[1]Jasmine J. Karagthala
[2]Dr. Vrushank Shah

Analyzing the recent advancements for Speech Recognition using Machine Learning: A Systematic Literature Analysis

Journal of Electrical Systems

Abstract: - Speech Recognition (SR) technology, empowered by Machine Learning (ML) and Deep Learning (DL), has revolutionized human-computer interaction by enabling accurate conversion of spoken language into text or commands. This advancement has found widespread application in consumer electronics, enhancing user engagement through voice commands on devices like smart speakers and smartphones. SR also improves accessibility in sectors such as healthcare and automotive industries, supporting tasks like medical transcription and in-car navigation. This bibliometric study employing the PRISMA model investigates the utilization of Speech Recognition and Machine Learning. Initially, a search yielded 170 results, which were then refined through filters to exclude non-article documents, reducing the collection to 105 articles. Subsequently, inaccessible papers were further removed, resulting in a final list of 35 papers included in the analysis. Ongoing research focuses on enhancing SR's capability to handle diverse accents and languages using advanced deep learning models like RNNs and transformers, aiming to create more intuitive and personalized user experiences through integration with Natural Language Processing (NLP). ML-driven SR continues to drive innovation in AI, promising enhanced efficiency and communication across various domains.

Keywords: Speech Recognition, Feature extraction, machine learning, deep learning, Optimization

I. INTRODUCTION

Effective communication is an essential and fundamental component of how people act. Human beings use natural languages for interacting with one other, both orally and in written form. The written language is a representation of human interaction in the form of spoken speech [1][2]. This will culminate in the advancement of voice recognition system technology, enabling machines to comprehend the semantic content that people speech. As a result, communication among computers and humans has grown more convenient, and technology systems have become more intuitive [3]. "Speech recognition" (SR) technology is crucial in current apps since it converts words spoken into writing or instructions, improving user engagement with gadgets and devices [4]. The applicability of this technology extends across multiple areas, including as telecommuting, medical services, automobiles, and electronic goods [5]. Furthermore, SR technology is utilized in electronic devices to empower AI assistants such as Alexa and Siri. SR technology is used to turn spoken words into written scripts [6]. An "Automated Speech Recognition" (ASR) system is used to identify the language being said, and then the parts of the user's speech are translated into corresponding units of text in the appropriate human language. This allows for the seamless operation of smart home and gadgets, providing a simple user interface [7].

A study on the characteristics of typical data retrieval requests has demonstrated that the discourse interface should adequately cater to needs such as speaker autonomy, smoothly delivered speech, and unregulated vocabulary [8]. The currently known SR systems are able to offer an approximation of these demands. In addition, these criteria necessitate a multifaceted strategy involving fields such as data processing, linguistic and pattern identification [9]. To facilitate the synchronization of these methods and to remain receptive to future advancements in these domains, a voice recognition framework has been implemented and will also be showcased [10].

Advanced abilities have been created to identify and understand handwriting and speech. "Machine learning" (ML) has emerged as a result of utilizing software programs that enable machines to acquire knowledge and skills based on previous experiences [11]. ML focuses on the self-acquiring and adjustment of knowledge from data, without requiring human involvement. In the field of speech recognition, many ML approach have been applied for acoustic modeling. In particular, Markov models have been utilized for prediction tasks, where they learn to

[1] [1]PhD Scholar Department of E& C, [2]Head of Department of E& C,

[2]Indus University, Ahmedabad, India [2]Indus University, Ahmedabad, India

Email Id: [1]jdaftary@gmail.com [2]ec.hod@indusuni.ac.in

associate acoustic signals with phonemes or syllables [12]. After receiving instruction, the model has the ability to forecast text based on fresh audio inputs as shown in figure 1. After processing approaches such as modeling languages or beam searching can enhance these predictions to achieve greater accuracy. ML is widely recognized as the most prevalent technology for SR and is regarded as important paradigm change in the field [13]. Statistical modeling approaches, specifically "Hidden Markov Models" (HMMs), have significantly contributed to the progress of the area [14]. In recent years, the discipline of voice processing has undergone a significant transformation through the integration of advanced methods, such as "Deep Learning" (DL).

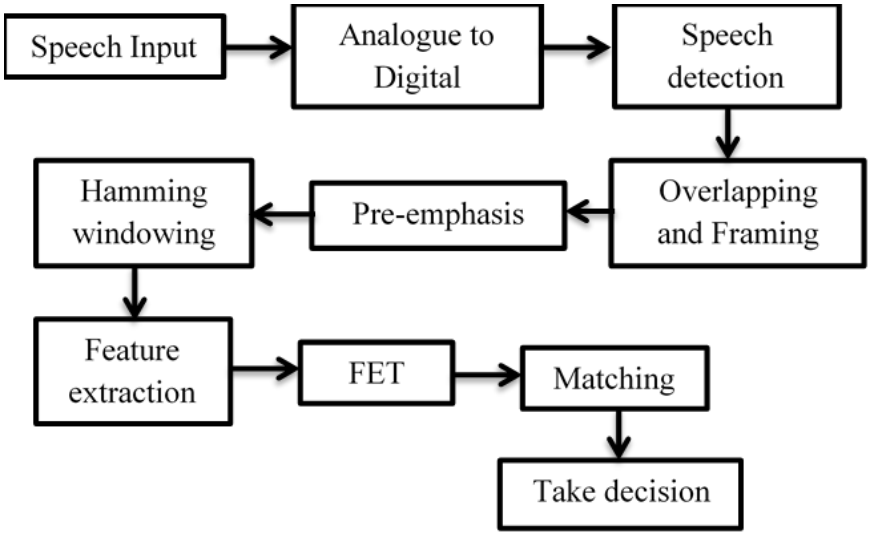

Fig 1. Speech Recognition process [15]

The figure illustrate SR process begins with voice signals being converted from analog to digital. Sections of speech are detected in the digital stream and divided into discrete frames, often overlapping, to capture temporal information. The signal is pre-emphasized to boost higher frequencies and signal-to-noise ratio. Frequency spectrum analysis is used to extract useful information from speech signals. The collected features are then matched to existing trends or systems to recognize and recognize what is said. Finally, the results are analyzed to determine the uttered phrases or words. This rigorous approach ensures voice signal recognition and comprehension [16].

Literature research has identified certain mediator characteristics, such as ML and DL that contribute to the development of SR stated by Vashisht et al. [17] and Padmanabhan et al. [18]. ML algorithms made significant contributions to the detection of discriminatory speech and the study of social media content in general M. A. Al-Garadi [19]. Morgan [20]conducted an analysis on the topic of SR using discriminatively learned feed-forward systems. Hemdal & Hughes [21] proposed that spoken language is composed of a limited set of distinct phonetic units. They identified and labeled these units through their analysis of sounds produced during speech.

Several inquiries were conducted in the field of SR. Mehrish et al. [22] accompanied a comprehensive analysis of advanced DL models, such as RNNs, converters, adherents, and diffusing models, and discussed their applications in speech-processing tasks. Furthermore, multiple end-to-end deep learning frameworks have been proposed by M. Sarma et al. and Fayek et al. [23] [24], which were capable of simultaneously performing feature extraction and categorization. In Schmidt et al. [25], thoroughly and critically analyzed the areas associated with the automated detection of harsh and critical language in the field of language processing. Mishaim et al. [26] and Y. Zhang [27] conducted an investigation of various techniques for extracting features, cutting-edge models for classification, and how these factors affected an Automatic SR system. DL methods primarily relied on data, so a wide range of voice datasets available on the internet were frequently utilized.

SR encounters various challenges, such as managing the diversity in speech patterns caused by pronunciations, ambient noise, and speaker attributes. Durable models must possess strong generalization capabilities across a wide range of datasets and be able to adjust to various speech styles. Effective models must possess strong

generalization capabilities across a wide range of datasets and be able to adjust to various speaking styles. The aim of SR is to create machines that can effectively process spoken information and generate appropriate replies depending on that input. These machine mimic human conversation and overcome these challenges in order to minimize this barrier through this research, which was created and implemented to create systems that can effectively assist individuals in sharing information by using voice commands to operate a computer. The study highlights the several research contributions mostly concentrate on enhancing system architectures such as "Deep Neural Networks (DNN)", refining ways of feature selection, and devising novel methods to effectively process uninterrupted voice streams. Tackling these obstacles and progressing in research leads to the development of more precise, dependable, and adaptable SR systems that can be used in a wider range of languages and settings. Careful preparation and well-defined SCOPUS database search tactics yielded considerable research contributions. There have been notable advancements in ML techniques for SR. The literature study reveals patterns in real-time data processing, multilingual identification systems, and dynamic and assistive technology.

The paper provides a comprehensive analysis of the literature, with a specific emphasis on SR recognition using machine learning. The report offers a thorough examination of the existing literature and presents a methodology that utilizes rigorous research techniques such as PRISMA to investigate and evaluate SR. The material also includes the findings and analysis derived from the investigations.

II. RESEARCH METHODOLOGY

The current work is aimed at making an extensive literature review for the research articles published within 2015-24, means the review is marked for previous 10 years research articles on speck recognition using machine learning. The methodological flow is being presented in the form of PRIMA model which shows the screening process for the entire literature analysis/systematic review. The initial phase of the systematic review is planning phase which includes making questions for the requirement of the review and also presenting a defined search strategy so that the defined database can be surfed. On the basis of search strategy, the literature search is conducted for 10 years for SCOPUS database for defined keywords as speech recognition and machine learning. On the basis of the search strategy and following all exclusion and inclusion criteria's the final data is retrieved. Next and last phase of the process is analysis phase where on the basis of clustering the literature is analyzed and grouped for some common categories.

