Revue d'Intelligence Artificielle
Vol. 38, No. 2, April, 2024, pp. 551-558
Journal homepage: http://iieta.org/journals/ria

ASER: An Exhaustive Survey for Speech Recognition based on Methods, Datasets, Challenges, Future Scope

Dharil Patel[1], Soham Amipara[1], Malay Sanaria[1], Preksha Pareek[1], Ruchi Jayaswal[1*], Shruti Patil[2]

[1] Department of Artificial Intelligence and Machine Learning, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, Maharashtra, India
[2] Symbiosis Centre for Applied Artificial Intelligence, India Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, Maharashtra, India

Corresponding Author Email: ruchi.jayaswal@sitpune.edu.in

https://doi.org/10.18280/ria.380218

Received: 28 August 2023
Revised: 1 December 2023
Accepted: 18 January 2024
Available online: 24 April 2024

Keywords:
speech emotion recognition, feature extraction, datasets, classification, emotions, machine learning, deep learning, hybrid methods

ABSTRACT

AI has been used to process the data for decision-making, problem-solving, interaction with humans and to understand human's feelings, emotions and their behavior. In today's world, communication between humans takes place digitally, so human's emotions play a very important role for communication as well as detection and analysis. Although there are many surveys related to emotions from speech already done, selecting appropriate datasets and methods are challenging tasks. This survey will primarily concentrate on efficient techniques, including Machine Learning, Deep Learning, and transformer-based approaches, while also providing brief descriptions of existing challenges and outlining future prospects. Additionally, this paper provides a comparative analysis of various datasets and techniques employed by researchers. After conducting the survey, we discovered that deep learning and transformer-based techniques are more effective and yield superior performance results.

1. INTRODUCTION

Natural Language Processing (NLP) [1] combined computational and linguistic approaches to help computers comprehend human languages and enable human interactions, has continued to use Artificial Intelligence (AI) [1]. The development of AI includes both the study of human feelings and emotions as well as two-way interactions between humans and machines. AI that monitors, comprehends, replicates, and responds to emotions of humans is referred to as "feeling AI" [2]. Emotions continue to be crucial in human-computer interaction because they help computers comprehend human behavior or become more sensitive to it. Therefore, technology that is emotion-driven is crucial to decision-making, which may be useful across a wider number of application areas. Management, Education, Healthcare, Finance, Public Monitoring, etc. are included in this domain. Compared to sending texts, speaking to another person is a quick and simple approach to convey views. Speech plays a crucial part in all forms of communication, thus the recent development apps that may take advantage of these technologies have created a huge potential for the extraction of emotions from speech.

The challenge of recognizing emotions is highly difficult for a variety of reasons. The first justification is to choose the right datasets and efficient techniques. The second purpose is to determine the speech's tone and pitch [2].

While numerous surveys have been conducted in the field of speech emotion recognition, we observed that not all available and efficient techniques for SER have been comprehensively addressed. Some authors have exclusively concentrated on either machine learning or deep learning techniques and specific datasets. This paper outlines the techniques currently available, different types of datasets utilized by researchers, and the corresponding performance results they have achieved. We have also addressed the key challenges faced by researchers and provide future directions to overcome these challenges.

Since many of the extracted features, such as pitch, energy, and tones, are directly influenced by these traits, the incorporation of many languages, accents, phrases, speaking styles, and speakers also makes the task more challenging. Using a suitable classifier that carefully extracts and chooses features, including prosodic and spectral parameters like pitch, intonation, and duration using the Mel-Frequency Cepstral Coefficient (MFCC) [3, 4], Perceptual Linear Prediction (PLP) [5], Relative Spectral Filtering (RSF), and Linear Prediction Cepstral Coefficient (LPCC) [6] The Support Vector Machine (SVM) [7, 8], Hidden Markov Model (HMM) [9, 10], Gaussian-Mixture Model (GMM) [11] are just a few of the numerous classifiers used for emotion recognition.

1.1 Search strategy

Bibliometrics is employed in SER study to gain a comprehensive overview of publications, utilizing quantitative methods for evaluating terms in research metrics. A visual representation of the search strategy for SER is depicted in Figure 1 [12]. Essential data, including research evaluation

metrics and information on SER, can be retrieved from reputable repositories such as "Scopus," introduced by Elsevier in 2004 [13], and "Web of Science," which contains abstract and citation databases of SCI and SSCI publications [14].

