Survey of Interrater Agreement in Automatic Affect Prediction for Speech Emotion Recognition

A Systematic review

Oscar Wezenaar
Supervisor: Bernd Dudzik
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 23, 2024

Name of the student: Oscar Wezenaar
Final project course: CSE3000 Research Project
Thesis committee: Bernd Dudzik, Catharine Oertel

Abstract

Emotional datasets for automatic affect prediction usually employ raters to annotate emotions or verify the annotations. To ensure the reliability of these raters some use interrater agreement measures, to verify the degree to which annotators agree with each other on what they rate. This systematic review explores what kind of interrater agreement measures are used in emotional speech corpora. The affective states, the affect representation schemes, and the collection method of the datasets as well as the popularity of these measures were investigated. Scopus, IEEE Xplore, Web of Science, and ACM digital library were used to extract papers that describe the creation of datasets. 45 papers were included in the review. The review concludes that the interrater agreement measures used, are highly dependent on the collection method for speech and the affect representation schemes. It was found that there is no standardized way to measure interrater agreement. Datasets that use actors to record emulated emotions mostly use recognition rate as their interrater agreement measures. Datasets that use a dimensional representation scheme often compute the mean agreement of the raters and the standard deviation of that measure to check the interrater agreement. Datasets that do not use actors nor are dimensional use a plethora of different measures such as probabilistic computing of agreement, or majority agreement measures, but a large amount use no measures at all.

1 Introduction

People use speech to express emotions, and correctly recognizing emotion helps navigate complex social environments [1]. However, different people might hear and recognize emotions differently, which makes its identification slightly subjective [2].

Speech affect prediction or the field of Speech Emotion Recognition (SER) tries to predict emotion or affect from speech. Affect, which is often used interchangeably with emotion, can be seen as the subjective and immediate effect of emotion or a subjective state of feeling [3]. SER has applications in aiding AI speech recognition machines, such as Alexa or Google Home [4], in diagnosing psychiatric problems [5, 6], and detecting depression [7].

To develop SER machines, the model is generally trained through supervised learning. This means that the datasets need to be labeled. However, two people might observe emotions differently. Some might confuse fear, with sadness for example. Therefore, labeling data on emotion is challenging. One solution to manage these differences of opinion is to devise agreement measures between the people who rate emotions. These measures are called interrater agreement measures and examples are kappa statistics or percentage agreement [8]. These measures enhance, and verify the reliability of the final labels chosen.

Existing reviews and research on emotional datasets for SER often focus on the emotions targeted, the collection method for the audio records, or the feature and classifier selection [9, 10, 11, 12, 13]. There is rarely any focus on the interrater agreement measures. Therefore, it is unclear what measures are being used in the creation of emotional speech corpora.

To research these measures this paper seeks to answer the following question: *"How do existing datasets for Speech Emotion Recognition differ concerning interrater agreement measures?"* To help answer this question several sub-questions have been identified as shown in table 1.

A systematic review will be conducted to assess the use of interrater agreement in datasets created for SER. A systematic review was chosen to give a reproducible and unbiased overview of the currently existing published works. In section 2 the methodology of the systematic review will be described in detail, with the results of the literature search. Section 3 will present the results found by the systematic review. In section 4, the ethical considerations surrounding this research will be examined. In section 5, the interpretation of the results will be discussed. Lastly, in section 6, a conclusion will be drawn from the findings of the research, and the possibilities of future work will be explored.

2 Methodology

This paper is structured according to PRISMA guidelines for report writing [14]. The survey will be performed according to the steps explained in the student guide to systemic reviews [15]. Firstly, the eligibility criteria for a paper to be included in the review are formulated. The search databases used to retrieve the literature are discussed after. Thirdly, the search strategy and the queries used to build including the constraints for feasibility are presented. Next, the selection process of retrieved papers is described, followed by a description of the data extraction process. Lastly, the results are described and presented.

Table 1: Sub-questions

Sub-question	Explanation
SQ1: What types of affective states have been targeted by datasets (e.g., only emotions or mood?)	To fully understand how the affect is modeled, it is important to know what affective states have been targeted. Is it only emotions, or is mood also targeted?
SQ2: What different affect representation schemes have been used in these datasets (and what is the specific motivation for using specific schemes)?	To understand how different raters have rated the affect in Speech, it is necessary to know how the affect is modeled. There might be a large difference in interrater agreement when using the six emotions in comparison to the Valence-Arousal model for example.
SQ3a: Do datasets collect multiple ratings for a record (and how many)?	To measure interrater agreement within a dataset, this is needed to see the differences.
SQ3b: If so, do datasets measure interrater agreement?	It is relevant to see if they also measure the agreement, instead of just using multiple raters.
SQ3c: What measures do they use for this (and what is the level of agreement)?	To further understand how interrater agreement is used, the measurement strategies need to be retrieved.
SQ3d: Do dataset creators use any strategies to facilitate/facilitate interrater agreement (and what are these)?	Lastly, if the interrater agreement is measured, do the creators use any strategies to enhance the agreement or ratings?
SQ4: Is there a change in how datasets measure interrater agreement over time?	We want to understand how the interrater agreement measures and usage have changed over time. Do newer researchers use it more, less, or the same? Did the strategies change over time?
SQ5: Is there a relationship between the affect representation scheme used by datasets and their interrater agreement?	As mentioned for SQ2, we want to see the difference a representation scheme can have on the interrater agreement, to see if some schemes might be more intuitive for raters to rate in.

2.1 Eligibility criteria

To ensure a systematic approach to the review, and consistency in selecting papers, inclusion and exclusion criteria need to be identified. Inclusion criteria describe the specific attributes a study or source must have to be included in the review. Exclusion criteria describe the attributes that disqualify a

study from being in the review. To identify these criteria sub-questions that will help answer the main research question are considered and defined in table 1. These sub-questions form the foundation of the eligibility criteria, and the necessity to answer the main question is further specified. The identified Inclusion Criteria are explained in table 2 and the exclusion criteria in table 3. Lastly, we identified known speech corpora following this review on SER [9].

Table 2: Inclusion Criteria

Inclusion Criteria	Motivation
The paper describes the creation of a dataset.	The scope of the research is finding the datasets to analyze. Therefore, only papers that describe these datasets are to be included.
The paper uses a scheme or a consistent way to represent emotion.	The emotion has to be represented consistently for all raters within the dataset. Raters should not be able to choose different representations.
The dataset is or can be used for SER.	SER is the subject of the research. A dataset can be used for SER if they save speech records with emotion labeled to these records.

Table 3: Exclusion criteria

Exclusion Criteria	Motivation
The paper is not written in English.	The paper should be readable.
The dataset does not use human labeling to label emotion.	The goal of this research is to survey the inter-rater agreement in these datasets. Therefore, human labeling is necessary. Note that datasets that use actors to emulate emotions will not be excluded, as the dataset receives the annotations from humans.
The dataset represents emotion in a binary manner.	Datasets can represent emotion in a binary manner, e.g. this speech is emotional or this speech is not emotional. These datasets are excluded, as these datasets do not target specific affective states.
The paper describes a multi-modal database and this dataset does not store speech with emotion labels separately.	Datasets with data of multiple different emotional inputs exist. If these datasets store speech as a distinct input with separate labels, they can be included. If these datasets store e.g. videos with speech, and that combination is labeled they should be excluded.
The dataset uses gibberish as emotional speech.	Datasets may focus on speech without any meaning with emotional intonations behind it. For this review these are excluded.