A. Planning Stage

Speech recognition is being considered one of the major application areas for AI machines where the human to machine interaction is the module which necessitates the concept and at the same time for many of the other application areas. Just because of the growing need for speech recognition, research in the field is taking a sharp growing track. Machine learning for speech recognition is still an unsolved domain which needs huge exploration to understand the applicability and solutions provided. The current is aimed at providing the subject or topic overview with the help of proper citations of recent publications and of good priority journals. The work explores the applicability of machine learning for speech recognition and shows the growing technical advancement in the field with some bibliometric trends for the domain. The presented work will certainly mark a positive presence to better understand the current, past and making valid prediction for the future timelines on the basis analysed records for the literature. The presented literature analysis marks the resolutions for the research questions provided below:

- What is the year wise trend of publication for speech recognition using machine learning?
- What is the country wise trend of publication for speech recognition using machine learning
- What are the various techniques used for the speech recognition?
- Which datasets are best suited for maximum accuracy for speech recognition?
- Which feature extraction technique best fit the process to present efficient framework for speech recognition?

The current research has used the key terms as speech recognition and machine learning for searching process and have counted 10 years' time line. The literature search is conducted for SCOPUS database and keyword matching

is considered for work retrieval, reason being the keywords, machine learning and speech recognition are frequent and popular terms and have shown huge retrieved results.

B. Review Stage

The current stage is to mark the inclusion and exclusion criteria for literature search, the defined criteria are better shown as under:

- Search is made for 2015-24 for SCOPUS database,
- Excluding the papers showing document type other than article,
- Excluding the research papers whose publication stage is not final,
- Excluding the research papers which are not open access nature.

C. Analysis

In the final stage the remaining list of research articles were accessed and explored thoroughly to present a literature analysis for the fulfilment of the defined research questions. The figure below shows the PRISMA model for the literature search and screening.

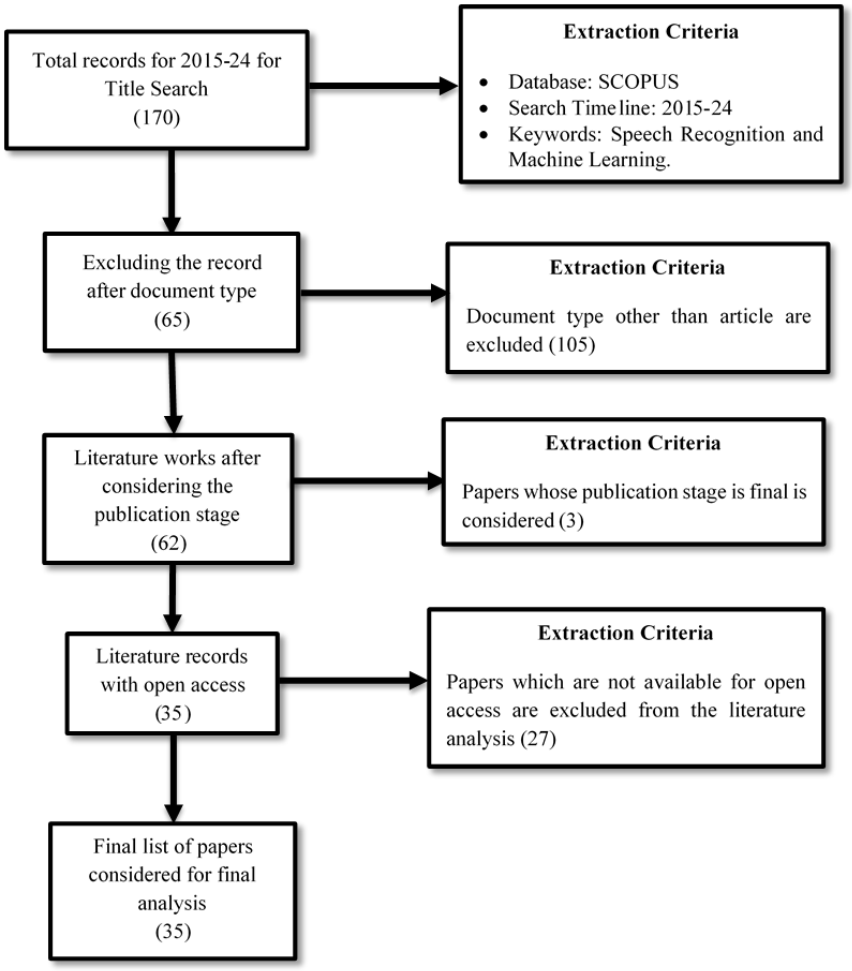

Fig 2. PRISMA Model for the literature analysis

Fig. 2 presents a systematic procedure for choosing publications to include in a final assessment, using particular criteria for retrieval. The initial phase entails conducting a search using certain keywords relevant to both SR and ML in the abstracts of articles that appeared in the SCOPUS repository between the years 2015 and 2024. The initial search produces a grand total of 170 results. This leads to an initial set of records. The starting filtering

process eliminates non-article documents, resulting in a reduction of the collection by 65 records, leaving a total of 105 articles. The subsequent stage focuses exclusively on papers that have reached the final publishing stage, further reducing the pool by 3 records. As a result, papers that are not accessible for free access are eliminated, resulting in the removal of an extra 27 files. After going through the procedure, a final list of 35 papers is generated. These papers match all the criteria and are being selected for the final evaluation.

III. Literature Analysis

As per the facts analyzed in the demonstrated literature various key points and dimensions have been seen towards Speech Recognition and Machine Learning. Dataset used, key techniques for feature extraction/selection, techniques used for preprocessing, featured techniques for optimization and recognition, also the literature have marked the channel towards the key consideration of the Deep learning and Machine learning techniques. Based on above shown variables or factors the retrieved literature is analyzed. The literature gathered from the defined themes is analyzed and assessed to provide a comprehensive analysis and a relevant recommendation for future study on the Impact of Speech Recognition and Machine Learning. The determinants of speech recognition in this study are perceived to be the quality of services and convenience of use. These predictors have been supported and examined by several literature studies, which are assessed below through thematic analysis and evaluation. The literature is also well explored based on some bibliometric data in graphical format and textual as well. Documents searched for subject base, year wise research publication, country wise publication trends for subject oriented research.

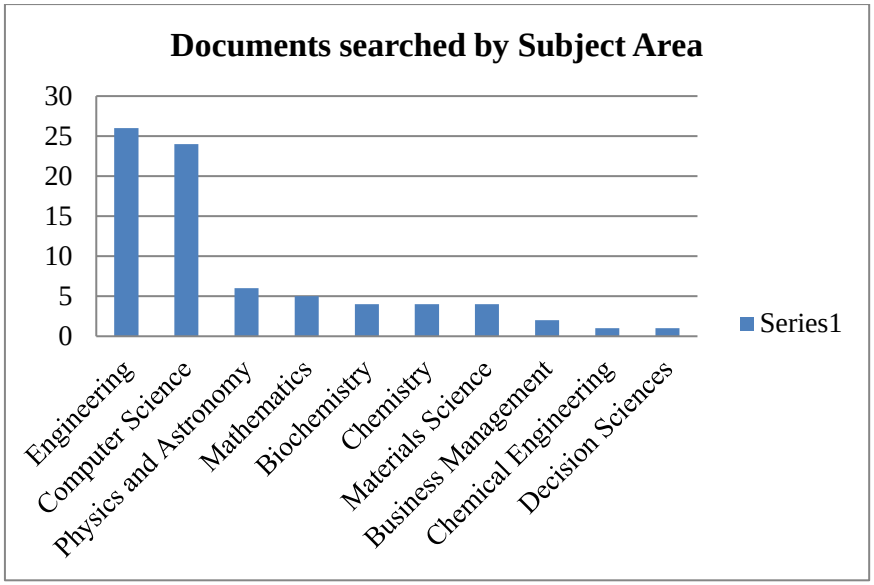

Fig 3. Subject based research paper retrieved

The figure presented is a bar chart titled "Documents searched by Subject Area," showcasing the distribution of document searches across various academic disciplines. The chart is laid out with the subject areas aligned along the horizontal axis and the number of documents searched represented on the vertical axis. The scale on the vertical axis ranges from 0 to 30. Each bar in the chart corresponds to a specific subject area. Engineering and Computer Science emerge as the most prominent fields, each with over 25 documents searched. These two disciplines significantly lead the chart, indicating a high volume of searches in these areas. Physics and Astronomy follow, albeit with a lower count of approximately 10 searches. Mathematics, Biochemistry, and Chemistry each have around 5 searches, showing moderate interest. Materials Science, Business Management, and Chemical Engineering have even fewer documents searched, with counts ranging from 2 to 4. Finally, Decision Sciences is the least searched subject area, with barely 1 or 2 documents.

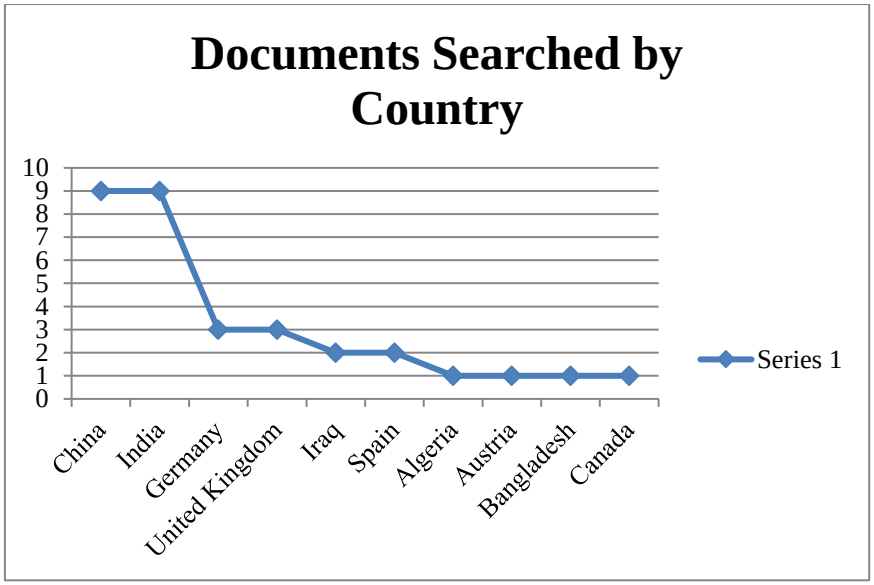

Fig 4. Country based research paper retrieved

The graph titled "Documents Searched by Country" shows the number of documents searched across ten different countries. China and India have the highest number of documents searched, with both countries reaching a peak of 9 documents each. Following these two countries, there is a notable decrease, with Germany having 3 documents searched. The United Kingdom comes next with 2 documents searched. The remaining countries, including Iraq, Spain, Algeria, Austria, Bangladesh, and Canada, each have 1 document searched. This distribution indicates a significant concentration of document searches in China and India, while other countries show much lower search activity. The sharp decline from Germany to the rest of the countries highlights the disparity in document searches among these nations.