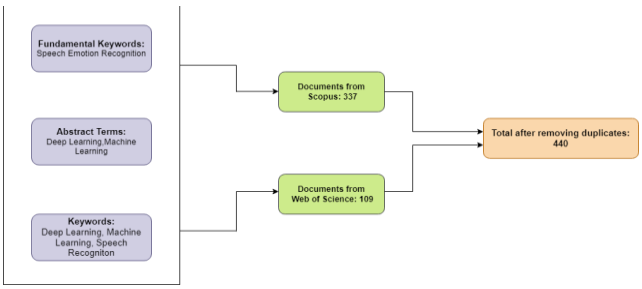

Figure 1. Search strategy

Query in Scopus:

(TITLE ("Speech Emotion Recognition") AND 'TITLE-ABS-KEY ("Machine Learning")' OR 'TITLE-ABS-KEY ("Deep Learning")' AND 'TITLE-ABS-KEY ("Speech Recognition"))'.

Query in Web of Science:

("Speech Emotion Recognition" (Topic) AND "Machine Learning" (Title) OR "Deep Learning" (Title) AND "Speech Recognition" (Title)).

1.2 Quantitative analysis

Here, we have shown quantitative analysis in which Figure 2 shows the Co-occurrences of SER keywords and Figure 3 shows the Countries/Regions wise publication done for SER which we have extracted from "Web of Science" [14].

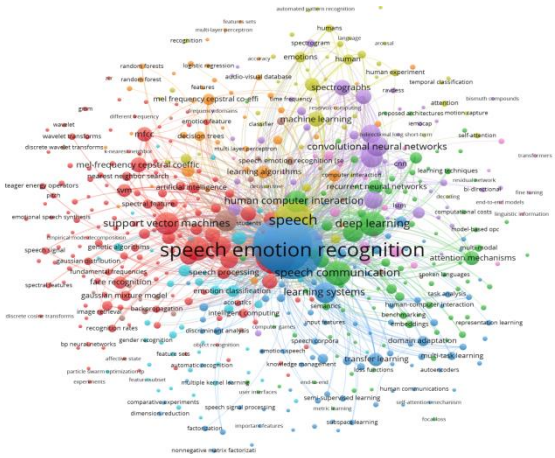

Figure 2. Co-Occurrence of SER

Figure 3. Countries/Regions wise publications in WOS

1.3 Motivation

There are many surveys done on speech emotion recognition but there is no proper survey existing that describes the proper techniques, datasets and application domains, comparative analysis of deep learning and machine learning technique, detailed information and future usage in different domains.

The systematic review of this paper shows the insufficiency in the surveys about the datasets and classification techniques. Even though similar datasets of this nature exist, each one has a lack of information and few target labels that are associated with emotions. So, it can't be generalized.

This systematic survey of speech emotion recognition is highly dependent on different AI techniques, datasets and challenges.

1.4 Evolution of speech emotion recognition

Emotion [15] was used by the scientist in his work, "The Expression of the in Animals and Man" [16]. It was the first hypothesis to identify human emotion and expression. There are eight primary and fundamental emotions included in this: grief, surprise, joy, trust, disgust, and anticipation [15].

First facial expression-based emotion identification system was created in 1980 [16]. The Emotion Facial Action Coding System (EFACS) [17] was created by researchers [17].

Traditional strategies, such as HMM paired with feed-forward Artificial Neural Network (ANN), continued to rule speech recognition from 2002 to 2004 [18]. However, a deep learning technique known as Long Short-Term Memory (LSTM) [18, 19], a Recurrent Neural Network (RNN) [20, 21] described by researchers, has already taken over many elements of voice recognition [22].

During 2011-2022, there were many research has been done on SER, also there were lots of new technology has arrived in the world, In 2018, Microsoft introduced "Cortona" and amazon gives the "Echo", a voice controlled speaker powered by Alexa [23]. In 2019, voice activated digital assistants come into the world which are primarily based on a speaker dependent system and speaker independent system [23]. In 2021, Vehicle speech recognition has come into the world, where drivers can use the default word "Hey BMW" to assist for a more personalized experience [24].

1.5 Prior research

The writers provided a comprehensive discussion of the many methods for drawing out emotion from speeches. The characteristics, strategies, procedures, and datasets used in this speech article emotion extraction are discussed [25]. Problems with dataset quality and difficulties are also addressed, which might help in determining the subject areas where sentiment analysis research is needed [25]. In their survey study, the authors analyzed several speeches and emotions recognition system-based methodologies. The performance of the classifier, features, recognition rate, and datasets are compared by the authors. The authors also investigated the various corpora and lexicons that are available for speech emotion recognition.

Here, we covered some of the surveys on SER. The task of SER is not brand-new, and literature has long examined it. Early projects based on Speech Emotion Recognition (SER) [25] were created using HMM, according to the author's

summary, but the recent introduction of deep neural networks to the field has also substantially enhanced the performance [25].

Authors have also used that auto encoders were employed to complete the job, and all the techniques specified in the survey study were investigated, such as Tensor Fusion Networks (TFN) [25] and Low-Rank Matrix Multiplication (LRMM) [25] in place of trivial concatenation. Random forest (RF) [25], Gradient Boosting (XGB), SVM, Multinomial Naive Bayes (MNB) [25], as well as Multi-Layer Perceptron (MLP) [25] and LSTM were cited as ways employing both ml and dl [25].

