A preliminary Survey of the Literature on Multilingual, Multimodal Speech Emotion Detection using Bibliometric and Co-Occurrence Analysis

Sudipta Bhattacharya[1], Brojo Kishore Mishra[1], and Samarjeet Borah[2]

[1]Department of Computer Science and Engineering, GIET University, Gunupur, India

[2]Department of Computer Applications, SMIT, Sikkim Manipal University, Sikkim, India

sudipta.bhattacharya@giet.edu, bk.mishra@giet.edu, samarjeet.b@smit.smu edu.in

Article Info
Page Number: 1962-1971
Publication Issue:
Vol. 71 No. 3 (2022)

Abstract
Emotion recognition across languages is a rapidly growing field of study. Many different techniques have been developed to determine the emotions of a speaker's words as interest in the study of speech signals has grown. The research that has been done on speech emotion identification makes use of a wide range of methods, including traditional speech analysis and classification strategies, in order to decode the feelings included within signals (SER). This study surveys the field of multilingual and multimodal speech emotion detection. Review topics include speech emotion detection limitations, speech emotion extraction methods, speech emotion databases used, and speech emotion extraction contributions. The purpose of these publications was to evaluate the multilingual and multimodal speech emotion detection in Scopus in the context of research trends and top nations. Based on voice emotion recognition in multiple languages and communication modes, the search syntax. In our essay, we provide an overview of the ten years' worth of study on this subject. The Scopus database was searched using the keywords "voice emotion detection" and "multilingual and multimodal" to find 706 records between January 1, 2013, and October 23, 2022. These records were then assessed using the VOSviewer software.

Keywords: - Speech Emotion Detection; Publications; Multimodal; Multilingual; Bibliometric Analysis.

Article History
Article Received: 15 January 2022
Revised: 24 February 2022
Accepted: 18 March 2022

1. Introduction

The way we most naturally communicate is through speech. One of the most important subfields of signal processing is speech processing. Numerous applications rely on this method of signal processing, including those dealing with voice commands, interactive voice communications, medical developments, emotion detection, contact centre automation, virtual assistants, robots, and more. The advent of cutting-edge AI, ML, and signal-processing technologies has made this a reality. Emotional analysis of speech has been a hard area of study for researchers because of its many practical applications. One of the challenges faced by spoken emotion recognition systems is the lack of sufficient emotional databases, as well as finding the relevant feature vector and choosing the right classifiers.

Research into the ability to recognise emotions dates back quite some time. The concept that emotions can be deduced from a person's facial expressions laid the groundwork for the field of emotion detection research [1]. There has been a lot of study on the topic of emotion detection from vocal signals in recent years. The emotional link between people and machines is crucial [2]. Recently, there has been a rise in interest in speech emotion recognition, often known as SER. This research endeavours to decipher emotional states by analysing speech patterns. Nevertheless, one of the most challenging challenges is to successfully extract effective emotional aspects from SER.[16-20]

In [3], we see a proposal for a multi-modal approach to SER that incorporates both text and speech. Mel frequency cepstral coefficients are used as a representation of the sound mode in this method. [14],[15] (MFCCs, popular feature-set for representing sound) (MFCCs, common feature-set for representing sound). RNN-based architectures are used on both the audio and textual data to determine which emotions are most commonly found in published works. Among these are anger, sadness, apathy, joy, and contentment (mixed with excited for the purpose of balancing out the unequal labels). Anywhere from 68.8 to 71.8% precision might be assigned to a Weighted Average Precision (WAP) score. Further in [6], authors describe a convolutional neural network (CNN), attentional network (ANN), and bidirectional long short-term memory (LSTM)-based architecture that runs on input characteristics extracted from the 3-dimensional LogMel spectrum. IEMOCAP (UA 69.32% on Angry, Happy, Sad, and Neutral on the improvised sessions) and emoDB (85.39% on all seven emotions) are used for the assessment. The research also performs cross-corpus tests by training on the IEMOCAP dataset and then evaluating on the emoDB dataset for four emotions (happiness, sadness, anger, and neutrality). With an accuracy of 63.84 percent, the results showed that the two languages share certain similarities in the presentation of emotional states. [16,21]