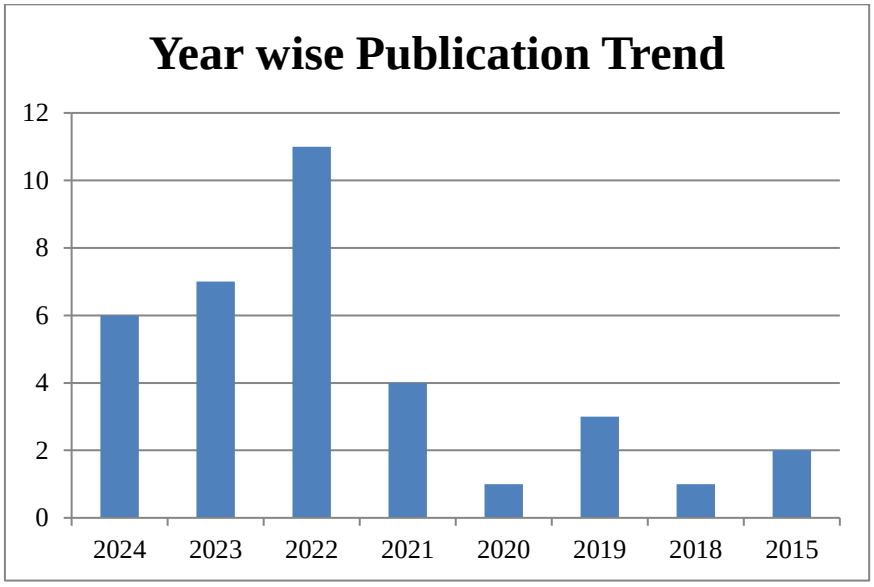

Fig 5. Year wise research paper retrieved.

The figure is a bar chart titled "Year wise Publication Trend," which depicts the number of publications per year over a range of years. The horizontal axis lists the years, which include 2024, 2023, 2022, 2021, 2020, 2019, 2018, and 2015. The vertical axis represents the number of publications, with values ranging from 0 to 12. In the year 2024, there are 6 publications. This number remains consistent with 6 publications in 2023 as well. The year 2022 shows a significant increase, peaking at 11 publications, the highest in the given timeframe. The number of publications drops to 5 in 2021. There is a sharp decline in 2020, with only 1 publication. In 2019, the number of publications slightly increases to 4. The year 2018 sees a further drop with 1 publication, and 2015 matches this with 1 publication as well. The chart illustrates a clear upward trend in publications starting from 2018, peaking in

2022, followed by a noticeable decline. This trend may indicate a surge in research activity or data availability in the years leading up to 2022, followed by a decrease in subsequent years. The chart makes it easy to see how the number of publications has changed over time, showing times when less or more study papers were published.

A. Literature depiction datasets for speech recognition and machine learning

The review of the literature covers a wide spectrum of works aiming on many facets of speech and emotion recognition applied using different datasets and approaches. Table 1 summarizes various datasets available in speech recognition domain. In Costantini et al. (2022) [28] investigated emotional voice detection using the Emofilm database. Jiao et al. (2021) [29] created a spoken English teaching system emphasizing useful applications in language learning environments. Emphasizing emotional speech and databases tailored to Telugu, Raghu et al. (2022) [30] investigated utilizing many datasets including RAVDESS, IIITH-TEMD, and DETL. Oukas et al. (2024) [31] most likely for Arabic speech analysis, made contributions using the ArabAlg set. Emo-DB and IEMOCAP datasets were used by Guo et al. (2019) [32] to examine emotional speech recognition across several emotional expressions. For voice recognition applications in Zhang et al. (2015) [33], demonstrated TI46 speech corpus collection. Badr et al. (2021) [34] used the aGender dataset presumably for gender categorization in voice analysis. Yang et al. (2023) [35] using an English text classification corpus dataset concentrated on textual content classification instead of speech. German voice recognition skills were evaluated by Schädler et al. (2018) [36] using the German matrix sentence recognition test. Ramyasree et al. (2023) [37] used RAVDESS, SAVEE, EMOVO, and URDU databases for a thorough investigation spanning several languages and emotional contexts. Muthusamy et al. (2015) [38] examined emotional speech across three distinct datasets. For their evaluation of speech recognition models, Wojnar et al. (2024) [39] creatively used 141 randomly chosen YouTube videos. For thorough emotional recognition experiments in Koti et al. (2024) [40]asserted the Berlin Database of Emotional Speech (EMO-DB). Possibly included gathering and evaluating acoustic data for different uses in Wang et al. (2022) [41], an acoustic emission (AE) monitoring instrument. Garcia-Cueste et al. (2024) [42] provided EmoMatchSpanishDB and EmoSpanishDB for emotional recognition in Spanish speech. Dealing with particular difficulties experienced by users of such devices, Ameen et al. (2023) [43] investigated electrolarynx device speech for phoneme categorization. Investigating developments in emotional speech recognition systems, Daqrouq et al. (2024) [44] went back over the EMO-DB dataset. Mashhadi et al. (2023) [45] enhanced a small dataset of speech emotion identification with extra audio recordings with an aim to improve model resilience and accuracy. In order to investigate machine learning methods in emotional detection, Doğdu et al. (2022) [46] concentrated on the Berlin Database of Emotional Speech Using datasets of ten female and ten male speech samples, Prasetio et al. (2022) [47] investigated stress and objective speech categorization. the use of many datasets, tools, and other approaches to raise the accuracy, dependability, and applicability of speech recognition systems throughout several fields and real-world situations.

Authors name	Dataset	Benefits	Applications
Costantini et al., (2022) [28]	Emofilm database	Designed for emotion recognition research, it provides a controlled environment for studying emotional speech variability across different stimuli and subjects.	Emotion recognition in films or video content for content analysis, sentiment analysis in movie reviews, or emotion-aware recommendation systems.
Jiao et al., (2021) [29]	Spoken English teaching system	Tailored for educational purposes, this dataset supports the development of spoken English teaching systems by providing authentic speech samples for language learning and pronunciation practice.	Speech emotion recognition for enhancing spoken English teaching systems, assessing emotional engagement or stress levels in learners, or providing feedback on pronunciation and emotion in language learning apps.
Raghu et al., (2022) [30]	RAVDESS, IIITH-TEMD, and DETL	Combines multiple databases to offer a diverse range of emotional speech data in	Multi-modal emotion RAVDESS, IIITH-TEMD, and DETL for applications

		different languages (English and Telugu), supporting cross-cultural emotion recognition and language-specific emotion studies.	such as human-computer interaction, virtual reality experiences, or sentiment analysis in customer service.
Oukas et al., (2024) [31]	ArabAlg dataset	Focuses on Arabic speech commands, facilitating the development of Arabic-specific speech recognition systems and applications.	Emotion recognition specifically tailored for Arabic speakers, applications in Arabic language sentiment analysis, emotion-aware content recommendation in Arabic media, or personalized digital assistants.
Guo et al., (2019) [32]	Emo-DB and IEMOCAP	Offers emotional speech data from actors and naturalistic interactions, respectively, supporting emotion recognition research in controlled and natural settings.	Emo-DB, IEMOCAP for applications in call center analytics, emotion-aware dialogue systems, or emotion-driven virtual agents.
Zhang et al., (2015) [33]	TI46 speech corpus dataset	Provides a corpus of speech data in Mandarin Chinese, valuable for speech recognition and synthesis research specifically in Mandarin.	Speech emotion recognition using the TI46 corpus, potentially for improving emotional interaction in assistive technologies or virtual agents.
Badr et al., (2021) [34]	aGender dataset	Focuses on gender-related speech variations, aiding studies on gender-specific speech characteristics and recognition algorithms.	Gender classification from speech for applications in voice-based authentication systems or gender-specific marketing analytics.
Yang et al., (2023) [35]	English text classification corpus dataset	Provides textual data for sentiment analysis and emotion classification tasks, enhancing research in natural language processing and affective computing.	Text classification for sentiment analysis or topic detection in English-language content, enhancing content recommendation systems.
Schädler et al., (2018) [36]	German matrix sentence recognition test	Designed for assessing speech recognition in noise, it contributes to the development of robust speech recognition systems in challenging acoustic environments.	Assessing speech recognition and processing algorithms in German, crucial for improving speech-to-text systems and hearing aid technologies
Ramyasree et al., (2023) [37]	RAVDESS, SAVEE, EMOVO, and URDU	Combines multiple datasets to enrich emotion recognition research with a wide variety of emotional expressions and linguistic backgrounds.	Multi-dataset emotion recognition in speech for cross-cultural emotional analysis or emotion-driven content creation in entertainment.
Muthusamy et al., (2015) [38]	Three different emotional speech databases	Provides access to three distinct databases for emotion recognition research,	Emotional speech synthesis or emotion recognition in diverse emotional contexts,