1.6 Contributions of the work

The following are the contributions of this comprehensive literature review:

- ❖ We have provided a thorough review of the existing research on SER with an emphasis on methods, datasets, operations, related challenges, and potential future directions.
- ❖ We have explored around and attempted to examine the methods and concepts employed, as well as how they are improving SER exploration.
- ❖ We have also provided a comparative study of several publicly accessible datasets that might aid with SER exploration.
- ❖ We have discussed various challenges with SER, such as the dataset, the efficacy of current methods, and the control of data quality.

2. LITERATURE SURVEY

The SER system is a set of strategies that will process, classify, and detect the emotions. Figure 4 indicates that, while we take a huge perspective, we're dividing it into the various classes' emotion models, datasets, methods, application domains, preprocessing strategies [25], feature extraction [25] strategies, and classifiers are shown in Figure 4 [25]. It will be beneficial to get a deeper information of emotions to decorate the classification process. Modeling feelings contains lots of techniques as well as processes. There are specific techniques used for extraction of features, which includes MFCC [26], LPCC [26] etc. For Classification, numerous algorithms were proposed which includes in this survey [26]. More recently, DL, ML and transformer-based techniques have normally been used. We have given a comparative analysis of these techniques within the coming sections.

2.1 Background

Implementing SER requires following steps. Figure 4 shows a model based on SER. The first step is to get data from available sources. Next, we need to perform data preprocessing. Data Preprocessing steps includes removing ambient unwanted noise from the speech signal, recognizing speech frequency and adjusting a vocal tract length [27]. The next phase is features analysis, which includes extraction of features and selection of features [27]. Classification is the final stage, assigning the information to one of the valid classes. classification techniques are used to recognize emotions [28]. For example, there are many ML, DL and transformer-based classifiers which are included in this literature.

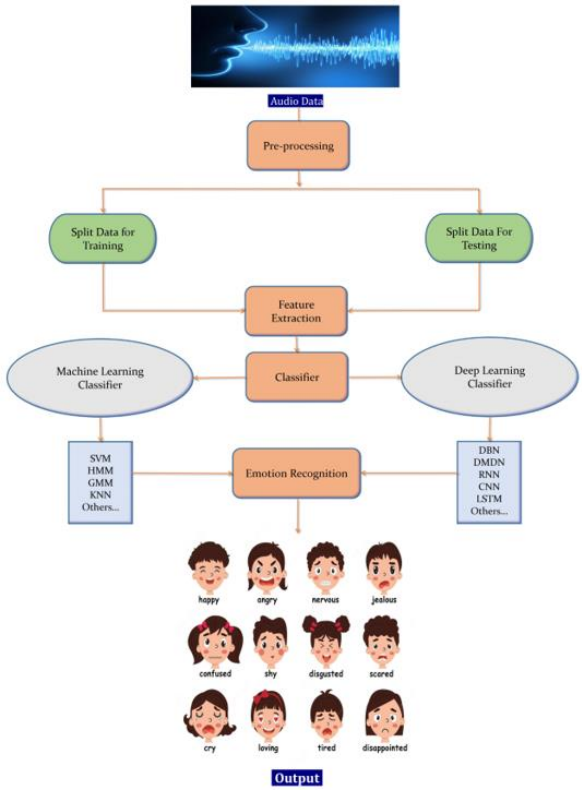

Figure 4. Block Diagram for speech emotion recognition

2.2 Research gaps

- The majority of studies focus on recognizing basic emotions, neglecting the complexity of mixed emotions and subtle variations in emotional states that occur in natural conversations.
- Limited exploration of real-time and continuous emotion recognition in dynamic social interactions poses a significant gap in SER.
- The impact of variations in speech characteristics, such as accents, speech disorders, and environmental noise, on the accuracy of emotion recognition models remains insufficiently explored, limiting the robustness of these systems.
- Limited research exists on the integration of contextual information, such as facial expressions or contextual cues, into SER systems, hindering their ability to accurately capture the complexity of emotional expression in multimodal communication.

2.3 Techniques for data pre-processing

Pre-processing of data [29] is necessary for feature engineering, feature selection, and data cleaning before applying any model. For applying any AI techniques, numerical data is necessary to train the model. If we do not apply data processing, the model will provide inaccurate results that are not in line with our expectations. For carrying out this, a variety of techniques and built-in libraries are available. We must apply data Pre-processing to the available speech or audio data to recognize speech emotions. The following strategies for data Pre-processing used by researchers are shown in Figure 5:

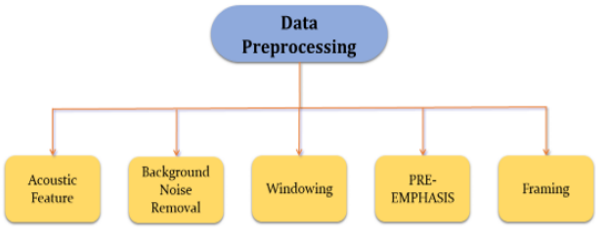

Figure 5. Data Pre-processing methods

In the realm of speech emotion recognition, effective data preprocessing techniques play a pivotal role in enhancing the accuracy and reliability of the models. Acoustic feature extraction enables the representation of relevant characteristics from speech signals, contributing to the discrimination of emotional cues. Background noise removal is crucial for isolating the desired speech signal from unwanted environmental sounds, ensuring a cleaner dataset. Windowing and framing techniques aid in segmenting speech signals into manageable frames, facilitating subsequent analysis. Pre-emphasis helps in emphasizing high-frequency components, enhancing the signal's clarity. Altogether, these preprocessing steps collectively contribute to the robustness and efficacy of speech emotion recognition systems, as evidenced by findings in existing surveys.

2.4 Feature extraction methods

We have used the MFCC, Mel-Spectrogram, and Chroma [30] as three essential components from the audio data in our study. The literature indicates that Chroma has also been considered by researchers. So, the researcher has evaluated fundamental modeling using MFCC, Mel and chroma [30].

Table 1 presents a comparison of feature extraction techniques utilized by researchers, accompanied by performance metrics, specifically in terms of accuracy, reported by different studies.

Table 1. Comparison of feature extraction with different classifier

No.	Year	Features	Dataset	Classifier	Accuracy
1	2017 [31]	Pitch, Format, Phoneme	RAVDESS	SVM, HMM	77%
2	2017 [32]	MFCC Acceleration, Velocity	Berlin Emotion Speech Dataset (Emo-DB)	CNN, LSTM	80%
3	2018 [33]	Mel Spectrogram, Harmonic Percussive	RAVDESS	DNN, KNN	90%
4	2019 [34]	MFCC	RAVDESS, eNTERFace	SNN	83%

2.5 Classification methods

There have been several suggested classification algorithms throughout the past few decades. The choice of the optimal approach for our model is crucial. Here, we have listed a few of the categorization techniques in the following manner:

2.5.1 Machine learning classifier

Different classification methods are employed to model emotional states, but the most used ones include SVM [35, 36], HMM [36], KNN [37], and GMM [38].

Figure 6 shows the major techniques of machine learning used in SER by most of the authors. We have conducted a comparative analysis of various machine learning classifier as shown in Table 2.

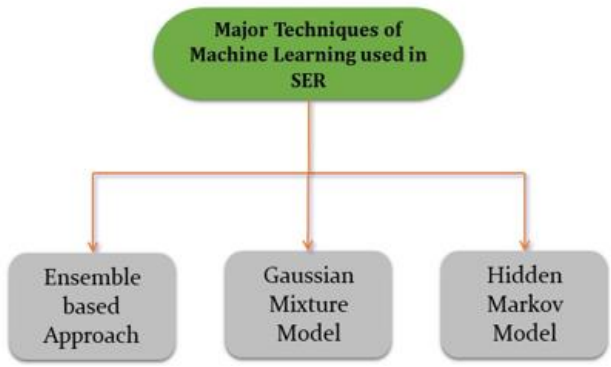

Figure 6. Major techniques of machine learning used in SER

Table 2. Comparison of machine learning classifier

No.	Year	Preprocessing	Feature Extraction	ML Classifier	Dataset	Metric Value (%)
1	2003 [39]	Analog to Digital Conversion, Noise Reduction	MFCC, LPCC	HMM, LPCC	Self-Made Burme Selang	Accuracy (96%)
2	2020 [40]	-	MFCC, LPCC	SVM	LDC, UGA	Accuracy (72.7%)
3	2015 [41]	Acoustic Feature, Pre-emphasis	MFCC, Wavelet features	GMM, KNN	Berlin Emotional Speech Database	Accuracy (76%)
4	2021 [42]	Pre-emphasis, Framing	MFCC	KNN, Medium KNN, Cosine KNN	Berlin Dataset	Accuracy (90.1%)

2.5.2 Deep learning classifier

In Recent years, there has been a lot of interest in the newly developing machine learning research subject known as deep learning. A few scientists have employed DNNs to train the various SER models they are using. Multiple nonlinear components that carry out computation in parallel make up deep learning algorithms [43]. To get around the drawbacks of previous approaches, these solutions must be more elaborately built with deeper layers of design. RNN, Deep Belief Network (DBN) [43], CNN, LSTM, and One of the basic deep learning methods utilized for SER is Auto Encoder (AE) [44], which considerably improves the developed system's overall performance [44].

Figure 7 shows the major techniques of deep learning used in SER by most of the authors. We have conducted a comparative analysis of various deep learning classifier as shown in Table 3.