Experiments are performed on emoDB [7], Enterface [8], and AFEW 4.0 [9] in a number of different languages in [6]. Two of the experiments are conducted on emoDB, while the remaining four are conducted on the other databases; the model is trained on one (the source dataset) and evaluated on the other (the target dataset). For example, on emoDB, performance is 52.27 percent Weighted Average Recall (WAR) (five emotional labels) when eNTERFACE is used as the source set, but it drops to 47.80 percent when AFEW 4.0 is utilised. There are six different types of courses. Their strategy is based on a modification of the least-squares regression for use in their particular field (DaLSR). Using emoDB, eNTERFACE, and FAU Aibo, as well as a technique called non-negative matrix factorization, [6] conducts a battery of cross-corpusm tests. Similarities can be seen between the tests presented here and those described in [7]. Their results, employing eNTERFACE as a training set, show a 52.10 percent accuracy across all five emotions when examined on emoDB (Anger, Disgust, Fear, Happiness and Sadness). emoDB, eNTERFACE, and FAU Aibo served as the key data sources for the cross-corpus investigations reported in reference [8], which also presents a transfer linear subspace learning framework as their suggested technique. They show 53.85% accuracy using emoDB as the evaluation dataset and eNTERFACE for training (five emotions). [11], [12], [13]

Through bibliometric and co-occurrence analysis, the literature on multilingual and multimodal speech emotion recognition in terms of speech was critically reviewed in this study. Co-occurrence analysis is essentially the counting of paired data inside a collection unit, and bibliometric analysis is essentially a method to measure the relationships and effects of publications within a certain field of study. This study looked at roughly 706 "speech emotion detection" data from January 1, 2013, to October 23, 2022. The terms "multilingual and multimodal speech emotion recognition" and "analysed with VOSviewer software" were used to locate these records in the Scopus database. The ethics statement, study design, data collection, and visualisation are all included in part 2 of this article, which also discusses multilingual and multimodal speech emotion detection analysis. Section 3 of this analysis discusses the results. Excel and the VOSviewer tool were used to analyse the information that was taken from the Scopus database. Based on these numbers, we may deduce that researchers in the fields of voice processing in Japan, China, India, the United Kingdom, and the United States are the most prolific. Our data shows that conference presentations account for 73% of all voice-processing research papers published in the last five years. Section 4 describes the topic and the driving force behind the study. Section 5 provides evidence for the conclusion.

2. Methods

2.1 Design

A descriptive and bibliometric analysis of a literature database served as the foundation for the current study [9].

2.2 Collecting Information

The retrieved Scopus data spans from January 1, 2013, through October 23, 2022. When the parameters "TITLE-ABS-KEY ((speech AND emotion AND detection) AND multilingual AND multimodal)" were input into the search bar, the results page displayed the title "Multilingual and Multimodal Speech Emotion Detection. "Utilizing a CSV file format and VOSviewer, data were analysed. Before performing Bibliometric analysis, the data was checked for reliability and consistency (e.g., inconsistency in country names and titles, which occasionally contained abbreviations, acronyms, etc.) [10].

2.3 Visual Representation

The data gathered from the Scopus database was examined using Excel and the VOSviewer programme.

3. Result

Search results for "Multilingual and Multimodal Speech Emotion Detection" between January 1, 2013, and October 23, 2022, revealed that 706 archives have been published (Figure 1). It's fascinating to watch how much progress has been made in the field of recognising emotions conveyed through speech over the past few years.