2.2 Search databases

The databases used for the collection of papers are Scopus, Web of Science, IEEE Explore, and ACM Digital Library. These databases were chosen, as the TU Delft library allows access to all these databases, and the results will not only offer a large recall for the area of Computer Science but also fields such as psychology, which is prevalent in this research.

2.3 Search Strategy

For the project, the necessary papers are papers that describe emotional speech datasets. Furthermore, these datasets are to be used for Speech Emotion Recognition. For these reasons, the following keywords were identified: Speech, Affect, Recognition, and Dataset, see table 4. Narrowing keywords were also identified to further specify the search, as shown in table 5. To verify the accuracy of our query, 8 datasets with their corresponding papers were identified and used for validation of the query. The scoped datasets should be returned in the query. These datasets can be found in Appendix A.2. Lastly, to attain the final query, shown in table 6, a feasibility constraint was applied, such that the main keywords were restrained to only the title. This resulted in the loss of two scoped papers in the query search, but this was found acceptable. If no feasibility constraints were in play, the final search query shown in Appendix A.3 is preferred.

Table 4: Keywords

Affect	Speech	Recognition	Dataset
Affect*	Speech	recogni*	dataset
Emotion*	Speaker	predict*	database
mood*	Spoken input	detect*	corpus
feeling	audio	analysis	
		classification	
		represent*	
		annotat*	
		label*	

Table 5: Narrowing Keywords

Classification	Creation
classification	creat*
represent*	develop*
analysis	construct
label*	produc*
annotat*	design*
	collect*
	build

2.4 Selection Strategy

After generating the lists of papers, and the removal of duplicate papers, they need to be further screened on the inclusion and exclusion criteria to ensure their validity. That will be done in the following manner.

Table 6: Final Search Query (Scopus)

Query	(TITLE-ABS-KEY ((represent* OR "classification" OR "analysis" OR label OR annotat*) AND (predict* OR recogni* OR detect* OR identif* OR interpret* OR measur) AND (develop* OR create* OR construct* OR produc* OR design OR collect* OR buil*)) AND TITLE (emotion* OR affect*) AND TITLE ("Speech" OR "Spoken input" OR "audio" OR "speaker") AND TITLE ("dataset" OR "database" OR "corpus"))
Result	165
Intersection	6/8

1. *Filtering by title:* The first step of the filtering process. If it is concluded from the title that the paper is on another subject and/or does not adhere to the eligibility criteria, they are excluded.
2. *Filtering by abstract:* The first step is repeated, but now the abstract will be analyzed. This will likely be done simultaneously with the title filtering step.
3. *Full-text filtering:* The papers that remain, will all be obtained for full-text screening. In this stage of screening, the papers are read fully, to again check if they meet the eligibility criteria.

The entire identification of papers is summarized in the PRISMA flow diagram as shown in figure 1.

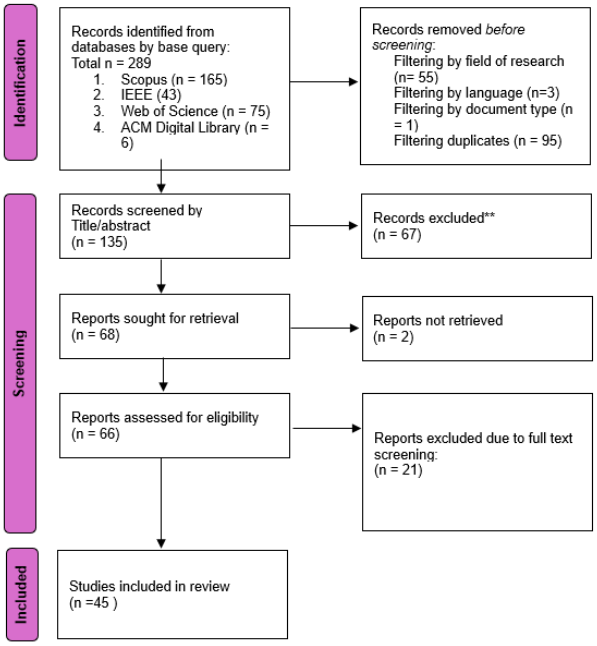


Figure 1: PRISMA flow diagram

2.5 Feasibility Filtering

Due to a large recall of papers, it was deemed infeasible to analyze all papers on their eligibility criteria for a 10-week project. Therefore, feasibility criteria were constructed to address this issue. This filtering is done in addition to the filtering operations mentioned in section 2.4. In this filtering process, it is eligible papers will be excluded, but due to time constraints, this is unavoidable. This filtering process will be done as follows:

1. *Title constraints in the Base Query:* As mentioned in section 2.3, the main keywords were constrained to be only in the title. Generally, widening the keywords to abstract is safer, as there is more room for error within the chosen keywords, and in title naming by the authors. However, as only 2 papers were lost from this constraint, it was deemed acceptable.
2. *Filtering by Research Topic:* Scopus and Web of Science have built-in filtering options, that allow for filtering on research topics. To narrow results in these two databases only papers within this field will be retrieved.

2.6 Search Results

The protocol, as described above, resulted in a set of 45 papers that are included in the review. These papers and their properties are shown in chapter 3 in table 7.

2.7 Data Extraction

The papers that are selected after the screening stages are used for data extraction. The information necessary to answer the research question, derived from the sub-questions, shown in table 1 is listed below. This is ultimately also the information that is retrieved from the datasets.

- Affective states targeted (SQ1)
- Affect representation scheme/model (SQ2 + SQ5)
- Source of emotional speech (SQ5)
- Annotating strategy (SQ3 + SQ5)
- Ratings per record (SQ3)
- Interrater agreement, if applicable (SQ3 + SQ5)
  - Measuring method
  - Level of agreement
  - Facilitation strategy
- Interrater strategies (SQ3)
- Publication year (SQ4)

3 Results

In this section, the data extracted from the 45 papers is presented. Table 7 shows the included datasets with the properties relevant to this review. Datasets that did not have a defined name are indicated with (P) in the table. Subsection 3.1 describes the affective states targeted by the datasets, and the schemes used to represent the emotions, as shown by the second and third column of table 7. Secondly, subsection 3.2 will describe how the audio records for the datasets are collected