		facilitating comparative studies and algorithm development.	benefiting virtual agents or therapeutic applications.
Wojnar et al., (2024) [39]	YouTube videos (141 randomly selected)	Offers a diverse collection of spontaneous speech from YouTube, reflecting real-world variability and supporting research on unscripted speech recognition and understanding.	Emotion analysis in real-world video content for understanding viewer engagement or sentiment monitoring in social media.
Koti et al., (2024) [40]	Berlin Database of Emotional Speech (EMO-DB)	Established benchmark for emotional speech analysis, aiding research in emotional speech recognition and synthesis.	Emotional speech recognition for enhancing human-computer interaction, improving customer service automation, or emotion-aware virtual assistants.
Wang et al., (2022) [41]	Acoustic emission (AE) monitoring tool	Provides data for monitoring acoustic emissions, supporting applications in health monitoring and industrial quality control.	AE monitoring for industrial applications such as fault detection in machinery or quality control in manufacturing processes.
Garcia-Cuesta E et al., (2024) [42]	EmoSpanishDB and EmoMatchSpanishDB	Focuses on emotional speech in Spanish, supporting studies on cultural and language-specific emotion recognition algorithms.	Emotional speech recognition in Spanish for personalized educational tools, sentiment analysis in Spanish media, or emotion-driven marketing strategies.
Ameen Z.J et al., (2023) [43]	electrolarynx device	Offers speech data from users of electrolarynx devices, supporting research in assistive technology and speech rehabilitation.	Speech enhancement and communication aids for laryngectomy patients, improving speech intelligibility and quality of life.
Daqrouq K et al., (2024) [44]	EMO-DB dataset	A subset or variation of the Berlin Database of Emotional Speech (EMO-DB), providing additional data for emotion recognition studies.	Emotion recognition in speech using EMO-DB for applications in affective computing, personalized mental health monitoring, or emotion-aware virtual environments.
Mashhadi et al., (2023) [45]	Limited dataset of speech emotion recognition augmented with additional audio files	A subset or variation of the Berlin Database of Emotional Speech (EMO-DB), providing additional data for emotion recognition studies.	Enhanced emotion recognition using augmented datasets, potentially improving accuracy and robustness of emotion-sensitive applications.
Doğdu C et al., (2022) [46]	Berlin Database of Emotional Speech	Widely used dataset for emotional speech analysis, facilitating benchmarking and comparison of emotion recognition algorithms.	Emotional speech analysis for understanding affective states in human-computer interaction, improving emotional intelligence in AI systems.

Prasetio B et al., (2022) [47]	10 female and 10 male discourse datasets	Provides gender-specific discourse data, supporting studies on gender differences in speech patterns and recognition tasks.	Gender-specific discourse analysis for studying gender differences in communication styles or enhancing natural language processing models with gender-aware features.

Table 1. Various speech recognition datasets available.

B. Literature depiction SR techniques for speech recognition and machine learning

Investigating approaches for identifying emotions from speech signals, Costantini et al. (2022) [28] and Guo et al. (2019) [32] both concentrate on Speech Emotion Recognition (SER). Jiao et al. (2021) [29] evaluate speech processing algorithms or evaluate MATLAB. Working on Arabic Speech Command Recognition (ASCR), a system designed to interpret spoken commands in Arabic, Oukas et al. (2024) [31]. Wang et al. (2021) [48] advance the area of Silent Speech recognition Systems (SSRS), which enable speech detection without audible speech production. a method wherein semantic comprehension is obtained by analyzing speech streams, Semantic Stream Processing is investigated in Rajarajeswari et al. (2021) [49] Including voice-controlled automation by means of speech recognition, Peña-Cáceres et al. (2022) [50] create a spoken command-based Smart Home. The Random Subspace Method (RSM) is used by Pagidirayi et al. (2022) [51] to combine many classifiers or features thereby improving speech recognition accuracy. Zhang et al., (2015) [33], investigate isolated word recognition using bioinspired digital LMS. Focusing on Automatic Speech Recognition (ASR) technologies, Schädler et al. (2018) [36] and Mendiratta et al. (2019) [38] address both the evolution of speech recognition systems and approaches for maximising feature extraction and selection.

Ravenscroft et al. (2022) [52] use graphene-based strain gauge sensors to capture physical parameters associated to speech production. The use of prosodic features, wavelet transformations, and spectral features in speech recognition tasks is investigated by Ramyasree et al., (2023) [37]. Using the features to improve speech recognition accuracy and robustness, Olatinwo et al. (2023) [53] use machine learning and deep learning methods. Xu et al. (2019) [54]investigate graph embedding-based subspace learning and extreme learning machines (ELM) with an aim to improve feature representation and classification in audio recognition tasks. Using emotional features collected from fundamental frequency and resonance degree, an SVM classifier developed by Zhang et al. (2022) [27] mapped to high-level feature space using KCCA. In Muthusamy et al., (2015) [55], feature enhancement using Gaussian mixture model (GMM), classification using Extreme Learning Machine (ELM), and k-Nearest Neighbor (kNN). Emphasizing developments and comparative assessments in speech recognition technologies, Wojnar et al. (2024) [39] evaluate state-of-the-art ASR machine learning models. Dwivedi et al. (2022) [56] explore speech analytics and synthesis with an emphasis on converting speech signals into usable data and synthesizing speech from digital data sources. Koti et al. (2024) [40], Wang et al. (2022) [41], and Garcia-Cuesta et al. (2024) [42] use various machine learning models including Neural Networks (NN), Convolutional Neural Networks (CNN), Recurrent Neural Network (RNN), Random Forest, XGI For phoneme recognition, Ameen Z.J. et al., (2023) [43] combine CNN, RNN, ANN, Random Forest, XGBoost, and LSTM among other machine learning models.

Daqrouq K et al., (2024) [44] SER, specifically modeling emotions using speech signals. Deep Neural Network (DNN) is used for English voice feature recognition in Chen et al., (2022) [57]. One-dimensional Convolutional Neural Network (conv1D) and Random Forest (RF) models for speech emotion classification in Mashhadi et al.,(2023) [45]. Saini A et al., (2023) [58] Multinomial Naive Bayes (MNB), Logistic Regression (LR), and Linear Support Vector Machine (LSVM) for emotion classification in speech. Stacked Convolutional Autoencoder (SCAE) for feature extraction and Sequence of Multi-Label Extreme Learning Machines (MLELM) for classification in Hossain P et al., (2022) [59]. Doğdu C et al., (2022) [46] Multilayer Perceptron Neural Network (MLP), J48 Decision Tree (DT), Support Vector Machine with Sequential Minimal Optimization (SMO), Random Forest (RF), k-Nearest Neighbor (KNN), Simple Logistic Regression (LOG), and Multinomial Logistic Regression (MLR) to compare the performance of machine-learning algorithms. Neural Arrange, k-Nearest Neighbour, Bolster Vector Machine, combination of NN-k-NN and combination of NN-SVM are used for classification in Prasetio B et al., (2022) [47]. Du S et al., (2020) [60] developing an English speech recognition system based on machine learning. Table 2 below summarizes various techniques along with their advantages.

Authors	SR techniques	Benefits	Outcomes
Costantini et al., (2022) [28]	SER	Recognition of emotions from voice signals helps affective computing, human-computer interaction, and psychological investigations.	Cross-gender tests are harder than two-language activities, demonstrating bigger variations in male and female emotions than across languages.
Jiao et al., (2021) [29]	MATLAB	MATLAB is a powerful numerical computation and algorithm development environment with signal processing, machine learning, and data analysis capabilities.	The solution performed well, met design objectives, and met user demand.
Oukas et al., (2024) [31]	ASCR	Recognizes Arabic instructions to improve usability in smart assistants, automated systems, and voice-controlled products for Arabic speakers.	The dataset is useful for training and assessing Arabic-speaking Machine Learning and Deep Learning ASCR systems.
Wang et al., (2021) [48]	SSRS	SSRS permits quiet communication utilizing electromyography (EMG) signals or other non-auditory ways in healthcare, assistive technology, and covert communication.	Experiments demonstrate that the silent speech recognition system (SSRS) can withstand ~45% face deformation using tattoo-like electrodes and recognize up to 110 words from daily vocabularies with an average accuracy of 92.64% using small-sample machine learning.
Rajarajeswari et al., (2021) [49]	Semantic stream process	Uses semantic information from voice streams to improve natural language processing (NLP), automated transcription, and intelligent system comprehension and context.	The research employs machine learning to improve call identification systems by reliably recognizing emotional states and voice tones from speech, which is essential for increasing human-computer interaction in numerous applications.
Peña-Cáceres et al., (2022) [50]	Smart home based on spoken commands	Introduces spoken commands for smart home device control, making them more accessible and usable for disabled and hands-free users.	The categorization model predicts categories, subcategories, and actions from sentences with 82.99%, 76.19%, and 90.28% accuracy, respectively.
Pagidirayi et al., (2022) [51]	RSM	RSM randomizes feature subsets to increase classification accuracy, reduce overfitting, and strengthen speech and pattern recognition ML models.	Compared to SVM and bagged trees, the proposed technique evaluates accuracy, PPV rate, and training time on male and female speech signals.
Guo et al., (2019) [32]	SER	SER methods can identify and classify emotional states from speech signals in healthcare, customer service, and affective computing.	The findings also demonstrate that merging auditory-based and spectrogram-based variables might increase performance over traditional techniques.

