IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 14, No. 4, August 2025, pp. 3421∼3434
ISSN: 2252-8938, DOI: 10.11591/ijai.v14.i4.pp3421-3434

Exploring bibliometric trends in speech emotion recognition (2020-2024)

Yesy Diah Rosita[1,2], Muhammad Raafi'u Firmansyah[2], Annisaa Utami[2]
[1]Center of Excellence for Human Centric Engineering, Institute of Sustainable Society, Telkom University, Main Campus, Bandung City, Indonesia
[2]Informatics Engineering Study Program, Telkom University, Purwokerto Campus, Banyumas City, Indonesia

Article Info

Article history:

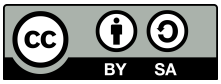
Received Apr 21, 2024
Revised Jun 12, 2025
Accepted Jul 10, 2025

Keywords:

Audio features
Classification model
Emotions
Preprocessing
Speech emotion recognition

ABSTRACT

Speech emotion recognition (SER) is crucial in various real-world applications, including healthcare, human-computer interaction, and affective computing. By enabling systems to detect and respond to human emotions through vocal cues, SER enhances user experience, supports mental health monitoring, and improves adaptive technologies. This research presents a bibliometric analysis of SER based on 68 articles from 2020 to early 2024. The findings show a significant increase in publications each year, reflecting the growing interest in SER research. The analysis highlights various approaches in preprocessing, data sources, feature extraction, and emotion classification. India and China emerged as the most active contributors, with external funding, particularly from the National Natural Science Foundation of China (NSFC), playing a significant role in the advancement of SER research. Support vector machine (SVM) remains the most widely used classification model, followed by K-nearest neighbors (KNN) and convolutional neural networks (CNN). However, several critical challenges persist, including inconsistent data quality, cross-linguistic variability, limited emotional diversity in datasets, and the complexity of real-time implementation. These limitations hinder the generalizability and scalability of SER systems in practical environments. Addressing these gaps is essential to enhance SER performance, especially for multimodal and multilingual applications. This study provides a detailed understanding of SER research trends, offering valuable insights for future advances in speech-based emotion recognition.



Corresponding Author:

Yesy Diah Rosita
Informatics Engineering Study Program, Telkom University, Purwokerto Campus
Jln. D.I. Panjaitan No. 128, Purwokerto, Banyumas City, 53147, Indonesia
Email: yesydr@telkomuniversity.ac.id

1. INTRODUCTION

Hate speech is often driven by strong negative emotions, such as hatred or anger, which can trigger social conflicts and escalate tensions between individuals or groups [1], [2]. In this context, speech emotion recognition (SER) emerges as a technology capable of detecting and interpreting emotions from human speech. By identifying negative emotions such as anger in speech, SER can be utilized to support online content moderation, enhance hate speech detection systems, and analyze social interactions to prevent conflict escalation.

SER was first introduced by Rosalind W. Picard and her team at the MIT Media Laboratory in the early 2000s [3]. Since then, this field has experienced rapid growth, with advancements in feature extraction, classification models, and multimodal approaches. Over the past decade, the rise of deep learning and the

availability of larger speech datasets have significantly improved the accuracy of SER systems. Today, this technology is widely applied in various domains, including healthcare, human-computer interaction, and digital security.

This study aims to provide a bibliometric analysis of 68 articles on SER published between 2020 and early 2024. This timeframe was selected due to the increasing number of publications in recent years, reflecting the growing interest in SER research, particularly following the COVID-19 pandemic, which accelerated the adoption of voice-based technologies in communication and emotion analysis. The analysis is conducted by collecting data from the Scopus database, covering key trends in SER, methodological developments, and research collaborations among scholars from various countries.

Bibliometrics is a statistical analysis technique used to understand the historical development of a scientific field [4]. This method helps uncover collaboration patterns in multidisciplinary research [5], identify trends in scientific publications, and analyze inter-article relationships [6]. Additionally, bibliometric analysis enables the evaluation of research impact and the mapping of scientific structures using various statistical indicators [7]. Research collaboration tends to enhance the influence of a study compared to individual research efforts [8], particularly when it involves multiple relevant disciplines.

SER has become a rapidly evolving research field, employing various approaches to recognize emotions in human speech [9]. In several studies, SER has been applied in sentiment analysis [8] and human-computer interaction [10], helping to identify dominant topics in scientific publications. Moreover, its application in speech and video data analysis demonstrates significant potential for understanding emotional dynamics across different contexts. Although SER research has seen substantial growth in the past five years, several key challenges remain. These include difficulties in collecting and analyzing accurate speech data, the complexity of understanding and interpreting human emotions through speech, and limitations in handling linguistic and cultural variations. This study aims to identify SER's contributions and impacts on other fields while highlighting areas that require further exploration. One of the primary challenges in this study is ensuring the accuracy and consistency of the collected data from various sources. A rigorous data-cleaning process and manual review are necessary to ensure that the analyzed articles are relevant and of high quality. Additionally, the complexity of interpreting bibliometric analysis results presents another challenge. While SER has been widely explored, bibliometric studies specifically mapping the distribution of classification models, the interdisciplinary collaborations, and cross-cultural gaps in emotion recognition remain limited. This study seeks to fill these gaps by providing a comprehensive overview of trends, collaborations, and underexplored areas in SER literature between 2020 and 2024.

This research advances existing methodologies by integrating SER models with a comprehensive review of relevant literature. Data collection is based on titles, abstracts, and the full content of selected articles, followed by a manual review to ensure relevance to the research topic. The main objectives of this study are: i) to provide an overview of SER research trends using the 5W+1H approach (what, who, when, why, where, and how); and ii) to identify potential research subtopics that warrant further exploration.

The structure of this article is organized as follows. Section 2 explains the data collection methodology. Section 3 presents the research findings related to SER. Lastly, section 4 provides the study's conclusions.

2. METHOD

Scientific articles are obtained from the Scopus database, where some articles can be accessed directly, while others are closed access. The availability of access to scientific articles can influence the ease of obtaining relevant information and literature. Additionally, access-restricted articles may require additional efforts to gain full access, such as through an institutional library or database subscription service. In the context of academic research, it is important to consider the available information sources and the access methods that can be used to optimize the use of existing information resources. Data collection was a crucial step in this research, ensuring that the study could be replicated. To ensure that this research can be replicated, a well-defined query strategy was implemented, which had been tested for effectiveness in retrieving relevant articles from the Scopus database. This structured approach helped obtain consistent and high-quality data for bibliometric analysis. Moreover, the user-friendly interface of the Scopus database facilitated the data retrieval process, enabling researchers to focus more on data interpretation and analysis. The article selection and data analysis were primarily carried out using spreadsheet software, which allowed for efficient organization, filtering, and summarization of the dataset. This approach was chosen to maintain flexibility in the review process and

to adapt to the evolving nature of the research. The article search was conducted using an advanced search query applied to titles, abstracts, and keywords, with additional filters based on publication year, article source, publication stage, and document type. The last data collection was performed on February 2, 2024, resulting in an initial set of 80 articles. Since these articles originated from diverse sources, a rigorous data-cleaning process was undertaken to remove duplicates and inconsistencies. The names of authors, publishers, journals, and research funders were cross-checked for duplication. The workflow for data collection and analysis is illustrated in Figure 1 (data collection flowchart) and Figure 2 (query instruction). The data collection steps were carried out as follows:

- Search by query: articles were retrieved using an advanced search query applied to title, abstract, and keywords, with filters based on publication year, article source, publication stage, and document type.
- Data cleaning: the names of authors, publishers, journals, and funding institutions were checked to eliminate duplication.
- Article numbering: each article was assigned a unique identification number to facilitate tracking during the review and analysis process.
- Manual review: a detailed manual review of each article was conducted by a team of three independent reviewers to ensure relevance to the research topic, verify the adequacy of the information presented, and assess the quality of the journal. Discrepancies in the selection of articles were resolved through discussion.
- Reviewed summary: compile a summary of each article that has been reviewed to provide reference material in writing a literature review.
- Nomenclature: compile a nomenclature or list of terms used in the articles to facilitate readers' understanding.
- Dataset source: compile the dataset source or list of data sources in the articles.
- List document: compile a list of articles that will be reviewed, based on certain criteria to be used as a basis for compiling a literature review.

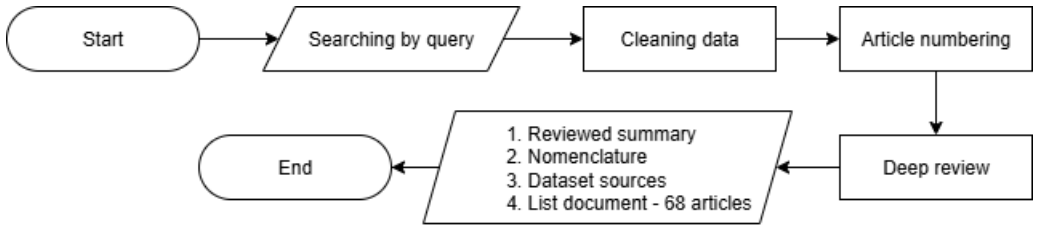


Figure 1. The flowchart for collecting data

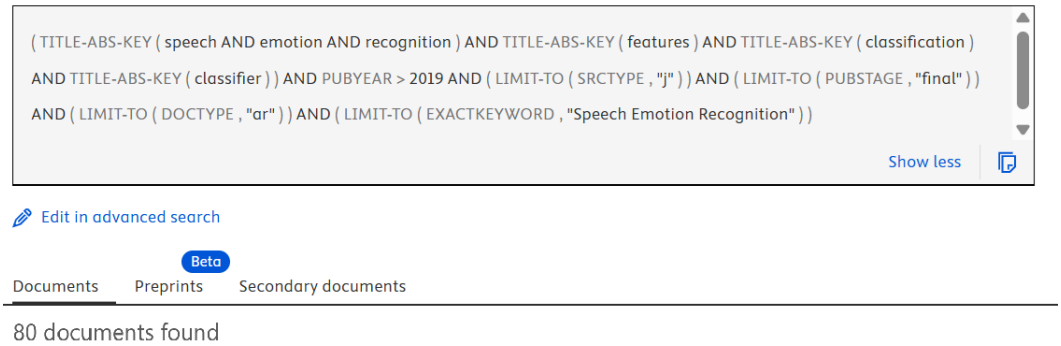


Figure 2. The query for collecting data

From the 80 initially retrieved articles, 68 (85%) were deemed relevant and included in the final dataset. The selection process was carried out through a structured screening procedure to ensure that only the

most pertinent and impactful studies were retained. This process involved a thorough review of each article's abstract, keywords, and full text when necessary to determine its suitability for inclusion. The selection process was based on the following criteria:

- Relevance to SER: articles that explicitly discuss SER methodologies, datasets, or applications were prioritized.
- Citation impact: articles with significant citations (when available) were given preference to ensure academic influence.
- Publication year: articles published between 2020 and early 2024 were included to reflect recent developments in SER research.

The bibliometric analysis revealed a significant increase in SER-related research over the past five years, indicating growing interest and impact in this domain. However, the number of publications alone does not necessarily correlate with high citation counts. Some journals with a high number of publications had relatively low citation averages, while others with fewer articles had a substantial citation impact. Table 1 shows the summary per provenance as a result of that query. Based on it, there are variations in the number of articles published and the average citations per year for each journal. In general, journals that publish more articles tend to have a higher average of citations per year. However, the correlation between the number of articles and the average citations per year is not always the case. For instance, the Multimedia Tools and Applications journal, despite having a high number of articles (13.24%), has a relatively low average of citations per year (7 citations per year); the International Journal of Advanced Computer Science and Applications, with only 2.94% of 68 articles, has a very high average of citations per year (59 citations per year). This suggests that other factors, such as article quality, research novelty, and journal indexing, also contribute to citation impact.

Table 1. The summary per provenance

Journal	Number of articles	Percentages %	Average number of citations/years
Multimedia Tools and Applications [3]–[7], [9]–[12]	9	13.24	8 (8 citations)
IEEE Access [13]–[19]	7	10.29	20 (20 citations)
Applied Acoustics [20]–[24]	5	7.35	39.8 (40 citations)
International Journal of Speech Technology [25]–[27]	3	4.41	7.4 (7 citations)
Journal of Supercomputing [28]–[30]	3	4.41	1.4 (1 citation)
Signal, Image and Video Processing [31], [32]	2	2.94	0.8 (1 citation)
Electronics (Switzerland) [33], [34]	2	2.94	2.8 (3 citations)
Sensors (Switzerland) [35], [36]	2	2.94	1.4 (1 citation)
IEEE/ACM Transactions on Audio Speech and Language Processing [37], [38]	2	2.94	2.2 (2 citations)
Journal of Ambient Intelligence and Humanized Computing [39], [40]	2	2.94	4.4 (4 citations)
International Journal of Advanced Computer Science and Applications [41], [42]	2	2.94	59.2 (59 citations)
The journals that have only 1 article [8], [43]–[70]	29	42.65	2.83 (3 citations)
Total	68	100	-

The review technique that will be carried out by applying the 5W+1H concept (what, who, where, when, why, and how) shows the following:

- What: data sources used, features, types of emotions. Information on data sources used in research can be obtained from the data/material section, while the method for extracting voice characteristics/features and determining emotions as output is obtained from the method section.
- Where: country of origin of the main researcher and correspondent. Information on the author's country of origin can be obtained on the first page, commonly below the author's name.
- When: year of publication. Information regarding the year of publication of the article can be obtained from the first page. It is generally put before the abstract; it states the time of submission, revision time, time of acceptance, and time of publication of the article in the journal.
- Who: research funding agent. Information about the institution that funded the research was obtained from the acknowledgment section as a form of the researchers' thanks. Some articles did not mention the institution that provided the research funding, which could mean that the research was funded independently.

– Why: the root of the problem. Reviewing the root of the problem is carried out by observing the background, including the problem formulation defined by the researcher.
– How: classifier model used. An overview of the classifier models used by researchers can be found in the method section. Several researchers compared various classifier methods; others applied a single method but refined the model's architectural configuration.