Table 7: Datasets included in the review and their properties

Dataset Name	Emotion model	Affective states targeted	Speech	Raters	Interrater measurement	Year
DES [16]	Categorical	Happy, angry, sad, surprise, neutral	1	20	Recognition rate 67%	1997
MASC [17]	Dimensional	Happy, angry, sad, fear, neutral	1	-	-	2006
VAM [18]	Dimensional	Valence, Arousal, Dominance	3	17	Self Assessment Manikin	2008
IITKGP-SESC [19]	Categorical	Happy, anger, sad, fear, disgust, surprise, compassion, sarcastic	1	25	Recognition rate 64%	2009
Polish Emotional Speech Database [20]	Categorical	Happy, anger, sad, fear, disgust, surprise, neutral	1	202	Recognition rate 57%	2009
The New Italian Audio and Video Emotional Database [21]	Categorical	Happy, anger, sad, fear, surprise, sarcasm	1	40	Recognition rate 50%	2010
KEG [22]	Categorical + Dimensional	Happy, Anger, Sadness, Fear, Neutral + VA	3	5	Majority agreement 74%	2011
IITKGP-SEHSC [23]	Categorical	Happy, anger, sad, fear, disgust, surprise, sarcastic, neutral	1	25	Recognition rate 72%	2011
Tamil Corpus [24]	Categorical	Happy, anger, sad, fear, neutral	3	5	-	2014
(P) [25]	Categorical	Happy, anger, sad, fear, disgust, boredom, neutral	1	10	Recognition rate	2014
EmoLUKS [26]	Categorical	Happy, anger, sad, fear, disgust, surprise, neural	3	5	Majority vote 89% included	2015
REGIM_TES [27]	Categorical	neutral, sadness, happiness, anger, fear	1	10	Recognition rate 67%	2016
TED-LIUM [28]	Categorical	Angry, happy, sad, neutral + 8 secondary emotions	3	7 + 3 crowdsourced	Majority agreement 82% included	2017
(P) [29]	Categorical	Happy, angry, sad, neutral	1	0	-	2017
Simulated emotion speech database [30]	Categorical	Happy, angry, sad, neutral	1	0	-	2017
Phonetically and Prosodically Balanced Database [31]	Categorical	Happy, angry, sad, neutral	1	3	Average degree of emotion	2017
KES [32]	Categorical	Happy, angry, sad, fear, neutral	1	25	Recognition rate 95%	2018
(P) [33]	Categorical	Happy, angry, sad, fear, neutral	3	50	Mean Opinion Score	2018
MSP-Podcast [34]	Categorical + Dimensional	happy, angry, sad, fear, disgust, surprise, contempt, neutral + Valence Arousal Dominance	3	5	SAMs, Cohen's Kappa, Krippendorff's alpha $\alpha \approx 0.4$	2019
Emotional Speech, Video and Gestures database [35]	Categorical	happy, anger, sad, fear, disgust, surprise	1	12	Recognition rate 80%	2019
TaMaR-EmoDB [36]	Categorical	Happy, anger, sad, anxiety, happy, neutral	1	20	Recognition rate 85%;95%;80%	2019
Dataset for depression detection [37]	Categorical	Happy, anger, sad, fear, surprise, neutral, depressed	1	-	-	2019
Urdu-Sindhi Speech Emotion Corpus [38]	Categorical	Happy, anger, sadness, disgust, surprise, sarcasm, neutral	1	-	-	2020
DEMoS [39]	Categorical	Happy, anger, sadness, fear, disgust, surprise, guilt	2	3	Majority agreement 23% excluded	2020
(P) [40]	Dimensional	Happy, anger, sadness, fear, disgust, surprise, trust, anticipation	3	3	average of intensity values	2020
Adult Emotional Speech Corpus [41]	Dimensional	Happy, angry, sadness, neutral + Valence Arousal Dominance	2	-	Mean accuracy	2020
Emotional Gujarati Speech Corpus [42]	Categorical	happy, angry, sad, surprise, disgust, fear	1	-	-	2020
LSSED [43]	Categorical	Happy, anger, sad, fear, disgust, surprise, disappointed, bored, excited, neutral	2	1	-	2021
Saudi Dialect Corpus [44]	Categorical	anger, happy, sad, neutral	3	1	-	2021
Urdu Emotional Speech Corpus [45]	Categorical	Happy, angry, sad , digust, neutral	1	1 psychologist + 15 students	Recognition rate 85%	2022
MES-P [46]	Categorical + Dimensional	Happy, anger, sad, neutral + Valence Arousal	1	7	Mean recognition + Cohen's Kappa 87%	2022
HI,KIA [47]	Categorical	Happy, angry, sad, neutral	1	8	Recognition rate avg $\approx 65\%$	2022
CADKES [48]	Categorical	Happy, Angry, Sad, fear, boredom, neutral	1	25	Recognition rate + Fleiss' Kappa 80%;$\kappa \approx 0.7$	2022
GreThe [49]	Dimensional	Valence and Arousal	3	4	Mean deviation 0.5	2022
Quechua Collao speech corpus [50]	Categorical + Dimensional	happy, angry, sad, fear, bored, excited, sleepy, neutral + Valence Arousal Dominance	1	4	Self Assessment Manikin + Cronbach's alpha $\alpha \approx 0.55$	2022
PEMO [51]	Categorical	happy , angry, sad, neutral	3	3	Full agreement	2022
NITK-KLESC [52]	Categorical	happy, angry, sad, fear, neutral	1	-	-	2023
DESCU [53]	Categorical	Happy, angry, sad, neutral	1	8	Recognition rate + Randolph's free-marginal kappa 73%;$\kappa = 0.64$	2023
Dusha [54]	Categorical	Happy, angry, sad, neutral	1 + 3	at least 3	Dawid-Skene algorithm	2023
BMISEC [55]	Categorical	Happy, angry, fear, sad, disgust, surprise, neutral	3	2	-	2023
Punjabi Audio Emotional Dataset [56]	Categorical	Happy, angry, sad, fear, surprise, neutral	1 + 3	0	-	2023
EAED [57]	Categorical	Happy, angry, sad, fear, surprise, neutral	3	4 + 1 for tiebreak	Majority vote	2024
JVNV [58]	Categorical	Happy, angry, sad, fear, disgust, surprise	1	-	-	2024
MEACorpus [59]	Categorical	Happy, anger, sad, fear, disgust, neutral	3	3	-	2024
EmoMatchSpanishDB [60]	Categorical	Happy, angry, sad, disgust, fear, surprise, neutral	1	3 to 10	Majority vote	2024

and annotated. The numbers in the speech column of table 7, refer to the three methods described in this section. Furthermore, in subsection 3.3 the interrater agreement measures and strategies are presented. Lastly, in subsection 3.4 the changes in interrater measurement over time are shown.

3.1 Affective States

All of the papers collected only targeted emotions for affective states. Mood was mentioned as a possibility in future work and recommendations in a few papers, but ultimately never used for the creation of the datasets.

As shown in table 7 most papers used a categorical representation scheme. The number of emotions targeted, and the specific emotions do differ. The emotions happy, angry, and sad are used in all datasets, and most datasets use some variation of Ekman's big six emotions, happy, angry, sad, fear, disgust, and surprise, as shown by table 8. Datasets that used simpler schemes to model emotions were mainly focused on emotions easy to emulate or understand for raters, or their higher frequency of appearance in natural emotion. Some databases also targeted topic-specific emotions. For example [37], a dataset created for depression detection, based their model on Ekman's big six emotions, but instead of using disgust, they opted for depressed. TED-LIUM [28] used happy, angry, sad and neutral as their primary emotions, but used ted-talk or presentation-specific secondary emotions such as leadership, friendly, or arrogant. The last three rows of table 8 depict the dimensional emotional models, which only count 9 datasets.

