoll. Audiovisual behavior modeling by combined feature spaces.

[68] O Martin, I Kotsia, B Macq, and I Pitas. The enterface'05 audio-visual emotion database, 2006.

[69] Siddique Latif, Adnan Qayyum, Muhammad Usman, and Junaid Qadir. Cross lingual speech emotion recognition: Urdu vs. western languages. 12 2018.

[70] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed Abdelwahab, Najmeh Sadoughi, and Emily Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8:67–80, 1 2017.

[71] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014 - 3d dimensional affect and depression recognition challenge. pages 3–10. Association for Computing Machinery, 11 2014.

[72] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schröder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3:5–17, 1 2012.

[73] Angeliki Metallinou, Zhaojun Yang, Chi chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan. The usc creativeit database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. *Language Resources and Evaluation*, 50:497–521, 9 2016.