3. RESULTS AND DISCUSSION

After conducting a thorough review of the literature, the researchers decided to include only 68 articles in the final analysis, revealing the results of the processed data. This selection provides a comprehensive overview of various aspects of recognizing emotions in speech and offers valuable insights into the current state of research in this field.

3.1. What

In the review of the articles, four key aspects were identified as crucial components within the "What" category of SER. These aspects include the preprocessing stages used in the analysis, the data sources employed for training the models, the types of features extracted from the speech signals, and the emotions that serve as the target or class for emotion detection in speech. These four factors play an essential role in shaping the methodologies used to recognize and classify emotions in speech and have a significant impact on the accuracy and applicability of the models.

Figure 3 presents a detailed map of the data, illustrating the distribution of articles that discuss each of these four critical aspects. The map categorizes the number of articles into seven distinct ranges based on the frequency with which each aspect is covered. These ranges are as follows: 6> articles, 6-10 articles, 11-20 articles, 21-30 articles, 31-40 articles, 41-50 articles, 51-60 articles, and >60 articles. This classification allows for a better understanding of which aspects of emotion recognition in speech are most frequently addressed in the literature, highlighting the areas of the field that are receiving the most attention and those that may require further exploration.

Figure 3. Distribution of review data based on the concepts of 'What'

3.1.1. Preprocessing

Preprocessing, or the preprocessing stage, is a critical step in processing speech signals for the recognition of emotions in speech. This stage aims to improve the quality of the sound signal before further analysis is carried out. In this research, preprocessing includes three main stages: silence removal, noise removal, and unspecified. The results of the analysis show that researches involving silence and noise removal processes are only 7 articles [9], [23], [25], [36], [39], [57], [61] and studies examining preprocessing of silence removal only are 2 articles [21], [65] and others focusing noise removal only are 23 articles [3], [6], [7], [11]-[13], [15],

[16], [20], [30], [32], [40], [42], [46], [48]-[55], [66], [69]. However, most of the articles (32 articles) did not specifically mention the preprocessing steps they used. A summary of the types of preprocessing. There is still variation in the preprocessing approaches used in SER research. Most researchers did not provide specific details about the preprocessing step they undertook. The main challenge in this stage is to ensure that the resulting sound signal is free from interference and ready for further analysis. Therefore, further researches need to explore various preprocessing methods that can improve the quality of speech signals and the accuracy of emotion recognition in speech.

3.1.2. Data sources

Data sources are an important component in research into emotion recognition in speech, as the quality and representativeness of the data can have a major impact on the results of the analysis. In this research, there are variations in the data sources used by researchers. Berlin database of emotional speech (EMO-DB) is the most commonly used data source, with 31-40 articles using data from it [4]-[7], [9]-[14], [18], [20]-[23], [26], [28], [31], [32], [34], [35], [37], [38], [42], [45], [49], [50], [52], [56]-[60], [63]-[65]. There are also other popular data sources such as the interactive emotional dyadic motion capture (IEMOCAP) taken by 21-30 articles [8], [11], [14], [15], [17]-[22], [28], [29], [33], [35]-[38], [43], [49], [52], [54]-[56], [58], [59], [62], [65], [68], ryerson audio-visual database of emotional speech and song (RAVDESS) [5], [6], [8], [10], [11], [23], [27], [30], [31], [35]-[37], [39], [40], [42], [43], [45], [52], [53], [58], [59], [64], [65], [69], [70] and surrey audio-visual expressed emotion (SAVEE), fairly common data source, is used in 11-20 articles [3], [10], [13], [18], [21], [23], [31], [34], [35], [39], [42], [45], [50]-[53], [58], [60], [63], [69].

Meanwhile, toronto emotional speech set (TESS) only becomes sources in less than 11 articles [3], [6], [25], [37], [38], [40], [53], [69]. In addition to this main data source, there are also other data sources used by fewer than 6 articles, which are included in the "Others" category. The variations in data sources indicate that researchers have diverse choices in selecting data for their research. This also shows the importance of having good access to a variety of relevant data sources to ensure the representativeness of research results. In the context of SER research, it is important to select data sources that are appropriate to the research objectives and capable of representing a variety of different emotional states. Parameters that can influence data quality include the distance from the recorder to the transmitter of respondents, the specifications of the equipment used, the recording duration of the recording, and the significance of the emotions given by the respondents.

Despite the frequent use of well-known datasets such as EMO-DB and IEMOCAP, this analysis reveals a lack of diversity in the selection of data sources, particularly those that capture spontaneous emotional expressions or represent non-Western cultural contexts. This suggests a research gap in cross-cultural emotional representation and real-world data variability, which may limit the generalizability of current SER models. By identifying this gap through bibliometric mapping, this study encourages future research to explore and develop more inclusive, diverse, and naturalistic datasets to enhance the robustness of SER systems.

3.1.3. Features

The features used in speech analysis play an important role in the recognition of emotions in speech. In this research, the mel-frequency cepstral coefficients (MFCC) feature is the most commonly used feature, with more than 41 articles using it [3], [4], [6], [7], [9], [10], [12], [13], [15], [16], [19], [21]-[23], [25]-[30], [34], [36], [38]-[40], [42], [43], [45], [46], [48], [49], [51]-[53], [57], [60], [61], [63], [67], [68], [70]. Besides, pitch is also a popular feature, found in 12 articles [6], [7], [9], [14], [21], [25], [27], [29], [34], [46], [68], [70]. In addition, there are several other features used by 6-10 articles, including mel-spectrogram [3], [5], [10], [46], [48], [51], [54], [58], linear predictive coding (LPC) [6], [9], [13], [26], [29], [40], [61], formant [9], [14], [27], [46], [57], [59], energy [6], [9], [29], [46], [51], and chroma [25], [28], [46], [48], [51], [61]. These features reflect variations in speech analysis approaches used to identify emotional patterns in speech. Apart from these main features, there are also other features used in fewer than six articles, which fall into the "Others" category. The variation shows that researchers have applied varied approaches in analyzing sound signals for emotion recognition, with each feature having its advantages and disadvantages. Therefore, selecting appropriate features is a critical step in the development of an effective emotion recognition system. As in previous research, the use of the dominant weight normalization feature selection algorithm also has an influence on the level of accuracy, which shows sufficient transmission with a relatively small amount of data. This research shows that with 300 data points, it is able to show an accuracy rate of 86%, so that this algorithm can be used as a consideration for use in developing SER research [71].

3.1.4. Emotions

Analysis of emotions in speech involves identifying the different types of emotions that can be expressed through sound. In this study, the emotion "Happy" was found out to be the most commonly investigated emotion, sometimes referred to as "Joy/Joyful", with more than 62 articles (91.18%). In addition, the emotions "Angry" 60 articles (88.24%), "Sad" 60 articles (88.23%), and "Neutral" 59 articles (86.76%) are also equally chosen. The emotions "Fear" 47 articles (69.12%) and "Disgust" 46 articles (67.65%) are also quite commonly used. Apart from that, there were also "Surprise" emotions in 36 articles (52.94%), and "Boredom" emotions in 24 articles (35.29%). In addition to them, there are also others used by fewer than 11 articles in the category 'Others'. The variation in the types of emotions shows that researchers have varied interests in understanding and identifying different types of emotions in speech. Research classified these emotions into 3 types, namely positive, negative, and neutral [35]. It also reflects the complexity of human emotional expression and the challenges in developing systems capable of recognizing emotions with high accuracy in a variety of contexts and situations.