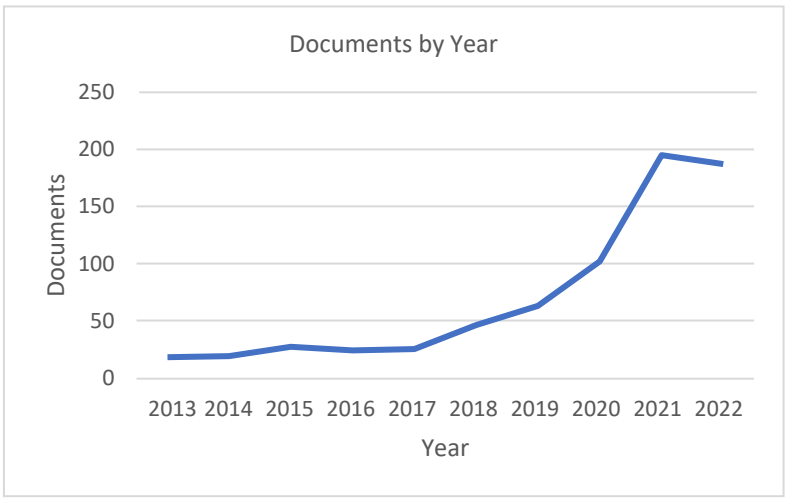

Figure 1. Documents in Scopus Arranged by Year

China, the US, India, the UK, Spain, Italy, Germany, Singapore, Canada, and Australia have reported significant work (66% of the total reported activity). In addition, the research on Multilingual and Multimodal Speech Emotion Detection received significant contributions from Japan, Ireland, and France (Figure 2).

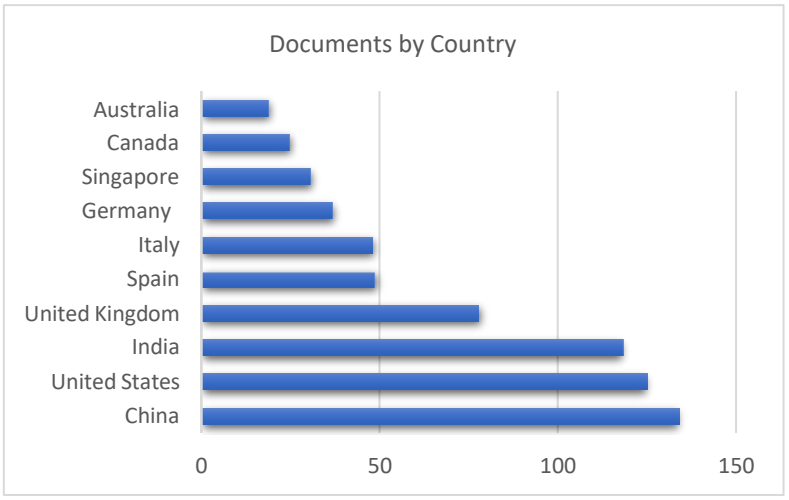

Figure 2. Scopus Database Documents, Segmented by Country/Territory (2013-October 2022)

Result articles make up 53% of the documents we reviewed, conference papers make up 30% and only book chapters and reviews make up the remaining 6% and 11%. (Fig. 3). On the subject of multilingual and multimodal speech emotion detection, no notable major study has been published as a book. Most of the research has been conducted in the domains of engineering (13%), computer science (40%), and the social sciences, mathematics, arts and humanities, decision sciences, and the like (figure. 4).

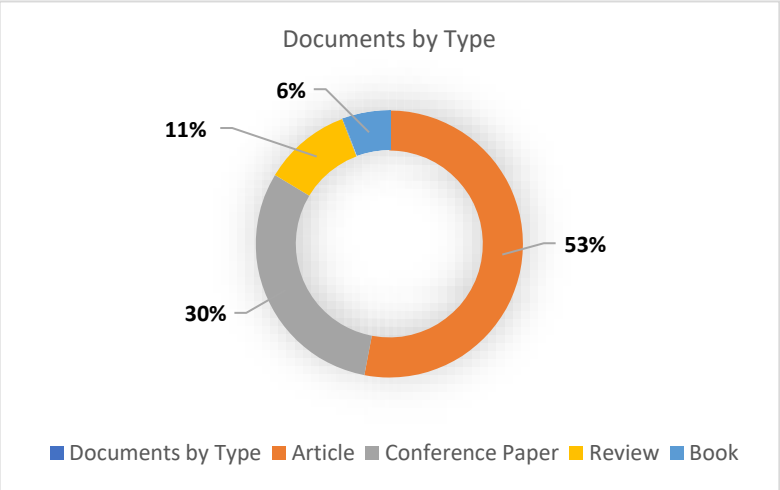

Figure 3: Scopus Database Documents, Classified (2013- October 2022)