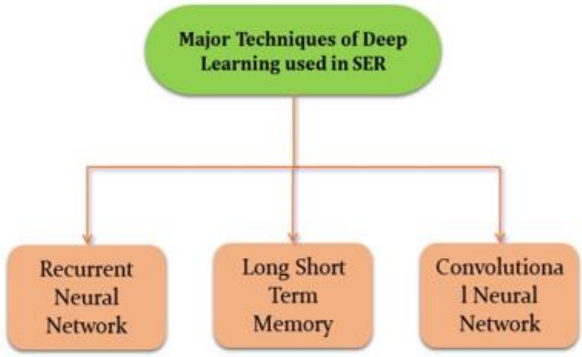

Figure 7. Major techniques of deep learning used in SER

2.5.3 Transformer-based techniques

Transformer models have demonstrated remarkable performance in various natural language processing tasks and have been adapted for speech-related tasks as well. Here, Table 4 contains the comparison of transformer-based techniques for SER.

2.6 Datasets

There are several datasets available for SER. Some of the Datasets contain gender specific and some do not. Table 5 gives the comparison analysis of various datasets which is used by authors:

Table 3. Comparison of deep learning classifier

No.	Year	Preprocessing	Feature Extraction	DL Classifier	Dataset	Metric Value (%)
1	2003 [45]	Pre-emphasis, Analog to Digital Conversion	Pitch & Energy contour	HMM	Speech Corpus	Word Recognition (86.8%)
2	2014 [46]	Pre-emphasis, Framing	MFCC	ANN, CNN	TIMIT phone recognition dataset	Word Error Rate (37.1%)
3	2018 [47]	Analog to Digital Conversion, Pre-emphasis, Framing	MFCC	RNN	Google Speech Command Dataset	Accuracy (96.6%)
4	2019 [48]	Framing, Noise Reduction, Pre-emphasis	MFCC	CNN, BLSTMCTC	Mandarin speech corpus AISHELL-1	Word Error Rate (19.2)

Table 4. Comparison of transformer-based techniques

No.	Year	Preprocessing	Feature Extraction	Transformer Technique	Dataset	Metric Value (%)
1	2023 [49]	Removal of Noise, Data Augmentation	MFCC, FFT	Multi-head Attention	EMO-DB, SAVEE, EMOVO	Accuracy (95%)
2	2022 [50]	-	Local Attention Module	Local Attention Speech Transformer	LibriSpeech	Word Error Rate 2.3/5.5%
3	2022 [51]	-	Stack of CNN Layers	Pre-trained XLSR Model	Tamil Conversational Speech	Word Error Rate (39.65%)
4	2023 [52]	Remove non-speech segments, Adaptive thresholding	MFCC, Chroma	2D-CNN, BiLSTM-Transformer	EMO-DB, RAVDESS	Average Recognition Rate (89.06%)

Table 5. Dataset information

No.	Dataset name	Language	Application	Link
1	IEMOCAP	English, German	Recognition & Analysis of Emotional Expression	https://sail.usc.edu/iemocap
2	PF-STAR	British English	Children Speech Identification	http://www.thespeechark.com/pf-star-page.html
3	Urdu Dataset	Urdu	Emotional Recognition for Urdu	https://ieeexplore.ieee.org/document/8978091
4	The Berlin Database of Emotional Speech (EMO-DB)	German	Emotional Recognition for German	http://emodb.bilderbar.info/start.html
5	TIMIT Acoustic Phonetic Continuous Speech Corpus	English	Development of ASR Systems	https://catalog.ldc.upenn.edu/LDC93s1
6	CHIME	English	Noise-Robust Speech Processing Research	https://archive.org/details/chime-home
7	RAVDESS	American English	Speech Emotion Recognition	https://zenodo.org/record/1188976#.Y4xtAXZBxEa

3. CHALLENGES

There are some challenges of SER which need to be solved. Challenges in terms of performance, insufficient data, different languages, high cost these need to be solved.

Some of the challenges of SER are mentioned below:

3.1 Availability of different Languages

There are many datasets available for speech but very few

of the datasets are available in different languages. It is a task to express emotion in specific languages.

Emotions are complex and culturally influenced, leading to variations in expression across languages and communities. The inherent subjectivity arises from the interpretation of emotional content, which can differ based on linguistic nuances and cultural contexts.

Emotions are multifaceted and culturally dependent, making it challenging to create universally applicable datasets that capture the diversity of emotional expression in speech across various languages. Researchers must grapple with these subjective and cultural intricacies, hindering the development of comprehensive and universally applicable speech emotion recognition models.

3.2 Emotional subjectivity and fuzziness

The interpretation of emotional cues in speech is highly subjective, making it challenging to establish clear-cut boundaries for emotional categories. Additionally, the fuzziness of emotional expressions further complicates the recognition process.