3.2. Where

The discussion regarding "Where" stems from an in-depth analysis of the countries of origin and the institutions affiliated with the first author and the corresponding author, as illustrated in Figure 4. This aspect of the research is crucial for understanding the geographical and institutional spread of the studies on SER. By examining whether the first and corresponding authors come from the same or different countries, we can gain insights into the international collaboration patterns within the field of emotion detection in speech. A significant number of articles authored by researchers from multiple countries suggests a broad, global network of researchers engaged in voice emotion identification. Conversely, studies with authors from a single country or institution may indicate more localized research efforts. Figure 4 shows the number of authors from countries, either as the first author or corresponding author, who have published on SER with more than three contributing authors. Others come from Taiwan, Turkey, Pakistan, Australia, Indonesia, Japan, Egypt, France, Iraq, Italy, Kazakhstan, London, Portugal, Saudi Arabia, Vietnam, Bhutan, and Malaysia.

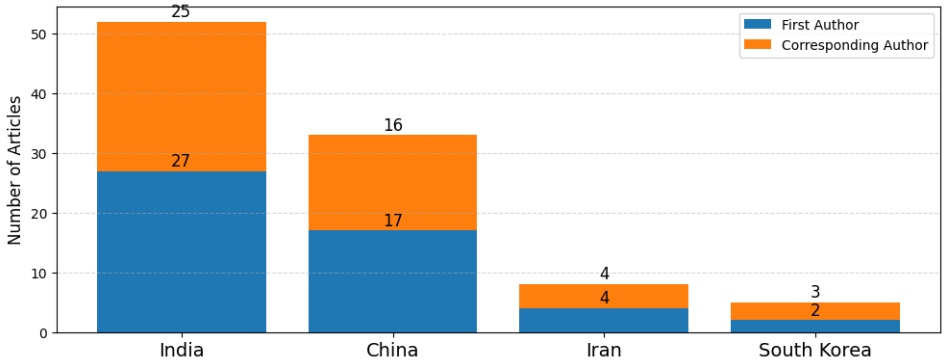


Figure 4. Top 4 countries by number of authors

Moreover, this geographical analysis indicates the level of global interest and involvement in SER research, reflecting how research in this domain is distributed across different regions. It also allows for the identification of leading countries or institutions that are driving innovation and contributing to advancements in this field. Understanding the "Where" thus highlights not only the scope of international collaboration but also the potential for future networking opportunities and the sharing of knowledge across borders.

3.2.1. The first author

The first author of a study often reflects the institution or country where the research was conducted. In this study, the first authors came from countries around the world. The countries contributing the most first authors are India, with 27 articles [3], [5]-[8], [10], [11], [13], [20], [23], [25], [27], [30], [31], [39], [40], [43], [46], [48], [50], [52], [53], [56], [60], [65], [66], [70] and China, with 17 articles [15], [16], [19], [32], [33], [34], [38], [44], [45], [49], [54], [55], [59], [61], [62], [67], [68]. Apart from that, there are several other contributing countries with much fewer articles such as Iran (4 articles) [9], [21], [28], [57], South Korea

(2 articles) [29], [36], Pakistan (2 articles) [33], [39], Taiwan (2 articles) [14], [58], Turkey (2 articles) [22], [24] and several other countries with an article including Egypt, France, Indonesia, Iraq, Italy, Japan, Kazakhstan, London, Portugal, Saudi Arabia, and Vietnam. Among the authors, Banusree Yalamanchili from India was the most active by publishing 3 articles over the last 5 years as first author.

3.2.2. The corresponding author

Corresponding authors often have an important role in research, especially in terms of communication with journal editors and other researchers. In this research, they come from various countries of origin. A similar figure is seen like the first author trend. Again India is the country with the most contributions for its corresponding author, with 25 articles [3], [5]-[8], [10], [11], [20], [23], [25], [27], [31], [39], [40], [43], [46], [48], [50], [52], [53], [56], [60], [65], [66], [70], followed by China with 16 articles [15], [16], [19], [30], [32]-[34], [38], [44], [45], [49], [54], [59], [61], [67], [68]. Apart from that, several other countries also contribute, such as Iran (4 articles) [9], [21], [28], [57], South Korea (3 articles) [29], [35], [36], Australia (2 articles) [12], [62], Indonesia (2 articles) [26], [42], Japan (2 articles) [17], [55], Taiwan (2 articles) [14], [58], Turkey (2 articles) [22], [24], and several other countries have only an article such as Bhutan, Egypt, France, Iraq, Italy, Kazakhstan, London, Malaysia, Pakistan, and Portugal. Among the corresponding authors, 43 authors are also the first authors. This shows that they have a significant role in the research carried out, both as the main initiator and as the person responsible for communication and coordination with other parties, such as journal editors and other researchers. This also shows the high level of involvement and contribution of these researchers in the development and dissemination of knowledge in the field of SER.

3.3. When

The distribution of articles about SER by year shows an interesting trend in the last five years as shown in Figure 5. In 2020, 14 articles [12]-[14], [20], [21], [26], [29], [35], [36], [41], [54], [57], [61], [62] were published, indicating a moderate level of research activity in this area. The following year, in 2021, the number of articles increased slightly to 11 articles [7], [15], [22], [23], [30] [33], [34], [42], [58], [59], [64] showing a temporary increase in research output. A higher increase is also seen in 2022 with 17 articles [10], [11], [17], [24], [27], [32], [40], [43], [46], [51], [53], [55], [56], [62], [63], [65], [67], showing renewed interest in SER research. This trend continues in 2023, with the highest number of articles reaching 22 articles [3]-[5], [8], [9], [16], [18], [19], [25], [28], [37]-[39], [44], [45], [48], [50], [52], [60], [66], [68], [70] indicating continued growth in research activity and perhaps also maturity of the field.

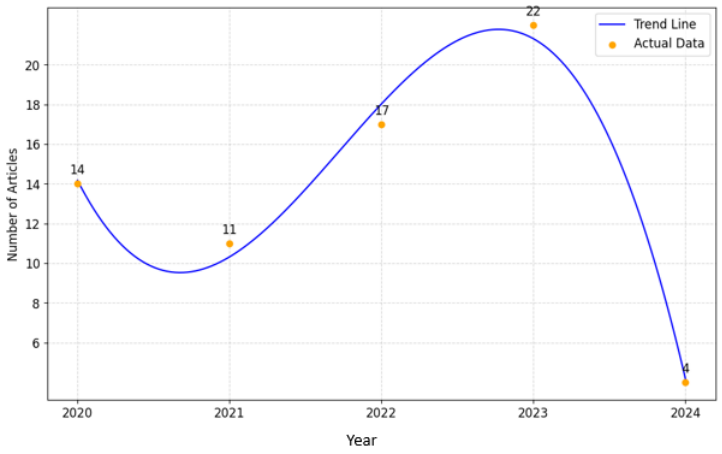


Figure 5. Number of articles published by year

As of February 2024, there are four articles [6], [47], [50], [69] that prove research in this field remains sustainable, despite a dramatic downturn. During this period, some articles begin discussing calm SER, and it is possible that by the end of 2024, there will be a significant rise compared to 2023. A clearer trend of the last five years in the number of publications regarding SER articles. It shows that emotion recognition in speech remains a relevant and interesting topic, and it can be expected that further research will continue

to be conducted to expand understanding of the technologies that can be used to detect and interpret human emotions.