Table 8: Affective states targeted in different datasets

Affective States	**Papers**	**Number of Papers**
Primary Emotions (happy, angry, sad)	[28], [30], [29], [31], [41], [44], [45], [46], [47], [51], [53], [54]	12
Primary emotions with fear	[22], [24], [27], [32], [33], [36], [52]	7
Variation of Ekman's six emotions (happy, angry, sad, fear, digust, surprise)	[18], [19], [20] [21], [25], [26], [34], [35] [37], [38], [39] ,[42], [43], [48], [50], [55], [56], [57] [58], [59], [60]	21
Neutral	[16], [17], [20], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [36], [37], [38], [41], [43], [44], [45], [46], [47], [48], [50], [51], [52], [53], [54], [55], [56], [57], [59], [60]	34
Extended emotions (including boredom, sarcasm, etc.)	[19], [21], [23], [28], [37], [34], [39], [43]	8
Plutchik's wheel of emotion	[17], [40]	2
Valence, Arousal	[21], [22], [49]	3
Valence, Arousal, Dominance	[18], [34], [41], [50]	4

3.2 Speech Collection and Annotation Strategies

When creating a speech corpus there are 3 common ways to collect the audio records.

1. Hiring and recording actors emulating specified emotions.
2. Creating a controlled environment where emotion is induced in participants
3. Collecting audio records from real-life scenarios or films and using annotators to identify emotions.

These different methods will determine the way these datasets will be annotated and use raters to create the dataset. The datasets using the first method do not use raters to annotate the records, as the actors already emulate, and thus annotate the necessary emotions. However, they often use raters to increase the reliability of the actors' emulations; this will be further discussed in section 3.3. The second approach generally requires manual annotators. Even though emotion is induced in its participants, the validity of these induced emotions cannot be ensured. Therefore, annotators are used to establish some level of correctness, in the induced emotions. In the last approach, annotators are an absolute necessity, as the first-person emotions are unknown.

Table 9: Collection method for audio records of datasets

Collection method	**Papers**	**Number of papers**
Actors	[16], [17], [20], [19], [21], [23], [25], [27], [29], [30], [31], [32], [35], [36], [37], [38], [42], [45], [46], [47], [48], [50], [52], **[54]**, **[56]**, [58], [60]	28
Inducing emotions	[39], [41], [43]	3
External audio records	[18], [22], [24], [26], [28], [33], [34], [40], [44], [49], [51], **[54]**, [55], **[56]**, [57], [59]	16

As table 9 shows, actors are the most common approach for creating audio for an emotional speech corpus. Even though hiring actors can be an expensive approach, it is a reliable way to create a large amount of emotional utterances. Inducing emotion, while it is the best at eliciting natural emotion, is less popular as it requires more time and work in comparison with the alternatives. Lastly, using external audio is used frequently, as it is the cheapest choice. [54] and [56] both appear twice because they are created by combining recordings collected from both methods.

3.3 Interrater Agreement Measures and Strategies

The interrater agreement measures are different for all three annotation strategies and are different for categorical and dimensional emotion schemes. This also means that some measures cannot be applied to some databases. However, datasets with similar structures often use similar measures for interrater agreement. The frequency of the measures is shown in figure 2. The methods shown in the graphs will be explained in the subsections 3.3.1 to 3.3.5.

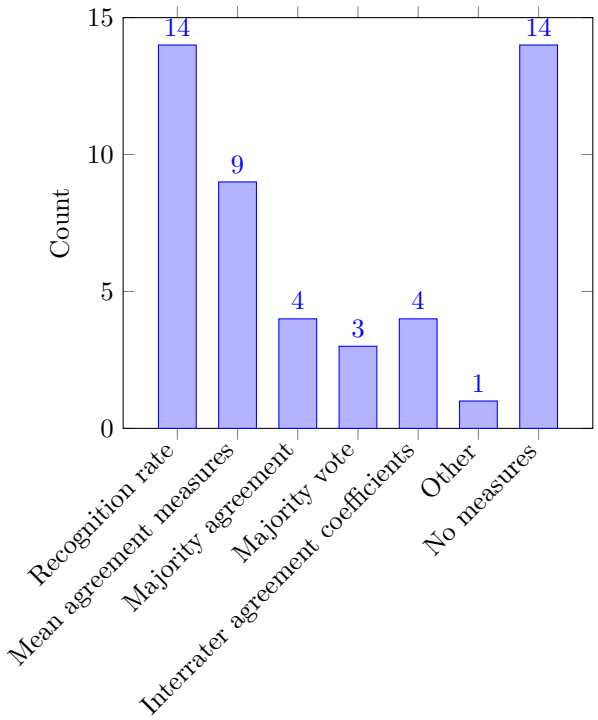
3.3.1 Recognition Rate

When speakers are recorded emulating desired emotions, these emotions are known, and therefore it can be argued that raters are unnecessary. However, raters can be employed to verify if the speakers conveyed these emotions properly, and if they are recognizable to people. This is the most common measure of interrater agreement in datasets that use actors, used in the following datasets: [16], [19], [20], [21], [23], [25], [27], [32], [35], [36], [45] [47], [48], [53]. The recognition rate is the percentage at which records of certain emotions get recognized correctly by the raters. For example, if five raters rate an utterance emulating anger, and four correctly label the utterance as anger, the recognition rate is 80%. This is a form of interrater agreement, as the recognition rate also reflects the rate at which raters agree on the emulated emotion.

To further the quality of the databases, some databases decided to exclude speech excerpts with low recognition rates [27] [45] [47]. Other datasets may only use this measurement as a subjectivity analysis for completeness.

Lastly, figure 3 shows the relation between the number of emotions targeted and the number of raters with the recognition rate.

Figure 2: Frequency of Interrater agreement measures



3.3.2 Mean agreement measures

For dimensional or numerical representations of emotion, mean agreement measures are a common measure to label the emotion. Often the standard deviation is calculated to find the interrater agreement.

