

A Comprehensive Review and Analysis of Speech Recognition Techniques

Hardik Jain¹, Yash Khapke² and Deepali Naik³

¹⁻³Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India
Email: hardik.jain20@pccoepune.org, yash.khapke20@pccoepune.org, deepali.naik@pccoepune.org

Abstract— Due to the increase in technology and the evolution of computers, speech recognition technology plays a vital role in our daily lives. Communicating with the computer through speech is more feasible than communicating through text. Speech recognition is the process of conversion of human speech to readable text to perform various operations. India has the most amazing diverse, multilingual community. Currently, there are 22 recognized languages with 12 scripts in the constitution. On the other hand, 1900+ dialects are divided into four language groups. This literature review aims to comprehensively analyse the automatic speech recognition system based on its history, types, application and problems in existing systems. Furthermore, the state-of-art speech recognition approaches are examined to lead a comparative study. As a result of an extensive literature review, future research directions are highlighted in terms of practical and relevant methods.

Index Terms— Speech Recognition, NLP, HMM, Hybrid Models.

I. INTRODUCTION

Communication is an essential part of human life. It has evolved from face-to-face conversations to various electronic devices and platforms like phones, WhatsApp, Telegram, Facebook, Twitter, and SMS services provided by different service providers. Communication can be done through text or speech, depending on the medium used.

Speech recognition has become increasingly important in the era of computers and advancing technology. It is the process of converting spoken language into readable text that can be used to perform various tasks[1]. In a developing country like India, many communities of people are illiterate and unable to write. For such people, speech-based communication is considerably more convenient than textual communication.

Speech recognition technology allows computers to interpret human language and convert it into a computer-understandable format to perform various operations [2]. Technology has advanced to the point where people can now give voice instructions to machines in regional languages, which is much more convenient and faster than typing commands manually. Speech recognition technology is especially beneficial for individuals who are not proficient in typing, have difficulty with language characters, have physical disabilities, or are illiterate [3]. It is a valuable alternative to traditional text-based input methods to improve the accessibility of communication for everyone.

Thus, a system is required to convert voice instructions into readable text [4] efficiently. Moreover, produce data in the text format to perform further operations such as recommendation and detection[5]. In the recent literature, Hidden Markov Model (HMM) and Artificial Neural Networks (ANN) are famous transcript- techniques [3]

utilised by various researchers. Keeping the base of recent literature in mind, authors have focused on HMM and ANN techniques for detailed study. This paper emphasises the importance of speech recognition technology and its potential to improve communication between individuals and machines. Also, the various applications are examined to evaluate the benefits of voice input over traditional text-based input methods[6]. Finally, the future directions are derived based on the analysis summary.

A. Proposed Model

Fig 1 demonstrates the overall architecture of HMM and ANN-based speech recognition system. Speech to Text conversion system involves various steps. Firstly, the audio is captured through a microphone. The audio is then converted to digital format using a feature extraction technique like Mel Frequency Cepstral Coefficient [7]. Next, the system uses an acoustic model to identify various sounds and speech patterns[8]. It involves breaking the digital audio into smaller units called phonemes. A language model is used to identify the most probable word or sentence. A pronunciation model differentiates between two similar words based on their pronunciation. The voice processing models are applied, like Hidden Markov Model and Artificial Neural Network Model, to generate a transcription of the audio input [9]. These algorithms analyse probabilities of phonemes and word sequences through the acoustic, language, and pronunciation models.

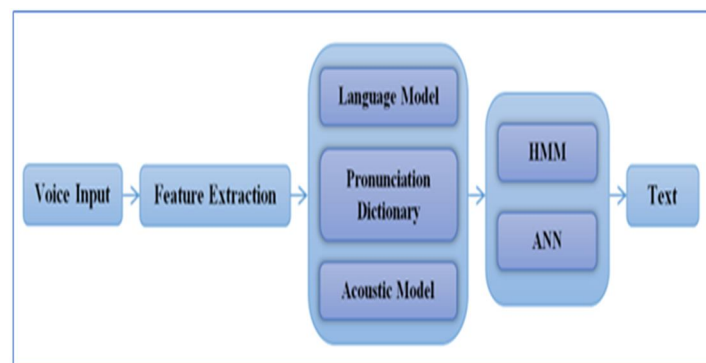


Figure 1. Overall Architecture of HMM and ANN-based Speech Recognition System

The rest of the paper is organised as follows. The methodology and framework exerted to extract scholarly articles to carry out analysis have been discussed in Section 2. Section 3.1 focuses on the history of the speech recognition system. Section 3.2 describes the feature extraction techniques of the proposed system. Section 3.3 illustrates various voice processing models. Section 4 represents a discussion of a thorough analysis of the literature review. Section 5 exchanges views on concluding remarks with future work.

II. METHODOLOGY FRAMEWORK FOR LITERATURE REVIEW

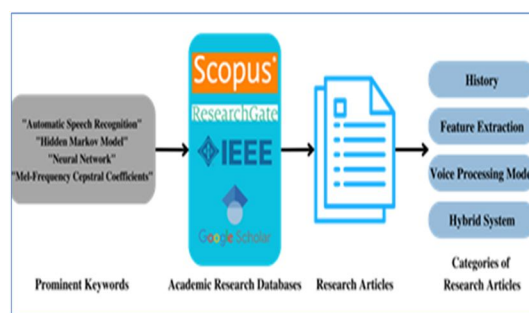


Figure 2. Methodology Framework for Literature Review

Fig 2 represents the methodology framework applied to extract relevant scholarly articles and conduct a literature review. Definite scholarly articles were collected from various repositories, including Scopus, ResearchGate, IEEE Xplore, ACM, and Google Scholar. We designed a search query using prominent keywords such as "Automatic Speech Recognition," "Hidden Markov Model," "Neural Network," and "Mel-Frequency Cepstral Coefficient,". On execution of the search query, many articles were extracted from all the repositories. For selecting relevant articles for study few criteria were considered, such as 1) use of the techniques, 2) publication year, 3) appearance of keywords 4) duplicate entries in the different repositories. After applying filters, 30 potential articles were

finally selected for comprehensive study and analysis. Also, various authorised web links were referred to gather more information about speech recognition systems.

The research papers and relevant material were studied carefully and classified into history, feature extraction, and voice processing algorithms. All these mentioned topics are discussed in detail in the upcoming sections.

III. LITERATURE REVIEW

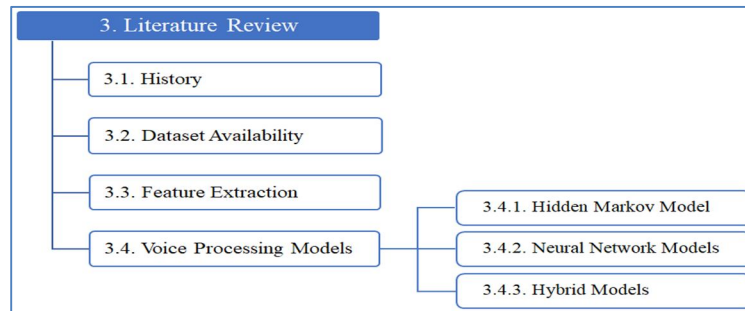


Figure 3. The Flow of Literature Review

Fig 3 represents the flow of the literature review. The literature review initiated with categorising extracted articles based on the evolution of techniques, feature extraction techniques, and voice processing algorithms referred in existing research on speech recognition systems. Specifically, the discussion is performed on immensely used voice processing algorithms such as Hidden Markov Model (HMM), Neural Networks (NN), and hybrid models that combine multiple algorithms with improving accuracy. At the end of the analysis, a detailed study of existing literature is summarised with the help of a list of limitations in studied literature, followed by the drafting of future directions for upcoming researchers in the stated research field.

A. Proposed Model

In 2008, O'Shaughnessy [4] studied Automatic Speech Recognition (ASR) from the perspective of pattern recognition. The authors examined topics related to ASR, including speech signals and data reduction. Moreover, they compared ASR with other Pattern Recognition techniques. The paper also discussed the accomplishments and failures of ASR and contributed towards an overview of future developments in the speech recognition research field. In 2016, Rajyaguru [8] explored various voice and speech recognition techniques. Researchers also discussed the need for data conversion followed by processing converted data. In addition to this author also discussed the techniques that could be used to solve specific problems or manage hardware interfaces that handle the external environment are explored. In the research [10], Shafran and Rose focused on developing an algorithm for robust speech detection that could be used to convert speech to text. The author's work involved non-parametric estimation of the noise spectrum of the surrounding noise using little statistics of smoothed Short Term Fourier Transform (STFT). The algorithm proved efficiency in fluctuating signal-to-noise ratios and had latency and a computation cost of as little as 2%. The results were calculated during experimentation based on two criteria: Speech Presence Probability Estimation (SPINE) and How Many Inputs Hear You (HMIHY). These two criteria vary depending on the speech style, types of background noise, and wavelength used in the experiment. The authors claimed that the proposed system is suitable for real-time application. Researchers G. and Sinha [11] aimed to address the challenges faced by end-to-end ASR systems over a certain amount of code-switching data. They explored sound similarities and context-dependent target-to-word[12] (T2W) conversion schemes to reduce the end-to-end (E2E) target set. The paper also discussed using physical features to enhance the modelling system in cases of fluctuating data. Furthermore, authors [11] performed experimentation over a new Hindi-English code-switching corpus, whereas the old target set system was processed for detailed examinations.

Table I describes the evolution of Voice-to-text in the studied literature. Voice-to-text technology has evolved significantly, with several key organizations contributing to its development. In the early days, Bell Labs [4] conducted early experiments in speech recognition systems, using acoustic and statistical language models to transcribe speech. Later, Dragon NaturallySpeaking used Hidden Markov Models (HMMs) for desktop computers, while mobile devices began using Automatic Speech Recognition (ASR) algorithms to transcribe speech into text. Apple's Siri, Google Now, Amazon's Alexa, and Samsung's Bixby use a combination of ASR, natural language processing (NLP), and machine learning to understand and respond to user requests. Today, [15] advanced voice-to-text technology uses deep learning algorithms such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to achieve high levels of accuracy in transcription and understanding of speech.

TABLE I. EVOLUTION OF VOICE TO TEXT

Ref	Year	Organization	Devices	Methodology
[4]	1950	Bell Labs	Early experiments in speech recognition systems	Used acoustic models and statistical language models to recognize and transcribe speech
[4]	1990	Dragon	Dragon NaturallySpeaking software for desktop computers	Used Hidden Markov Models (HMMs) to recognize and transcribe speech
[4]	2000s	Various	Voice-to-text technology integrated into mobile devices	Used Automatic Speech Recognition (ASR) algorithms, which convert speech to text by breaking it down into phonemes or units
[13]	2011	Apple	Siri virtual assistant on iPhone and iPad	Used a combination of ASR, natural language processing (NLP), and machine learning to understand and respond to requests
[1]	2012	Google	Google Now voice-activated personal assistant	Used a combination of ASR, NLP, and machine learning to understand and respond to requests
[14]	2014	Amazon	Alexa virtual assistant on Amazon Echo smart speaker	Used a combination of ASR, NLP, and machine learning to understand and respond to requests
[14]	2017	Samsung	Bixby virtual assistant on Samsung flagship smartphones	Used a combination of ASR, NLP, and machine learning to understand and respond to requests
[15]	2020s	Various	Advanced voice-to-text technology using machine learning and AI	Use deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to achieve high levels of accuracy in transcription and understanding of speech

B. Dataset availability

Quality dataset availability is essential when training machine learning or deep learning models to perform operations. In the studied literature, many researchers utilised various datasets during their experimentation. In this literature review, we have collected dataset details and summarised them under the name of the dataset, link to access, size and type of dataset. Table iii presents the details of various datasets used in the studied literature. Collecting dataset information will be helpful for upcoming researchers in the speech recognition domain. Table ii lists various speech datasets used in the studied literature. The table lists the name of the dataset, the type of dataset (speech corpus or image corpus), a link or resource where the dataset can be accessed, and the size of the dataset. The datasets listed in the table include timits and aits, the wall street journal (wsj) corpus, aishell-1 mandarin, imagenet, google speech recognition, rt-03 mde, aishell -1 and -2, ted-lium-v2, dysarthric dataset, and the google voice dataset. These datasets have varying sizes and are used for different purposes, such as speech recognition, command speech recognition, metadata extraction, and dysarthric speech recognition[16][17]. The table provides a comprehensive list of various datasets that can be used for speech recognition, along with links to access the dataset.

TABLE II. DATASETS USED IN SPEECH RECOGNITION LITERATURE REVIEW

Ref	Dataset	Type	Link/Resource	Size
[18]	TIMITS and AITS	Speech Corpus	https://www.kaggle.com/datasets/nltkdata/timitcorpus	6300 sentences
[19]	Wall street journal (WSJ) corpus	Speech Corpus	https://paperswithcode.com/dataset/sms-wsj	73 hours of speech
[13]	AISHEEL-1 Mandarin	Speech Corpus	https://www.openslr.org/33/	520 hours of speech
[6]	ImageNet	Image corpus	https://www.image-net.org/download.php	1.2 million images
[1]	Google speech recognition	Command speech corpus	https://www.tensorflow.org/datasets/catalog/speech_commands	65,000 utterances
[20]	RT-03 MDE	Metadata Extraction corpus	https://catalog.ldc.upenn.edu/LDC2004S08	7861 recordings
[2]	AISHELL -1 AISHELL -2	Speech Corpus	https://www.openslr.org/33/	178 hours of speech
[3]	TED-LIUM-v2	Speech Corpus	https://www.openslr.org/51/	19 TED talks
[17]	Dysarthric Dataset	Speech Corpus	http://www.isle.illinois.edu/sst/data/UASpeech/	21386 audio recordings
[16]	Google voice dataset	Command speech corpus	http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz	65,000 utterances
[18]	TIMITS and AITS	Speech Corpus	https://www.kaggle.com/datasets/nltkdata/timitcorpus	6300 tences

C. Feature Extraction: Analog-to-Digital Conversion

In the speech-to-text conversion process, first, audio must be captured through a microphone in Analog format. The audio is then converted to digital format using a Mel Frequency Cepstral Coefficient feature extraction technique. In 2008, Niewiadomy and Pelikant [7] worked on Mel Frequency Cepstral Coefficient (MFCC) vector generation. Also, they pointed out the issues of acquiring data from voice input, parameterisation, and other techniques to gather helpful information regarding input sound. In addition, the authors focused on the current possible uses of the MFCs for querying data in Query-based voice and sound algorithms. The authors discussed how faster modern computers have enabled the development of new valuable applications, overcoming the limitations of older computing systems. (Byunggyu Lee ,Jinho Yang, Jun soo Cho and Suhwan Kim)[13] Proposed a new keyword spotting technique in 2022, which uses a "ring-oscillator-based" time domain processing method for Analog feature extractor (FEx). This technique uses time-based encoders to process the Analog voice input in a time-based manner, except for the voltage-to-time conversion level. This allows for a new technique for spotting keywords. It shows better performance than the old voltage-domain design. (Kwantae Kim, Chang Gao, Rui Graca, Ilya Kiselev, Hoi-Jun Yoo, Tobi Delbruck and Shih-Chii Liu) [21] Presented a Micro-Electro-Mechanical Systems (MEMS) in which mic was working on triple sampling ADC (Analogue to digital convertor) instead of the old Programmable Gain Amplifier (PGA) in 2022. [22]In the ADC, a half-period delay input with a timed input can be included in a copy of the original signal to enhance the integrator. Although, this technique can have a degrading effect due to oversampling. But, design results show an increase in Signal Noise Ration due to increasing the magnitude of the signal strength.

D. Voice Processing Models

Table III evaluates various speech recognition techniques and their performance in the studied literature. The techniques compared include Hidden Markov Models, Neural Networks, LSTM, and Transformer-based models. The table also notes relevant insights and limitations of each technique, such as requiring extensive training data, decreasing accuracy with an increase in vocabulary size, and ambiguity in output for certain command words. The [3]LSTM with HMM model performed better than the traditional model. The[23] Transformer-based model works better than the pipeline models. The [2] LASO model using NN showed 50 times more performance than autoregressive models. The [24] deep neural network model showed inaccurate results due to less research in the Malayalam language. The[1] proposed multilingual system can produce a model for illiterate people to communicate. The results suggest that adaptation algorithms work better than hybrid and E2E models.

TABLE III. EVALUATION OF DIFFERENT SPEECH RECOGNITION TECHNIQUES

Ref.	Year	Methodology	Performance	Relevant Insights
[25]	1988	Hidden Markow Model	Accuracy: 98.1%	Not feasible as it requires extensive training data
[18]	2019	Hidden Markow Model	95% precision, -70 words 100% precision – 30 words	Anomaly in results when tested with different vocabulary sizes
[26]	2020	Hidden Markow Model	Accuracy:95.0%	More Vocabulary size, the lesser the accuracy
[19]	2021	Neural Networks Model	WER rate more for MVDR-2	Ambiguity in results for small angles in interference.
[27]	2021	Recurrent Fusion Method	68.67% noise reduction- MCT 54.88 %- normal MCT	More the enhancement, the lesser the distortion produced
[6]	2021	Hidden Markow Model, Neural Networks Model	11.8 % more accuracy than traditional systems	Adaptation algorithms work better than hybrid and E2E models.
[1]	2022	Multilingual System	-	Able to produce a model for illiterate people to communicate
[20]	2022	Transformer Based Model	Word Error rate- 13.26 %	Transformer based model works better than the pipeline models
[2]	2022	LASO using NN	50 times more performance than autoregressive models	Cross-modal improves the performance of speech model
[3]	2020	LSTM using HMM	Word Error Rate- 0.9% than the traditional model.	LSTM transformer is better than the latent transformer
[24]	2020	Deep Neural Networks	Word Error Rate-3.01 % Sentence Error Rate-14.59%	Inaccurate results due to less research in the Malayalam language

1. Hidden Markov Model

Automatic Speech Recognition (ASR) is a technology that aims to convert spoken language into written text. Several techniques were used in ASR, including the Hidden Markov Model (HMM) and Artificial Neural Network (ANN) models. In a study by Cox (1988), [14] the Hidden Markow Model was found to be one of the most important algorithms for ASR, as it captured various patterns and analysed them using statistical modelling. HMM,

the algorithm provided an accuracy of nearly 98.1%; when the speaker was independent and digits were used as vocabulary. However, this model required much data to train and was not feasible for large vocabularies. As a result, the authors use small-sized vocabulary to get the speech recognised efficiently and accurately. A study by Kumar and Ankit (2019) focused on [26] continuous Hindi speech recognition using mono phone-based acoustic modelling, which was done acoustically using HMM and ANN techniques. The authors utilised a database of 70 Hindi words and tested the model for speaker-dependent speech recognition. During experimentation, the analysis was done with different vocabulary sizes; also, testing was conducted with the help of 10 speech files, each containing 7-12 continuous words. The combined use of HMM and ANN provides an accuracy of around 95% with a dictionary size of 70 words and 100% accuracy with up to 30 words. Based on the comparative study of results derived over Mel Frequency Cepstral Coefficient (MFCC) and Perceptual Linear Prediction (PLP) feature extraction techniques, authors concluded that MFCC achieves improved results than PLP.

Another study by (Choudhary, Dua, and Mohit) (2020) used [18] Triphone-based acoustic modelling on Hindi ASR systems and compared the performance of MFCC and PLP feature extraction techniques. The authors used a 5-state HMM model for all mono phones of the Hindi language and achieved an accuracy of around 95% with a vocabulary size of 100 words. A comparison was carried out between isolated, linking, and continuous words. Observation says that the system's accuracy increased when the vocabulary size decreased. According to the authors, system performance could be enhanced with trial and error with different Gaussian Mixture Model (GMM) states. The authors emphasize the significance of selecting an appropriate ASR technique for various applications and demonstrate the strengths and limitations of each method. It is essential to conduct further research to enhance ASR system performance and explore new techniques for better accuracy and efficiency.

(Moncy, Ashana Mariam, Athira, M.Jasmin, Hanna, Rajan, Rajeev) [2] Presented an automatic speech recognition system for the Malayalam language using deep neural networks. The authors applied the Mel-frequency algorithm for feature extraction and created acoustic models through Gaussian HMM and deep neural networks. The authors trained the proposed model using the Kaldi open-source toolkit and tested it on a speaker-independent system. The database contained speech corpus data of 43 male and female speakers, with approximately 9 hours of duration. They tested the system on a sizeable word-size dictionary of the Malayalam language on a speaker-independent system. The experiments present that the proposed model achieved a word error rate of 3.01% and a sentence error rate of 14.59%. Furthermore, the authors concluded that the hybrid model results were much better than the traditional Gaussian hidden Markov model-based systems. (Aswin Shanmugam Subramanian, Chao Weng, Meng Yu, Shi-Xiong Zhang, Yong Xu, Shinji Watanabe, Dong Yu)[19] proposed a target speech multichannel extraction method with the help of an end-to-end automatic speech recognition system. A target speech requires pre-saved user speech to resolve ambiguity. Also, they demonstrated the system by including noise as a disturbance medium. Furthermore, a hybrid model was proposed that could improve performance by taking both the location and the user's speech. A performance analysis was done on four beam formers with a steering vector as input. The study found that using the Minimum Variance Distortionless Response (MVDR) with two former beams resulted in a higher word error rate (WER) than the location-based system. Additionally, the system that only used the user's speech performed worse than the location-based system. The authors suggested that further work should be done to enhance the system's performance when the distinction between the user's speech and external interference is minimal.

Desai and Dabhi [28] conducted a survey in 2021 that provides an overview of the resources and applications needed by Hindi Language Natural Language Processing (NLP). The survey discussed the benefits and drawbacks of Hindi NLP in different systems, trending topics and applications such as sentiment analysis and machine translation. The survey also provided insights on the future scope of Hindi NLP in domains like information extraction and recommendation systems. In 1990, Murveit and Moore introduced a new idea for speech systems by integrating natural language constraints into Hidden Markov Model (HMM)-based speech recognition. [29] This approach utilised the syntactic and semantic constraints given by the NLP component of the speech understanding task, and the results showed potential for extensive vocabulary systems. However, HMM recognition required high computational requirements with computers having high-speed microprocessors or specially designed speech recognition hardware.

2. Neural Networks Model

In the comparative study, authors (H. Ibrahim and A. Varo)(2020) showed that [30] the Artificial Neural Network model was more suitable for ASR due to its ability to process speech and its flexibility over other similar models. Additionally, ANN can adapt various sensations in a human voice, such as accent, tones, and disturbance. ANN requires a massive amount of data for voice processing. ANN can recognise various types of phonemes, which was impossible with the Hidden Markow model. In a review article by De Mori [9], a comparison study between end-to-end (E2E) and hybrid models for ASR was conducted. The attention encoder-decoder (AED) and recurrent

neural network (RNN) transducer were the main E2E models focused on in the paper. Although the E2E models were trained for multiple languages and required only paired speech and text data, training an extensive dataset into a single E2E model wasn't easy. On the other hand, hybrid models used paired data for the acoustic model, making it easier to train the model. The authors predicted that the E2E models would replace the hybrid models in the future.

(Alim Misbullah, Hai-Hsing Lin, Chia-Yuan Chang, Hsiu-Wei Yeh, Ko-Cheng Weng) [3] Demonstrated a speech recognition model for dysarthric speakers using deep neural networks with a time delay to accept speech longer than usual. Dysarthric means a person's inability to speak correctly due to weak vocal muscles. The database was collected from dysarthric patients from a Taiwan hospital for research purposes. The model was trained using both dysarthric and non-dysarthric utterances combined with the English Mozilla voice dataset. The combination of the dysarthric and Mozilla datasets shows an improvement of 58% in the results, which was more significant than state-of-art models. In their study, authors (Yang Xuebin, Jia Lei, and Yu Hongzhi) [24] proposed a speech recognition method using convolutional neural networks (CNN) for command words from Google voice data corpus. According to the authors, increasing the accuracy of commonly used words in regular speech recognition can improve the system. For experimentation, the Google voice corpus was resampled to the command words, and observation from the results concludes that the method releases better performance in convolutional neural networks than in deep neural networks. However, research on perfect and imperfect command words corpus is identified as a demanding future work. In 2022, (Ye Bai, Jianguan Yi, Jianhua Tao, Zhengkun Tian, Zhengqi Wen, and Shuai Zhang) [1] proposed a non-autoregressive model for speech recognition called Listen attentively and spell once (LASO), which was based on neural networks. The speech features were represented in a hidden format for each token. The authors also proposed a cross-model with the help of the text model to improve the model's speech performance. In the research paper, experiments were conducted on two datasets: AISHELL 1 & 2 and cross model, which achieved 50 times better results than autoregressive models. In addition, the authors also concluded that the cross-model improves the performance of the speech recognition system.

3. Hybrid Model

In speech recognition, hybrid models combine multiple algorithms used to improve the system's accuracy and efficiency. The most common hybrid models in speech recognition systems are a combination of Hidden Markov Models (HMMs) and Neural Networks (NNs). In the comprehensive survey, authors Rista and Kadriu (2020) [25] compared the HMM ANN and an end-to-end speech recognition system. The authors proposed a hybrid model that combined the HMM and ANN models to achieve excellent accuracy, where the ANN model compensates for the HMM model's weaknesses and vice versa. The authors also suggested that end-to-end speech recognition systems were more flexible, satisfactory, and cost-effective than hybrid models, as hybrid models take longer processing time. (Peter Bell, Joachim Feinberg, Ondrej Klejch, Jinyu Li, Steve Renals, Pawal Swietajanski)[15] Proposed a hybrid model that combined neural networks, hidden Markov models, and adaptation algorithms for speaker, domain, and accent adaptation. In the paper, both the hybrid and end-to-end models performed well for acoustic and language models. During the research, adaptation algorithms were applied to limited data conditions, and the performance was calculated using relative error rate reductions. With this proposed hybrid model, the authors could improve by 11.8% accuracy as compared to traditional end-to-end systems. Also, a reduction in relative error rate is observed by 7.2% and 5.0. Based on the analysis, the authors concluded that speech recognition systems with adaptation algorithms work better for both the hybrid and end-to-end models. (Parnia Bahar, Nikita Makarov, Albert Zeyer, Ralf Schluter and Hermann Ney)[3] Explored a zero-order direct hidden Markov model (HMM) based on latent attention for ASR. The authors applied the transformer and long short-term memory (LSTM) model and experimented on the LibriSpeech version 1000h, Switchboard, and TED-LIUM datasets. The results showed that the latent transformer performed worse than the proposed LSTM transformer, with a 0.9% word error rate. (Cunhang Fang, Jianhua Tao, Zhengkun Tian, Bin Liu, Zhengqi Wen)[27] Proposed a joint enhancement and speech transformer method for training with Gated Recurrent fusion to solve the problem of speech distortion in end-to-end speech recognition. The problem of speech distortion was addressed using a gated recurrent fusion algorithm, which included the noisy medium. The data set was a Mandarin speech transcription called the AISHELL 1, consisting of 400 speakers and 170 hours of audio data. The noisy data set was generated by selecting each record from the AISHELL dataset. The analysis was conducted with and without enhancement extraction features. The E2E-ASR version MCT achieved 68.67% noise reduction, compared to 54.88% for standard E2E-ASR systems, using enhanced speech to remove the noisiest data.

The proposed system depends on the enhancement features, as the system is bound to be affected by distortion; the more the enhancement, the lesser the distortion. In 2020, (Shahana Bano, Pavuluri Jithendra, Gorsa Niharika, and Yalavarthi Sikhi) [6] implemented a system that used Google Speech recognition with multi-language features to convert speech to text in a particular language. Google Speech Recognition was used by adding multi-language

features. [31]Pyaudio and Tkinter packages were imported in a command prompt interface. These packages helped in the conversion of audio to text. The text was passed through a Google translation engine which translated the text into the desired language of the user's choice. The authors aimed to make a model to make it easier for illiterate people to communicate using their regional language with the computer. In 2021, (Junwei Liao, Yu Shi, and Yong Xu) [32]developed an automatic speech recognition system that transformed inaccurate and noisy output into a more easily understandable format for humans. In the experimentation, the Metadata Extraction corpus was utilised to construct task-driven datasets and employed a hidden Markov model, which was data augmented to create a grammatically correct dataset. They evaluated the model using the Bilingual Evaluation Study (BLEU method), which showed a margin of 17.53 and a Word Error rate of 13.26 compared to traditional methods. Authors used RoBERT's language model to translate the text output directly to error-free, human-understandable language. Based on detailed analysis authors concluded that the proposed system performed better than the pipeline-based approach for speech-to-text conversion.

IV. DISCUSSION ON STUDIED EXISTING LITERATURE

Speech recognition systems have come a long way since their early days when they could only convert a limited set of words to text using dynamic time-wrapping algorithms. While later upgrades to automatic speech recognition using hidden Markov models significantly improved accuracy for independent speakers. The field has now shifted towards end-to-end automatic speech recognition using neural networks, which can process a more extensive set of spoken words. Hybrid models that integrate the hidden Markov model with artificial neural network models have also been developed to address accent and background noise limitations.

A. Limitations

1. Regional language-specific Speech recognition systems are unavailable in the existing literature. Thus, we must develop a system that can work on every language spoken worldwide.
2. The existing speech recognition model's accuracy is impeding due to different accents, background noise, depth, tone, and speech volume.
3. No efficient algorithm can pre-process the noisy voice and convert it effectively. For example, hidden Markov models are ineffective in reducing background noise or handling accents, while neural networks require a large amount of data to be trained effectively. Hybrid models can solve some of these problems but require much processing time.
4. Few systems are speaker dependent. Such systems must train a specific speaker's voice to work effectively.
5. Existing Speech recognition systems struggle to accurately recognise speech in specific contexts, such as in noisy environments and different accents.

B. Future Directions

Future work in the speech recognition domain can be derived from the limitations identified during this literature survey. The forthcoming researchers from the speech recognition domain can mainly concentrate on finding the solution to overcome the limitations stated in section A. The foremost solution to overcome the majority of limitations is to develop a multilingual-based speech recognition system that can competently understand the voice instructions specified in any human language. Imminent researchers can also work to improve accuracy by developing a system that can adequately handle background noise and accent issues. Furthermore, enhanced processing time in hybrid model training could make them more practical for real-time applications. Finally, integrating other technologies, such as natural language processing and machine translation, will improve speech recognition systems' accuracy and usefulness.

V. CONCLUSION

For a human being, giving voice instructions are more comfortable than typing instructions. Here the demand for speech recognition comes into the picture. In the existing literature, various speech-to-text algorithms are used by different researchers. But they face difficulties in accuracy because of the quality of input voice, accent problems and variation in the vocabulary of different languages.

The literature review focused on studying the evolution of technologies in the ASR domain. In the study of individual models, hybrid models and end-to-end models are Based on the analysis, it is determined that end-to-end speech recognition, which works on neural network concepts, is becoming more critical than hybrid models. After the meticulous analysis of extracted literature, the authors suggest using integrated end-to-end models with hybrid models to develop a more effective speech recognition system.

REFERENCES

- [1] S. Bano, P. Jithendra, G. L. Niharika, and Y. Sikhi, "Speech to Text Translation enabling Multilingualism," 2020 IEEE Int. Conf. Innov. Technol. INOCON 2020, pp. 20–23, 2020, doi: 10.1109/INOCON50539.2020.9298280.
- [2] Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen, and S. Zhang, "Fast End-to-End Speech Recognition Via Non-Autoregressive Models and Cross-Modal Knowledge Transferring from BERT," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 29, pp. 1897–1911, 2021, doi: 10.1109/TASLP.2021.3082299.
- [3] P. Bahar, N. Makarov, A. Zeyer, R. Schlüter, and H. Ney, "Exploring a zero-order direct HMM based on latent attention for automatic speech recognition," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, pp. 7854–7858, 2020, doi: 10.1109/ICASSP40776.2020.9054545.
- [4] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," Pattern Recognit., vol. 41, no. 10, pp. 2965–2979, 2008, doi: 10.1016/j.patcog.2008.05.008.
- [5] J. Wu, "Introduction to Convolutional Neural Networks," Introd. to Convolutional Neural Networks, pp. 1–31, 2017.
- [6] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals, and P. Swietojanski, "Adaptation Algorithms for Neural Network-Based Speech Recognition: An Overview," IEEE Open J. Signal Process., vol. 2, no. August 2020, pp. 33–66, 2021, doi: 10.1109/OJSP.2020.3045349.
- [7] D. Niewiadomy and A. Pelikant, "Implementation of MFCC vector generation in classification context," J. Appl. Comput. Sci., vol. 16, no. 2, pp. 55–65, 2008.
- [8] V. C. Rajyaguru, "Different Methods Used In Voice Recognition Techniques," Int. Res. J. Eng. Technol., pp. 699–703, 2016.
- [9] R. De Mori, "Recent Advances in Automatic Speech Recognition.," no. February, pp. 106–124, 1979, doi: 10.1561/116.00000050.
- [10] I. Shafran and R. Rose, "Robust speech detection and segmentation for real-time ASR applications," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 1, pp. 432–435, 2003, doi: 10.1109/icassp.2003.1198810.
- [11] G. Sreeram and R. Sinha, "Exploration of End-to-End Framework for Code-Switching Speech Recognition Task: Challenges and Enhancements," IEEE Access, vol. 8, pp. 68146–68157, 2020, doi: 10.1109/ACCESS.2020.2986255.
- [12] A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, "Hate and Offensive Speech Detection in Hindi and Marathi," CEUR Workshop Proc., vol. 3159, pp. 239–247, 2021.
- [13] K. Kim et al., "A 23- μ W Keyword Spotting IC With Ring-Oscillator-Based Time-Domain Feature Extraction," IEEE Journal of Solid-State Circuits, vol. 57, no. 11, pp. 3298–3311, 2022, doi: 10.1109/JSSC.2022.3195610.
- [14] A. Rista and A. Kadriu, "Automatic Speech Recognition: A Comprehensive Survey," SEEU Rev., vol. 15, no. 2, pp. 86–112, 2020, doi: 10.2478/seeur-2020-0019.
- [15] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis," IEEE Access, vol. 10, pp. 47628–47642, 2022, doi: 10.1109/ACCESS.2022.3170836.
- [16] X. Yang, H. Yu, and L. Jia, "Speech recognition of command words based on convolutional neural network," Proc. - 2020 Int. Conf. Comput. Inf. Big Data Appl. CIBDA 2020, pp. 465–469, 2020, doi: 10.1109/CIBDA50819.2020.00110.
- [17] A. Misbullah, H. H. Lin, C. Y. Chang, H. W. Yeh, and K. C. Weng, "Improving Acoustic Models for Dysarthric Speech Recognition using Time Delay Neural Networks," Proc. Int. Conf. Electr. Eng. Informatics, vol. 2020-Octob, no. 1, pp. 5–8, 2020, doi: 10.1109/ICELTICS50595.2020.9315506.
- [18] Kumar, "Continuous hindi speech recognition using monophone based acoustic modeling," Int. J. Comput. Appl., pp. 163–167, 2014.
- [19] S. Subramanian et al., "Far-Field Location Guided Target Speech Extraction Using End-to-End Speech Recognition Objectives," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, pp. 7299–7303, 2020, doi: 10.1109/ICASSP40776.2020.9053692.
- [20] J. Liao, Y. Shi, and Y. Xu, "Automatic Speech Recognition Post-Processing for Readability: Task, Dataset and a Two-Stage Pre-Trained Approach," IEEE Access, vol. 10, no. October, pp. 117053–117066, 2022, doi: 10.1109/access.2022.3219838.
- [21] Lee, J. Yang, J. S. Cho, and S. Kim, "A Low-Power Digital Capacitive MEMS Microphone Based on a Triple-Sampling Delta-Sigma ADC With Embedded Gain," IEEE Access, vol. 10, pp. 75323–75330, 2022, doi: 10.1109/ACCESS.2022.3188661.
- [22] P. Upadhyaya, O. Farooq, M. R. Abidi, and Y. V. Varshney, "Continuous Hindi speech recognition model based on Kaldi ASR toolkit," Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017, vol. 2018-January, no. March, pp. 786–789, 2018, doi: 10.1109/WiSPNET.2017.8299868.
- [23] Y. Wang et al., "Transformer-Based Acoustic Modeling for Hybrid Speech Recognition," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, pp. 6874–6878, 2020, doi: 10.1109/ICASSP40776.2020.9054345.
- [24] A. M. Moncy, M. Athira, H. Jasmin, and R. Rajan, "Automatic Speech Recognition in Malayalam Using DNN-based Acoustic Modelling," 2020 IEEE Recent Adv. Intell. Comput. Syst. RAICS 2020, pp. 170–174, 2020, doi: 10.1109/RAICS51191.2020.9332493.
- [25] S. J. Cox, "Hidden Markov Models for Automatic Speech Recognition: Theory and Application.," Br. Telecom Technol. J., vol. 6, no. 2, pp. 105–115, 1988.
- [26] A. Kuamr, M. Dua, and A. Choudhary, "Implementation and performance evaluation of continuous Hindi speech recognition," 2014 Int. Conf. Electron. Commun. Syst. ICECS 2014, no. 1, 2014, doi: 10.1109/ECS.2014.6892777.

- [27] C. Fan, J. Yi, J. Tao, Z. Tian, B. Liu, and Z. Wen, "Framework for Robust End-to-End," vol. 29, pp. 198–209, 2021.
- [28] N. P. Desai, Prof., and V. K. Dabhi, "Taxonomic survey of Hindi Language NLP systems," 2021.
- [29] T. Kumar, K. O. Villalba-Condori, D. Arias-Chavez, R. K., K. C. M., and D. S. R. S., "An Evaluation on Speech Recognition Technology based on Machine Learning," *Webology*, vol. 19, no. 1, pp. 646–663, 2022, doi: 10.14704/web/v19i1/web19046.
- [30] H. Ibrahim and A. Varol, "A Study on Automatic Speech Recognition Systems," 8th Int. Symp. Digit. Forensics Secur. ISDFS 2020, no. August, 2020, doi: 10.1109/ISDFS49300.2020.9116286.
- [31] R. P. Dias, C. S. L. Vidanapathirana, R. Weerasinghe, A. Manupiya, R. M. S. J. Bandara, and Y. P. H. W. Ranasinghe, "Automated use case diagram generator using NLP and ML," 2023, [Online]. Available: <http://arxiv.org/abs/2306.06962>
- [32] C. Tran, M. K. A. Ahamed Khan, and S. Sridevi, "On the Training and Testing Data Preparation for End-to-End Text-to-Speech Application," 2020 11th IEEE Control Syst. Grad. Res. Colloquium, ICSGRC 2020 - Proc., no. August, pp. 73–75, 2020, doi: 10.1109/ICSGRC49013.2020.9232605.
- [33] A. Velankar, H. Patil, A. Gore, S. Salunke, and R. Joshi, "Hate and Offensive Speech Detection in Hindi and Marathi," *CEUR Workshop Proc.*, vol. 3159, pp. 239–247, 2021.
- [34] P. Upadhyaya, O. Farooq, M. R. Abidi, and Y. V. Varshney, "Continuous Hindi speech recognition model based on Kaldi ASR toolkit," *Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017*, vol. 2018-January, no. March, pp. 786–789, 2018, doi: 10.1109/WiSPNET.2017.8299868.
- [35] T. Kumar, K. O. Villalba-Condori, D. Arias-Chavez, R. K., K. C. M., and D. S. R. S., "An Evaluation on Speech Recognition Technology based on Machine Learning," *Webology*, vol. 19, no. 1, pp. 646–663, 2022, doi: 10.14704/web/v19i1/web19046.