3.4. Who

An analysis of funding sources for research into emotion recognition in speech shows the diverse origins of funds used to support this research. This diversity is reflected in several patterns identified in the dataset. From the data, these situations are identified:

- An institution funds research, totaling thirteen studies [12], [13], [17], [18], [20], [29], [32], [37], [45], [62]-[64], [68];
- A funding institution provides the funds for many studies, such as the National Natural Science Foundation of China (NSFC), supporting seven projects [32], [38], [45], [54], [55], [59], [61];
- A research is funded by many institutions; one research was funded by five institutions [15], [69], four institutions [61], three institutions [55], and two institutions [35], [58], [68];
- Others are self-funded research.

Funding plays a crucial role in research, with the NSFC reflecting China's commitment. However, many studies do not list funding sources, suggesting a combination of external, internal, or independent funding. Interestingly, the most cited studies are independent, particularly in the journal Multimedia Tools and Applications.

3.5. Why

The analysis of research methods behind emotion recognition in speech is visualized in Figure 6. Most research in this area is driven by three main reasons: classification model selection, feature selection, and implementation in other cases. There are studies (7 articles) discussing not only classification models but also the selection of voice features [3], [4], [17], [32], [44], [52], [70].

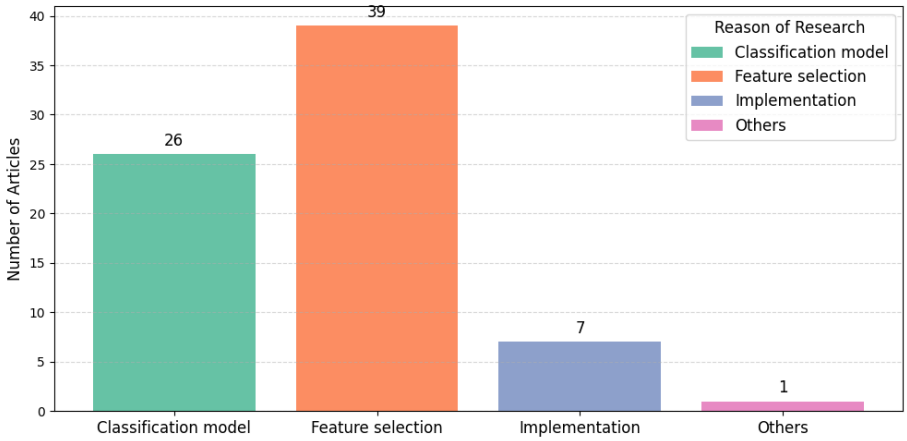


Figure 6. Distribution of research reasons in emotion recognition

- Classification model: a total of 26 articles use the selection of a classification model which is only used as the sole reason behind the methods of previous studies [3], [4], [10], [11], [14], [15], [17], [25]-[27], [32]-[34], [41], [43]-[45], [47]-[49], [52], [54], [60], [66], [68], [70]. This reflects the importance of selecting an appropriate and effective classification model for building an emotion recognition system in speech. Classification models have become the main choice for researchers to classify emotions in speech with high accuracy;
- Feature selection: a total of 39 articles use feature selection as the sole main reason behind their research methods [3]-[6], [8], [9], [12], [13], [17], [18], [20]-[24], [29]-[32], [35], [37], [39], [40], [44], [46], [50], [51]-[53], [56]-[59], [61], [62], [64], [65], [67], [70]. Appropriate and representative features are essential in building a reliable emotion recognition system. Features such as MFCC, Pitch, Mel-spectrogram, and others have become the focus of research to extract important information from sound signals that can be used to identify emotions;

– Implementation: a total of 7 articles used implementation in other cases as the rationale behind their research methods [16], [28], [38], [42], [55], [63], [69]. This suggests that some researchers have applied their approach in a broader application context, beyond emotion recognition in speech. This approach may involve the use of emotion recognition technology for such purposes as sentiment analysis, human-computer interaction, or psychological research.

Although most of the research was driven by these reasons, some research has used other reasons beyond the categories [59]. This shows that there is still variation in research motivations and approaches in emotion recognition in speech, and there is potential for further exploration in developing more innovative methods and techniques.

3.6. How

This research notes variations in the use of classifier models to identify emotions in human speech. More than 20 articles use support vector machine (SVM) as the main model, showing its popularity and effectiveness in emotion classification. Meanwhile, around 14 articles employ K-nearest neighbors (KNN) and 1D convolutional neural networks (CNN) 12 articles, while about 5-10 articles apply approaches such as decision tree (DT), deep neural network (DNN), long short-term memory (LSTM), multi layer perceptron (MLP), and random forest (RF) in more detail is shown in Table 2. The use of various classifier models shows an effort to explore various approaches in facing the challenge of emotion classification in human speech. In addition, there are also several other approaches used in smaller numbers, demonstrating the diversity in strategies and techniques used in SER researches.

Table 2. The summary of model classifiers

Number of articles	Model classifiers
> 20	SVM
10–15	KNN, 1D CNN
5–10	DT, DNN, LSTM, MLP, RF

A clear trend shows that deep learning models, particularly CNN and LSTM, are gaining popularity due to their ability to capture the complexity of speech signals and outperform traditional models like SVM, especially with large datasets. These models automatically learn features from raw data, offering better generalization in noisy environments. In contrast, while traditional models like SVM work well with smaller, structured datasets, they struggle with raw audio data, where deep learning excels. Therefore, deep learning models are becoming more prevalent in SER research due to their higher accuracy and adaptability.

4. CONCLUSION

This research presents a bibliometric analysis of 68 articles on SER published between 2020 and early 2024. There have been significant developments in SER research in the last five years, and India being the top contributor. The exploration of research topics provides a comprehensive overview of developments and trends in this field. The use of preprocessing techniques, such as silence removal and noise removal, is the main focus. The most commonly used data sources are EmoDB, IEMOCAP, and RAVDESS, while features such as MFCC and pitch are the most frequently used in the analysis. More diverse data sources, including real-world noisy data, can significantly improve SER models. By integrating datasets that reflect real-world conditions, including a broader range of emotional variations and loud environments, SER models can be trained to be more resilient to the challenges faced in everyday situations. This will help address current limitations, such as inconsistent data quality and a lack of emotional diversity in datasets, thereby enhancing the accuracy and generalizability of models in practical applications. Based on these findings, further research is suggested to develop multimodal approaches that integrate acoustic features with non-auditory data, such as facial expressions, body movements, or physiological signals. The combination of multimodal features can capture a more holistic representation of emotions, overcoming the limitations of single-voice-based systems susceptible to environmental noise or ambiguous contexts. The most frequently analyzed emotions are happy, angry, sad, neutral, fear, disgust, and surprise. In terms of classification modeling, SVM is the most widely used model, followed by KNN, 1D CNN, and several other approaches. Overall, this study provides an in-depth understanding of SER research trends and the techniques most commonly used in this analysis. It is recommended to develop more sophisticated pre-processing techniques and classification models that are more

efficient in classifying emotions in human speech, particularly in real-world, noisy environments. Moreover, the application of SER technologies in fields like healthcare, customer service, and mental health shows significant promise, offering potential improvements in emotional state monitoring and interaction. Looking toward the future, SER research will likely move toward developing real-time emotion recognition systems and addressing the challenges of cross-lingual and cross-cultural emotion recognition. These advancements will help create more adaptable and globally relevant SER applications.