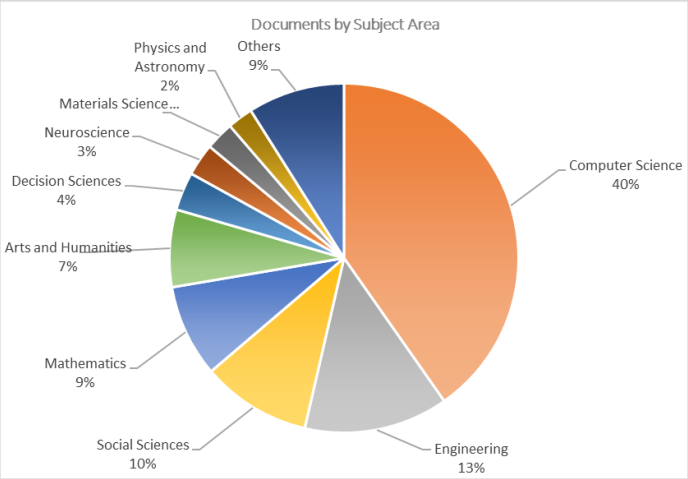

Figure 4.: Scopus Database Documents, Grouped by Subject (2013- October 2022)

Between January 1, 2013, and October 23, 2022, a total of 706 scientific studies on multilingual and multimodal speech emotion recognition trends were published (Figure 5).

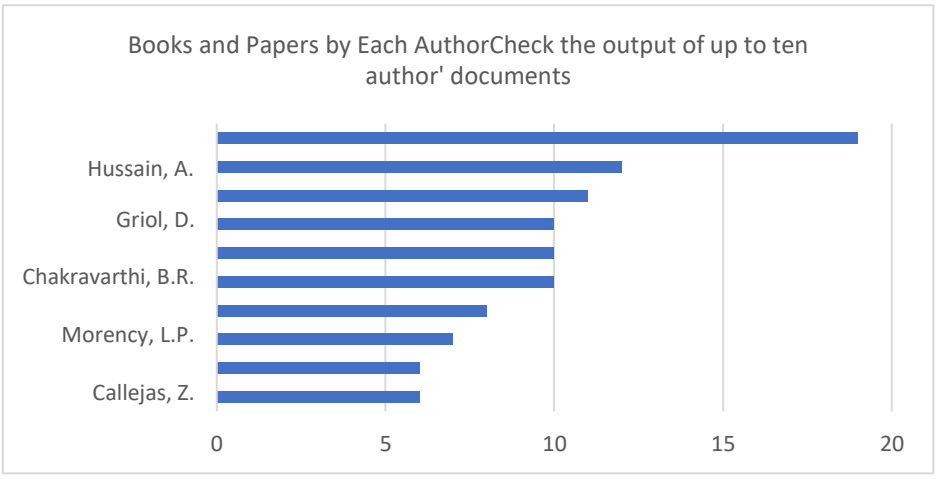

Figure 5: A Total of 706 Scientific Studies on Multilingual and Multimodal Speech Emotion Recognition Trends were Published

3.1 Network Analysis

Network analysis is a crucial part of bibliometric analysis and is more regularly used to assess papers on multilingual and multimodal speech emotion detection [3]. There are two options available. The first one is a co-occurrence analysis of keywords, and the second one is an analysis of co-authorship. 706 research articles published between January 1, 2013, and October 23, 2022, in total, were evaluated to determine the research trend in voice emotion recognition.