Overcoming this challenge requires a nuanced understanding of the subjective and fuzzy nature of emotions and the development of adaptable models capable of accommodating this inherent variability.

3.3 Emotional feature extraction and selection problems

Emotions are intricate phenomena with multifaceted expressions in speech, and identifying relevant features to characterize these expressions is a complex task.

The inherent subjectivity in emotional interpretation further complicates this challenge, as different individuals may prioritize distinct features when perceiving and expressing emotions. The process of selecting features that effectively capture the diverse range of emotional expressions across speakers and cultures becomes intricate due to the subjective nature of emotional experiences.

Addressing this challenge involves a comprehensive understanding of the subjective nature of emotions, the development of feature extraction techniques that accommodate this subjectivity, and the selection of features that can generalize across diverse emotional expressions in speech.

3.4 Lack of tagging data

The scarcity of adequately tagged data poses a significant challenge in the domain of speech emotion recognition. While substantial progress has been made in the field, the limited availability of labeled datasets hampers the training and evaluation of models. Deep learning methods can attain high performance levels, but they necessitate a substantial amount of high-quality labeled data.

Tagging speech emotion is a time-consuming and labour-intensive task due to the subjective and fuzzy nature of emotions. Additionally, it requires a substantial number of professionals. The urgent challenge in the field of Speech Emotion Recognition (SER) lies in efficiently collecting a large volume of emotion tagging information.

Researchers must contend with the difficulties of obtaining large, accurately annotated datasets that encompass the diverse and nuanced nature of emotional expressions in speech across various contexts and cultures.

3.5 High cost

The need for extensive and diverse datasets, coupled with the requirement for meticulous annotation by human experts, contributes significantly to the overall expenses.

Performance and outcomes of the emotion sensing machine rely upon the accuracy of the sensors which includes cameras, thermal picture sensors, facial reputation set of rules used and so on. A highly correct machine could be highly priced because of the usage of pricey components.

Addressing this challenge involves exploring cost-effective strategies for data collection and annotation, fostering collaborative efforts within the research community, and investigating alternative approaches that mitigate the financial constraints associated with obtaining high-quality labeled datasets.

4. CONCLUSION

This study offers a comprehensive overview of SER up to recent years. By conducting this survey, we have identified the current research gaps by providing comparative analysis of different methods and given the overview of the different types of datasets used by researchers. We have discussed the existing key challenges of SER and given a future direction to overcome these challenges. Our research indicates that incorporating attention mechanisms enhances the performance of SER systems.

Moving forward, our focus is to explore advanced techniques within the realm of transformers. In future, the work can be extended to gather real-time speech data with facial expressions to capture the complexity of emotional expression in multimodal fashion, employing these datasets for training and evaluation. This comprehensive survey aims to guide researchers in selecting appropriate methods, datasets and addressing existing challenges in the field of SER.

REFERENCES

[1] Basharirad, B., Moradhaseli, M. (2017). Speech emotion recognition methods: A literature review. In AIP conference proceedings. AIP Publishing. https://doi.org/10.1063/1.5005438

[2] Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Vora, D., Pappas, I. (2022). A review on text-based emotion detection--techniques, applications, datasets, and future directions. arXiv preprint arXiv:2205.03235. https://doi.org/10.48550/arXiv.2205.03235

[3] Dellaert, F., Polzin, T., Waibel, A. (1996). Recognizing emotion in speech. In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96, 3: 1970-1973. https://doi.org/10.1109/ICSLP.1996.608022

[4] Harshawardhan S. Kumbhar, Sheetal U. Bhandari. (2019). Speech emotion recognition using MFCC features and LSTM network. 2019 5th International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India. https://doi.org/10.1109/ICCUBEA47591.2019.9129067

[5] Dave, N. (2013). Feature extraction methods LPC, PLP and MFCC in speech recognition. International Journal for Advance Research in Engineering and Technology, 1(6): 1-4.