FUNDING INFORMATION

This work was financially supported by the Research and Community Service Unit of Telkom University, Purwokerto Campus with grant ID IT Tel9463/LPPM-000/Ka.LPPM/XII/2023 through funding for publication and research incentives.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Yesy Diah Rosita	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Muhammad Raafi'u Firmansyah			✓	✓	✓	✓		✓	✓	✓		✓		✓
Annisaa Utami				✓	✓	✓	✓		✓	✓			✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study were obtained from Scopus and are subject to license restrictions. Data are available from the authors upon reasonable request and with permission from Scopus.

REFERENCES

[1] M. Hayaty, A. D. Laksito, and S. Adi, "Hate speech detection on indonesian text using word embedding method-global vector," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 12, no. 4, pp. 1928-1937, 2023, doi: 10.11591/ijai.v12.i4.pp1928-1937.

[2] H. EL-Zayady, M. S. Mohamed, K. Badran, and G. Salama, "A hybrid approach based on personality traits for hate speech detection in Arabic social media," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 2, pp. 1979-1988, Apr. 2023, doi: 10.11591/ijece.v13i2.pp1979-1988.

[3] S. K. Panda, A. K. Jena, M. R. Panda, and S. Panda, "Speech emotion recognition using multimodal feature fusion with machine learning approach," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 42763–42781, 2023, doi: 10.1007/s11042-023-15275-3.

[4] M. J. Al Dujaili and A. Ebrahimi-Moghadam, "Automatic speech emotion recognition based on hybrid features with ANN, LDA and KNN classifiers," *Multimedia Tools and Applications*, vol. 82, no. 27, pp. 42783–42801, 2023, doi: 10.1007/s11042-023-15413-x.

[5] A. Marik, S. Chattopadhyay, and P. K. Singh, "A hybrid deep feature selection framework for emotion recognition from human speeches," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11461–11487, Mar. 2023, doi: 10.1007/s11042-022-14052-y.

[6] B. Paul, S. Bera, T. Dey, and S. Phadikar, "Machine learning approach of speech emotions recognition using feature fusion technique," *Multimedia Tools and Applications*, vol. 83, no. 3, pp. 8663–8688, Jan. 2024, doi: 10.1007/s11042-023-16036-y.

[7] M. D. Pawar and R. D. Kokate, "Convolution neural network based automatic speech emotion recognition using Mel-frequency Cepstrum coefficients," *Multimedia Tools and Applications*, vol. 80, no. 10, pp. 15563–15587, Apr. 2021, doi: 10.1007/s11042-020-10329-2.

[8] Y. Bhanusree, S. S. Kumar, and A. K. 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*, vol. 22, no. 1, pp. 49–76, Jan. 2023, doi: 10.32890/jict2023.22.1.3.