Zhang et al., (2015) [33]	LSM	LSM is a bioinspired computational model that can interpret temporal patterns and dynamic inputs, making speech recognition systems more resilient and efficient.	The results show that the proposed digital LSM outperforms all other reported recognizers, including LSM and neural network-based ones, in isolated word recognition using the TI46 speech corpus.
Schädler et al., (2018) [36]	ASR and SRTs	ASR systems automate speech-to-text translation, making transcription easier and faster, whereas SRTs quantify auditory sensitivity for hearing tests and device calibration.	A competing talker condition clearly showed one limitation of current ASR technology, as the empirical performance with SRTs lower than −20 dB could not be predicted.
Ravenscroft et al., (2022) [52]	Graphene-based strain gauge sensor	Uses graphene-based sensors to detect speech-related vibrations or physiological signals, improving speech recognition and allowing new wearable technology and healthcare monitoring applications.	The findings show that such sensors can anticipate spoken words. The word dataset has 55% word accuracy and the motion dataset 85%.
Ramyasree et al., (2023) [37]	Prosodic features, Wavelet, and Spectral features	These methods improve speech processing tasks like emotion detection and speaker identification by capturing intonation, rhythm, and frequency content.	It employs nonlinear feature selection via Fisher Criterion and achieves high recognition rates of around 79.66% to 95.78% across various emotion databases (RAVDESS, SAVEE, EMOVO, URDU) using Support Vector Machine and Decision Tree classifiers.
Mendiratta et al., (2019) [38]	Pre-processing, feature extraction and feature selection and classification	Optimizes speech signal analysis with extensive preprocessing and feature extraction and selection to improve machine learning models in numerous applications.	Results from SVM, ANN with Cuckoo search methods, and ANN with back propagation classifier techniques.
Olatinwo Det al., (2023) [53]	ML and DL algorithms	For speech recognition and natural language comprehension, uses powerful ML and DL algorithms to process and analyze speech signals with high accuracy and scalability.	The experiments showed that one of the suggested models outperformed the previous model with 98% accuracy.
Xu et al., (2019) [54]	Graph embedding-based subspace learning and ELM	Inserting graph structures into subspace learning and using ELM for speech signal analysis improves feature representation and classification.	Multiple paralinguistic corpora demonstrate that the methods boost performance.
Zhang et al., (2022)	SVM	SVMs can classify speech signal data based on extracted characteristics with excellent accuracy and generalization.	The testing findings reveal that this approach recognizes more than 90% of emotions and 95% of

			anger, fear, happiness, and sorrow.
Muthusamy et al., (2015) [55]	GMM and ELM and kNN	These speech signal feature representation, classification, and pattern recognition models are flexible and adaptable for many applications.	The suggested approaches considerably improved speech emotion recognition compared to published studies.
Wojnar et al., (2024) [39]	State-of-the-art ASR machine learning models	Improves speech recognition accuracy and efficiency with cutting-edge ASR models and real-time machine learning.	The findings show that YouTube may be used to test voice recognition algorithms for resilience, accuracy, and adaptation to different languages and acoustic situations.
Dwivedi A.K et al., (2022) [56]	Speech analytics and speech synthesis	Allows personal assistants, interactive systems, and automated speech-based services to comprehend and synthesize speech signals using speech analytics.	It emphasizes AI's contribution to improving service delivery and operational efficiency across various domains, facilitated by advancements in technology and machine learning platforms.
Koti et al., (2024) [40]	ML algorithms	Optimizes speech processing using ML algorithms to improve speech recognition, emotion detection, and other applications.	Achieving an accuracy of 74.33% on the EMO-DB dataset, this method demonstrates computational efficiency and promising results for applications like virtual assistants and emotional analysis in various fields, including psychotherapy and call center analytics.
Wang et al., (2022) [41]	ASR technology	ASR technology accurately and reliably converts speech to text, making transcription and spoken language interpretation easier and faster.	The findings reveal that AIEFPC can forecast the harmful status of the coal sample at any time using the MFCC of the 40 ms AE segment with >85% accuracy.
Garcia-Cuesta E et al., (2024) [42]	ML	Develops strong speech signal processing models using ML to provide scalable solutions for speech recognition and associated applications.	The findings for EmoMatchSpanishDB are better than EmoSpanishDB, hence we suggest using the emotional database technique.
Ameen Z.J et al., (2023) [43]	ML models including CNN, RNN, ANN, Random Forest, XGBoost, and LSTM	Used a variety of ML models to capture temporal and hierarchical relationships in speech signals to achieve state-of-the-art speech processing performance.	The findings reveal that the ANN machine learning approach beat other methods with 75% accuracy, 77% precision, and 21.85% phoneme error rate (PER).
Daqrouq K et al., (2024) [44]	Speech signals	Signal processing allows extensive analysis and feature extraction for machine learning applications from raw speech signals.	KNN and SVM classifiers accurately distinguished melancholy from other emotions.
Chen et al., (2022) [57]	DNN	DNNs excel in speech recognition and natural language	Experimental validation confirms the effectiveness of the algorithm

		processing by learning complicated speech patterns and representations.	in improving speech recognition performance amidst challenging acoustic conditions.
Mashhadi et al., (2023) [45]	conv1D and RF	Conv1D learns spatial characteristics and RF selects them, providing a strong framework for speech emotion identification and classification with better accuracy.	RF with feature selection demonstrated greater average accuracy (69%) than conv1D, the best precision for fear (72%), and the highest recall for calm (84%).
Saini A et al., (2023) [58]	MNB, LR, and LSVM	These speech classifiers are simple, efficient, and interpretable, making them dependable in diverse applications.	The experiment showed that LSVM outperformed the other two machine learning methods.
Hossain P et al., (2022) [59]	SCAE and a Sequence of MLELM	SCAE unsupervised feature learning and MLELM multi-label classification achieve solid performance in complicated speech recognition and classification challenges.	By adding age class, categorization accuracy rises from 85% to 95%. The model can recognize synthetic Bangla speech labels with 95% accuracy and applies to English speech datasets.
Doğdu C et al., (2022) [46]	MLP, J48 DT, Support Vector Machine with SMO, RF, KNN, Simple LOG, and MLR	Compares ML systems for speech recognition and emotion classification, revealing their strengths and weaknesses.	The emobase feature set performs best, suggesting clinical diagnosis, intervention, and HCI applications.
Prasetio B et al., (2022) [47]	KNN, Bolster Vector Machine, combination of NN-k-NN and combination of NN-SVM	Uses several classification methods to improve voice recognition and classification accuracy and flexibility, meeting varying application needs and data characteristics.	Combining Neural Organize with Back Vector Machine yields the best classification technique with 85% stretch recognition rate.
Du S et al.,(2020) [60]	ML-based approach	A holistic ML approach to speech recognition integrates numerous approaches and models for strong performance across speech processing tasks.	The research shows that the proposed algorithm has certain performance and can provide theoretical reference for related research. © 2020, CAD Solutions, LLC

Table 2. ML-based Speech recognition techniques.

C. Literature depiction Optimization techniques for speech recognition and machine learning

Costantini et al. (2022) [28] employ Kononenko's discretization and correlation-based feature selection methods to enhance their SER system. Jiao et al. (2021) [29] utilize a deep belief network (DBN) combined with a support vector machine (SVM) model, leveraging the hierarchical feature learning capabilities of DBN and the robust classification power of SVM for their spoken English teaching systems. Wang et al. (2021) [48] use general machine-learning algorithms to optimize their silent speech recognition system, Rajarajeswari et al. (2021) [49]explore Multiple Classifier Systems (MCSs), to enhance overall classification performance, potentially leveraging ensemble learning methods. Peña-Cáceres et al. (2022) [50] apply various machine learning techniques

to develop a smart home system based on spoken commands, Pagidirayi et al. (2022) [51] utilize Subspace-kNN (S-kNN), a variant of k-nearest neighbor (kNN) classification that operates on subspace projections of the feature space. Guo et al. (2019) [32] employ kernel extreme learning machine (KELM), an enhancement of extreme learning machines (ELM). Zhang et al. (2015) [33] use a bioinspired spike-based learning algorithm, mimicking biological neural networks to optimize their speech recognition systems.

Badr et al. (2021) [34] utilize CatBoost, a gradient boosting framework that excels in handling categorical features and producing accurate predictions. Yang et al. (2023) [35] employ a k-nearest neighbor graph approach, which uses graph-based representations to model relationships between instances. Ramyasree et al. (2023) [37] focus on selecting significant features based on nonlinear statistics utilize the most informative features from speech signals for emotion recognition. Mendiratta et al. (2019) [38] utilize the cuckoo search algorithm combined with backpropagation, optimizing neural network training parameters to improve the accuracy of their ASR systems. Olatinwo et al. (2023) [53] explore various optimization strategies and regularization techniques, likely focusing on improving the generalization ability and robustness. Zhang et al. (2022) [27] employ genetic algorithms, which use evolutionary principles to optimize feature selection and model parameters. Koti et al. (2024) [40] utilize Extreme Machine Learning (EML) with the Gaussian Mixture Model (GMM) algorithm, to enhance the performance of their speech emotion recognition system. Wang et al. (2022) [41] and Garcia-Cuesta et al. (2024) [42] utilize general machine learning (ML) techniques, applying a range of algorithms to optimize their systems for various speech recognition applications.

Chen et al. (2022) [57] employ a two-way search method for feature selection and a quick correlation filter for feature reduction, focusing on efficiently selecting and reducing features to improve the performance of their speech recognition models. Mashhadi et al. (2023) [45] utilize random forest (RF)-based feature selection, to identify and prioritize the most informative features for emotion classification. Du et al. (2020) [60] employ sub-word modeling, a technique aimed at improving the robustness and efficiency of language models. These studies illustrate the diverse array of optimization techniques applied in speech recognition research, each tailored to enhance specific aspects of system performance such as accuracy, efficiency, robustness. Table 3 presents optimization techniques used by authors with their speech recognition systems.

Authors name	**Optimization technique**	**Benefits**	**Drawbacks**
Costantini et al., (2022) [28]	Kononenko's discretization and correlation-based feature selection	Transforming continuous data into categorical ones and picking relevant variables based on target variable correlation improves classification accuracy.	May oversimplify data relationships and miss complex patterns, potentially leading to information loss in feature selection.
.Jiao et al., (2021) [29]	DBN- SVM model	DBNs improve pattern recognition by capturing hierarchical data representations, whereas SVMs classify high-dimensional voice data robustly.	DBNs can be computationally intensive and require large amounts of data for effective training, making them less practical for small datasets.
Wang et al., (2021) [48]	ML algorithm	Generic machine-learning principles provide speech recognition and processing flexibility and adaptation.	Generally, ML algorithms can overfit with noisy data or struggle with interpretability.
Rajarajeswari et al., (2021) [49]	MCSs	Integrates multiple classifiers to enhance overall system accuracy by leveraging diverse classification strategies.	Integration of multiple classifiers can increase computational complexity and require careful tuning to avoid conflicting outputs.
Peña-Cáceres et al., (2022) [50]	ML techniques	Optimizes speech recognition system performance using a variety of machine learning methods.	Generally, ML algorithms can overfit with noisy data or struggle with interpretability.