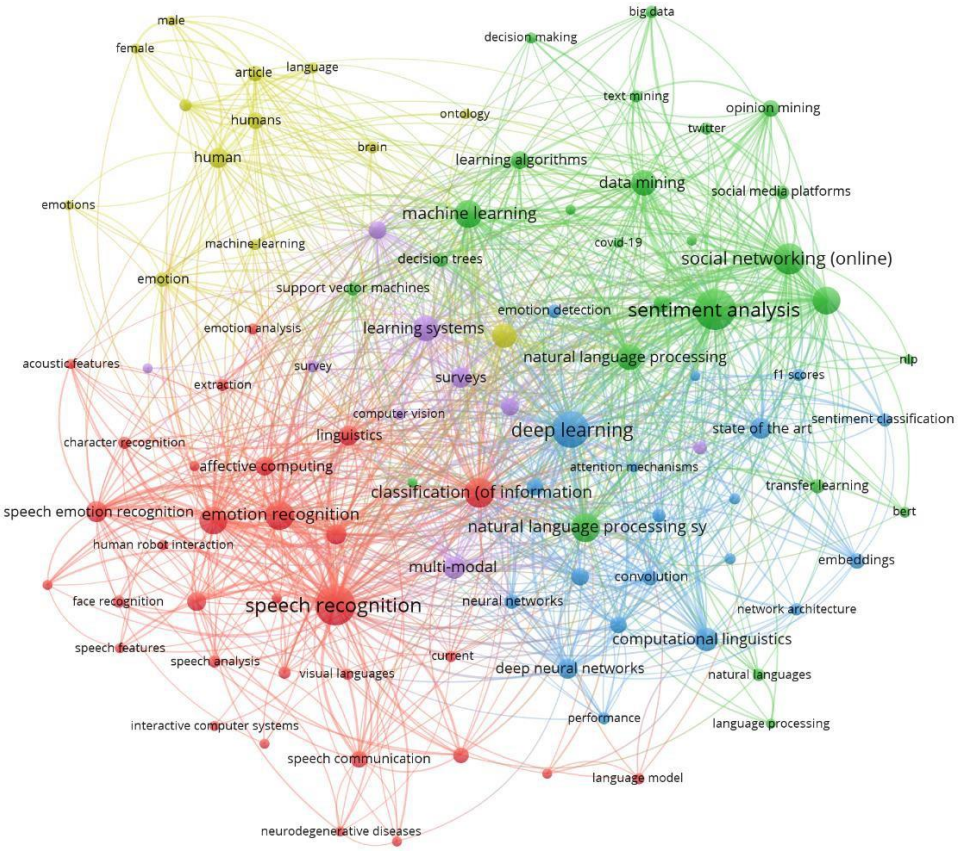

Figure 6. Network visualisation of title co-occurrences

3.1.1 Co-occurrence Network Analysis

This study's database (bibliographic database files (Scopus)) was analysed using the VOSviewer application. We choose to construct a word co-occurrence map based on text data (Figure 6.). Only 99 of the 4242 total words were sufficient for the analysis. There must be ten minimum repetitions of each phrase (10). Figure 6 features portraits of all 99 words. A relevancy score has been established for the visualisation of titles co-occurrence network. The most relevant phrases will be chosen based on this score. The 60% of terms that are most relevant are selected by this software's default settings. For the network visualisation of this study, 99 keywords from item titles were eliminated by the VOSviewer application. These 99 terms are broken up into 5 clusters, each of which includes a particular group of terms (total link 2473 and link strength 8106).

3.1.2 Co-authorship Network Analysis

Figure 7 shows the author collaboration network in terms of publications. The nodes stand in for the names of the writers, the linkages for the relationships of co-authorship between various authors, and the sizes of the nodes for the number of publications for each author. According to the results of the co-authorship network analysis, Hussain, A., and Dashtipour, K., are the authors with the greatest overall connection strength. Word co-authorship maps can be generated for this research if the requisite textual data is present. The decision on the threshold value was therefore necessary to consider. Out of 908 words, only 23 met the analysis's cut off point. The bare minimum frequency any phrase appears is (5). This software's default configuration chooses 60% of the most pertinent phrases. In order to visualise the co-authorship network for this study, the VOSviewer software removes 20 authors from the list of all authors (items). Each of these 20 terms belongs to one of four groups (total links = 36, link strength = 76).

In this sense, a connection is just the occurrence of two terms together (keywords). According to the VOSviewer user guide, a positive numerical number represents the strength of a connection. If this number is high, then the correlation is quite strong. Overall, the strength of the links between compounds indicates how often they are found in the same scientific papers.