[9] A. Bastanfard and A. Abbasian, "Speech emotion recognition in Persian based on stacked autoencoder by comparing local and global features," *Multimedia Tools and Applications*, vol. 82, no. 23, pp. 36413–36430, Sep. 2023, doi: 10.1007/s11042-023-15132-3.
[10] K. Chauhan, K. K. Sharma, and T. Varma, "A method for simplifying the spoken emotion recognition system using a shallow neural network and temporal feature stacking & pooling (TFSP)," *Multimedia Tools and Applications*, vol. 82, no. 8, pp. 11265–11283, Mar. 2023, doi: 10.1007/s11042-022-13463-1.
[11] B. Yalamanchili, K. R. Anne, and S. K. Samayamantula, "Speech emotion recognition using time distributed 2D-convolution layers for CAPSULENETS," *Multimedia Tools and Applications*, vol. 81, no. 12, pp. 16945–16966, May 2022, doi: 10.1007/s11042-022-12112-x.
[12] A. Bakhshi, S. Chalup, A. Harimi, and S. M. Mirhassani, "Recognition of emotion from speech using evolutionary cepstral coefficients," *Multimedia Tools and Applications*, vol. 79, no. 47–48, pp. 35739–35759, Dec. 2020, doi: 10.1007/s11042-020-09591-1.
[13] A. Dey, S. Chattopadhyay, P. K. Singh, A. Ahmadian, M. Ferrara, and R. Sarkar, "A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition," *IEEE Access*, vol. 8, pp. 200953–200970, 2020, doi: 10.1109/ACCESS.2020.3035531.
[14] T. W. Sun, "End-to-end speech emotion recognition with gender information," *IEEE Access*, vol. 8, pp. 152423–152438, 2020, doi: 10.1109/ACCESS.2020.3017462.
[15] L. Yang, K. Xie, C. Wen, and J. B. He, "Speech emotion analysis of netizens based on bidirectional LSTM and PGCDBN," *IEEE Access*, vol. 9, pp. 59860–59872, 2021, doi: 10.1109/ACCESS.2021.3073234.
[16] L. Yunxiang and Z. Kexin, "Design of efficient speech emotion recognition based on multi task learning," *IEEE Access*, vol. 11, pp. 5528–5537, 2023, doi: 10.1109/access.2023.3237268.
[17] B. T. Atmaja and A. Sasou, "Evaluating self-supervised speech representations for speech emotion recognition," *IEEE Access*, vol. 10, pp. 124396–124407, 2022, doi: 10.1109/ACCESS.2022.3225198.
[18] A. Mukhamediya, S. Fazli, and A. Zollanvari, "On the effect of log-mel spectrogram parameter tuning for deep learning-based speech emotion recognition," *IEEE Access*, vol. 11, pp. 61950–61957, 2023, doi: 10.1109/ACCESS.2023.3287093.
[19] Z. Kexin and L. Yunxiang, "Speech emotion recognition based on transfer emotion-discriminative features subspace learning," *IEEE Access*, vol. 11, pp. 56336–56343, 2023, doi: 10.1109/ACCESS.2023.3282982.
[20] S. R. Bandela and T. K. Kumar, "Unsupervised feature selection and NMF de-noising for robust speech emotion recognition," *Applied Acoustics*, vol. 172, Jan. 2021, doi: 10.1016/j.apacoust.2020.107645.
[21] F. Daneshfar, S. J. Kabudian, and A. Neekabadi, "Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and gaussian elliptical basis function network classifier," *Applied Acoustics*, vol. 166, Sep. 2020, doi: 10.1016/j.apacoust.2020.107360.
[22] S. Yildirim, Y. Kaya, and F. Kılıç, "A modified feature selection method based on metaheuristic algorithms for speech emotion recognition," *Applied Acoustics*, vol. 173, Feb. 2021, doi: 10.1016/j.apacoust.2020.107721.
[23] J. Ancilin and A. Milton, "Improved speech emotion recognition with mel frequency magnitude coefficient," *Applied Acoustics*, vol. 179, Aug. 2021, doi: 10.1016/j.apacoust.2021.108046.
[24] D. Tanko, S. Dogan, F. B. Demir, M. Baygin, S. E. Sahin, and T. Tuncer, "Shoelace pattern-based speech emotion recognition of the lecturers in distance education: ShoePat23," *Applied Acoustics*, vol. 190, 2022, doi: 10.1016/j.apacoust.2022.108637.
[25] N. Barsainyan and D. K. Singh, "Optimized cross-corpus speech emotion recognition framework based on normalized 1D convolutional neural network with data augmentation and feature selection," *International Journal of Speech Technology*, vol. 26, no. 4, pp. 947–961, Dec. 2023, doi: 10.1007/s10772-023-10063-8.
[26] K. Jermsittiparsert *et al*., "Pattern recognition and features selection for speech emotion recognition model using deep learning," *International Journal of Speech Technology*, vol. 23, no. 4, pp. 799–806, Dec. 2020, doi: 10.1007/s10772-020-09690-2.
[27] T. Jha, R. Kavya, J. Christopher, and V. Arunachalam, "Machine learning techniques for speech emotion recognition using paralinguistic acoustic features," *International Journal of Speech Technology*, vol. 25, no. 3, pp. 707–725, Sep. 2022, doi: 10.1007/s10772-022-09985-6.
[28] B. Nasersharif, M. Ebrahimpour, and N. Naderi, "Multi-layer maximum mean discrepancy in auto-encoders for cross-corpus speech emotion recognition," *Journal of Supercomputing*, vol. 79, no. 12, pp. 13031–13049, Aug. 2023, doi: 10.1007/s11227-023-05161-y.
[29] S. Byun, S. Yoon, and K. Jung, "Comparative studies on machine learning for paralinguistic signal compression and classification," *Journal of Supercomputing*, vol. 76, no. 10, pp. 8357–8371, Oct. 2020, doi: 10.1007/s11227-020-03346-3.
[30] V. Gupta, S. Juyal, and Y. C. Hu, "Understanding human emotions through speech spectrograms using deep neural network," *Journal of Supercomputing*, vol. 78, no. 5, pp. 6944–6973, Apr. 2022, doi: 10.1007/s11227-021-04124-5.
[31] S. P. Mishra, P. Warule, and S. Deb, "Speech emotion recognition using MFCC-based entropy feature," *Signal Image Video Process*, vol. 18, no. 1, pp. 153–161, Feb. 2024, doi: 10.1007/s11760-023-02716-7.
[32] L. Sun, Y. Huang, Q. Li, and P. Li, "Multi-classification speech emotion recognition based on two-stage bottleneck features selection and MCJD algorithm," *Signal Image Video Process*, vol. 16, no. 5, pp. 1253–1261, Jul. 2022, doi: 10.1007/s11760-021-02076-0.
[33] Y. Ying, Y. Tu, and H. Zhou, "Unsupervised feature learning for speech emotion recognition based on autoencoder," *Electronics (Switzerland)*, vol. 10, no. 17, Sep. 2021, doi: 10.3390/electronics10172086.
[34] S. Huang, H. Dang, R. Jiang, Y. Hao, C. Xue, and W. Gu, "Multi-layer hybrid fuzzy classification based on SVM and improved PSO for speech emotion recognition," *Electronics (Switzerland)*, vol. 10, no. 23, Dec. 2021, doi: 10.3390/electronics10232891.
[35] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. Bin Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors (Switzerland)*, vol. 20, no. 21, pp. 1–18, Nov. 2020, doi: 10.3390/s20216008.
[36] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors (Switzerland)*, vol. 20, no. 1, Jan. 2020, doi: 10.3390/s20010183.
[37] E. Guizzo, T. Weyde, S. Scardapane, and D. Comminiello, "Learning speech emotion representations in the quaternion domain," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 31, pp. 1200–1212, 2023, doi: 10.1109/TASLP.2023.3250840.
[38] S. Li, P. Song, and W. Zheng, "Multi-source discriminant subspace alignment for cross-domain speech emotion recognition," *IEEE/ACM Trans Audio Speech Lang Process*, vol. 31, pp. 2448–2460, 2023, doi: 10.1109/TASLP.2023.3288415.

[39] P. Tiwari, H. Rathod, S. Thakkar, and A. D. Darji, "Multimodal emotion recognition using SDA-LDA algorithm in video clips," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, pp. 6585–6602, Jun. 2023, doi: 10.1007/s12652-021-03529-7.

[40] N. Patel, S. Patel, and S. H. Mankad, "Impact of autoencoder based compact representation on emotion detection from audio," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 2, pp. 867–885, Feb. 2022, doi: 10.1007/s12652-021-02979-3.

[41] M. Iqbal, S. A. Raza, M. Abid, F. Majeed, and A. A. Hussain, "Artificial neural network based emotion classification and recognition from speech," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, pp. 434–444, 2020, doi: 10.14569/IJACSA.2020.0111253.

[42] O. U. Kumala and A. Zahra, "Indonesian speech emotion recognition using cross-corpus method with the combination of MFCC and Teager energy features," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 4, pp. 163–168, 2021, doi: 10.14569/IJACSA.2021.0120422.

[43] B. Yalamanchili, S. K. Samayamantula, and K. R. Anne, "Neural network-based blended ensemble learning for speech emotion recognition," *Multidimensional Systems and Signal Processing*, vol. 33, no. 4, pp. 1323–1348, Dec. 2022, doi: 10.1007/s11045-022-00845-9.

[44] Y. Dong and X. Yang, "A hierarchical depression detection model based on vocal and emotional cues," *Neurocomputing*, vol. 441, pp. 279–290, Jun. 2021, doi: 10.1016/j.neucom.2021.02.019.

[45] K. Mao, Y. Wang, L. Ren, J. Zhang, J. Qiu, and G. Dai, "Multi-branch feature learning based speech emotion recognition using SCAR-NET," *Connection Science*, vol. 35, no. 1, 2023, doi: 10.1080/09540091.2023.2189217.

[46] K. Kaur and P. Singh, "Impact of feature extraction and feature selection algorithms on Punjabi speech emotion recognition using convolutional neural network," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 21, no. 5, Apr. 2022, doi: 10.1145/3511888.

[47] M. A. Kanaan, J. F. Couchot, C. Guyeux, D. Laiymani, T. Atechian, and R. Darazi, "Combining a multi-feature neural network with multi-task learning for emergency calls severity prediction," *Array*, vol. 21, Mar. 2024, doi: 10.1016/j.array.2023.100333.

[48] S. Ganesan, "Deep learning model for identification of customers satisfaction in business," *Journal of Autonomous Intelligence*, vol. 7, no. 1, 2024, doi: 10.32629/jai.v7i1.840.