Pagidirayi et al., (2022) [51]	Subspace-kNN (S-kNN),	Reduces computational complexity and enhances classification accuracy by operating within a reduced feature subspace.	Sensitive to the choice of k and the quality of the subspace selection, which can affect classification accuracy and robustness.
Guo et al., (2019) [32]	KELM	Achieves high-speed learning and prediction through efficient kernel methods, suitable for real-time speech processing applications.	Limited interpretability due to the black-box nature of kernel methods, making it challenging to understand how decisions are made.
Zhang et al., (2015) [33]	Bioinspired spike-based learning algorithm	Mimics biological neural processing, offering potential for efficient and parallelized computations in speech recognition tasks.	High complexity in implementation and tuning, limited by current hardware capabilities and requiring specialized knowledge for optimization.
Badr et al., (2021) [34]	CatBoost machine	Optimizes gradient boosting models for categorical data, enhancing accuracy and speed in speech command recognition.	Can be slower to train compared to other boosting algorithms due to its handling of categorical variables, affecting real-time applications.
Yang et al., (2023) [35]	K-nearest neighbor graph	Facilitates similarity-based classification, useful for identifying patterns in speech data based on nearest neighbors in a graph representation.	Sensitive to the choice of k and distance metric, potentially leading to poor performance in high-dimensional spaces or with noisy data.
Ramyasree et al., (2023) [37]	Selecting significant features based on nonlinear statistics	Selects significant features using nonlinear statistical methods, improving the robustness and relevance of features in speech recognition models.	May not capture all relevant features in complex datasets, potentially leading to suboptimal feature selection and reduced model performance.
Mendiratta et al., (2019) [38]	Cuckoo search algorithm and backpropagation	Integrates optimization techniques with neural network training to enhance learning and generalization capabilities in speech analytics.	Cuckoo search's stochastic nature can lead to convergence issues or suboptimal solutions, requiring careful parameter tuning.
Olatinwo Det al., (2023) [53]	Optimization strategies and regularization techniques	Implements strategies to prevent overfitting and improve generalization performance in machine learning models applied to speech data.	Complexity in selecting appropriate strategies and techniques can lead to overfitting or underfitting in model training.
Zhang et al., (2022) [27]	Genetic algorithm	Optimizes feature selection and model parameters by simulating evolution, achieving robust solutions for speech recognition tasks.	Convergence to suboptimal solutions and computational overhead due to its iterative nature, especially with large search spaces.
Koti et al., (2024) [40]	EML with the GMM algorithm	Combines the strengths of extreme learning machines with Gaussian mixture models to capture complex patterns and distributions in speech data.	Sensitivity to the quality of feature selection and the assumption of Gaussian distributions, which may not always reflect the true data distribution.

Wang et al., (2022) [41]	ML	Leverages generic ML techniques to enhance speech recognition accuracy and efficiency through adaptive learning from data.	Generally, ML methods can be sensitive to noisy data, require large datasets for training, and may lack interpretability.
Garcia-Cuesta E et al., (2024) [42]	ML	Applies various ML models (CNN, RNN, ANN, etc.) tailored to the specifics of speech data, optimizing performance in different recognition tasks.	---
Chen et al., (2022) [57]	Two-way search method for feature selection and quick correlation filter for feature reduction	Efficiently selects relevant features and reduces computational load using a combination of search and filter methods tailored for speech recognition.	Risk of discarding potentially relevant features, especially in datasets with complex interdependencies, and may not scale well with large datasets.
Mashhadi et al., (2023) [45]	RF-based feature selection	Uses random forest algorithms to select important features, improving classification accuracy and reducing overfitting in speech emotion recognition.	Depending on hyperparameter tuning, can be prone to overfitting or underfitting, requiring careful validation to achieve optimal results.
Du S et al., (2020) [60]	Sub-word modeling	Modeling sub-word units improves accuracy and efficiency in large-vocabulary speech recognition systems affected by sparsity and robustness.	Accuracy highly dependent on the quality and depth of linguistic analysis, potentially leading to misinterpretation or misclassification of subtle linguistic cues.

Table 3. Summary of Optimization techniques in speech recognition systems.

D. Literature Analysis for Feature Extraction techniques used for speech recognition and machine learning

Costantini et al. (2022) [28] utilize RASTA, fundamental frequency (F0), Mel-frequency cepstral coefficients (MFCC), and spectral energy focus on capturing both spectral and temporal characteristics of speech signals. Raghu et al. (2022) [30] employ Mel-frequency cepstral coefficients (MFCC) along with their first and second derivatives (ΔMFCC and ΔΔMFCC) used in speech processing to capture the spectral dynamics and changes over time. Wang et al. (2021) [48] research tattoo-like electronics imperceptibly, applying novel techniques likely related to wearable or embedded technologies. Rajarajeswari et al. (2021) [49] apply the XGBoost model, which is typically a machine learning algorithm rather than a feature extraction technique. Peña-Cáceres et al. (2022) [50] use CountVectorizer, a technique commonly used in natural language processing (NLP) for converting text data into numerical features. Pagidirayi et al. (2022) [51], Koti et al. (2024) [40] and Wang et al. (2022) [41]employ Mel Frequency Cepstral Coefficients (MFCC), a standard technique in speech processing known for its effectiveness in capturing speech features in a compact representation suitable for machine learning algorithms. Guo et al. (2019) [32] utilize auditory-based empirical features and spectrogram-based statistical features capturing both physiological aspects of human auditory perception and statistical properties of speech signals, Badr et al. (2021) [34] use the quantile technique, which involves segmenting data based on statistical distributions. Ramyasree et al. (2023) [37] employ MFCC, Formants, and Long-Term Average Spectrum (LTAS), formants capture the resonant frequencies of the vocal tract. Zhang et al. (2022) [27]focus on fundamental frequency (F0) and resonance degree, which are fundamental aspects of speech production and vocal tract resonance. Daqrouq et al. (2024) [44]combine discrete wavelet transform (DWT) with linear prediction coding (LPC), to capture both time-domain and frequency-domain characteristics of speech signals. Chen et al. (2022) [57] employ complex cepstrum domain techniques, which involve analyzing the logarithm of the spectrum to extract features.

Mashhadi et al. (2023) [45] utilize a comprehensive set of features including MFCC, chromogram, Mel-scale spectrogram, spectral contrast feature, Tonnetz representation, and zero-crossing rate to enhance the discrimination power for emotion recognition. Doğdu et al. (2022) [46] use OpenSMILE feature sets such as IS-09, emobase, GeMAPS, and eGeMAPS. OpenSMILE is a widely used tool for extracting a variety of low-level descriptors from speech signals. Prasetio et al. (2022) [47] focus on speech energy and frequency, which are fundamental features in speech signal analysis. These techniques highlight a diverse range of feature extraction techniques applied in speech recognition and emotion recognition research, each tailored to capture specific aspects of speech signals that are critical for accurately identifying emotional states, enhancing system performance, and supporting various applications in human-computer interaction and beyond.

Authors name	**Technique used for feature extraction**	**Benefits**	**Drawbacks**
Costantini et al., (2022) [28]	RASTA, F0, MFCC and spectral energy	RASTA filtering decreases noise, F0 catches pitch, MFCC represents spectral characteristics, and spectral energy offers amplitude details, improving speech recognition accuracy.	RASTA may filter out relevant information along with noise, impacting feature quality. F0 extraction can be sensitive to noise and speaker variations.
Raghu et al., (2022) [30]	MFCC and a combination of features (MFCC + ΔMFCC + ΔΔMFCC).	MFCCs collect spectrum characteristics, while their deltas (ΔMFCC) and double deltas (ΔΔMFCC) capture dynamics, enhancing robustness and temporal information for emotion recognition.	Combining MFCC derivatives may increase computational complexity and require careful tuning to avoid overfitting.
Wang et al., (2021) [48]	Tattoo-like electronics imperceptibly	integrating imperceptible, lightweight electronics with wearable speech recognition devices for comfort and accuracy in continuous monitoring.	May lack robustness in noisy environments and require specialized equipment for implementation.
Rajarajeswari et al., (2021) [49]	XGBoost model	For speech analytics with complicated feature spaces, uses gradient boosting to automatically choose important features and enhance classification accuracy.	Limited in feature interpretation due to its black-box nature, making it challenging to understand which features contribute most to classification.
Peña-Cáceres et al., (2022) [50]	CountVectorizer	Converts text data into numerical feature vectors, enabling efficient analysis and modeling of spoken commands in smart home systems.	May oversimplify linguistic features and fail to capture nuanced semantic information, especially in complex texts.
Pagidirayi et al., (2022) [51]	MFCC	Accurate speech recognition across settings requires comprehensive speech feature representation that captures frequency and amplitude variations.	Sensitive to noise and speaker variability, requiring careful preprocessing and parameter tuning for optimal performance.
Guo et al., (2019) [32]	Auditory-based empirical features and Spectrogram-based statistical features	Integrates auditory perception models and statistical spectrogram features to enhance speech recognition robustness and accuracy in noisy conditions.	Empirical features may oversimplify auditory perception, while spectrogram-based features may be computationally

			intensive and sensitive to noise.
Badr et al., (2021) [34]	Quantile technique	Utilizes quantile normalization to preprocess speech data, ensuring uniform feature distributions and improving model performance in diverse acoustic environments.	May not capture all statistical properties of the data distribution, potentially missing important features in skewed datasets.
Ramyasree et al., (2023) [37]	MFCC, Formants, and LTAS	Includes formants and LTAS alongside MFCCs to capture vocal tract characteristics and long-term spectral patterns, enhancing emotion and speech recognition accuracy.	Formant estimation can be sensitive to noise and speaker characteristics, impacting accuracy in real-world environments.
Zhang et al., (2022) [27]	Fundamental frequency and resonance degree	Focuses on fundamental frequency (F0) and resonance properties to capture vocal timbre variations, improving speaker identification and emotion recognition.	Extraction can be challenging in noisy environments and may require robust preprocessing techniques.
Koti et al., (2024) and [40]. Wang et al., (2022) [41]	MFCCs	Extracts MFCCs to capture speech dynamics and spectral features, optimizing performance in automatic speech recognition systems.	Sensitive to noise and speaker variability, requiring careful preprocessing and parameter tuning for optimal performance.
Daqrouq K et al., (2024) [44]	DWT with LPC	Integrates wavelet analysis for time-frequency localization and LPC for spectral envelope estimation, enhancing speech feature extraction accuracy in noisy environments.	Complex implementation and tuning required for optimal wavelet selection and LPC coefficients, impacting computational efficiency.
Chen et al., (2022) [57]	Complex cepstrum domain techniques	Applies complex cepstrum analysis to capture phase information alongside magnitude, improving discrimination of speech patterns in noisy and reverberant conditions.	High computational cost and complexity in interpretation, requiring specialized knowledge for effective implementation.
Mashhadi et al., (2023) [45]	Mel-frequency cepstral coefficients, chromogram, Mel-scale spectrogram, spectral contrast feature, Tonnetz representation and zero-crossing rate	Integrates diverse feature sets to capture broad spectral and temporal characteristics, enhancing robustness and accuracy in speech emotion recognition.	High-dimensional feature space may lead to overfitting, requiring careful regularization and feature selection.
Doğdu C et al., (2022) [46]	OpenSMILE feature sets (i.e., IS-09, emobase, GeMAPS and eGeMAPS)	Utilizes IS-09, emobase, GeMAPS, and eGeMAPS feature sets to capture a wide range of speech attributes, improving emotion recognition and natural language understanding tasks.	May include redundant or irrelevant features, requiring additional feature selection techniques for optimal performance.