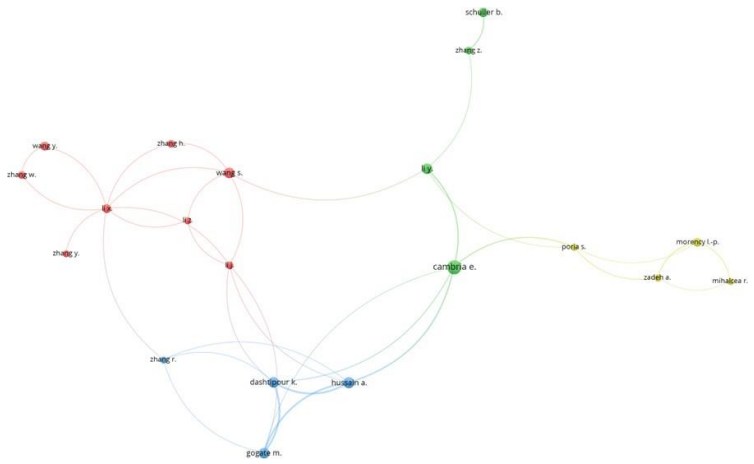

Figure 7. The Author Co-Authorship Network.

Note: VOSviewer presents a network view. There are 20 nodes, 4 clusters, and 36 links in the entire network. The value of the total link strength is 76.

According to the study's findings, the quantity of publications in which two or more items appear correlates with the overall link strength (co-incidence connection between objects or keywords**).**

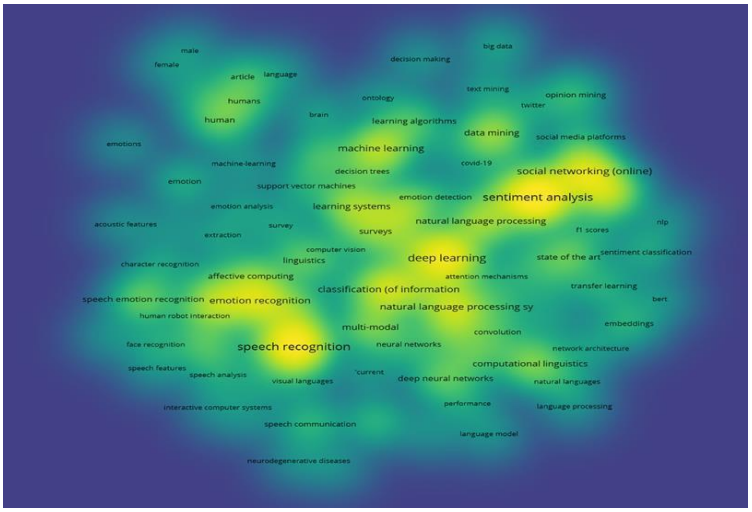

Figure 8. The Network of Title Co-Occurrences, Visualised By Density.

simultaneously. We discovered that deep learning, sentiment analysis, and speech recognition have better connection strengths than the other components in our experiment. Figure 8 displays the density visualisation based on the overall hyperlink strength.

4. Discussion

This work demonstrates how multilingual and multimodal speech emotion identification has recently attracted a lot of research attention. Speech emotion recognition technology is advancing quickly in multilingual and multimodal settings. The most notable effects of speech recognition, sentiment analysis, and deep learning were discovered in this voice emotion detection analysis study. Voice recognition systems in the subject of speech emotion detection generally try to imitate how people communicate in order to solve users' queries and concerns. Speech is the most prevalent and important form of human communication [4]. For this reason, the study of speech emotion recognition has recently become a more important research subject. [5]. Speech emotion recognition offers a wide range of real-world uses. The crime investigation department would benefit from the emotional analysis of criminals' telephone conversations, interactive movies, narratives, and e-tutoring apps would be more beneficial if they could adapt to the emotional states of listeners or pupils, and call centre conversations can be used to examine call attendants' interactions with customers. Additionally, it has been noted that conversations with robot companions and humanoid partners will be more pleasurable and realistic if they can comprehend and express human-like emotions [6, 7]. The range of multilingual and multimodal speech emotion detection will expand as long as there is continuing research and development in several disciplines of speech emotion identification.