[49] L. Yi and M. W. Mak, "Improving speech emotion recognition with adversarial data augmentation network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 1, pp. 172–184, Jan. 2022, doi: 10.1109/TNNLS.2020.3027600.

[50] S. P. Mishra, P. Warule, and S. Deb, "Improvement of emotion classification performance using multi-resolution variational mode decomposition method," *Biomed Signal Process Control*, vol. 89, 2024, doi: 10.1016/j.bspc.2023.105708.

[51] A. A. Alnuaim *et al*., "Human-computer interaction with detection of speaker emotions using convolution neural networks," *Computational Intelligence and Neuroscience*, vol. 2022, 2022, doi: 10.1155/2022/7463091.

[52] S. Murugaiyan and S. R. Uyyala, "Aspect-based sentiment analysis of customer speech data using deep convolutional neural network and BiLSTM," *Cognitive Computation*, vol. 15, no. 3, pp. 914–931, May 2023, doi: 10.1007/s12559-023-10127-6.

[53] S. Jothimani and K. Premalatha, "MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," *Chaos, Solitons & Fractals*, vol. 162, 2022, doi: 10.1016/j.chaos.2022.112512.

[54] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Communication*, vol. 120, pp. 11–19, Jun. 2020, doi: 10.1016/j.specom.2020.03.005.

[55] L. Tan *et al*., "Speech emotion recognition enhanced traffic efficiency solution for autonomous vehicles in a 5G-enabled space-air-ground integrated intelligent transportation system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2830–2842, Mar. 2022, doi: 10.1109/TITS.2021.3119921.

[56] 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 Process*, vol. 41, no. 11, pp. 6113–6135, Nov. 2022, doi: 10.1007/s00034-022-02068-6.

[57] S. Langari, H. Marvi, and M. Zahedi, "Improving of feature selection in speech emotion recognition based-on hybrid evolutionary algorithms," *International Journal of Nonlinear Analysis and Applications*, vol. 11, no. 1, pp. 81–92, Dec. 2020, doi: 10.22075/ijnaa.2020.4227.

[58] A. Amjad, L. Khan, and H. T. Chang, "Effect on speech emotion classification of a feature selection approach using a convolutional neural network," *PeerJ Computer Science*, vol. 7, 2021, doi: 10.7717/PEERJ-CS.766.

[59] Z. T. Liu, A. Rehman, M. Wu, W. H. Cao, and M. Hao, "Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence," *Information Sciences*, vol. 563, pp. 309–325, Jul. 2021, doi: 10.1016/j.ins.2021.02.016.

[60] P. Vasuki, "Design of hierarchical classifier to improve speech emotion recognition," *Computer Systems Science and Engineering*, vol. 44, no. 1, pp. 19–33, 2022, doi: 10.32604/csse.2023.024441.

[61] Z. T. Liu, A. Rehman, M. Wu, W. H. Cao, and M. Hao, "Speech personality recognition based on annotation classification using log-likelihood distance and extraction of essential audio features," *IEEE Trans Multimedia*, vol. 23, pp. 3414–3426, 2021, doi: 10.1109/TMM.2020.3025108.

[62] W. Wei, X. Cao, H. Li, L. Shen, Y. Feng, and P. A. Watters, "Improving speech emotion recognition based on acoustic words emotion dictionary," *Natural Language Engineering*, vol. 27, no. 6, pp. 747–761, Nov. 2021, doi: 10.1017/S1351324920000339.

[63] R. Elbarougy, N. M. El-Badry, and M. N. Elbedwehy, "An improved speech emotion classification approach based on optimal voiced unit," *Information Sciences Letters*, vol. 11, no. 4, pp. 1001–1011, Jul. 2022, doi: 10.18576/isl/110401.

[64] G. Assunção, P. Menezes, and F. Perdigão, "Speaker awareness for speech emotion recognition," *International journal of online and biomedical engineering*, vol. 16, no. 4, pp. 15–22, 2020, doi: 10.3991/ijoe.v16i04.11870.

[65] P. Singh, S. Waldekar, M. Sahidullah, and G. Saha, "Analysis of constant-Q filterbank based representations for speech emotion recognition," *Digital Signal Processing*, vol. 130, Oct. 2022, doi: 10.1016/j.dsp.2022.103712.

[66] S. N. Padman and D. Magare, "Multi-modal speech emotion detection using optimised deep neural network classifier," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 11, no. 5, pp. 2020–2038, 2023, doi: 10.1080/21681163.2023.2212082.

[67] S. Zhang and C. Li, "research on feature fusion speech emotion recognition technology for smart teaching," *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/7785929.
[68] C. Fu, C. Liu, C. T. Ishi, and H. Ishiguro, "An adversarial training based speech emotion classifier with isolated gaussian regularization," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2361–2374, Jul. 2023, doi: 10.1109/TAFFC.2022.3169091.
[69] E. Mancini, A. Galassi, F. Ruggeri, and P. Torroni, "Disruptive situation detection on public transport through speech emotion recognition," *Intelligent Systems with Applications*, vol. 21, Mar. 2024, doi: 10.1016/j.iswa.2023.200305.
[70] N. Choudhury and U. Sharma, "Enhanced emotion recognition from spoken assamese dialect: A machine learning approach with language-independent features," *Traitement du Signal*, vol. 40, no. 5, pp. 2147–2160, 2023, doi: 10.18280/ts.400532.
[71] H. Heriyanto, T. Wahyuningrum, and G. F. Fitriana, "Classification of Javanese script hanacara voice using Mel frequency cepstral coefficient MFCC and selection of dominant weight features," *Jurnal INFOTEL*, vol. 13, no. 2, pp. 84–93, May 2021, doi: 10.20895/infotel.v13i2.657.

BIOGRAPHIES OF AUTHORS

Yesy Diah Rosita received a Master of Computer from Surabaya Institute of Science and Technology, Surabaya City, Indonesia. Her thesis about the classification of infants' cry sound using artificial neural network was published in IEEE, 2016. Now, she is a lecturer at Telkom University, that has been located in Purwokerto, Central Java Province, Indonesia. She is interested in speech recognition, deep learning, and decision support systems. She is also a contributor to Mathworks, which is a company that specializes in mathematical computing software. She is also an IEEE member since 2024 with ID member 100463417. She can be contacted at email: yesydr@telkomuniversity.ac.id.

Muhammad Raafi'u Firmansyah received a Master of Engineering from Gadjah Mada University, Yogyakarta City, Indonesia in 2022. His thesis is about identifying speakers using artificial neural networks and was published in IEEE, 2021. He is currently a lecturer at the Faculty of Informatics in Telkom University, Purwokerto. His research areas of interest include speech recognition, machine learning, and data mining. He has also worked on several projects related to computer vision and IoT. He can be contacted at email: raafiu@telkomuniversity.ac.id.

Annisaa Utami received a Master's Program in Computer Science from Gadjah Mada University, Yogyakarta City, Indonesia. Her thesis is about how the recommendation process is carried out by calculating the value of closeness or similarity between new cases and old cases stored on a case basis using the nearest neighbor method and Manhattan distance. She is interested in data mining, artificial intelligence, and decision support systems. She can be contacted at email: annisaau@telkomuniversity.ac.id.