Prasetio B et al., (2022) [47]	Speech energy and frequency	Detects essential acoustic signals like energy and frequency to improve speech segmentation and recognition.	Oversimplifies speech characteristics, potentially missing subtle variations crucial for accurate classification

Table 4.

IV. DISCUSSION

Recent advancements in the systematic analysis on Speech Recognition (SR) reflects a significant evolution driven by advancements in Machine Learning (ML) and Deep Learning (DL) techniques, as highlighted by several key studies. Vashisht et al. [17] and Padmanabhan et al. [18] highlight the importance of ML and DL in improving SR capabilities, especially when it comes to talking recognition systems. ML algorithms for SER, specifically looking at cross-linguistic and cross-gender elements. On the Emofilm database, MLP classifier outperformed SVM and Naïve Bayes, achieving an accuracy of over 90% for tasks involving a single language and over 80% for tasks involving several languages. Significant gender-based expression differences were brought to light by the fact that cross-gender identification was more difficult [28]. Another study used a MATLAB-based system that utilized a DBN-SVM model to improve spoken English learning for Chinese learners. There has been encouraging progress in real-time responsiveness and pronunciation accuracy thanks to its ability to categorize mistakes, rate quality, and provide feedback [29]. The idea put out by Hemdal and Hughes [21] is that, at its core, language is made up of separate phonetic units that may be discovered by careful study of individual speech sounds. This realization has played a crucial role in improving SR systems' ability to understand and properly transcribe spoken language. In their thorough examination of sophisticated DL structures including RNNs, Transformers, and diffusion models, Mehrish et al. [22] emphasize the essential significance that DL models have played. More resilient and adaptable systems are now possible because to these models' enhanced SR feature extraction and classification capabilities.

End-to-end DL frameworks, which were suggested by researchers such as M. Sarma et al. and Fayek et al. [23][24], have also simplified the SR workflow by combining the steps of feature extraction and classification, which makes it more efficient and effective. Machine learning is used to combine speech detection with smart house technologies. It creates a model to understand spoken orders as actions by humans. This model is very good at identifying smart home job categories, subcategories, and actions, which is a step forward for natural language processing in AI-driven smart homes [50]. Mishaim et al. [26] and Y. Zhang [33] go into more detail about feature extraction methods and cutting-edge classification models. They show how these improvements have made Automatic SR systems more useful. Importantly, DL's reliance on large datasets that are easily accessible on the internet has made it easier to test and compare SR models with each other, which has led to ongoing improvements in accuracy and usability across a wide range of speech recognition tasks. In a fourth study, Speech Emotion Recognition (SER) is improved by using MFCC and the Random Subspace Method (RSM) along with kNN (S-kNN). This method improves the accuracy of mood detection, especially for male and female speaking patterns, showing better performance measures than standard models like SVM [51]. This thorough study makes a collective contribution to the continuous development of SR technology, demonstrating its increasing importance in diverse industries, including social media analysis and real-time speech processing applications.

V. CONCLUSION

In conclusion, Speech Recognition (SR) technology has revolutionized human-computer interaction, enabling accurate conversion of spoken language into text or commands through ML and DL advancements. SR systems are integral across sectors like telecommunications, healthcare, and consumer electronics, facilitating hands-free operations and enhancing user engagement with devices. Ongoing research aims to enhance real-time processing, expand vocabulary recognition, and improve natural language understanding, promising more intuitive and effective voice interfaces. SR continues to drive innovation in AI, shaping a future where seamless human-machine interaction is ubiquitous and transformative. The current research has used the key terms as speech recognition and machine learning for searching process and have counted 10 years' time line. The literature search is conducted for SCOPUS database and keyword matching is considered for work retrieval, reason being the keywords, machine learning and speech recognition are frequent and popular terms and have shown huge retrieved results. This bibliometric study examines the deployment of Speech Recognition and Machine Learning. The initial search

produces a grand total of 170 results. This leads to an initial set of records. The starting filtering process eliminates non-article documents, resulting in a reduction of the collection by 65 records, leaving a total of 105 articles. As a result, papers that are not accessible for free access are eliminated, resulting in the removal of an extra 27 files. After going through the procedure, a final list of 35 papers is generated.

Speech Recognition (SR) driven by Machine Learning (ML) has changed the way people and computers talk to each other by making exchanges more accurate and reliable. It's widely used in consumer goods for things like voice requests on smart speakers and smartphones, which makes things easier and faster. SR also makes things easier to access in the education, healthcare, and auto industries by helping with things like medical transcription and guidance in the car. In the future, improvements will be made to SR's ability to handle different accents and languages by using more advanced deep learning models, such as RNNs and transformers. This will make user experiences more natural and customized by combining SR with NLP. ML-driven SR offers more improvements in ease, communication, and effectiveness across many fields.

REFERENCES

[1] P. Sanderson, "Cognitive work analysis and the analysis, design, and evaluation of human-computer interactive systems," in Proceedings 1998 Australasian Computer Human Interaction Conference. OzCHI'98 (Cat. No.98EX234), 1998, pp. 220–227. doi: 10.1109/OZCHI.1998.732218.

[2] U. S. Tiwary and T. Siddiqui, Natural Language Processing and Information Retrieval. USA: Oxford University Press, Inc., 2008.

[3] M. Gavrilescu, "Improved automatic speech recognition system using sparse decomposition by basis pursuit with deep rectifier neural networks and compressed sensing recomposition of speech signals," in 2014 10th International Conference on Communications (COMM), 2014, pp. 1–6. doi: 10.1109/ICComm.2014.6866711.

[4] S. Mendiratta, N. Turk, and D. Bansal, "A robust isolated automatic speech recognition system using machine learning techniques," Int J Innov Technol Exploring Eng, vol. 8, pp. 2278–3075, 2019.

[5] M. A. M. Hasan and S. Ahmad, "PredSucc-site: Lysine succinylation sites prediction in proteins by using support vector machine and resolving data imbalance issue," Int J Comput Appl, vol. 182, no. 15, pp. 8–13, 2018.

[6] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," J Acoust Soc Am, vol. 24, no. 6, pp. 637–642, 1952.

[7] M. A. M. Abushariah, R. N. Ainon, R. Zainuddin, M. Elshafei, and O. O. Khalifa, "Natural speaker-independent Arabic speech recognition system based on Hidden Markov Models using Sphinx tools," in International Conference on Computer and Communication Engineering (ICCCE'10), 2010, pp. 1–6. doi: 10.1109/ICCCE.2010.5556829.

[8] H. Isyanto, A. S. Arifin, and M. Suryanegara, "Performance of Smart Personal Assistant Applications Based on Speech Recognition Technology using IoT-based Voice Commands," in 2020 International Conference on Information and Communication Technology Convergence (ICTC), 2020, pp. 640–645. doi: 10.1109/ICTC49870.2020.9289160.

[9] A. Izbassarova, A. Duisembay, and A. P. James, "Speech Recognition Application Using Deep Learning Neural Network," in Deep Learning Classifiers with Memristive Networks: Theory and Applications, A. P. James, Ed., Cham: Springer International Publishing, 2020, pp. 69–79. doi: 10.1007/978-3-030-14524-8_5.

[10] S. Das, "Speech recognition technique: A review," Int J Eng Res Appl, vol. 2, no. 3, pp. 2071–2087, 2012.

[11] G. Salton, "Introduction to modern information retrieval," McGraw-Hill, 1983.

[12] C. Gerber, "A general approach to speech recognition," in Proceedings of the Final Workshop on Multimedia Information Retrieval (Miro'95), BCS Learning & Development, 1995.

[13] X. Cai and W. Li, "Ranking through clustering: An integrated approach to multi-document summarization," IEEE Trans Audio Speech Lang Process, vol. 21, no. 7, pp. 1424–1433, 2013.

[14] J. M. Baker et al., "Developments and directions in speech recognition and understanding, Part 1 [DSP Education]," IEEE Signal Process Mag, vol. 26, no. 3, pp. 75–80, 2009.

[15] V. Vashisht, A. Pandey, and S. Yadav, "Speech Recognition using Machine Learning," IEIE Transactions on Smart Processing & Computing, vol. 10, pp. 233–239, Jun. 2021, doi: 10.5573/IEIESPC.2021.10.3.233.

[16] A. A. Varghese, J. P. Cherian, and J. J. Kizhakkethottam, "Overview on emotion recognition system," in 2015 international conference on soft-computing and networks security (ICSNS), IEEE, 2015, pp. 1–5.

[17] V. Vashisht, A. K. Pandey, and S. P. Yadav, "Speech recognition using machine learning," IEIE Transactions on Smart Processing & Computing, vol. 10, no. 3, pp. 233–239, 2021.

[18] J. Padmanabhan and M. J. Johnson Premkumar, "Machine Learning in Automatic Speech Recognition: A Survey," IETE Technical Review, vol. 32, no. 4, pp. 240–251, Jul. 2015, doi: 10.1080/02564602.2015.1010611.