5. Conclusion

The study of emotion identification has been ongoing for some time. The fundamental model for studying emotion detection was developed by recognising emotions from facial expressions. The subject of recognising emotions from speech signals has been the subject of a great deal of research in recent years. This paper aims to do two things: present a brief outline of existing multilingual and multimodal speech emotion identification applications and provide

some recommendations for further study. Many contemporary technologies and methods of communication rely heavily on the processing, interpretation, and understanding of spoken language. The purpose of this research is to give a thorough evaluation of speech emotion detection, covering the years 2013 through October of this year. It was shown that deep learning, sentiment analysis, and voice recognition had the most influence on this investigation. These three topics are where the bulk of the study's focus lies. It seems that this area of signal processing will continue to expand substantially in the years ahead. Scholars agree that this rule is noteworthy due to the technological influence it has had on society. Researchers will eventually be able to create systems powerful enough to evaluate natural speech as technology progresses. The findings of this investigation should encourage more research and development in this area.

References:

[1] F. W. Smith and S. Rossit, ``Identifying and detecting facial expressions of emotion in peripheral vision," PLoS ONE, vol. 13, no. 5, May 2018, Art. no. e0197160, doi: 10.1371/journal.pone.0197160.

[2] M. Chen, P. Zhou, and G. Fortino, ``Emotion communication system," IEEE Access, vol. 5, pp. 326 337, 2017, doi: 10.1109/ACCESS.2016.2641480.

[3] S. Yoon, S. Byun, and K. Jung, “Multimodal speech emotion recognition using audio and text,” in 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018, pp. 112–118.

[4] H. Meng, T. Yan, F. Yuan, and H. Wei, “Speech emotion recognition from 3d log-mel spectrograms with deep learning network,” IEEE access, vol. 7, pp. 125 868–125 881, 2019.

[5] Y. Zong, W. Zheng, T. Zhang, and X. Huang, “Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression,” IEEE signal processing letters, vol. 23, no. 5, pp. 585–589, 2016.

[6] O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The enterface’05 audiovisual emotion database,” in 22nd International Conference on Data Engineering Workshops (ICDEW’06). IEEE, 2006, pp. 8–8.

[7] Dhall, R. Goecke, J. Joshi, K. Sikka, and T. Gedeon, “Emotion recognition in the wild challenge 2014: Baseline, data and protocol,” in Proceedings of the 16th international conference on multimodal interaction, 2014, pp. 461–466.

[8] S. Steidl, Automatic classification of emotion related user states in spontaneous children’s speech. Logos-Verlag, 2009. [Page 24.]

[9] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, “Crosscorpus speech emotion recognition based on transfer non-negative matrix factorization,” Speech Communication, vol. 83, pp. 34–41, 2016.

[10] P. Song, “Transfer linear subspace learning for cross-corpus speechemotion recognition.” IEEE Trans. Affect. Comput., vol. 10, no. 2, pp. 265–275, 2019.

[11] Bhattacharya, S., Das, N., Sahu, S., Mondal, A., & Borah, S. (2021, March). Deep Classification of Sound: A Concise Survey, In Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR 2020 (Vol. 169, p. 33). Springer Nature.

[12] Bhattacharya, S., Borah, S., Mishra, B. K., & Das, N. (2022, September). Deep Analysis for Speech Emotion Recognization. In 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA) (pp. 1-6). IEEE.

[13] Bhattacharya, S., Borah, S., Mishra, B. K., & Mondal, A. (2022). Emotion detection from multilingual audio using deep analysis. Multimedia Tools and Applications, 1-30.

[14] Mathur, M., Samiulla, S., Bhat, V., & Jenitta, J. (2020, October). Design and Development of Writing Robot Using Speech Processing. In 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC) (pp. 1-4). IEEE.

[15] Handoko, L. H. (2021). COVID-19 research trends in the fields of economics and business in the Scopus database in November 2020. Science editing, 8(1), 64-71.

[16] Das, N., & Padhy, N. (2021, August). A Bibliometric and Co-occurrence Analysis of Speech Processing Literature Published from the Year 2015 to mid of June 2021. In Proceedings of the International Conference on Data Science, Machine Learning and Artificial Intelligence (pp. 237-243).

[17] Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. International Journal of Computer Applications, 10(3), 16-24.

[18] Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: a review. International journal of speech technology, 15(2), 99-117.

[19] Thapliyal, N., & Amoli, G. (2012). Speech based emotion recognition with gaussian mixture model. Int. J. Adv. Res. Comput. Eng. Technol, 1(5), 65-69.

[20] Vogt, T., André, E., & Wagner, J. (2008). Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation. Affect and emotion in human-computer interaction, 75-91.