[6] Gupta, H., Gupta, D. (2016). LPC and LPCC method of feature extraction in speech recognition system. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence), Noida, India, pp. 498-502. https://doi.org/10.1109/CONFLUENCE.2016.7508171
[7] Zhang, W., Zhao, D., Chai, Z., Yang, L.T., Liu, X., Gong, F., Yang, S. (2017). Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services. Software: Practice and Experience, 47(8): 1127-1138. https://doi.org/10.1002/spe.2487
[8] Shen, P., Changjun, Z., Chen, X. (2011). Automatic speech emotion recognition using support vector machine. In Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology, 2: 621-625. https://doi.org/10.1109/EMEIT.2011.6023178
[9] Schuller, B., Batliner, A., Steidl, S., Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, 53(9-10): 1062-1087. https://doi.org/10.1016/j.specom.2011.01.011
[10] Maseri, M., Mamat, M. (2020). Performance analysis of implemented MFCC and HMM-based speech recognition system. In 2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, pp.1-5. https://doi.org/10.1109/IICAIET49801.2020.9257823
[11] Cheng, X., Duan, Q. (2012). Speech emotion recognition using gaussian mixture model. In 2012 International Conference on Computer Application and System Modeling, Atlantis Press, pp. 1222-1225. https://doi.org/10.2991/iccasm.2012.311
[12] Bidwe, R.V., Mishra, S., Patil, S., Shaw, K., Vora, D.R., Kotecha, K., Zope, B. (2022). Deep learning approaches for video compression: A bibliometric analysis. Big Data and Cognitive Computing, 6(2): 44. https://doi.org/10.3390/bdcc6020044
[13] https://www.elsevier.com/en-in/solutions/scopus.
[14] https://www.webofscience.com/wos/woscc/basic-search.
[15] https://en.wikipedia.org/wiki/The_Expression_of_the_Emotions_in_Man_and_Animals.
[16] Al-onazi, B.B., Nauman, M.A., Jahangir, R., Malik, M.M., Alkhammash, E.H., Elshewey, A.M. (2022). Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion.Applied Sciences, 12(18): 9188. https://doi.org/10.3390/app12189188
[17] Ekman, P., Freisen, W.V., Ancoli, S. (1980). Facial signs of emotional experience. Journal of personality and social psychology, 39(6): 1125-1134. https://doi.org/10.1037/h0077722
[18] Zhao, J., Mao, X., Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47: 312-323. https://doi.org/10.1016/j.bspc.2018.08.035
[19] https://medium.com/nerd-for-tech/what-is-lstm-peephole-lstm-and-gru-77470d84954b.
[20] Lim, W., Jang, D., Lee, T. (2016). Speech emotion recognition using convolutional and Recurrent Neural Networks. 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, Korea (South). https://doi.org/10.1109/APSIPA.2016.7820699
[21] Li, D.D., Liu, J.L., Yang, Z., Sun, L.Y., Wang, Z. (2021) Speech emotion recognition using recurrent neural networks with directional self-attention. Expert Systems with Applications, 173(3): 114683. https://doi.org/10.1016/j.eswa.2021.114683
[22] Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J. (2001). Gradient flow in recurrent nets: The difficulty of learning long-term dependencies.
[23] https://www.verizon.com/articles/speech-recognition-technology.
[24] https://summalinguae.com/language-technology/the-present-and-future-of-in-car-speech-recognition/.
[25] Sahu, G. (2019). Multimodal speech emotion recognition and ambiguity resolution. arXiv preprint arXiv:1904.06022. https://doi.org/10.48550/arXiv.1904.06022
[26] Madhavi, A., Priya Valentina, A., Mounika, K., Rohit, B., Nagma, S. (2021). Comparative analysis of different classifiers for speech emotion recognition. In Proceedings of International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2020, Springer Singapore, pp. 523-538. https://doi.org/10.1007/978-981-15-9293-5_48
[27] Efremova, K.O., Frey, R., Volodin, I.A., Fritsch, G., Soldatova, N.V. and Volodina, E.V. (2016). The postnatal ontogeny of the sexually dimorphic vocal apparatus in goitred gazelles (Gazella subgutturosa). Journal of Morphology, 277: 826-844. https://doi.org/10.1002/jmor.20538
[28] Basak, S., Agrawal, H., Jena, S., Gite, S., Bachute, M., Pradhan, B., Assiri, M. (2023). Challenges and limitations in speech recognition technology: A critical review of speech signal processing algorithms, tools and systems. CMES-Computer Modeling in Engineering & Sciences, 135(2): 1053-1089. https://doi.org/10.32604/cmes.2022.021755
[29] de Paula, P.O., da Silva Costa, T.B., de Faissol Attux, R. R., Fantinato, D.G. (2023). Classification of image encoded SSVEP-based EEG signals using Convolutional Neural Networks. Expert Systems with Applications, 214: 119096. https://doi.org/10.1016/j.eswa.2022.119096.
[30] Chroma. (2019). En.wikipedia.org. https://en.wikipedia.org/wiki/Chroma.
[31] Likitha, M.S., Gupta, S.R.R., Hasitha, K., Raju, A.U. (2017). Speech based human emotion recognition using MFCC. In 2017 international conference on wireless communications, signal processing and networking (WiSPNET), Chennai, India, pp. 2257-2260. https://doi.org/10.1109/WiSPNET.2017.8300161.
[32] Basu, S., Chakraborty, J., Aftabuddin, M. (2017). Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. In 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, pp. 333-336. https://doi.org/10.1109/CESYS.2017.8321292
[33] Tarunika, K., Pradeeba, R.B., Aruna, P. (2018). Applying machine learning techniques for speech emotion recognition. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bengaluru, India, pp. 1-5. https://doi.org/10.1109/ICCCNT.2018.8494104
[34] Mansouri-Benssassi, E., Ye, J. (2019). Speech emotion recognition with early visual cross-modal enhancement using spiking neural networks. In 2019 International joint