[19] M. A. Al-Garadi et al., "Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges," IEEE Access, vol. 7, pp. 70701–70718, 2019, doi: 10.1109/ACCESS.2019.2918354.

[20] N. Morgan, "Deep and Wide: Multiple Layers in Automatic Speech Recognition," IEEE Trans Audio Speech Lang Process, vol. 20, no. 1, pp. 7–13, 2012, doi: 10.1109/TASL.2011.2116010.

[21] J. F. Hemdal and G. W. Hughes, "A feature based computer recognition program for the modeling of vowel perception," Models for the Perception of Speech and Visual Form, Wathen-Dunn, W. Ed. MIT Press, Cambridge, MA, 1967.

[22] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," Information Fusion, vol. 99, p. 101869, 2023.

[23] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion Identification from Raw Speech Signals Using DNNs.," in Interspeech, 2018, pp. 3097–3101.

[24] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," Neural Networks, vol. 92, pp. 60–68, 2017.

[25] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proceedings of the fifth international workshop on natural language processing for social media, 2017, pp. 1–10.

[26] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," Multimed Tools Appl, vol. 80, pp. 9411–9457, 2021.

[27] Y. Zhang and G. Srivastava, "Speech emotion recognition method in educational scene based on machine learning," EAI Endorsed Transactions on Scalable Information Systems, vol. 9, no. 5, pp. e9–e9, 2022.

[28] G. Costantini, E. Parada-Cabaleiro, D. Casali, and V. Cesarini, "The emotion probe: On the universality of cross-linguistic and cross-gender speech emotion recognition via machine learning," Sensors, vol. 22, no. 7, p. 2461, 2022.

[29] F. Jiao, J. Song, X. Zhao, P. Zhao, and R. Wang, "A spoken English teaching system based on speech recognition and machine learning," International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 14, pp. 68–82, 2021.

[30] K. Raghu and M. Sadanandam, "Emotion recognition from speech utterances with hybrid spectral features using machine learning algorithms," Traitement du Signal, vol. 39, no. 2, p. 603, 2022.

[31] N. OUKAS, S. HABOUSSI, C. MAIZA, and N. BENSLIMANE, "ArabAlg: A new Dataset for Arabic Speech Commands Recognition for Machine Learning Purposes," International Journal of Computing and Digital Systems, vol. 15, no. 1, pp. 989–1005, 2024.

[32] L. Guo, L. Wang, J. Dang, Z. Liu, and H. Guan, "Exploration of Complementary Features for Speech Emotion Recognition Based on Kernel Extreme Learning Machine," IEEE Access, vol. 7, pp. 75798–75809, 2019, doi: 10.1109/ACCESS.2019.2921390.

[33] Y. Zhang, P. Li, Y. Jin, and Y. Choe, "A Digital Liquid State Machine With Biologically Inspired Learning and Its Application to Speech Recognition," IEEE Trans Neural Netw Learn Syst, vol. 26, no. 11, pp. 2635–2649, 2015, doi: 10.1109/TNNLS.2015.2388544.

[34] A. A. Badr and A. K. Abdul-Hassan, "CatBoost Machine Learning Based Feature Selection for Age and Gender Recognition in Short Speech Utterances.," International Journal of Intelligent Engineering & Systems, vol. 14, no. 3, 2021.

[35] L. Yang, "Unsupervised machine learning and image recognition model application in English part-of-speech feature learning under the open platform environment," Soft comput, vol. 27, no. 14, pp. 10013–10023, 2023.

[36] M. R. Schädler, A. Warzybok, and B. Kollmeier, "Objective prediction of hearing aid benefit across listener groups using machine learning: Speech recognition performance with binaural noise-reduction algorithms," Trends Hear, vol. 22, p. 2331216518768954, 2018.

[37] K. Ramyasree and C. S. Kumar, "Multi-Attribute Feature Extraction and Selection for Emotion Recognition from Speech Through Machine Learning," Traitement du Signal, vol. 40, no. 1, p. 265, 2023.

[38] S. Mendiratta, N. Turk, and D. Bansal, "A robust isolated automatic speech recognition system using machine learning techniques," Int J Innov Technol Exploring Eng, vol. 8, pp. 2278–3075, 2019.

[39] T. Wojnar, J. Hryszko, and A. Roman, "Mi-Go: tool which uses YouTube as data source for evaluating general-purpose speech recognition machine learning models," EURASIP J Audio Speech Music Process, vol. 2024, no. 1, p. 24, 2024.

[40] V. M. Koti, K. Murthy, M. Suganya, M. S. Sarma, G. V. S. S. S. Kumar, and N. Balamurugan, "Speech Emotion Recognition using Extreme Machine Learning," EAI Endorsed Transactions on Internet of Things, vol. 10, 2024.

[41] H. Wang, Z. Li, D. Song, X. He, and M. Khan, "Applying Machine Learning and Automatic Speech Recognition for Intelligent Evaluation of Coal Failure Probability under Uniaxial Compression," Minerals, vol. 12, no. 12, p. 1548, 2022.

[42] E. Garcia-Cuesta, A. B. Salvador, and D. G. Pãez, "EmoMatchSpanishDB: study of speech emotion recognition machine learning models in a new Spanish elicited database," Multimed Tools Appl, vol. 83, no. 5, pp. 13093–13112, 2024.

[43] Z. J. M. Ameen and A. A. Kadhim, "Machine learning for Arabic phonemes recognition using electrolarynx speech," International Journal of Electrical and Computer Engineering, vol. 13, no. 1, p. 400, 2023.

[44] K. Daqrouq, A. Balamesh, O. Alrusaini, A. Alkhateeb, and A. S. Balamash, "Emotion Modeling in Speech Signals: Discrete Wavelet Transform and Machine Learning Tools for Emotion Recognition System," Applied Computational Intelligence and Soft Computing, vol. 2024, no. 1, p. 7184018, 2024.

[45] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest," PLoS One, vol. 18, no. 11, p. e0291500, 2023.

[46] C. Doğdu, T. Kessler, D. Schneider, M. Shadaydeh, and S. R. Schweinberger, "A comparison of machine learning algorithms and feature sets for automatic vocal emotion recognition in speech," Sensors, vol. 22, no. 19, p. 7561, 2022.

[47] B. H. Prasetio, E. R. Widasari, and F. A. Bachtiar, "A Study of Machine Learning Based Stressed Speech Recognition System.," International Journal of Intelligent Engineering & Systems, vol. 15, no. 4, 2022.

[48] Y. Wang et al., "All-weather, natural silent speech recognition via machine-learning-assisted tattoo-like electronics," npj Flexible Electronics, vol. 5, no. 1, p. 20, 2021.

[49] P. RAJARAJESWARI and O. ANWAR BÉG, "AN EXECUTABLE METHOD FOR AN INTELLIGENT SPEECH AND CALL RECOGNITION SYSTEM USING A MACHINE LEARNING-BASED APPROACH," J Mech Med Biol, vol. 21, no. 07, p. 2150055, Sep. 2021, doi: 10.1142/S021951942150055X.

[50] O. Peña-Cáceres, H. Silva-Marchan, M. Albert, and M. Gil, "Recognition of Human Actions through Speech or Voice Using Machine Learning Techniques," CMC-COMPUTERS MATERIALS & CONTINUA, vol. 77, no. 2, pp. 1873–1891, 2023.

[51] A. K. Pagidirayi and A. Bhuma, "Speech Emotion Recognition Using Machine Learning Techniques," Revue d'Intelligence Artificielle, vol. 36, no. 2, p. 271, 2022.

[52] D. Ravenscroft, I. Prattis, T. Kandukuri, Y. A. Samad, G. Mallia, and L. G. Occhipinti, "Machine learning methods for automatic silent speech recognition using a wearable graphene strain gauge sensor," Sensors, vol. 22, no. 1, p. 299, 2021.

[53] D. D. Olatinwo, A. Abu-Mahfouz, G. Hancke, and H. Myburgh, "IoT-enabled WBAN and machine learning for speech emotion recognition in patients," Sensors, vol. 23, no. 6, p. 2948, 2023.

[54] X. Xu, J. Deng, E. Coutinho, C. Wu, L. Zhao, and B. W. Schuller, "Connecting subspace learning and extreme learning machine in speech emotion recognition," IEEE Trans Multimedia, vol. 21, no. 3, pp. 795–808, 2018.

[55] H. Muthusamy, K. Polat, and S. Yaacob, "Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals," Math Probl Eng, vol. 2015, no. 1, p. 394083, 2015.

[56] A. K. Dwivedi, D. Virmani, A. Ramasamy, P. B. Acharjee, and M. Tiwari, "Modelling and analysis of artificial intelligence approaches in enhancing the speech recognition for effective multi-functional machine learning platform–A multi regression modelling approach," Journal of Engineering Research-ICMET Special Issue, pp. 4–6, 2022.

[57] Y. Chen and B. Martinuzzi, "Machine Learning for Predictive Analytics in the Improvement of English Speech Feature Recognition," Mobile Information Systems, vol. 2022, no. 1, p. 3541667, 2022.

[58] A. Saini, A. R. Khaparde, S. Kumari, S. Shamsher, J. Joteeswaran, and S. Kadry, "An investigation of machine learning techniques in speech emotion recognition," Indonesian Journal of Electrical Engineering and Computer Science, vol. 29, no. 2, pp. 875–882, 2023.

[59] P. S. Hossain, A. Chakrabarty, K. Kim, and M. J. Piran, "Multi-label extreme learning machine (MLELMs) for bangla regional speech recognition," Applied Sciences, vol. 12, no. 11, p. 5463, 2022.

[60] S. Du, "Optimization of speech recognition system of english education industry based on machine learning," Computer-Aided Des Appl, vol. 17, no. 1, pp. 124–136, 2019.

[61] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," Foundations and Trends® in Signal Processing, vol. 1, no. 3, pp. 195–304, 2008.

[62] C. Suneetha and R. Anitha, "Advancements in Speech-Based Emotion Recognition and PTSD Detection through Machine and Deep Learning Techniques: A Comprehensive Survey".