One example that facilitates this, is the emotional speech evaluation method using Self Assessment Manikins (SAMs). This was proposed by the creators of the VAM database in earlier work to evaluate emotion in dimensional space [61]. SAMs offers an image array of 5 images for the three emotions in the Valence, Arousal, and Dominance model. Raters are to rate valence arousal and dominance of a speech record based on these images, which reflect a scale from 1 (low intensity) to 5 (high intensity). To evaluate true emotion and to minimize outside influences and biases an additive Gaussian noise signal is assumed. The method uses a Maximum Likelihood Estimator (MLE) to evaluate the mean true emotion of all raters. The interrater agreement for a single audio record can then be calculated by the standard deviation $d^{(i)}$ where:

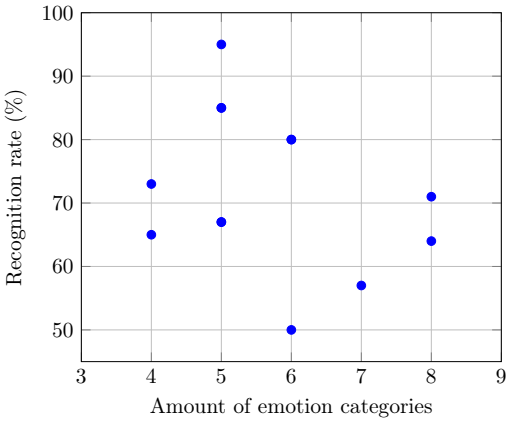
$$d^{(i)} = \sqrt{\frac{1}{(K-1)} \sum_{k=1}^{K} \left(x_k^{(i)} - x^{MLE,(i)}\right)^2}$$

where $i = \{V, A, D\}$, $K = number\ of\ raters$, and $x = value\ of\ emotion$. To produce the final annotations an Evaluator Weighted Estimator is used which takes the average of annotations weighted by the annotator's confidence score. The confidence score is estimated by differences in an annotator's ratings of multiple records and the MLE for those records.

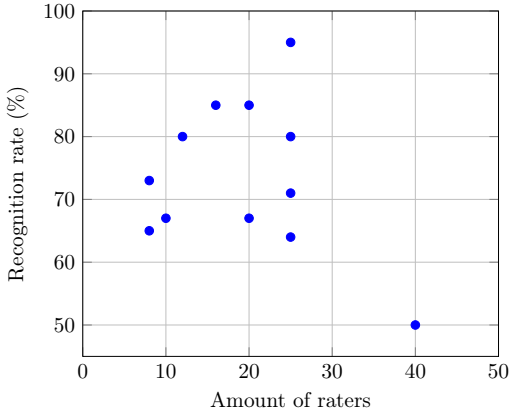
Evaluation with SAMs is used by three datasets [18] [34] [50]. [41] uses the same measurement and process, but does not mention the use of images such as SAMs. Similar measures of mean deviations are also used in [46] and [49].

Figure 3: Recognition rate with affective states targeted and amount of raters

(a) Recognition rate with number of emotion categories



(b) Recognition rate with amount of raters



3.3.3 Majority Agreement and Voting

As shown by table 7 majority agreement and voting measures are mostly used in datasets that represent emotions categorically and need raters to label these emotions. For a majority agreement measure, more than half of the annotators should label a record the same emotion. This measure is used by [22], [28], [39]. When a record does not reach this majority agreement, it is excluded from the dataset. For majority voting the emotion, which most raters choose, gets used as the label. Even if less than half of the raters agree. This measurement is used by three other datasets [26] [57] [60], where [60] is the main outlier, as it uses actors to collect speech. Lastly, [51] uses a full agreement measurement with three raters, where a record is only labeled and included if all raters agree.

These measures are not true interrater agreement measures. They do not calculate certain agreement levels between multiple raters. They, however, enforce percentage agreement on the speech clips used. Percentage agreement is to what percentage raters agree with each other [8]. Majority agreement, for example, enforces a minimum of 50% percentage agreement, as it screens the records where this is not reached. Therefore the average agreement in the dataset, which is often not mentioned in these papers, is also above 50%.

3.3.4 Interrater Agreement Coefficients

Kappa coefficients, κ, are statistical measurements that quantify interrater agreement. Kappa can be calculated as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o refers to the observed agreement probability and P_e refers to the prior probability. Due to the use of the prior probability the chance of raters accidentally agreeing with each other is taken into account. Different kappa measurements might calculate P_e and P_o differently, but most use this base to calculate κ.

Cohen's kappa, which is used in [46], calculates the reliability between two raters. It ranges from 0 to 1, and any value of κ above 0.6 can be considered substantial.

Fleiss' kappa, used in [48], is an interrater agreement measurement between more than two raters. It calculates the degree to which raters agree in their assessment of the records regarding the emotions

they could choose. Fleiss' κ, which ranges between -1 and 1, has positive values for slight to perfect agreements and negative values for anything worse. Anything above 0.6 can again be considered substantial.

Randolph's free-marginal kappa is an alternative to Fleiss' kappa for agreement measurements for multiple raters. For Fleiss' kappa P_e is assumed to be known to a certain extent. Randolph's free-marginal kappa is the alternative when P_e is either unknown or untrustworthy [62]. [53] uses Randolph's free-marginal kappa.

Another alternative, used by [34] is Krippendorff's alpha coefficient:

$$\alpha = 1 - \frac{D_o}{D_e}$$

where D_o is the observed disagreement and D_e the expected disagreement by chance. Krippendorff's alpha is a more generalized method that can be used for every number of raters, different models and scales, etc. [63].

3.3.5 No Interrater agreement Measures

Noticeable in figure 2 is the high number of datasets that have no interrater agreement measures. For some datasets, they may use some sort of interrater agreement or strategy, but this is not mentioned in the papers [43], [44]. [43] mentions it uses teamwork annotations, but it is not mentioned what that means. [44] mentions very little from the annotation process, so it is unlikely they used agreement measures, but inconclusive. [30] argues against the use of raters, as they use emotionally biased text that would influence human raters. [59] is a dataset focused on the feature extraction of emotional speech and therefore does not focus on raters. The ones not mentioned all use actors to emulate emotions and do therefore not use extra human raters.

3.4 Interrater Agreement Measures Over Time

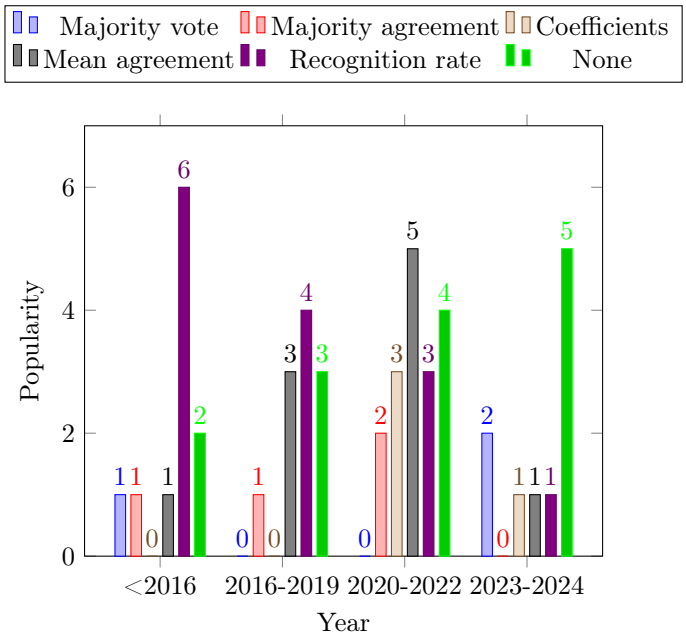
In this section the popularity of the interrater agreement measures over time will be discussed. The measures of all 45 papers are divided into roughly equal groups by years published of all papers as shown in figure 4. The groups contain 11, 11, 14, and 9 datasets respectively. Datasets that use multiple measures appear in the graph multiple times.

Figure 4 shows that the recognition rate seems to have declined slightly. while using no measures has increased over time. Mean agreement measures spike in 2020-2022, which is in line with a relatively high amount of dimensional datasets during that time frame. The use of coefficients showed growth in 2022 but lessened in 2023 until now. Majority vote and agreement have not been popular overall, and therefore do not show clear trends.

4 Responsible Research

This paper describes a systematic review. This was chosen to ensure reproducibility and impede possible biases. As described in section 2, to ensure a systematic approach, eligibility criteria were identified to ensure clear rules for including and excluding papers. The search databases used to find papers for the review were named. The search queries used to retrieve papers are also shared to ensure different researchers can retrieve the same papers. Additionally, the student's guide to systematic reviews [15], and PRISMA guideline for writing systematic reviews [14], were used. By following known procedures reproducibility is guaranteed to a certain extent. However, the review is carried out by a bachelor student with no prior experience with systematic reviews of this magnitude. Therefore, possible errors may arise from inexperience. Although procedures were followed, this should remain a consideration.

Figure 4: Popularity of Measuring Types Over the Years



Another ethical consideration regarding SER can be overreliance in the medical field. SER has applications in the detection of depression [7] and psychiatric problems [5, 6]. This can be very helpful, but the medical specialist should always carry the main responsibility of diagnosing these disorders. Care should be taken in using these measures. Furthermore, the subjectivity in rating could lead to biased datasets and SER systems. Creating these datasets with diversity in raters lessens possible biases. To address inescapable biases these datasets should also be made transparent, such that users of these systems are aware of the possible subjectivity.

5 Discussion

Certain limitations should be discussed when analyzing the interrater agreement measures between the datasets. Firstly, what has barely been discussed is the different goals the datasets were created for. Not every paper that created a dataset, had the creation of the dataset as its main goal. Many used it to test various other applications such as feature extraction or SER algorithms. When the goal is not the dataset it is logical to spend less time in its creation.

Furthermore, the method of collection matters a lot in how raters are used. Datasets that use actors do not need raters for labeling. In general, they see it as a good measure to have them, but slightly as an afterthought. If they do use raters, calculating the recognition rate is most common. This method cannot be used for other collection methods.

Thirdly, the measures used are also dependent on the affect representation scheme type. Categorical schemes heavily outnumber the dimensional ones. However, most dimensional sets use some interrater agreement measure. Most use standard deviation, which is also often used to calculate the final label. In this case, having an interrater agreement measure is part of the annotation process, and thus necessary. Categorical representation does not have this requirement for interrater agreement measures. There is more freedom to choose, but also to ignore.

Lastly, the datasets are made for different languages. There might be differences in how emotional a language is in general and that might affect interrater agreement. This is not taken into account

when measuring.

There are also some limitations in the process of the review that need to be discussed. Firstly, the feasibility criteria that were set up excluded some eligible papers that otherwise should have been included in the review. An example of one that was excluded due to feasibility was the German EmoDB [64]. This database was not only part of the verification process described in 2.3 but is also cited by most papers included in the review.

Secondly, the final search query used is relatively precise. Eligible datasets will have been filtered out due to this precise query.

For interpreting the results it is hard to describe a common theme for interrater agreement measures in Speech corpora, mostly due to the limitations presented above. Most of the methods are highly dependent on earlier choices made in the creation process. Most datasets that use actors, often use a recognition rate that is exclusive to this method. The other speech collection methods, which need raters to label emotions, are often content with checking which emotion is chosen most. Dimensional datasets do often use interrater agreement measures, as they often calculate the standard deviation. However, those measures are only valid for a dimensional or numeric representation of emotion.

The coefficients explained in 3.3.4 however, could be used by every dataset that uses multiple raters, as they compare the probability of raters choosing particular emotions. These are however hardly using these measures. This can either be explained by the use of different easier measures that were mentioned for reliability or the papers' low priority for high interrater agreement.

6 Conclusions and Future Work

Datasets were reviewed on their interrater agreement measures. Measuring recognition rate was used most commonly in total, and was used in almost all datasets that used actors. For the other collection methods, the used measures were more diverse. Most datasets with dimensional representation schemes, used mean agreement measures while categorically represented datasets enjoyed more flexibility. 14/45 of the datasets used no interrater agreement measures. Overall there is no standard interrater agreement measure that all datasets use. Datasets generally use the interrater agreement measure that best fits the context of how they created the dataset, if they implement any. Interrater agreement measures that can be used universally are not used often.

For future work, it would be beneficial to look at how the use of interrater agreement measures affects the empirical performance of the SER model trained on the datasets. Additionally, research could be done on the effect other factors like age, gender, and language have on the creation of speech datasets for affect prediction. Research on the imbalance between dimensional emotion and categorical emotion can further our understanding of the creation of speech datasets. Lastly, to elevate the comparability between different datasets regarding the interrater agreement, it would be recommended to investigate the possibility of standardizing the interrater agreement measures in emotional speech datasets. This can elevate the comparability between different datasets regarding the raters.

A Appendix

A.1 Base search queries

Table 10: First Search Query (Scopus)

Query	TITLE-ABS-KEY (("Speech" OR "Spoken input" OR "speaker" OR "audio") AND (affect* OR emotion* OR "mood" OR "feeling") AND (represent* OR "classification" OR "analysis" OR label* OR annotat* OR predict* OR recogni* OR detect*) AND ("dataset" OR "database" OR "corpus"))
Result	13,277
Intersection	8/8

Table 11: Search Query Scopus (executed on 17-5-2024)

Query	(TITLE-ABS-KEY ((represent* OR “classification” OR “analysis” OR label OR annotat*) AND (predict* OR recogni* OR detect* OR identif* OR interpret* OR measur) AND (develop* OR create* OR construct* OR produc* OR design OR collect* OR buil*)) AND TITLE (emotion* OR affect*) AND TITLE (“Speech” OR “Spoken input” OR “audio” OR “speaker”) AND TITLE (“dataset” OR “database” OR “corpus”))
Result	165

Table 12: Search Query Web Of Science (executed on 17-5-2024)

Query	(((((TI=((“Speech" OR “Spoken input" OR “audio" OR “speaker"))) AND TI=((emotion* OR affect*))) AND AB=((represent* OR “classification" OR “analysis" OR label* OR annotat*))) AND AB=((predict* OR recogni* OR detect* OR identif* OR interpret* OR measur*))) AND AB=((develop* OR create* OR construct* OR produc* OR design* OR collect* OR buil*))) AND TI=((“dataset" OR “database" OR “corpus"))
Result	75

Table 13: Search Query IEEE (executed on 17-5-2024)

Query	(((“Document Title”:“Speech" OR “Document Title": “speaker" OR “Document Title": “spoken input" OR “Document Title": “audio") AND (“Document Title": emotion* OR “Document Title": affect*) AND (“Abstract": represent* OR “Abstract": “classification" OR “Abstract": “analysis" OR “Abstract": label* OR “Abstract": annotat*) AND (“Abstract": predict* OR “Abstract": recogni* OR “Abstract": detect* OR “Abstract": identif* OR “Abstract": interpret OR “Abstract": measure) AND (“Abstract": develop OR “Abstract": create OR “Abstract": construct OR “Abstract": produc OR “Abstract": design OR “Abstract": collect OR “Abstract": build OR “Abstract": “built") AND (“Document Title": “dataset" OR “Document Title": “database" OR “Document Title": “corpus")))
Result	43

Table 14: Search Query ACM (executed on 17-5-2024)

Query	[[Title: “speech"] OR [Title: “spoken input"] OR [Title: “audio"] OR [Title: “speaker"]] AND [[Title: emotion*] OR [Title: affect*]] AND [[Abstract: represent*] OR [Abstract: “classification"] OR [Abstract: “analysis"] OR [Abstract: label*] OR [Abstract: annotat*]] AND [[Abstract: predict*] OR [Abstract: recogni*] OR [Abstract: detect*] OR [Abstract: identif*] OR [Abstract: interpret*] OR [Abstract: measur*]] AND [[Abstract: develop*] OR [Abstract: create*] OR [Abstract: construct*] OR [Abstract: produc*] OR [Abstract: design*] OR [Abstract: collect*] OR [Abstract: buil*]] AND [[Title: “dataset"] OR [Title: “database"] OR [Title: “corpus"]]
Result	6

A.2 Validation Query

Table 15: Scoped emotional speech databases for validation

Database
Berlin Emotional Database (EmoDB) [64]
The Interactive EmotionalDyadic Motion CaptureDatabase (IEMOCAP) [65]
Chinese Natural Emotional Audio–Visual Database(CHEAVD) [66]
Danish Emotional Speech Database (DES) [16]
Italian Emotional Speech Database(EMOVO) [67]
RECOLA Speech Database [68]
Tamil Malayalam Ravula - Emotion DataBase (TaMaR-EmoDB) [36]
The Vera am Mittag German audio-visual emotional speech database [18]

A.3 Recommended Query

Table 16: Recommend Search Query Scopus without feasibility constraints

Query	(TITLE-ABS-KEY(("Speech" OR "Spoken input" OR "audio" OR "speaker") AND (emotion* OR affect*) AND (represent* OR "classification" OR "analysis" OR label* OR annotat*) AND (predict* OR recogni* OR detect* OR identif* OR interpret* OR measur*) AND (develop* OR create* OR construct* OR produc* OR design* OR collect* OR buil*)) AND TITLE ("dataset" OR "database" OR "corpus"))
Result	537

References

[1] J. Hall, S. Andrzejewski, and J. Yopchick, "Psychosocial correlates of interpersonal sensitivity: A meta-analysis," *Journal of Nonverbal Behavior*, vol. 33, pp. 149–180, 2009.

[2] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[3] M. Alpert and A. Rosen, "A semantic analysis of the various ways that the terms "affect," "emotion," and "mood" are used," *Journal of Communication Disorders*, vol. 23, no. 4, pp. 237–246, 1990.

[4] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[5] D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Tröster, and C. Haring, "Activity and emotion recognition to support early diagnosis of psychiatric diseases," 2008, pp. 9–10.

[6] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," 2011.

[7] L.-S. Low, N. Maddage, M. Lech, L. Sheeber, and N. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, pp. 574–586, 2011.

[8] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[9] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639319302262

[10] "Databases, features and classifiers for speech emotion recognition: a review," *International Journal of Speech Technology*, vol. 21, pp. 93–120, 2018.

[11] M. Al-Dujaili and A. Ebrahimi-Moghadam, "Speech emotion recognition: A comprehensive survey," *Wireless Personal Communications*, vol. 129, pp. 2525–2561, 2023.

[12] S. Madanian, T. Chen, O. Adeleye, J. Templeton, C. Poellabauer, D. Parry, and S. Schneider, "Speech emotion recognition using machine learning â a systematic review," *Intelligent Systems with Applications*, vol. 20, 2023.

[13] A. Hashem, M. Arif, and M. Alghamdi, "Speech emotion recognition approaches: A systematic review," *Speech Communication*, vol. 154, 2023.

[14] (2020) PRISMA 2020 checklist. [Online]. Available: https://www.prisma-statement.org/

[15] A. Boland, M. G. Cherry, and R. Dickson, Eds., *Doing a Systematic Review: A Student's Guide*, 2nd ed. SAGE, 2017.

[16] I. Engberg, A. Hansen, O. Andersen, and P. Dalsgaard, "Design, recording and verification of a danish emotional speech database," 1997, pp. 1695–1698.

[17] T. Wu, Y. Yang, Z. Wu, and D. Li, "Masc: A speech corpus in mandarin for emotion analysis and affective speaker recognition," 2006.

[18] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," 2008, pp. 865–868.

[19] S. Koolagudi, S. Maity, V. Kumar, S. Chakrabarti, and K. Rao, *IITKGP-SESC: Speech database for emotion analysis*, 2009, vol. 40, uses people to play out emotions. Talks about subjectivity analysis and how accurate raters are after.

[20] P. Staroniewicz and W. Majewski, *Polish emotional speech database - Recording and preliminary validation*, 2009, vol. 5641 LNAI.

[21] A. Esposito and M. T. Riviello, "The new italian audio and video emotional database," A. Esposito, N. Campbell, C. Vogel, A. Hussain, and A. Nijholt, Eds., vol. 5967, 2010, pp. 406–422, 2nd COST 2102 International Training School on Development of Multimodal Interfaces, Dublin, IRELAND, MAR 23-27, 2009.

[22] B. Dropuljić, M. Chmura, A. Kolak, and D. Petrinović, "Emotional speech corpus of croatian language," 2011, pp. 95–100.

[23] S. Koolagudi, R. Reddy, J. Yadav, and K. Rao, "Iitkgp-sehsc : Hindi speech corpus for emotion analysis," 2011.

[24] C. Joe, "Developing tamil emotional speech corpus and evaluating using svm," 2014.

[25] B.-C. Chiou and C.-P. Chen, "Speech emotion recognition with cross-lingual databases," 2014, pp. 558–561.

[26] T. Justin, V. Štruc, J. Žibert, and F. Mihelič, *Development and evaluation of the emotional Slovenian speech database - EmoLUKS*, 2015, vol. 9302.

[27] M. Meddeb, H. Karray, and A. Alimi, "Automated extraction of features from arabic emotional speech corpus," *International Journal of Computer Information Systems and Industrial Management Applications*, vol. 8, pp. 184–194, 2016.

[28] D. Bertero, F. Siddique, and P. Fung, "Towards a corpus of speech emotion for interactive dialog systems," 2017, pp. 241–246.

[29] D. Pravena and D. Govind, "Development of simulated emotion speech database for excitation source analysis," *International Journal of Speech Technology*, vol. 20, pp. 327–338, 2017.

[30] D. Pravena, S. Nandhakumar, and D. Govind, "Significance of natural elicitation in developing simulated full blown speech emotion databases," 2017, pp. 261–265.

[31] E. Takeishi, T. Nose, Y. Chiba, and A. Ito, "Construction and analysis of phonetically and prosodically balanced emotional speech database," 2017, pp. 16–21.

[32] A. Geethashree and D. Ravi, *Kannada emotional speech database: Design, development and evaluation*, 2018, vol. 14.

[33] S. Koolagudi, Y. Murthy, and S. Bhaskar, "Choice of a classifier, based on properties of a dataset: case study-speech emotion recognition," *International Journal of Speech Technology*, vol. 21, pp. 167–183, 2018.

[34] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, pp. 471–483, 2019.

[35] *Multimodal Database of Emotional Speech, Video and Gestures*, 2019, vol. 11188 LNCS.

[36] R. Rajan, U. Haritha, A. Sujitha, and T. Rejisha, "Design and development of a multi-lingual speech corpora (tamar-emodb) for emotion analysis," vol. 2019-Septe, 2019, pp. 3267–3271.

[37] L. Mar, W. Pa, and T. Nwe, "Dataset for depression detection from speech emotion recognition," 2019, pp. 101–106.

[38] Z. S. Syed, S. A. Memon, M. S. Shah, and A. S. Syed, "Introducing the urdu-sindhi speech emotion corpus: A novel dataset of speech recordings for emotion recognition for two low-resource languages," *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, vol. 11, pp. 805–810, 4 2020.

[39] E. Parada-Cabaleiro, G. Costantini, M. B. Anton, Schmitt, and B. W. Schuller, "Demos: an italian emotional speech corpus elicitation methods, machine learning, and perception," *LANGUAGE RESOURCES AND EVALUATION*, vol. 54, pp. 341–383, 6 2020.

[40] R. Sato, R. Sasaki, N. Suga, and T. Furukawa, "Creation and analysis of emotional speech database for multiple emotions recognition," 2020, pp. 33–37.

[41] N. Jia, C. Zheng, and W. Sun, "Design and evaluation of adult emotional speech corpus for natural environment," vol. 1, 2020, pp. 53–56.

[42] V. Tank and S. Hadia, "Creation of speech corpus for emotion analysis in gujarati language and its evaluation by various speech parameters," *International Journal of Electrical and Computer Engineering*, vol. 10, pp. 4752–4758, 2020.

[43] "Lssed: A large-scale dataset and benchmark for speech emotion recognition," vol. 2021-June, 2021, pp. 641–645.

[44] R. Aljuhani, A. Alshutayri, and S. Alahdal, "Arabic speech emotion recognition from saudi dialect corpus," *IEEE Access*, vol. 9, pp. 127 081–127 085, 2021.

[45] A. Asghar, S. Sohaib, S. Iftikhar, M. Shafi, and K. Fatima, "An urdu speech corpus for emotion recognition," *PeerJ Computer Science*, vol. 8, 2022.

[46] Z. Xiao, Y. Chen, W. Dou, Z. Tao, and L. Chen, "Mes-p: An emotional tonal speech dataset in mandarin with distal and proximal labels," *IEEE Transactions on Affective Computing*, vol. 13, pp. 408–425, 2022.

[47] T. Kim, S. Doh, G. Lee, H. Jeon, J. Nam, and H.-J. Suk, "Hi,kia: A speech emotion recognition dataset for wake-up words," 2022, pp. 1590–1595.

[48] Y. Nam and C. Lee, "Chung-ang auditory database of korean emotional speech: A validated set of vocal expressions with different intensities," *IEEE Access*, vol. 10, pp. 122 745–122 761, 2022.

[49] M. Moutti, S. Eleftheriou, P. Koromilas, and T. Giannakopoulos, "A dataset for speech emotion recognition in greek theatrical plays," 2022, pp. 1040–1046, 13th International Conference on Language Resources and Evaluation (LREC), Marseille, FRANCE, JUN 20-25, 2022.

[50] R. Paccotacya-Yanque, C. Huanca-Anquise, J. Escalante-Calcina, W. Ramos-Lovón, and A. Cuno-Parari, "A speech corpus of quechua collao for automatic dimensional emotion recognition," *Scientific Data*, vol. 9, 2022.

[51] C. Singla and S. Singh, "Pemo: A new validated dataset for punjabi speech emotion detection," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 10, pp. 52–58, 2022.

[52] S. Tomar, P. Gupta, and S. Koolagudi, "Nitk-klesc: Kannada language emotional speech corpus for speaker recognition," 2023.

[53] M. Qasim, T. Habib, S. Urooj, and B. Mumtaz, "Descu: Dyadic emotional speech corpus and recognition system for urdu language," *Speech Communication*, vol. 148, pp. 40–52, 2023.

[54] V. Kondratenko, A. Sokolov, N. Karpov, O. Kutuzov, N. Savushkin, and F. Minkin, "Hybrid dataset for speech emotion recognition in russian language," vol. 2023-Augus, 2023, pp. 4548–4552.

[55] L. Mar, W. Pa, and T. Nwe, "Bmisec:corpus of burmese emotional speech," vol. 2023-Febru, 2023, pp. 248–253.

[56] K. Kaur and P. Singh, *Extraction and Analysis of Speech Emotion Features Using Hybrid Punjabi Audio Dataset*, 2023, vol. 1788 CCIS.

[57] S. Safwat, M.-M. Salem, and N. Sharaf, *Building an Egyptian-Arabic Speech Corpus for Emotion Analysis Using Deep Learning*, 2024, vol. 14327 LNAI.

[58] D. Xin, J. Jiang, S. Takamichi, Y. Saito, A. Aizawa, and H. Saruwatari, "Jvnv: A corpus of japanese emotional speech with verbal content and nonverbal expressions," *IEEE Access*, vol. 12, pp. 19 752–19 764, 2024.

[59] R. Pan, J. García-Díaz, M. Rodríguez-García, and R. Valencia-García, "Spanish meacorpus 2023: A multimodal speech-text corpus for emotion analysis in spanish from natural environments," *Computer Standards and Interfaces*, vol. 90, 2024.

[60] E. Garcia-Cuesta, A. Salvador, and D. PÃ£ez, "Emomatchspanishdb: study of speech emotion recognition machine learning models in a new spanish elicited database," *Multimedia Tools and Applications*, vol. 83, pp. 13 093–13 112, 2024.

[61] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," 2005, pp. 381–385.

[62] J. Randolph, "Free-marginal multirater kappa (multirater κfree): An alternative to fleiss fixed-marginal multirater kappa," vol. 4, 01 2010.

[63] K. Krippendorff, "Computing krippendorff's alpha-reliability," 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:59901023

[64] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," 2005, pp. 1517–1520.

[65] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[66] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "Cheavd: a chinese natural emotional audioâvisual database," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, pp. 913–924, 2017.

[67] G. Costantini, I. Iadarola, A. Paoloni, and M. Todisco, "Emovo corpus: An italian emotional speech database," 2014, pp. 3501–3504.

[68] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," 2013.