conference on neural networks (IJCNN), Budapest, Hungary, pp. 1-8. https://doi.org/10.1109/IJCNN.2019.8852473
[35] Cao, H., Verma, R., Nenkova, A. (2015). Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. Computer speech & language, 29(1): 186-202. https://doi.org/10.1016/j.csl.2014.01.003
[36] Lee, C.C., Mower, E., Busso, C., Lee, S., Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. Speech Communication, 53(9-10): 1162-1171. https://doi.org/10.1016/j.specom.2011.06.004
[37] Chen, L., Mao, X., Xue, Y., Cheng, L.L. (2012). Speech emotion recognition: Features and classification models. Digital signal processing, 22(6): 1154-1160. https://doi.org/10.1016/j.dsp.2012.05.007
[38] Yeh, J.H., Pao, T.L., Lin, C.Y., Tsai, Y.W., Chen, Y.T. (2011). Segment-based emotion recognition from continuous Mandarin Chinese speech. Computers in Human Behavior, 27(5): 1545-1552. https://doi.org/10.1016/j.chb.2010.10.027
[39] Nwe, T.L., Foo, S.W., De Silva, L.C. (2003). Speech emotion recognition using hidden Markov models. Speech communication, 41(4): 603-623. https://doi.org/10.1016/S0167-6393(03)00099-2
[40] Jain, M., Narayan, S., Balaji, P., P, B.K., Bhowmick, A., R, K., Muthu, R.K. (2020). Speech emotion recognition using support vector machine. https://doi.org/10.48550/arXiv.2002.07590
[41] Lanjewar, R., Mathurkar, S., Patel, N. (2015) Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K- Nearest Neighbor (K-NN) Techniques. Procedia Computer Science 49(1): 50-57. https://doi.org/10.1016/j.procs.2015.04.226
[42] Venkata Subbarao, M., Terlapu, S.K., Geethika, N., Harika, K.D. (2021). Speech emotion recognition using k-nearest neighbor classifiers. In Recent Advances in Artificial Intelligence and Data Engineering: Select Proceedings of AIDE 2020, Singapore: Springer Singapore, pp. 123-131. https://doi.org/10.1007/978-981-16-3342-3_10
[43] Harby, F., Alohali, M., Thaljaoui, A., Talaat, A.S. (2024). Exploring sequential feature selection in deep Bi-LSTM models for speech emotion-recognition. Computers, Materials & Continua, 78(2): 2689-2719. https://doi.org/10.32604/cmc.2024.046623
[44] Khalil, R.A., Jones, E., Babar, M.I., Jan, T., Zafar, M.H., Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7: 117327-117345. https://doi.org/10.1109/ACCESS.2019.2936124
[45] Schuller, B., Rigoll, G., Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings, 2: II-1. https://doi.org/10.1109/ICME.2003.1220939
[46] Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., Yu, D. (2014). Convolutional neural networks for speech recognition. IEEE/ACM Transactions on audio, speech, and language processing, 22(10): 1533-1545. https://doi.org/10.48550/arXiv.1808.08929
[47] De Andrade, D.C., Leo, S., Viana, M.L.D.S., Bernkopf, C. (2018). A neural attention model for speech command recognition. arXiv preprint arXiv:1808.08929. https://doi.org/10.48550/arXiv.1808.08929
[48] Wang, D., Wang, X.D., Lv, S.H. (2019). End-to-end mandarin speech recognition combining CNN and BLSTM. Symmetry 11(5): 644. https://doi.org/10.3390/sym11050644
[49] Chen, W., Xing, X., Xu, X., Pang, J., Du, L. (2023). DST: Deformable speech transformer for emotion recognition. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1-5. https://doi.org/10.1109/ICASSP49357.2023.10096966
[50] Fu, P.B., Liu, D.X., Yang, H.R. (2022). LAS-transformer: An enhanced transformer based on the local attention mechanism for speech recognition. https://doi.org/10.3390/info13050250
[51] Suhasini, S., Bharathi, B. (2022). Transformer based approach for speech recognition for vulnerable individuals in Tamil. https://doi.org/10.18653/v1/2022.ltedi-1.23
[52] Kim, S., Lee, S.-P. (2023). A BiLSTM– transformer and 2D CNN architecture for emotion recognition from speech. Electronics, 12(19): 4034 https://doi.org/10.3390/electronics12194034

NOMENCLATURE

NLP	Natural Language Processing
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
SER	Speech Emotion Recognition
CNN	Convolutional Neural Network
HMM	Hidden Markov Model
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network
ANN	Artificial Neural Network
SVM	Support Vector Machine
KNN	K-Nearest Neighbor
MFCC	Mel Frequency Cepstral Coefficient
LPCC	Linear Prediction Cepstral Coefficient
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform