IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 20~43
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp20-43

Evolutionary trends in automatic speech recognition with artificial intelligence: a systematic literature review

Gabriel Oluwatobi Sobola[1], Emmanuel Adetiba[1,2], Olabode Idowu-Bismark[1], Abdultaofeek Abayomi[3,4], Raymond Jules Kala[5], Surendra Colin Thakur[6], Sibusiso Moyo[7]

[1]Department of Electrical and Information Engineering, Covenant Applied Informatics and Communication African Center of Excellence, Covenant University, Ogun State, Nigeria
[2]Faculty of Engineering and The Built Environment, Durban University of Technology, Durban, South Africa
[3]Faculty of Engineering, Built Environment and Information Technology, Walter Sisulu University, East London, South Africa
[4]Department of Information Technology, Summit University, Offa, Nigeria
[5]International University of Grand-Bassam, Grand-Bassam, Côte d'Ivoire
[6]Department of Computer Science, University of South Africa, Pretoria, South Africa
[7]Department of Mathematical Sciences, School for Data Science and Computational Thinking, Stellenbosch University, Stellenbosch, South Africa

Article Info

Article history:

Received Jul 29, 2024
Revised Oct 31, 2025
Accepted Nov 8, 2025

Keywords:

Artificial intelligence
Automatic speech recognition
Deep learning
Hidden Markov model
Language model
Machine learning

ABSTRACT

Human beings depend greatly on communication and continually seek ways to overcome language barriers. Automatic speech recognition (ASR) has emerged as a vital tool for enhancing human interaction. Early ASR research relied on probabilistic models, particularly the hidden Markov model (HMM) and Gaussian mixture model (GMM), with mel-frequency cepstral coefficients (MFCCs) for feature extraction, leading to the creation of Audrey at Bell Laboratories. Subsequently, artificial intelligence (AI) approaches, especially deep learning, have transformed ASR and produced systems such as Jasper, Whisper, Google Assistant, Microsoft Cortana, Apple Siri, and Amazon Alexa. This paper presents a systematic literature review that examines ASR's evolution, the AI architectures employed, their features, strengths and weaknesses, and the performance gains achieved since AI was integrated into probabilistic modelling. A snowballing approach was used to identify relevant studies from Google Scholar and Scopus to address five research questions, iterating through backward and forward searches until no new information was found. Findings reveal that ASR dates back to the 1920s with the Radio Rex toy and has since advanced through architectures including deep learning, recurrent neural networks (RNN), support vector machines (SVM), and transformers, all contributing to improved performance measured by reduced word error rates (WER).

Corresponding Author:

Emmanuel Adetiba
Department of Electrical and Information Engineering
Covenant Applied Informatics and Communication African Center of Excellence, Covenant University
Km. 10 Idiroko Road, Canaan Land, Ota, Ogun State, Nigeria
Email: emmanuel.adetiba@covenantuniversity.edu.ng

1. INTRODUCTION

The development of automatic speech recognition (ASR) for human to machine communication started over 40 years ago [1]–[3]. Then, it explored the probabilistic approaches such as the hidden Markov model (HMM), Gaussian mixture model (GMM) coupled with mel-frequency cepstral coefficients (MFCCs) for feature extraction [1]. The state of the art ASR systems which are HMM based with state probability distribution modelled using the GMM could not improve the performance of an ASR system [1], [4]. This can be enhanced via the artificial neural networks (ANN) and HMM hybrid system. The pre and post processing techniques involving feature extraction, language model, pronunciation model, and model adaptability involving speaker, channel and noise variations are the identified stages where performance improvement is needed [5].

The exploration of artificial intelligence (AI) for speech recognition started about five decades ago [2], [3] while the use of ANN involving deep neural networks (DNN) (a subset of machine learning) which are mostly unsupervised in nature for ASR development started in 2006 [1]. This was a shift in the development technique of ASR from the probabilistic approach. Machine learning (a subset of AI) is the ability of machines or computers to learn from input data without being explicitly programmed to do so. It has produced lots of technologies such as DNN for the development of speech recognition. There are recent advances in machine learning using DNN due to its ability to accommodate huge size of training data and its highly increased processing abilities of computer chips [1]. It comes in the form of recurrent neural network (RNN), convolutional neural network (CNN), and time delay neural network (TDNN) for the modelling of ASR, and sometimes coupled with the HMM. Research works have been carried out on different architectures of hybrid HMM and ANN to model ASR systems. These are the multi-band hybrid HMM+ANN, all-combinations multi-band HMM+ANN hybrid, all-combinations multi-stream HMM+ANN hybrid, the multi-stream tandem HMM+ANN hybrid, and the narrow-band tandem HMM+ANN hybrid, which are exploited over the probabilistic approach involving the conventional hybrid HMM and GMM [5] in order to improve the performance of ASR. Firms like OpenAI has produced the Whisper ASR that explored deep learning for modelling ASR [6], while NVIDIA produced the Jasper ASR that uses 54 convolutional layers for training [7]. Other deep learning techniques that have been used for ASR are the self-organizing map (SOM), radial basis function (RBF), multilayer perceptron (MLP), as well as support vector machine (SVM) [2].

The ASR system has found application in various fields including health, military, disabled people, and technologies [3]. For instance, in the health sector, it has been implemented for medical documentation, making treatment notes, and as a surgical assistant. In the military, it can be used to command an autopilot, set radio frequencies, control the parameters of weapons being released; while in technology, there are different ASR systems such as the Google Assistant, Microsoft Cortana, Apple Siri, and Amazon Alexa [8] that are being used by individual for personalized learning, interactive learning, adaptive learning, language learning; as well as aiding reading and engagement of disabled people with hearing and writing difficulties in discussion [3]. In addition, there are also other online ASR tool kits such as the HTK, Sphinx, Kaldi, CMU LM toolkit, and SRILM that are available to build a working ASR system [4].

The type of dataset and the availability of data is another factor that needs to be considered in ASR development [9]–[12]. There are lots of speech dataset like the Librespeech dataset, National Institute of Standards and Technology (NIST) speech dataset, Hub, Texas Instruments Massachusetts Institute of Technology (TIMIT), TED-LIUM, Common voice, The spoken Wikipedia, CSTR VCTK, Aishell-I, Persian consonant vowel combination (PCVC), and the Arabic Speech Corpus [13]–[16]. This indicates that lots of efforts have been put in place towards achieving the ultimate goal of an ASR system in performing as a human listener [2], [17]. The availability and size of data have a greater influence on the performance of an ASR system. It was shown in [4], that the performance of ASR system improved since 1976 via a reduction in the word error rate (WER) because of the availability of more data at present than the previous years.

Different performance measures have been exploited to evaluate the performance of an ASR system [18]–[23]. The WER is used to evaluate the accuracy, and the real time factor (RTF) to evaluate the speed of the ASR. Other precision measures include: command success rate (CSR), frame error rate (FER), NIST detection cost function (DCF), recognition rate (RR), concept value error rate (CVER), concept error rate (CER), phone recognition error rates (PRER), and frame classification (FRER), word recognition rate (WRR), phone error rate (PER), dB SIR gain, label error rate (LER), phoneme classification performance (PCP), root mean square error (RMSE), sentence accuracy, query error rate (QER), unweighted classification accuracy and, gain in dB [1], [2].

The remaining parts of this article is structured as follows: section 2 contains the methodology applied for the study. The results obtained are presented in section 3. The discussion as tailored to each research question in the study is contained in section 4. Finally, section 5 presents the conclusion.

2. METHOD

In a quest to carry out a thorough systematic review on the evolutionary trend on the ASR with AI, there is the need to address the topic by developing research questions to guide and serve as the focus of this paper. These sets of questions were developed to have deeper insight on the topic. Lots of research works have been studied in this area, and different researchers have approached the ASR system in different ways to achieve similar results in terms of generating the text equivalents of the speech signals. Hence 5 research questions were developed and are tailored towards the evolution of AI with ASR, the different types of AI that have been explored, their benefits and demerits, and performance enhancement obtained by researchers-these are in a quest to make findings in regard to ASR evolution with AI.

After developing these set of questions, the selection of papers was carried using the snowballing approach. Journal articles and review papers were sourced and selected from the databases of Scopus and Google Scholar. These set of papers formed the initial start set papers. The snowballing approach has the backward snowballing approach and forward snowballing approach which were explored for the selection of papers from the initial start sets. The backward approach involves selection of papers cited in the work while papers were selected from the citations of each paper on their Scopus database using the forward approach. The snowballing approach is illustrated in Figure 1. Papers were analyzed and selected or rejected using their titles, authors of the papers, the abstract, and years.

The selected papers were reviewed to answer the research questions as described in the research questions section. The snowballing approach was run through 3 iterations from the 13 start set papers selected from the initial 18 papers until no candidate paper was fit for selection. The steps taken to search for and select papers from Google Scholar and Scopus databases were in alignment with the snowballing approach laid down in [24].

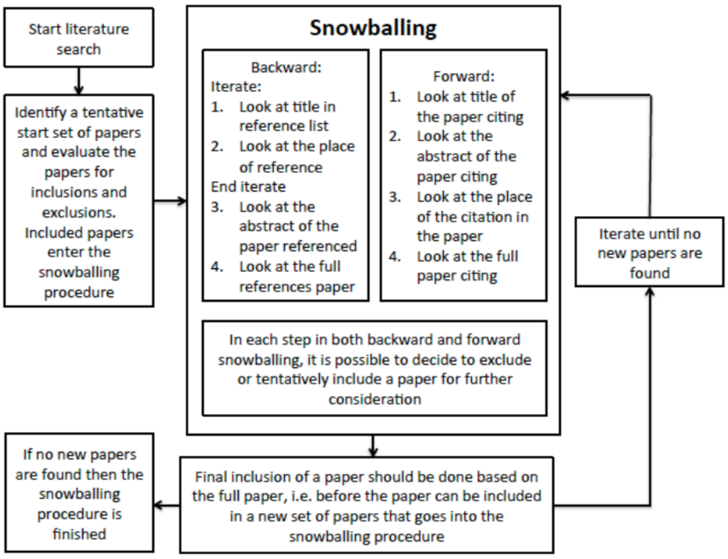

Start literature search
Identify a tentative start set of papers and evaluate the papers for inclusions and exclusions. Included papers enter the snowballing procedure
Snowballing
Backward:
Iterate:
1. Look at title in reference list
2. Look at the place of reference
End iterate
3. Look at the abstract of the paper referenced
4. Look at the full references paper
Forward:
1. Look at title of the paper citing
2. Look at the abstract of the paper citing
3. Look at the place of the citation in the paper
4. Look at the full paper citing
In each step in both backward and forward snowballing, it is possible to decide to exclude or tentatively include a paper for further consideration
Iterate until no new papers are found
If no new papers are found then the snowballing procedure is finished
Final inclusion of a paper should be done based on the full paper, i.e. before the paper can be included in a new set of papers that goes into the snowballing procedure

Figure 1. Snowballing approach [24]

2.1. Research questions

2.1.1. When did the development of ASR with AI start?

The development of speech recognition demands that aspiring and interested researchers have understanding and insights into when speech recognition started particularly with the exploration of AI. Before the integration of AI for the ASR, different statistical approaches have been utilized to model speech recognition. This question tends to dig deeper into date of various AI approaches for the development of ASR. The question provides various answers to when each of AI approaches were integrated for ASR modelling. For instance, it answers when DL started to gain popularity for the development of AI, the days of popularity in the utilization of RNN, and various robust feature extraction methods.

2.1.2. What types of AI techniques have been explored by various researchers to develop an ASR system?

Different types of AI have been utilized to model speech recognition. This question gives insight for the interested reader into the various types of AI that find application in ASR modelling. Different deep learning architectures such as the CNN, RNN, SVMs, SOM, TDNN, long short-term memory (LSTM) have been used to model ASR. Coupled with this is the utilization of different feature extraction techniques such as the wavelet transform, linear predictive coding (LPC), and relative spectral filtering (RASTA) to model ASR that can withstand the varying environmental condition, speaker variation, and intonation. Via this question, researchers and readers are able to have deep understanding of the different types of AI being utilized to the modelling of ASR.

2.1.3. What are the benefits and disadvantages of exploring each of those algorithms mentioned in 2?

Different AI algorithms have been explored for the modelling of ASR. Each of these algorithms comes with it is own benefits and disadvantages. This question addresses their benefits such as performance enhancement in terms of reduction in WER compared to the statistical approaches, ability to obtain optimized results; and limitations of AI usage such as their inflexibility to handle timing variability, inability to model ASR that can model noise variability. The question also explains that the choice of different AI approach comes with their benefits which might be a limitation in other AI approaches.

2.1.4. Does the exploration of AI for building an ASR system have performance improvement over the traditional probabilistic approaches?

The traditional probabilistic approaches involving the HMM, GMM, along with other algorithms like the forward algorithms, forward and reverse algorithm, Viterbi algorithm come with their benefits. However, since the exploration of AI, enhance performance have been achieved. AI approach involves training with large corpus of speech data and robust speech recognition model have been developed. The question provides answers to if the purpose of utilizing AI in performance enhancement in terms of WER and WRR have been achieved.

2.1.5. Has the introduction of AI in the development and modelling of ASR been able to achieve the ultimate goal of ASR of performing as a human listener?

Researchers have developed various speech recognition system to recognize human speech so that the output text will look like a human transcriber’s output. Various ASRs have been developed to achieve this, however, the question that calls for concern is to know if the ultimate goal of ASR to interact like human being has been achieved. Whisper for instance, was able to function above human transcriber in terms of the output generated text, but the result is not the ideal ASR result that fulfils the ultimate goal of an ASR system.

2.2. Start set

There were 5, 5, 6, 1, and 1 paper selected, which addressed research questions 1, 2, 3, 4, and 5 respectively. For research questions 1, 2, 4, and 5; papers were sourced from the Google Scholar database while those of research question 3 were obtained from the Scopus database. A total of 18 papers were selected. The papers are illustrated using the letter “Q”+a number+letter “P”+a number where “Q”+a number represents the question number being addressed, while “P”+a number represents paper number in relation to the research question number being addressed. The selection criteria for selecting papers from the Google Scholar for questions 1, 2, 4, and 5 were the relation of the paper via thorough reading of the abstract, the title of the paper, the number of citations, and year of publication mostly within 10 years. The number of citations of Q1P1, Q1P2, Q1P3, Q1P4, Q1P5, Q2P1, Q2P2, Q2P3, Q2P4, Q2P5, Q4P1, and Q5P1 are 924, 154, 42, 8, 210, 102, 78, 116, 500, 211, 49, and 111 respectively on Google Scholar as at the date of documentation (September 12, 2023). We did not search for papers published some 20-30 years ago for some research questions like Q1, and Q5 because some earlier review papers used in this current study already have answers to some of the questions needed. On the other hand, Q3 papers were sourced from Scopus in order to have a mixture of different databases for paper selections in this work.

The thorough searching commenced by entering the search words with the plus operator as: Automatic+speech+recognition. Then, some set of keywords as shown in Table 1 were specified to filter the searched results. The results of this, reduced the number of listed papers to 21,083. The subject area and document type were also specified as provided in Table 2 to filter the results to 18,167 documents. Next is the refined search with the search word automatic speech recognition+artificial intelligence which reduced the number to 7,020, while the language factor which involves limiting the search papers to English language gave 6,948 papers. This was followed by using the refined search with the key word “benefit” which gave a very anticipated reduced value of 359 papers and later 204 when the range was set between 2010 and 2020 so as to have that small number of 204 for selecting papers from and to have papers that have the latest and updated technology as well as latest research findings in ASR. The first paper, Q3P1 with 8 citations in Scopus and 15 citations in Google Scholar was selected after reading the abstract to ensure it contained the needed information. Other papers such as the Q3P2, Q3P3, and Q3P4 were also selected as detailed in Table 3. From the selection state of Q3P3, using keywords such as deep learning, feature extraction, DNN, neural network, RNN, CNN, SVM, Bayesian networks, ANN, temporal classification, multilayer neural networks, transfer learning, LSTM, and pattern recognition brought the number of papers down to 188 documents, among which Q3P5 was selected.

Finally, from the selection state of Q3P5, the use of the refined keyword machine learning led to the selection of Q3P6. A detailed analysis of the reduction process leading to the eventual selection of papers Q3P1, Q3P2, Q3P3, and Q3P4 are given in Tables 1 to 3 where the number of papers left at each stage of the filtering are indicated with an arrow pointing right. The selection processes of Q3 papers are illustrated in Figure 2.

Table 1. Search keywords for Q3 in Scopus (≥21,083 documents)

Acoustic model	Performance
Acoustic modelling	Reverberation
Artificial intelligence	Robust speech recognition
ASR automatic speech recognition	Speech enhancement
Audio signal processing	Speaker recognition
Automatic recognition	Speech recognition
Automatic speech recognition	Speech recognition software
Automatic speech recognition (ASR)	Speech recognition systems
Automatic speech recognition system	Speech recognizer
Continuous speech recognition	Speech signals
Language model	WER
Machine learning	

Table 2. Filtering of Q3 papers by subject area and document type in Scopus (≥18,167 documents)

Subject area	Document type
Computer science	Conference paper
Engineering	Article
	Review
	Conference review

Table 3. Searching with refined keywords for Q3P1, Q3P2, Q3P3, and Q3P4 after the steps used to obtain Tables 1, 2, and 3 in Scopus

Refined keywords ≥year range	Reduction in search results	Reduction in search results	Reduction in search results
Benefit ≥2010-2020	359 ≥204	NILL	NILL
Advantage ≥2010-2020	NILL	367 ≥185	367
Number of papers selected	Q3P1, Q3P2	Q3P3	Q3P4

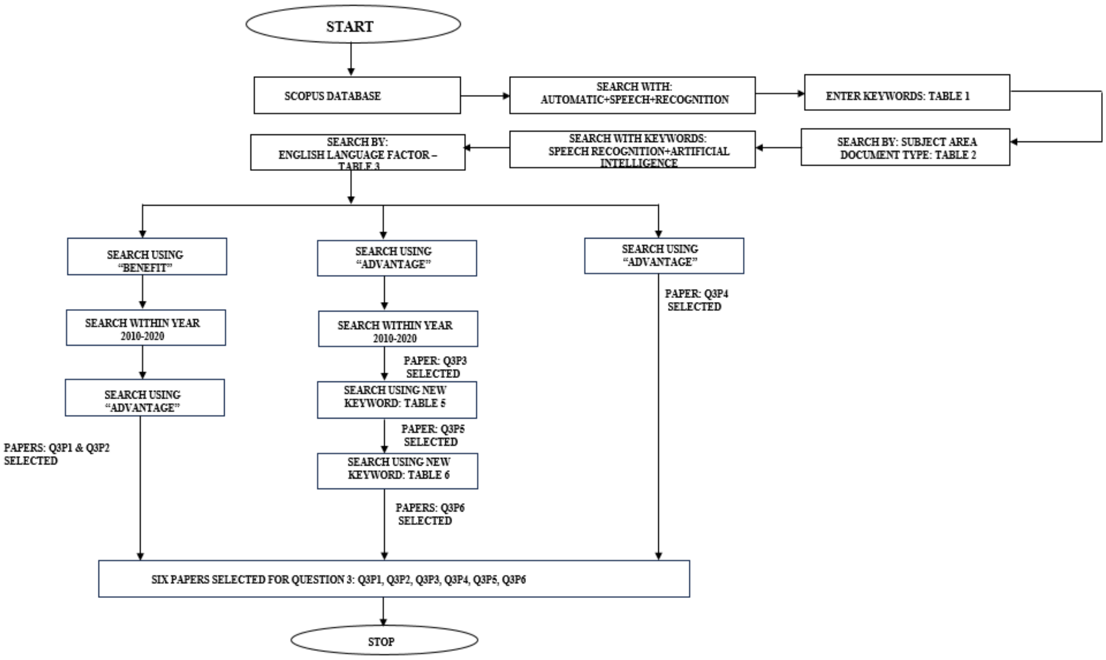

START
SCOPUS DATABASE
SEARCH WITH: AUTOMATIC+SPEECH+RECOGNITION
ENTER KEYWORDS: TABLE 1
SEARCH BY: SUBJECT AREA DOCUMENT TYPE: TABLE 2
SEARCH WITH KEYWORDS: SPEECH RECOGNITION+ARTIFICIAL INTELLIGENCE
SEARCH BY: ENGLISH LANGUAGE FACTOR – TABLE 3
SEARCH USING "BENEFIT"
SEARCH WITHIN YEAR 2010-2020
SEARCH USING "ADVANTAGE"
PAPERS: Q3P1 & Q3P2 SELECTED
SEARCH USING "ADVANTAGE"
SEARCH WITHIN YEAR 2010-2020
PAPER: Q3P3 SELECTED
SEARCH USING NEW KEYWORD: TABLE 5
PAPER: Q3P5 SELECTED
SEARCH USING NEW KEYWORD: TABLE 6
PAPERS: Q3P6 SELECTED
SEARCH USING "ADVANTAGE"
PAPER: Q3P4 SELECTED
SIX PAPERS SELECTED FOR QUESTION 3: Q3P1, Q3P2, Q3P3, Q3P4, Q3P5, Q3P6
STOP

Figure 2. Flowchart of selection of papers for question 3

The 18 papers initially selected are listed in Table 4. After inclusion/exclusion criteria such as the relevance of the papers to the particular question being addressed and, the relevance of the contents to the question, hence, the papers eventually selected were Q1P1, Q1P2, Q1P4, Q1P5, Q2P1, Q2P2, Q2P3, Q2P4, Q2P5, Q4P1, Q5P1, Q3P3, and Q3P6. From the original 18 papers selected, papers Q1P3, Q3P1, Q3P2, Q3P4, and Q3P5 were eventually discarded. These sets of papers were rejected due to their inability to properly answer the questions being addressed, or they contained repetitions of what had been gotten from the previous journals. For instance, paper Q1P3 was a review paper that contained the majority of the information that we've acquired from the previous papers, which eventually led to its rejection. Paper Q3P2 was also rejected because its abstract deviated from the questions of the research work being carried out. Hence a total of 13 papers were selected out of the original 18 papers sourced from Google Scholars and Scopus as the start set of papers for the first iteration stage. The selected papers are listed in Table 5. The whole selection processes of the 13 start set papers are illustrated in Algorithm 1.

Algorithm 1. Selection process for 13 initial papers selected from 18 initial papers selected

```
1.  BEGIN
2.  Initialize the original array as empty list
3.  Scan/read through the abstract / title of paper 1
4.  Iterate through papers 1 to 18:
5.      If first paper contains relevant information:
6.      Append it to the array
7.      Else reject the paper
8.      If second paper contains relevant information:
9.      Append it to the array
10.     Else Reject the paper
11.     Repeat steps 4 to 5 for paper 3 through 18
12. Print the array containing all the selected papers
13. END
```

Table 4. The 18 initial set of papers selected from Scopus and Google Scholar

Ref. no.	Paper identity code	Ref. no.	Paper identity code
[1]	Q1P1	[25]	Q2P5
[2]	Q1P2	[26]	Q3P1
[27]	Q1P3	[28]	Q3P2
[3]	Q1P4	[29]	Q3P3
[4]	Q1P5	[16]	Q3P4
[30]	Q2P1	[31]	Q3P5
[32]	Q2P2	[33]	Q3P6
[17]	Q2P3	[5]	Q4P1
[34]	Q2P4	[35]	Q5P1

Table 5. The 13 selected star set papers for the first iteration

Ref. no.	Paper identity code	Ref. no.	Paper identity code
[1]	Q1P1	[34]	Q2P4
[2]	Q1P2	[25]	Q2P5
[3]	Q1P4	[29]	Q3P3
[4]	Q1P5	[33]	Q3P6
[30]	Q2P1	[5]	Q4P1
[32]	Q2P2	[35]	Q5P1
[17]	Q2P3		

2.3. Iteration

2.3.1. Iteration 1

The 13 set of papers (start set) selected from the 18 initially set of papers sourced from the Google Scholar and Scopus were run through the forward and backward snowballing approaches until no paper was found to be included in the list of selected papers. The starting set of papers consisted of 13 papers (Table 5) selected from the 18 initially set of papers (Table 4) sourced from Google Scholar and Scopus. The selection of these 13 papers was based on some exclusion and inclusion criteria based on the relevance of information in relation to the research questions being addressed. These papers were selected taking into consideration the years of publications, authors, and publishers with a key focus on their relevance. Algorithm 1 shows the selection process of these 13 starts set papers.

2.3.2. Backward snowballing

From the first paper, Q1P1, candidates were selected for inclusion using backward snowballing, 4 papers were selected which were added to the start set. From paper Q1P2 backward snowballing approach produces 6 papers which were added to the start set papers. From paper Q1P4, 2 papers were selected via this approach. Q1P5 produces 6 papers. Q2P1 produces 2 papers, Q2P2 produces 1 paper. Q2P3 produces 3 papers. Paper Q2P4 produces 3 papers. Paper Q2P5 produces 3 papers. Q4P1 produces 6 papers, Q5P1 produces 7 papers, Q3P3 generates 3 papers, while the last paper (Q3P6) produces 2 papers. Then a total of 48 papers were selected from the backward snowballing approach.

2.3.3. Forward snowballing

For the set of 13 start set papers selected, through the forward snowballing approach, the analysis of the papers selected using the inclusion/exclusion criteria are detailed thus: paper Q1P1 produces 11 papers after using some criteria such as the abstracts of the papers, the titles, and the number of citations of the papers. They are listed as Q1P1F1 through Q1P1F12. Paper Q1P2 produces 9 papers. Some conditions for paper selection were: if the papers had been included as part of the papers in the other selections, it was rejected, the abstract of the paper, and the relevance of the papers to the question being considered. No paper was selected from Q1P4. It had just 10 citations and 7 of the papers do not show relationship to the questions being addressed. The remaining 3 are not included due to inability to access the materials. For paper Q1P5 about 10 papers were seen as potential papers in relation to the questions being addressed none of which contained something unique from the information initially gathered from the previous papers. Besides, some of the contents had been used in other papers. Hence, no paper was selected. Paper Q2P1 with 106 citations had 1 paper for inclusion but it was rejected because it has been earlier selected, hence no paper was selected. Other papers are either repeated papers or not sharing or addressing the questions in this review. For papers Q2P1 all through Q5P1, no paper was selected for each of them. The reason being that majorly all the research topics share similarities with already studied papers for this work, or the irrelevance of topic or abstract to the questions being addressed.

Therefore, a total of 20 papers were selected via the forward snowballing approach, which shows that a total of 68 papers has been selected for inclusion in this work due to relevance in addressing these questions in this review. Hence, it can be concluded that from the first iteration, candidate papers were selected via the backward and forward snowballing approach and included in the start set selected list of papers making a total of 13 (star set papers)+48 papers selected via backward snowballing+20 papers selected via forward snowballing. The 68 selected papers are as shown in Table 6.

Table 6. Iteration 1 selected list of papers (68)

Ref. no.	Paper identity code	Ref. no.	Paper identity code	Ref. no.	Paper identity code	Ref. no.	Paper identity code
[36]	Q1P1B1	[37]	QIP5B6	[38]	Q4P1B5	[10]	Q1P1F5
[39]	Q1P1B4	[40]	Q2P1B1	[41]	Q4P1B6	[42]	Q1P1F6
[43]	Q1P1B5	[22]	Q2P1B2	[44]	Q5P1B1	[14]	Q1P1F7
[45]	Q1P1B6	[46]	Q2P2B2	[47]	Q5P1B2	[48]	Q1P1F8
[49]	Q1P2B1	[50]	Q2P3B3	[51]	Q5P1B3	[15]	Q1P1F9
[52]	Q1P2B3	[53]	Q2P3B4	[54]	Q5P1B4	[8]	Q1P1F10
[55]	Q1P2B4	[56]	Q2P3B5	[57]	Q5P1B5	[58]	Q1P1F11
[18]	Q1P2B5	[59]	Q2P4B1	[60]	Q5P1B6	[61]	Q1P1F12
[62]	Q1P2B6	[63]	Q2P4B2	[64]	Q5P1B7	[9]	Q1P2F1
[65]	Q1P2B7	[66]	Q2P4B3	[67]	Q3P3B1	[21]	Q1P2F2
[68]	Q1P4B1	[19]	Q2P5B1	[69]	Q3P3B2	[70]	Q1P2F3
[13]	Q1P4B3	[71]	Q2P5B2	[72]	Q3P3B3	[73]	Q1P2F4
[74]	QIP5B1	[75]	Q2P5B3	[76]	Q3P6B1	[23]	Q1P2F5
[77]	QIP5B2	[78]	Q4P1B1	[79]	Q3P6B2	[80]	Q1P2F6
[81]	QIP5B3	[82]	Q4P1B2	[11]	Q1P1F2	[83]	Q1P2F7
[84]	QIP5B4	[85]	Q4P1B3	[20]	Q1P1F3	[86]	Q1P2F8
[87]	QIP5B5	[88]	Q4P1B4	[12]	Q1P1F4	[89]	Q1P2F9

2.3.4. Summary of iteration 1

From the 13-start set of papers used for this research work, a total of 48 papers have been selected via the backward snowballing approach, where papers Q1P1, Q1P2, Q1P4, Q1P5, Q2P1, Q2P2, Q2P3, Q2P4, Q2P5, Q4P1, Q5P1, Q3P3, and Q3P6 had 232, 179, 10, 43, 9, 24, 26, 51, 75, 86, 75, 51, and 135 references analyzed respectively which gives a total of at most 996 references for the 48 papers included in the start set of papers based on the inclusion/exclusion criteria. Using the forward snowballing approach, a total of 20 papers were selected for inclusion in the start set papers. Papers Q1P1, Q1P2, Q1P4, Q1P5, Q2P1, Q2P2, Q2P3, Q2P4, Q2P5, Q4P1, Q5P1, Q3P3, and Q3P6, with 982, 179, 10, 215, 106, 83, 125, 516, 213, 49, 118, 667, and 118 citations respectively showed that a total of at most 3,381 references were analyzed using exclusion/inclusion criteria for the forward snowballing approach. Hence, a total of at most 4,377 papers were analyzed and a total of 68 papers had been included in the start set of papers in the iteration 1. In the iteration 2, the 68 papers were also analyzed to see which papers could be extracted from each via the backward and forward snowballing approach.

2.3.5. Iteration 2

It should be noted that at the start of the iteration 2 and subsequent ones, the papers explored were limited to 2022 and 2023. This is due to the fact that the majority of the solutions to the research questions in this research work have been obtained; then there is the need to reduce the search papers particularly for papers that had large number of citations or references. Here each of the 68 papers were analyzed to make selections of papers in both approaches using the inclusion and exclusion criteria.

2.3.6. Backward snowballing

Through backward snowballing the first set of 48 papers had no paper for inclusion considering the years (2022, and 2023) being considered at the moment. Hence, papers Q1P1B1 through Q3P6B2 with 0 reference produces no paper. Q1P1F2, with 2 references, has no paper for inclusion due to the titles of the papers. Q1P1F3, with 11 references for just the year 2023, produces no paper. Meanwhile, papers Q1P1F4, Q1P1F5, Q1P1F6, Q1P1F7, Q1P1F8, Q1P1F9, Q1P1F10, Q1P1F11, Q1P1F12, Q1P2F1, and Q1P2F2 with 0, 12, 0,1, 0, 21, 0, 0, 0, 0, and 0 references respectively produce no paper for inclusion. In addition, papers Q1P2F3, Q1P2F4, Q1P2F5, Q1P2F6, Q1P2F8, and Q1P2F9 with 1, 1,367, 1, 0, and 0 references respectively produces no paper while paper Q1P2F7 with 2 references has only 1 paper for inclusion, but it was rejected because it has already been used, hence it produces no paper. Therefore, no paper was selected via this approach.

2.3.7. Forward snowballing

Through the forward snowballing approach, Q1P1B1 with 19 citations produces no paper. Papers Q1P1B4, Q1P1B5, Q1P1B6, and Q1P2B1 have 366, 19, 27 and 1 citations respectively produce no paper and have no candidate(s) for inclusion. Q1P2B3, was identified with 1 reference which was rejected due to the title; hence no paper was selected for inclusion. Q1P2B4, with 33 citations, produces no paper. Q1P2B5, has 1 citation which has already been included in the selected list of papers for this work. Hence no paper was selected from it. Q1P2B6 with 2 citations and Q1P2B7, with 10 citations, have no paper to be included based on the criteria. Also, Q1P4B1 with 22 citations within the year considered, produces no paper. Q1P4B3, with 3 papers on speech, had none to be selected for inclusion. QIP5B1 has 4 citations and QIP5B2 had 1140 citations but none was selected due to the fact that the majority of those papers had links or similarities to other included papers. Similarly, QIP5B3, QIP5B4, and QIP5B5 have 8, 8, and 18 citations respectively but had no candidate for inclusion while QIP5B6 with 6 citations, produces no paper as well. Q2P1B1, with 255 citations produce 1 paper indicated as PG1. Q2P1B2, with 2 citations, produces no paper. Q2P2B2, with 4 citations, produces no paper. No paper was produced by Q2P3B3 (14 citations), Q2P3B4 (1020 citations), Q2P3B5 (45 citations), Q2P4B1 (2 citations), Q2P4B2 (5 citations), Q2P4B3 (0 citations), Q2P5B1 (109 citations), Q2P5B2 (0 citation), Q2P5B3 (149 citations), Q4P1B1 (0 citation), Q4P1B2 (0 citation), Q4P1B3 (3 citations), Q4P1B4 (0 citation), Q4P1B5 (40 citations), Q4P1B6 (3 citations), Q5P1B1 (1 citation), Q5P1B2 (2 citations), Q5P1B3 (37 citations), Q5P1B4 (0 citation), Q5P1B5 (56 citations), while Q5P1B6 (48 citations) has one paper for inclusion but was rejected because it has already been used, hence, it produces no paper. In addition, Q5P1B7, with 11 citations, produces no paper as well as Q3P3B1 (539 citations), Q3P3B2 (55 citations), Q3P3B3 (33 citations), Q3P6B1 (20 citations), Q3P6B2 (0 citation), Q1P1F2 (35 citations), Q1P1F3 (21 citations), Q1P1F4 (6 citations), Q1P1F5 (1 citation), Q1P1F6 (13 citations) Q1P1F7 (2 citations), Q1P1F8 (6 citations), Q1P1F9 (4 citations), Q1P1F10 (6 citations), Q1P1F11 (1 citation), Q1P1F12 (0 citation), Q1P2F1 (17 citations), Q1P2F2 (2 citations), Q1P2F3 (3 citations), Q1P2F4 (4 citations), Q1P2F5 (2 citations), Q1P2F6 (0 citation), Q1P2F7 (0 citation), Q1P2F8 (0 citation), and Q1P2F9 (0 citation). The only paper (PGI) to be included with the start set is contained in Table 7.

Table 7. Iteration 2 selected paper (1)

Ref. no.	Paper identity code
[90]	PG1

2.3.8. Summary of iteration 2

Apart from the 13-start set of papers used for this research work, only 1 paper has been selected via the backward snowballing and forward snowballing approaches during the iteration 2, and this is the paper to be included in the already selected 68 papers. A total of 420 papers were analyzed via the backward snowballing approach and a total of 4,267 papers were analyzed via the forward snowballing approach. Thus, a total of 4,687 papers were analyzed in the second iteration. The total number of papers selected so far is 69. In the next iteration, the single paper was experimented with both the backward and forward snowballing approaches.

2.3.9. Iteration 3

In this iteration 3, the single paper PG1which was selected via the inclusion/exclusion criteria from the previous iteration was also analyzed using the forward and the backward snowballing approaches. The processes of this present stage went thus:

- Backward snowballing: for paper PG1 no paper was identified within the time frame of 2022 to 2023, hence, no paper was selected.
- Forward snowballing: paper PG1 had no citation, hence no paper was selected via this approach.

The summary of iteration 3 is that no paper was analyzed via the backward and forward snowballing approach in this iteration, hence, no selection was made, and no paper was analyzed in iteration 3. However, for the purpose of efficiency, a single paper (2013) was analyzed via the backward snowballing approach, however it was not selected because the contents were on feature extraction techniques which have already been studied in the previous literatures. Thus, a total of 69 papers has been selected via the forward and backward snowballing approaches after the third iteration and adding this to the start set list of papers (13) gives a total of 82 papers selected in this work as provided in Table 8. The authors of the selected papers utilized in this current study were not contacted to prevent delay of this study considering the fact that the reason for contacting, is to source for additional materials (in relation to this study) from them. There can also be an element of bias hence the reason for not contacting any of the authors.

Table 8. Total selected list of papers including the 13 start set papers (82)

Ref. no.	Paper identity code	Ref. no.	Paper identity code	Ref. no.	Paper identity code	Ref. no.	Paper identity code
[1]	Q1P1	[62]	Q1P2B6	[75]	Q2P5B3	[20]	Q1P1F3
[2]	Q1P2	[65]	Q1P2B7	[78]	Q4P1B1	[12]	Q1P1F4
[3]	Q1P4	[68]	Q1P4B1	[82]	Q4P1B2	[10]	Q1P1F5
[4]	Q1P5	[13]	Q1P4B3	[85]	Q4P1B3	[42]	Q1P1F6
[30]	Q2P1	[74]	QIP5B1	[88]	Q4P1B4	[14]	Q1P1F7
[32]	Q2P2	[77]	QIP5B2	[38]	Q4P1B5	[48]	Q1P1F8
[17]	Q2P3	[81]	QIP5B3	[41]	Q4P1B6	[15]	Q1P1F9
[34]	Q2P4	[84]	QIP5B4	[41]	Q5P1B1	[8]	Q1P1F10
[25]	Q2P5	[87]	QIP5B5	[47]	Q5P1B2	[58]	Q1P1F11
[29]	Q3P3	[37]	QIP5B6	[51]	Q5P1B3	[61]	Q1P1F12
[33]	Q3P6	[40]	Q2P1B1	[54]	Q5P1B4	[9]	Q1P2F1
[5]	Q4P1	[22]	Q2P1B2	[57]	Q5P1B5	[21]	Q1P2F2
[35]	Q5P1	[46]	Q2P2B2	[60]	Q5P1B6	[70]	Q1P2F3
[36]	Q1P1B1	[50]	Q2P3B3	[64]	Q5P1B7	[73]	Q1P2F4
[39]	Q1P1B4	[53]	Q2P3B4	[67]	Q3P3B1	[23]	Q1P2F5
[43]	Q1P1B5	[56]	Q2P3B5	[69]	Q3P3B2	[80]	Q1P2F6
[45]	Q1P1B6	[59]	Q2P4B1	[72]	Q3P3B3	[83]	Q1P2F7
[49]	Q1P2B1	[63]	Q2P4B2	[76]	Q3P6B1	[86]	Q1P2F8
[52]	Q1P2B3	[66]	Q2P4B3	[79]	Q3P6B2	[89]	Q1P2F9
[55]	Q1P2B4	[19]	Q2P5B1	[11]	Q1P1F2	[90]	PG1
[18]	Q1P2B5	[71]	Q2P5B2				

2.4. Efficiency of systematic literature review using snowballing

In terms of the number of papers utilized for this research, the efficiency of utilization can be computed as:

i) Start set: 18 papers (or candidates) for initially selected and 13 start set papers were included, i.e. $Efficiency = \frac{13}{18} = 72\%$.

ii) Iteration 1: 4,377 candidates from snowballing from the start set, and 68 papers were included, i.e. $Efficiency = \frac{68}{4377} = 1.6\%$.

iii) Iteration 2: 4,687 candidates for inclusion were generated in the backward and forward snowballing, 1 paper was included, i.e. $Efficiency = \frac{1}{4687} = 0.02\%$.

iv) Iteration 3: 1 candidate was examined, and no paper was included, and hence the efficiency becomes: 0%

The overall efficiency becomes: $Efficiency = \left[\frac{(13 + 68\ + 1 + 0)}{18 + 4377 + 4687 + 1}\right] \times 100 = 0.90\ \%$.

According to Wohlin [24], an overall efficiency of 3.7% was obtained. In the work, the number of papers analyzed were not as voluminous compared to what is available in this research work. Besides, if the year range within which papers were selected for the iteration 2 had been extended beyond the two years of 2022 to 2023, the efficiency would have increased beyond the 0.90% obtained. In addition to this, it was obvious that there was a surge in the number of papers selected from the start set, in the iteration 1, however, the number selected follows a descending trend as the number of iterations increases, and consequently the energy expended in searching and selecting papers drops, which implies an increased and optimized way of selecting papers for reviewing.

3. RESULTS AND DISCUSSION

3.1. Data extraction

After a thorough review of each of the 96 papers selected for this current study. Relevant information and findings regarding ASR in relation to each of the research questions were extracted. These are presented in Tables 9 to 13 in the appendix.

3.2. Results of extracted data from selected papers

The results of and findings from extracted data from selected papers regarding the research question:

i) Q1: when did the development of ASR with AI start? is presented in Table 9 in the appendix.
ii) Q2: what types of AI have been explored by various researchers to develop ASR? is presented in Table 10 in the appendix.
iii) Q3: what are the benefits and disadvantages of exploring each of those algorithms mentioned in 2? is presented in Table 11 in the appendix.
iv) Q4: does the exploration of AI for building an ASR system have performance improvement over the traditional probabilistic approaches? is presented in Table 12 in the appendix.
v) Q5: has the introduction of AI in the development and modelling of ASR been able to achieve its ultimate goal to perform as a human listener? is presented in Table 13 in the appendix.

Haven't done a thorough review on the research questions 1-5 as enumerated, the results of various studies surveyed were listed in Tables 9 to 13. It can be seen from the tables that various approaches and methodologies involving AI, particularly machine learning and deep learning, have been explored by various researchers to achieve the conversion of speech to text. The tables further illustrates other existing methods like the probabilistic approach [17], [25], [35], template based approach [91]. More research works are ongoing [6], [7] to develop a more robust speech recognition that can withstand various variabilities like the environment, noise, speaker and most importantly to achieve the ultimate goal of speech recognition. It can also be seen that for extracting features from speech signals, majority of the researchers explored the MFCCs [1], [32], but when variability is considered, then other methods of feature extraction such as the wavelet transform, RASTA, and zero crossing peak amplitude (ZCPA) are explored [17], [32].

According to Malik *et al.* [2], it was noted that hybrid MFCC, perceptual linear predictive (PLP), and LPC can be explored most importantly for noisy environments. In [1], it was shown that the WER reduced by 30% when AI is explored over the probabilistic approaches even though some studies still prefer to use the conventional HMM in extracting features and subsequent training using machine learning [1], [17], [35]. Other research studies explored deep networks like CNN, RNN for feature extraction and classification of their models [3]. In [2], it was shown that SVM can be a better algorithm to be used for speech recognition, since it has an accuracy of 77.6%, which is 4% better than the HMM [92].

3.3. Discussion: tailored to each research question

The focus of this research is to address the research questions in relation to the evolutionary trend in ASR systems. Based on the results and findings detailed in the Tables 9 to 13, the findings of each question are hereby addressed as follows:

3.3.1. Question 1

When did the development of ASR with AI start? ASR has been an area of research that started as early as the 1950s. In fact most researchers [2], [34], [93] in the field of speech recognition listed this period as the birth of speech recognition. In [3], [17], [25], it was stated that speech recognition started at Bell Laboratory and it was coined Audrey speech recognition [3], [91]. In [91], speech recognition started as early as the 1920s and, in 1920, the first machine to recognize speech, Radio Rex toy was manufactured. However, the incorporation of various AI with ASR varies in years depending on the algorithm used.

According to Anusuya and Katti [91], neural networks were first introduced in the 1950s, but they did not prove useful initially because they had many practical problems. Then the exploration of neural networks started in the 1980s [25]. With MLP, which is a type of deep learning, coming into existence in the 1990s [17], researchers began to explore hybrid HMM+ANN [2] and then the use of DNN in 2010 [4], and an end to end DNN in 2017 [3]. At present, researchers are exploring deep learning architectures like the hybrid LSTM+RNN [2] and transformers [94].

3.3.2. Question 2

What types of AI have been explored by various researchers to develop ASR? Different types of AI techniques have been explored so far to model speech recognition systems. Starting with machine learning architecture like the SVM, k-nearest neighbors (KNN) [95]–[97], to the well-known deep learning architectures like neural networks [6], CNN, RNN, LSTM, SOM, RBF, TDNN [2], transformers [29], SVM-one against one techniques, and SVM-one against all techniques [2]. There are also hybrid mixtures like the

DNN+MFCC and deep learning with backpropagation [1], ANN with SVM [33], CNN with bidirectional long short term memory (BLSTM) and HMM collectively called convolutional long short-term memory deep neural network (CLDNN), and CNN with MLP and HMM [35], HMM with ANN [4], [17], SOM with DWT, SVM with HMM [2], fuzzy neural networks (FNN) and artificial neural fuzzy inference system (ANFIS) [2]. In [97], machine learning models such as decision tree (DT), extra tree (ET), KNN, logistic regression, XGBoost classifier, random forest (RF), and SVM were used to model the bee sound recognition system.

3.3.3. Question 3

What are the benefits and disadvantages of exploring each of those algorithms mentioned in research question 2? There are many benefits and disadvantages of exploring AI. The benefits depend on the type of AI being explored. Researchers tend to explore the AI for classification of the speech features that have been obtained using the statistical approach so as to have an enhanced performance in terms of WER improvement. Others prefer to explore AI for both the feature extraction and classification in modelling ASR systems. For instance in [1], DNN with many hidden layers on HMM has performance improvement on the WER, but the disadvantage is that it is work is rarely successful when it comes to continuous speech signal due to its inability to model temporal dependencies. ANNs overtrain and face the local minima problem, and also ignore the time variability content of the speech signal [2]. However, integrating HMM with ANNs solves these local minima problem, and the system is able to adapt to variability in speech signal. MLP has the inability to handle dynamicity of the input speech signal because they only take input of fixed length. The algorithm can only deal with small vocabularies, which makes them not an efficient word recognizer but a good phoneme recognizer [2]. Deep MLP is good for speech emotion recognition, whereas SOM and DWT integration is good for vowel recognition. A hybrid RBF and wavelet transform is more robust and achieves better results than a system involving only RBF [2]. A transformer outperforms the DNN+HMM based system in large dataset, noisy dataset, low resource dataset, far-field dataset; and also outperforms the RNN-based end-to-end system, however, it is slow in decoding process, hence a faster decoding algorithm must be developed for transformer for its comparison with the DNN+HMM. RNN also outperforms HMM based systems [29]. Hybrid wavelet transform, continuous and discrete hidden Markov models (CDHMM), and FNN in comparison with CDHMM, was proven to be more successful in a noisy environment; by achieving 15.2% more accuracy [98]. SVM cannot take varying inputs as this is the case for speech recognition data, has high computational cost when classifying two classes whereas SVM with RBF kernel has less processing time in the training phase, and also achieves a higher accuracy in comparison to MLP [99]. Considering the one against one SVM, it requires less training data, lower computational cost [100]–[102], however, it develops a relatively high number of binary SVMs [2] but the SVM-one against all has been shown to perform better than the HMM [92]. Other benefits and disadvantages of different AI techniques in speech recognition [103] can be seen in Table 6.

3.3.4. Question 4

Does the exploration of AI for building an ASR system have performance improvement over the traditional probabilistic approaches? Considering the results highlighted in Table 7, it can be seen that the exploration of AI has performance improvement over the probabilistic approaches. Using MFCC [104] with DNN has the WER reduced by 30% in comparison to the state-of-the-art models based on Gaussian mixtures and there is also advancement in speech spectrogram features [1]. ANN with SVM can be employed independently or as a hybrid model with HMM to obtain optimal results of ASR [2]. CNN with MLP and HMM is robust to noise, thereby improving on one of the variabilities issues in ASR, it has performance improvement over both the probabilistic approach and the hybrid MFCCs/PLP+ANN+HMM [35], and DNN with HMM produces significant error reduction [4], [105].

3.3.5. Question 5

Has the introduction of AI in the development and modelling of ASR been able to achieve it is ultimate goal to perform as a human listener of ASR? Based on findings, lots of research works are ongoing, particularly those involving deep learning architecture aimed at making speech recognition system [106] to be more robust through development of techniques or using various speech feature extraction techniques like the wavelet transform, and RASTA filtering [107], [108]. Other speech extraction feature techniques developed for robust ASR systems are ZCPA, average localized synchrony detection (ALSD), perceptual minimum variance distortion less response (PMVDR), power-normalized cepstral coefficients (PNCC), invariant integration features (IIF), amplitude modulation spectrogram (AMS), Gammatone frequency cepstral coefficients (GFCC), sparse auditory reproducing kernel (SPARK), and Gabor filter bank features [17], [34], [109]. Also, CNN with many hidden layers has gain in ASR [110], [111] performance due to the large number of hidden layers explored [35]. In addition to all these is the exploration of transformers based architecture or integration of transformer with DNN [112] which is aimed at achieving the ultimate goal of ASR [6], [7].

4. CONCLUSION

This current study investigated the evolutionary trend of AI in the modelling of ASR systems. Thorough literature reviews by extracting information from various studies have brought light to the research questions being addressed in this work, starting from time when ASR came into limelight to the beginning of the era of the exploration of AI, and to the research question considering whether the ultimate goal of speech recognition systems has been achieved. The findings of this review show that robust systems are currently being developed to solve the noise variability in speech signals. Researchers have explored various approaches including using hybrid systems involving the well-known statistical approach like HMM, GMM particularly for feature extraction, and deep learning architectures for classification or training. Other researchers consider modelling to be completely DL based while some prefer to use deep learning like ANN for feature extraction and HMM for classification in extracting the word sequence equivalent of speech features, and some use transformer-based architectures. Lots of ASR systems like the Whisper, Jasper, and other convolutional network architecture-based speech recognition systems have been developed by various researchers to enhance the performance of the speech recognition system. With more speech data made available for the training of the systems, and development of various robust speech feature extraction techniques, robust speech recognition systems would be developed thereby achieving the ultimate goal of speech recognition system.

LIMITATIONS

The limitations encountered in this work are the inability to access some papers due to charges, especially the IEEE papers, and some other papers from other journal sites whose papers can only be accessed after payments. However, we are of the opinion that the 82 papers utilized for this study is considered adequate.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the Covenant University Center for Research, Innovation and Discovery (CUCRID) and Google for providing funds towards the publication of this study

FUNDING INFORMATION

The publication fee of this paper was paid by the Department of Mathematical Sciences, School for Data Science and Computational Thinking, Stellenbosch University, South Africa, while no further funding was spent for the review paper development.

AUTHOR CONTRIBUTIONS STATEMENT

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

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Gabriel Oluwatobi Sobola	✓	✓			✓	✓	✓	✓	✓	✓	✓			
Emmanuel Adetiba	✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓
Olabode Idowu-Bismark		✓		✓	✓	✓	✓			✓	✓	✓	✓	
Abdultaofeek Abayomi		✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	
Raymond Jules Kala		✓		✓	✓	✓	✓			✓	✓	✓	✓	
Surendra Colin Thakur		✓		✓	✓	✓	✓			✓	✓	✓	✓	
Sibusiso Moyo		✓		✓	✓	✓	✓			✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors hereby declare that there is no known conflict of interest concerning this review paper.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The authors wish to state that no further data is available to support the findings of this research work. The known data for this work had been provided in form of tables and figures.

REFERENCES

[1] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: a systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019, doi: 10.1109/ACCESS.2019.2896880.

[2] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 9411–9457, 2021, doi: 10.1007/s11042-020-10073-7.

[3] P. Dubey and B. Shah, "Deep speech based end-to-end automated speech recognition (ASR) for Indian-English accents," *arXiv:2204.00977*, 2022.

[4] X. Huang, J. Baker, and R. Reddy, "A historical perspective of speech recognition," *Communications of the ACM*, vol. 57, no. 1, pp. 94–103, 2014, doi: 10.1145/2500887.

[5] A. Hagen and A. Morris, "Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR," *Computer Speech and Language*, vol. 19, no. 1, pp. 3–30, 2005, doi: 10.1016/j.csl.2003.12.002.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of Machine Learning Research*, 2023, vol. 202, pp. 28492–28518.

[7] J. Li *et al.*, "Jasper: an end-to-end convolutional neural acoustic model," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 71–75, 2019, doi: 10.21437/Interspeech.2019-1819.

[8] M. Bansal and T. K. Thivakaran, "Analysis of speech recognition using convolutional neural network," *Journal of Engineering Sciences*, vol. 11, no. 1, pp. 285–291, 2020.

[9] J. L. K. E. Fendji, D. C. M. Tala, B. O. Yenke, and M. Atemkeng, "Automatic speech recognition using limited vocabulary: a survey," *Applied Artificial Intelligence*, vol. 36, no. 1, 2022, doi: 10.1080/08839514.2022.2095039.

[10] I. Kipyatkova and I. Kagirov, "Deep models for low-resourced speech recognition: Livvi-Karelian case," *Mathematics*, vol. 11, no. 18, 2023, doi: 10.3390/math11183814.

[11] S. Dua *et al.*, "Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network," *Applied Sciences*, vol. 12, no. 12, 2022, doi: 10.3390/app12126223.

[12] F. S. Al-Anzi and D. AbuZeina, "Synopsis on Arabic speech recognition," *Ain Shams Engineering Journal*, vol. 13, no. 2, 2022, doi: 10.1016/j.asej.2021.06.020.

[13] N. Z.-Morales, P. Pancardo, J. A. H.-Nolasco, and M. G.-Constantino, "Attention-inspired artificial neural networks for speech processing: a systematic review," *Symmetry*, vol. 13, no. 2, pp. 1–43, 2021, doi: 10.3390/sym13020214.

[14] S. Basak *et al.*, "Challenges and limitations in speech recognition technology: a critical review of speech signal processing algorithms, tools and systems," *CMES-Computer Modeling in Engineering and Sciences*, vol. 135, no. 2, pp. 1053–1089, 2023, doi: 10.32604/cmes.2022.021755.

[15] W. Du, Y. Maimaitiyiming, M. Nijat, L. Li, A. Hamdulla, and D. Wang, "Automatic speech recognition for Uyghur, Kazakh, and Kyrgyz: an overview," *Applied Sciences*, vol. 13, no. 1, 2023, doi: 10.3390/app13010326.

[16] H. Kheddar, Y. Himeur, S. Al-Maadeed, A. Amira, and F. Bensaali, "Deep transfer learning for automatic speech recognition: towards better generalization," *Knowledge-Based Systems*, vol. 277, 2023, doi: 10.1016/j.knosys.2023.110851.

[17] S. Karpagavalli and E. Chandra, "A review on automatic speech recognition architecture and approaches," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 9, no. 4, pp. 393–404, 2016, doi: 10.14257/ijsip.2016.9.4.34.

[18] D. K. Dansena and Y. Rathore, "A survey paper on automatic speech recognition by machine," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 3, 2015.

[19] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2006, doi: 10.1016/j.specom.2006.04.003.

[20] S. Latif *et al.*, "Transformers in speech processing: a survey," *arXiv:2303.11607*, 2025.

[21] A. V. H. and R. Marimuthu, "Speech recognition using Taylor-gradient descent political optimization based deep residual network," *Computer Speech and Language*, vol. 78, 2023, doi: 10.1016/j.csl.2022.101442.

[22] M. U. Nemade and P. S. K. Shah, "Survey of soft computing based speech recognition techniques for speech enhancement in multimedia applications," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 2, no. 5, pp. 2039–2043, 2013.

[23] M. U. Hadi *et al.*, "Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects," *TechRxiv*, pp. 1-54, Nov. 2023, doi: 10.36227/techrxiv.23589741.

[24] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," in *18th International Conference on Evaluation and Assessment in Software Engineering*, 2014, pp. 1–10, doi: 10.1145/2601248.2601268.

[25] D. O'Shaughnessy, "Invited paper: automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965–2979, 2008, doi: 10.1016/j.patcog.2008.05.008.

[26] V. T. Pham *et al.*, "Independent language modeling architecture for end-to-end ASR," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, 2020, pp. 7059–7063, doi: 10.1109/ICASSP40776.2020.9054116.

[27] S. Benkerzaz, Y. Elmir, and A. Dennai, "A study on automatic speech recognition," *Journal of Information Technology Review*, vol. 10, no. 3, pp. 77–85, 2019, doi: 10.6025/jitr/2019/10/3/77-85.

[28] Z. Yue, H. Christensen, and J. Barker, "Autoencoder bottleneck features with multi-task optimisation for improved continuous dysarthric speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2020, pp. 4581–4585, doi: 10.21437/Interspeech.2020-2746.

[29] S. Karita *et al.*, "A comparative study on transformer vs RNN in speech applications," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019*, 2019, pp. 449–456, doi: 10.1109/ASRU46091.2019.9003750.

[30] N. Desai, K. Dhameliya, and V. Desai, "Feature extraction and classification techniques for speech recognition : a review," *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 12, pp. 1–5, 2013.

[31] P. G. Noe, T. Parcollet, and M. Morchid, “CGCNN: complex gabor convolutional neural network on raw speech,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7724–7728, doi: 10.1109/ICASSP40776.2020.9054220.

[32] S. K. Saksamudre, P. P. Shrishrimal, and R. R. Deshmukh, “A Review on different approaches for speech recognition system,” *International Journal of Computer Applications*, vol. 115, no. 22, pp. 23–28, 2015, doi: 10.5120/20284-2839.

[33] M. Cutajar, E. Gatt, I. Grech, O. Casha, and J. Micallef, “Comparative study of automatic speech recognition techniques,” *IET Signal Processing*, vol. 7, no. 1, pp. 25–46, 2013, doi: 10.1049/iet-spr.2012.0151.

[34] S. K. Gaikwad, B. W. Gawali, and P. Yannawar, “A review on speech recognition technique,” *International Journal of Computer Applications*, vol. 10, no. 3, pp. 16–24, 2010, doi: 10.5120/1462-1976.

[35] D. Palaz, M. M.-Doss, and R. Collobert, “End-to-end acoustic modeling using convolutional neural networks for HMM-based automatic speech recognition,” *Speech Communication*, vol. 108, pp. 15–32, 2019, doi: 10.1016/j.specom.2019.01.004.

[36] H. Singh and A. K. Bathla, “A survey on speech recognition,” *International Journal of Advanced Research in Computer Engineering and Technology*, vol. 2, no. 2, pp. 2186–2189, 2013.

[37] J. B. Hampshire and A. H. Waibel, “A novel objective function for improved phoneme recognition using time-delay neural networks,” *IEEE Transactions on Neural Networks*, 1990, vol. 1, no. 2, pp. 216–228, doi: 10.1109/72.80233.

[38] R. Collobert and S. Bengio, “SVMTorch: support vector machines for large-scale regression problems,” *Journal of Machine Learning Research*, vol. 1, no. 2, pp. 143–160, 2001.

[39] L. Deng, G. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: an overview,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8599–8603, doi: 10.1109/ICASSP.2013.6639344.

[40] L. Muda, M. Begam, and I. Elamvazuthi, “Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques,” *Journal of Computing*, vol. 2, no. 3, 2010.

[41] L. T. Niles and H. F. Silverman, “Combining hidden Markov model and neural network classifiers,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing-Proceedings*, 1990, pp. 417–420, doi: 10.1109/icassp.1990.115724.

[42] S. Bhatt, A. Jain, and A. Dev, “Acoustic modeling in speech recognition: a systematic review,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 397–412, 2020, doi: 10.14569/IJACSA.2020.0110455.

[43] Y. Zhang, “Speech recognition using deep learning algorithms,” *Machine Learning Final Project,* vol. 229, 2013.

[44] D. Palaz, “Towards end-to-end speech recognition,” *Thesis*, Department of Electrical Engineering, Swiss Federal Institute of Technology Lausanne, Lausanne, Swiss, 2016.

[45] I. Shahin, A. B. Nassif, and S. Hamsa, “Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emotional talking environments,” *Neural Computing and Applications*, vol. 32, no. 7, pp. 2575–2587, 2020, doi: 10.1007/s00521-018-3760-2.

[46] S. S. Bhabad and G. K. Kharate, “An overview of technical progress in speech recognition,” *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 3, pp. 2277–128, 2013.

[47] D. Palaz, R. Collobert, and M. M. -Doss, “End-to-end phoneme sequence recognition using convolutional neural networks,” *Idiap Research Report*, pp. 1–8, 2013.

[48] L. Pipiras, R. Maskeliūnas, and R. Damaševičius, “Lithuanian speech recognition using purely phonetic deep learning,” *Computers*, vol. 8, no. 4, 2019, doi: 10.3390/computers8040076.

[49] M. T. Singh, A. R. Fayjie, and B. Kachari, “A survey report on speech recognition system,” *International Journal of Computer Applications*, vol. 121, no. 11, pp. 1–3, 2015, doi: 10.5120/21581-4672.

[50] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993, doi: 10.1201/9781003348689-5.

[51] D. Palaz, R. Collobert, and M. M.-Doss, “Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks,” *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1766–1770, 2013, doi: 10.21437/interspeech.2013-438.

[52] R. Mehla and R. Aggarwal, “Automatic speech recognition: a survey,” *International Journal of Advanced Research in Computer Science and Electronics Engineering*, vol. 3, no. 1, pp. 45–53, 2014.

[53] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978, doi: 10.1109/TASSP.1978.1163055.

[54] D. Palaz, M. M.-Doss, and R. Collobert, “Learning linearly separable features for speech recognition using convolutional neural networks,” in *3rd International Conference on Learning Representations, ICLR 2015-Workshop Track Proceedings*, 2015.

[55] E. Trentin and M. Gori, “A survey of hybrid ANN/HMM models for automatic speech recognition,” *Neurocomputing*, vol. 37, no. 1–4, pp. 91–126, 2001, doi: 10.1016/S0925-2312(00)00308-8.

[56] J. Baker, “The DRAGON system-an overview,” *IEEE Transactions on Acoustics, speech, and signal Processing*, vol. 23, no. 1, pp. 24–29, 1975.

[57] D. Palaz, M. M.-Doss, and R. Collobert, “Analysis of CNN-based speech recognition system using raw speech as input,” 2015, doi: 10.21437/interspeech.2015-3.

[58] V. Kherdekar and S. Naik, “Speech recognition of mathematical words using deep learning,” in *Communications in Computer and Information Science*, 2021, pp. 356–362, doi: 10.1007/978-981-16-0493-5_31.

[59] R. L. Klevans and R. D. Rodman, *Voice recognition*. Artech House, Inc., 1997.

[60] D. Palaz, M. M.-Doss, and R. Collobert, “Convolutional neural networks-based continuous speech recognition using raw speech signal,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 4295–4299, doi: 10.1109/ICASSP.2015.7178781.

[61] D. Joshi, P. Waso, R. Shelke, S. Jadhav, K. Bhale, and A. Padalkar, “Automatic speech recognition using acoustic modeling,” in *Advances in Data Science and Computing Technologies (ADSC 2022)*, 2023, pp. 109–119, doi: 10.1007/978-981-99-3656-4_11.

[62] K. R. Lekshmi and S. Elizabeth, “Automatic speech recognition using different neural network architectures – a survey,” *International Journal of Computer Science and Information Technologies*, vol. 7, no. 6, pp. 2422–2427, 2016.

[63] L. R. Bahl, P. F. Brown, P. V. De Souza, and R. L. Mercer, “Estimating hidden markov model parameters so as to maximize speech recognition accuracy,” *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 1, pp. 77–83, 1993, doi: 10.1109/89.221369.

[64] B. Yegnanarayana and R. N. J. Veldhuis, “Extraction of vocal-tract system characteristics from speech signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 313–327, 1998, doi: 10.1109/89.701359.

[65] X. Tang, “Hybrid hidden Markov model and artificial neural network for automatic speech recognition,” in *2009 Pacific-Asia Conference on Circuits, Communications and System, PACCS 2009*, 2009, pp. 682–685, doi: 10.1109/PACCS.2009.138.

[66] A. M. Peinado, J. C. Segura, A. J. Rubio, P. Garcfa, and J. L. Pérez, “Discriminative codebook design using multiple vector quantization in HMM-based speech recognizers,” *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 2, pp. 89–95, 1996, doi: 10.1109/89.486058.

[67] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2018, pp. 5884–5888, doi: 10.1109/ICASSP.2018.8462506.

[68] A. Agarwal and T. Zesch, "German end-to-end speech recognition based on DeepSpeech," in *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, 2019, pp. 111-119.

[69] S. M. Lakew, M. Cettolo, and M. Federico, "A comparison of transformer and recurrent neural networks on multilingual neural machine translation," *COLING 2018-27th International Conference on Computational Linguistics*, pp. 641–652, 2018.

[70] D. Wang, Y. Wei, K. Zhang, D. Ji, and Y. Wang, "Automatic speech recognition performance improvement for mandarin based on optimizing gain control strategy," *Sensors*, vol. 22, no. 8, 2022, doi: 10.3390/s22083027.

[71] C. H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Communication*, vol. 25, no. 1–3, pp. 29–47, 1998, doi: 10.1016/S0167-6393(98)00028-4.

[72] S. Zhou, L. Dong, S. Xu, and B. Xu, "Syllable-based sequence-to-sequence speech recognition with the transformer in Mandarin Chinese," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 791–795, doi: 10.21437/Interspeech.2018-1107.

[73] T. Yu *et al.*, "Automatic speech recognition datasets in cantonese: a survey and new dataset," *2022 Language Resources and Evaluation Conference, LREC 2022*, pp. 6487–6494, 2022.

[74] X. Chen, A. Eversole, G. Li, D. Yu, and F. Seide, "Pipelined back-propagation for context-dependent deep neural networks," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[75] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975, doi: 10.1109/TASSP.1975.1162641.

[76] M. A. Anusuya and S. K. Katti, "Front end analysis of speech recognition: a review," *International Journal of Speech Technology*, vol. 14, no. 2, pp. 99–145, 2011, doi: 10.1007/s10772-010-9088-7.

[77] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012, doi: 10.1109/MSP.2012.2205597.

[78] J. Barker, P. Green, and M. Cooke, "Linking auditory scene analysis and robust ASR by missing data techniques," *Proceedings-Institute of Acoustics*, vol. 23, no. 3, pp. 295–308, 2024, doi: 10.25144/21801.

[79] B. Gamulkiewicz and M. Weeks, "Wavelet based speech recognition," in *Midwest Symposium on Circuits and Systems*, 2003, vol. 2, pp. 678–681, doi: 10.1109/MWSCAS.2003.1562377.

[80] W. Phatthiyaphaibun, C. Chaksangchaichot, P. Limkonchotiwat, E. Chuangsuwanich, and S. Nutanong, "Thai Wav2Vec2.0 with CommonVoice V8," *arXiv:2208.04799*, 2022.

[81] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," in *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2012, vol. 1, pp. 10–13, doi: 10.21437/interspeech.2012-3.

[82] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *6th International Conference on Spoken Language Processing, ICSLP 2000*, 2000, pp. 373–376, doi: 10.21437/icslp.2000-92.

[83] P. K. Chen, H. M. Wang, B. J. Huang, C. T. Chen, and J. C. Wang, "Enhancing automatic speech recognition performance through multi-speaker text-to-speech," in *ROCLING 2023-Proceedings of the 35th Conference on Computational Linguistics and Speech Processing*, 2023, pp. 370–375.

[84] Z. J. Yan, Q. Huo, and J. Xu, "A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013, pp. 104–108, doi: 10.21437/interspeech.2013-47.

[85] S. Bengio, "An asynchronous hidden Markov model for audio-visual speech recognition," in *NIPS 2002: Proceedings of the 15th International Conference on Neural Information Processing Systems*, 2002, vol. 15, pp. 1213–1220.

[86] A. T. Ali, H. Abdullah, and M. N. Fadhil, "Speaker recognition system based on mel frequency cepstral coefficient and four features," *Iraqi Journal of Computer, Communication, Control and System Engineering*, vol. 21, no. 4, pp. 82–89, 2021, doi: 10.33103/uot.ijccce.21.4.8.

[87] N. Morgan and H. Bourlard, "Continuous speech recognition using multilayer perceptrons with hidden Markov models," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 1990, vol. 1, pp. 413–416, doi: 10.1109/icassp.1990.115720.

[88] S. Becker, S. Thrun, and K. Obermayer, *Advances in neural information processing systems 15 - Proceedings of the 2002 Conference, NIPS 2002*, vol. 15. MIT Press, 2003.

[89] M. Malik and R. Khanam, "The state of the art on ASR systems and feature extraction technique," in *7th International Conference on Computing in Engineering & Technology (ICCET 2022), IET*, 2022, doi: 10.1049/icp.2022.0594.

[90] A. Saxena, A. K. Sinha, S. Chakrawarti, and S. Charu, "Speech recognition using MATLAB," *International Journal of Advances in Computer Science and Cloud Computing*, vol. 1, no. 2, pp. 26–30, 2013.

[91] M. A. Anusuya and S. K. Katti, "Speech recognition by machine, a review," *arXiv:1001.2267*, Jan. 2010.

[92] J. Galić, B. Popović, and D. Š. Pavlović, "Whispered speech recognition using hidden Markov models and support vector machines," *Acta Polytechnica Hungarica*, vol. 15, no. 5, pp. 11–29, 2018, doi: 10.12700/APH.15.5.2018.5.2.

[93] Y. Kökver, H. M. Pektaş, and H. Çelik, "Artificial intelligence applications in education: natural language processing in detecting misconceptions," *Education and Information Technologies*, vol. 30, no. 3, pp. 3035–3066, 2025, doi: 10.1007/s10639-024-12919-1.

[94] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, 2023, doi: 10.1016/j.inffus.2023.101869.

[95] S. Watve, M. Patil, and S. Thuse, "Indian language recognition using SVM and KNN," in *2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET)*, Jan. 2025, pp. 1–5, doi: 10.1109/ICAET63349.2025.10932166.

[96] S. H. Gill *et al.*, "Prosodic information extraction and classification based on MFCC features and machine learning models," *Measurement and Control*, vol. 59, no. 1, Jan. 2025, doi: 10.1177/00202940251315031.

[97] T. T. H. Phan, D. N.-Doan, D. N.-Huu, H. N.-Van, and T. P.-Hong, "Investigation on new mel frequency cepstral coefficients features and hyper-parameters tuning technique for bee sound recognition," *Soft Computing*, vol. 27, no. 9, pp. 5873–5892, 2023, doi: 10.1007/s00500-022-07596-6.

[98] Y. Meyer and R. D. Ryan, "Wavelets: algorithms & applications," *Philadelphia: SIAM (Society for Industrial and Applied Mathematics*, 1993.

[99] Y. Wang, S. Wang, and K. K. Lai, "A new fuzzy support vector machine to evaluate credit risk," *IEEE Transactions on Fuzzy Systems*, vol. 13, no. 6, pp. 820–831, 2005, doi: 10.1109/TFUZZ.2005.859320.

[100] H. A. Mait and N. Aboutabit, “HMM-GMM acoustic modeling for Arabic speech recognition system,” *Proceedings of the Third ICMDS'24: Machine Learning, Inverse Problems and Related Fields (ICMDS 2024)*, pp. 1–13, 2025, doi: 10.1007/978-3-031-94802-2_1.

[101] D. A. Kumar, K. V. N. Nadh, and M. S. Chowdary, “The integration of target speaker voice activity detection with transformers and end-to-end neural networks,” in *Computer Vision and Robotics: Proceedings of CVR 2024*, 2025, pp. 41–54, doi: 10.1007/978-981-97-8868-2_4.

[102] S. Kamal *et al.*, “Vision sensor for automatic recognition of human activities via hybrid features and multi-class support vector machine,” *Sensors*, vol. 25, no. 1, Jan. 2025, doi: 10.3390/s25010200.

[103] I. Samuel, F. A. Ogunkeye, A. Olajube, and A. Awelewa, “Development of a voice chatbot for payment using amazon lex service with eyowo as the payment platform,” in *2020 International Conference on Decision Aid Sciences and Application, DASA 2020*, 2020, pp. 104–108, doi: 10.1109/DASA51403.2020.9317214.

[104] L. K. Ajayi, A. A. Azeta, I. A. Odun-Ayo, F. C. Chidozie, and A. E. Azeta, “Systematic review on speech recognition tools and techniques needed for speech application development,” *International Journal of Scientific and Technology Research*, vol. 9, no. 3, pp. 6997–7007, 2020.

[105] S. S. Reddy, S. K. A. Mnoj, and K. P. Rao, “DNNT (deep neural network for Telugu): a framework for speech recognition of Telugu language with parallel computing approach,” *International Journal of Speech Technology*, vol. 28, no. 2, pp. 341–349, Jun. 2025, doi: 10.1007/s10772-025-10186-0.

[106] A. Abdulkareem, T. E. Somefun, O. K. Chinedum, and F. Agbetuyi, “Design and implementation of speech recognition system integrated with internet of things,” *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, pp. 1796–1803, Apr. 2021, doi: 10.11591/ijece.v11i2.pp1796-1803.

[107] D. D. Armaisya, P. D. Pamungkasari, A. P. Rifai, I. D. Sholihati, and Gopal Sakarkar, “Comparison of feature extraction techniques for long short-term memory models in Indonesian automatic speech recognition,” *Green Intelligent Systems and Applications*, vol. 5, no. 1, pp. 74–92, Apr. 2025, doi: 10.53623/gisa.v5i1.605.

[108] S. Paul, V. Bhattacharjee, and S. K. Saha, “An end-to-end continuous speech recognition system in bengali for general and elderly domain,” *SN Computer Science*, vol. 6, no. 5, Jun. 2025, doi: 10.1007/s42979-025-04058-2.

[109] K. Murugesan *et al.*, “Homomorphic encryption, privacy-preserving feature extraction, and decentralized architecture for enhancing privacy in voice authentication,” *International Journal of Electrical and Computer Engineering*, vol. 15, no. 2, pp. 2150–2160, Apr. 2025, doi: 10.11591/ijece.v15i2.pp2150-2160.

[110] D. Jurafsky and J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition (3rd ed.)*, Pearson/Prentice Hall, 2020.

[111] D. Pandey, “Speech recognition using soft computing,” *International Journal of Advanced Research in Computer Science*, vol. 13, no. 1, pp. 11–15, 2020, doi: 10.26483/ijarcs.v13i1.6797.

[112] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5455–5516, 2020, doi: 10.1007/s10462-020-09825-6.

APPENDIX

Table 9. Results and findings regarding research questions 1

S/N	Authors and year of publication	Technology used at starts of various ASR technologies and AI	Starting year of each of the technology
1	Nassif *et al.* [1], 2019	- Deep learning for speech applications. - Speech received more interest right from the introduction of AI.	6 years Past 5 decades
2	Malik *et al.* [2], 2021	- Development of ASR started [93]. - Development of system to distinguish between words by Russian scientist [107]. - Introduction of n-gram model. - Hybrid HMM and a feed-forward ANN. - LSTM, a type of RNN plus various deep learning techniques are being used [108]	1950s 1970 Late 1980s Early 2000s Currently in use
3	Dubey and Shah [3], 2022	- Development of Audrey speech recognition system by Bell Laboratories Researchers. - Deep speech, a state-of-the-art speech recognition system developed using end-to-end deep learning.	1952 2017
4	Huang *et al.* [4], 2014	DNN.	2010
5	Saksamudre *et al.* [32], 2015	Research in various ways and means to make computer record, interpret and understand human speech.	Since 1960
6	Desai *et al.* [30], 2013	Development of ASR started.	Past 60 years
7	Karpagavalli and Chandra [17], 2016	- Great amount of research has been done on speech recognition and its application. - Pattern-matching approach (HMM and DTW) became predominant. - Popularity of MLP with the SoftMax nonlinear function at the final layer. - Acoustic modelling approach: neural networks trained by back-propagation error derivatives.	More than 3 decades. Past 6 decades. 1990s 1980s
8	Gaikwad *et al.* [34], 2010	- Deciphering of speech started. - Computer scientists have been researching ways on human speech recognition by computers. - First attempt in speech recognition at Bell Laboratories.	1980s 1960s 1950s

Table 9. Results and findings regarding research questions 1 *(continued)*

S/N	Authors and year of publication	Technology used at starts of various ASR technologies and AI	Starting year of each of the technology
9	O'Shaughnessy [25], 2008	- Small-vocabulary recognition for digits spoken over the telephone at the Bell Laboratory.	1952
		- Filter banks were combined with dynamic programming.	1960s
		- LPC.	1969
		- Dynamic time warping (DTW); HMM.	1970s; 1975
		- MFCC; language models.	1980; 1980s
		- Neural networks; wavelet transform.	1980s; 1990
		- Kernel-based classifiers.	1998
		- Dynamic Bayesian networks.	1999
10	Anusuya and Katti [91], 2009	- Machine recognition of speech came into existence.	Early 1920s
		- The first machine to recognize speech, Radio Rex (toy) was manufactured.	1920
		- Bell Labs demonstrated a speech synthesis machine (which simulates talking) at the World Fair in New York; Bell laboratories, built a system for isolated digit recognition for a single speaker.	1939; 1952
		- The pioneering research of Reddy in the field of continuous speech recognition by dynamic tracking of phonemes.	1960s
		- A shift from template-based approach to statistical modelling methods involving the HMM.	1980s
		- Neural networks were first introduced, but they did not prove useful initially because they had many practical problems.	1950s

Table 10. Findings of research questions 2

S/N	Authors and year of publication	Types of AI	Features of the AI	Applications
1	Nassif *et al.* [1], 2019	1. DNN, DNN+HMM; 2. DNN+MFCC; 3. MLP; 4. Deep learning with backpropagation; 5. MFCCs, residual connection, PLP, bark frequency cepstrum coefficients (BFCC), batch normalization, linear discriminate analysis (LDA) transform, heteroscedastic linear discriminant analysis transform (HLDA) transform, maximum likelihood linear transform (MLLT), short time Fourier transform (STFT), and log power spectral (LPS), LPC; 6. Hybrid DNN model involving unsupervised (generative model) and supervised (discriminative model); 7. Recursive neural network (RNN) with hessian free optimization; and 8. LSTM.	1. Large number of hidden layers.	1. Training of the network; 3. Speech spectrogram features; 5. Feature extraction from the speech signals; and 8. Ability to generate sequence of text characters.
2.	Malik *et al.* [2], 2021	1. Spectral features (MFCC, PLP, and LPC, linear predictive cepstral coefficients (LPCC)), temporal features (DWT or discrete wavelet packet transform (DWPT) or RASTA-PLP); 2. Hybrid models: - LPCC+RBF+HMM - CNN+BLSTM+batch normalization+ ReLU - FNN wavelet transforms - ANFIS and (SOM+vector quantization (VQ)) ≥for clustering - SVM+HMM [100], [101] - SVM (one-against all techniques)+RBF Kernel 3. Non hybrid models: - HMM+Wavelet Transform; fast Fourier transform (FFT)+MFCC+MLP, - MFCC+DWT+MLP; sparse MLP (SMLP) - SOM+DWT=Wavelet; SOM (WSOM), - MFCC+LPC+ SOM; RBF+LPCC - Wavelet transformation+RBF; - Temporal RBF; RNN or bidirectional RNN - LSTM; CNN - CNN+three different types of input, which - Included MFCC, power spectrum, and raw wave - Format; CDHMM; FNN; SVM [95] 4. SVM (one-against one techniques)+MFCC	1. Spectral features methods are suitable for clean data, while temporal feature methods are fit for noisy speech data. 2. Can withstand noisy speech signals, the membership function proves to be very useful to map speech signals, as they have no clear boundaries.	1. Feature extraction from speech signals. 2. For feature extraction and classification. 3. For feature extraction and classification.

Table 10. Findings of research questions 2 *(continued)*

S/N	Authors and year of publication	Types of AI	Features of the AI	Applications
3.	Dubey and Shah [3], 2022	1. RNN training system utilizing multiple graphical processing units (GPUs) [Deep Speech]; 2. HMM, N-Gram; 3. RNN; 4. CNN; and 5. ANN	2. The model's parameters consist of state transition probabilities along with the means, variances, and mixture weights that define the output distribution of each state.	5. Classification and recognition of static patterns is the main advantage of ANNs.
4	Huang *et al.* [4], 2014	1. HMM+Gaussian density; 2. HMM+DNN; and 3. FFT, filter bank: for feature extraction.		
5	Saksamudre *et al.* [32], 2015	1. Hamming rectangular, Blackman, Welch or Gaussian; 2. MFCC [36], LPCC, PLP, wavelet, RASTA-PLP (relative spectral transform), principal component analysis (PCA), LDA, Independent component analysis (ICA), LPC, filter bank analysis, kernel-based feature extraction method, cepstral mean subtraction (CMS); 3. HMM; and 4. Bi-gram, tri-gram, n-gram.		1. To perform window functions; 2. For feature extraction in speech; 3. For acoustic modelling; and 4. For language modelling.
6	Desai *et al.* [30], 2013	1. LPC, MFCC, AMFCC, PLP, PCA, cepstral analysis; and 2. HMM		1. For feature extractions; 2. Pattern recognition approach for classifying speech signal features.
7	Karpagavalli and Chandra [17], 2016	1. PCA, LDA, ICA, (LPC, cepstral analysis, mel-frequency scale analysis, filter-bank analysis, MFCC, kernel-based feature extraction, dynamic feature extraction, wavelet-based features, spectral subtraction, and CMS; 2. Noise robust speech recognition: ZCPA, ALSD, PMVDR, PNCC, IIF, AMS, GFCC, SPARK, and Gabor filter bank features are effectively applied [37]; 3. Segmental models, super-segmental models (including hidden dynamic models), neural networks, maximum entropy models, and (hidden) conditional random fields; 4. CMU statistical language modelling (SLM) toolkit, Stanford Research Institute Language Modelling toolkit; 5. Viterbi algorithm, beam search, extended Viterbi and forward-backward algorithms [110]; 6. HMM; 7. DTW; 8. HMM+GMM; 9. HMM-ANN: e.g. HMM+MLP; and 10. HMM+DNN e.g. deep auto-encoders, deep Boltzmann machine, sum-product networks, the original form of deep belief network (DBN) and its extension to the factored higher-order Boltzmann machine in its bottom layer, deep-structured CRF, tandem-MLP architecture, deep convex or stacking network and it is tensor version, and detection-based ASR architecture.	8. The HMM is characterized by the initial probability, transition probability and emission probability.	1. For feature extraction; 2. For feature extraction of noisy data; 3. For acoustic modelling of speech signals. 4. Toolkits for ASR system; 5. Decoding methods of the ASR system; 6. Stochastic pattern matching; 7. Deterministic pattern matching; 8. For a generative learning approach; 9. For a discriminative learning approach; and 10. For deep learning approach.
8	Gaikwad *et al.* [34], 2010	1. HMM+learning VQ (LVQ); 2. LPC; 3. RASTA filtering; 4. Spectral subtraction; 5. Wavelet; 6. PCA; 7. Cepstral analysis; and 8. Dynamic feature extraction: LPC, MFCCs	2. It has 10-16 lower order coefficients; 3. For noisy speech; 4. Robust feature extraction methods; 5. Better time resolution than Fourier transform; 6. Eigen vector based, fast, nonlinear feature extraction methods, it has a linear map; 7. Static feature extraction method; 8. Acceleration and delta coefficients (II, III) order derivatives of normal LPC and MFCCs coefficients.	1. It produces highly discriminative reference vectors for the classification of static patterns; 2. For feature extraction; 3. Same as above; 4. Same as above; 5. Same as above; 6. Same as above; 7. Same as above; and 8. Same as above.

Table 10. Findings of research questions 2 *(continued)*

S/N	Authors and year of publication	Types of AI	Features of the AI	Applications
9	O'Shaughnessy [25], 2008	1. Filter banks were combined with dynamic programming; 2. LPC; 3. DTW; 4. HMM; 5. Mel-frequency cepstrum; 6. Language models; 7. Neural networks; 8. Wavelet transform; 9. Kernel-based classifiers; and 10. Dynamic Bayesian networks		1. For feature extraction; 2. For feature extraction; 3. For classification; 4. For classification; 5. For feature extraction; 6. For stochastic description of text likelihood; 7. For classification; 8. For feature extraction; 9. For classification; and 10. For classification
10	Palaz *et al*. [35], 2019	1. CNN; 2. CNN+BLSTM+DNN (CLDNN); 3. MFCCs/PLP+ANN+HMM; 4. CNN+MLP+HMM; 5. Filter bank or critical band energies; 6. Short-term magnitude spectrum features; and 7. Improved MFFCs: MFCC+LDA+MLLT+FMLLR	1. It consists of several convolution layers and classifier stage consisting of MLP.	1. Used for feature extraction, classification; 2. Used ASR modelling; 3. For ASR modelling; 4. For ASR modelling; 5. For feature extraction; 6. Same as above; and 7. Same as above
11	Karita *et al*. [29], 2019	Transformer	Ability to withstand noise, and it can train well both in small and large resource data.	For developing ASR systems.
12	Cutajar *et al*. [33], 2013	1. HMM+GMM; and 2. ANN+SVM	1. They are generative approach; and 2. They are discriminative approach	1. For modelling ASR systems; and 2. For modelling ASR systems.
13	Pandey [111], 2022	Soft computing approach (GA)	For the preparation of ANN.	To obtain a more precise and optimal arrangement.
14	Anusuya and Katti [91], 2009	1. PCA; 2. Mel-frequency cepstrum (MFFCs); and 3. The maximum likelihood linear regression (MLLR), the model decomposition, parallel model composition (PMC), and the structural maximum a posterior (SMAP) method	1. Non-linear feature extraction method, Linear map; fast; eigenvector-based; and 2. Power spectrum is computed by performing Fourier analysis	1. Traditional, eigenvector based method, also known as Karhuneu-Loeve expansion; good for Gaussian data, for feature extraction; 2. For extracting feature vectors in speech signals; and 3. For robust speech recognition.
15	Murugesan [109], 2014	1. Feature-domain vs. model-domain compensation; 2. Compensation using prior knowledge about acoustic distortion; 3. Compensation with explicit. vs. implicit distortion modelling; 4. Disjoint vs. joint model training; and 5. Compensation with deterministic vs. uncertainty processing.		1. Techniques used for noise robust ASR.
16	Mehrish *et al*. [94], 2023	Conformers=hybrid CNNs and transformers.	It has input, convolutional, self-attention, feedforward, and output layers.	For classification in speech processing.
17	Phan *et al*. [97], 2023	1. DT; 2. XGBoost or extreme gradient boosting; 3. RF; 4. ET; 5. k-nearest neighbor (KNN); 6. Logistics regression; 7. SVM; and 8. MFCCs		1. For classification and regression; 2. For classification; 3. For classification; 4. For classification; 5. Same as above; 6. Same as above; 7. Classification or regression; and 8. For feature extraction

Table 11. Findings of research questions 3

S/N	Authors and year of publication	AI algorithms/technologies	Benefits	Disadvantages
1	Nassif *et al.* [1], 2019	1. DNN with many hidden layers on HMM; and 2. MFCC with DNN	1. WER improvement; and 2. Advancement in speech spectrogram features	1. Neural networks struggle with continuous speech due to poor temporal modeling.
2	Malik *et al.* [2], 2021	1. MFCC; 2. LPCC; 3. PLP; 4. RASTA-PLP; 5. DWT; 6. DWPT; 7. MFCC, PLP, and LPC; 8. DWT, LPCC, and WPT; 9. MFCC, PLP, and LPC with either DWT; 10. MFCC, PLP, and LPC with either DWPT; 11. HMM+Wavelet Transform= hidden Markov tree (HMT) model; 12. ANNs; 13. HMM+ANNs; 14. MLP; 15. Deep MLP; 16. SOM+DWT; 17. RBF+LPCC; 18. RBF+Wavelet transformation 19. Neural networks, both feed-forward and recurrent; 20. HMM+neural networks 21. FNN; 22. Hybrid wavelet transforms, CDHMM, and FNN; 23. SOM+VQ+ANFIS; 24. SVMs; 25. SVMs+RBF kernel; 26. One-against-one (SVM) method; and 27. One-against-one (SVM) classifier in combination with a majority voting technique	2. It is 10% more efficient and 5.5% faster than MFCC; 3. It achieved 0.2% more accuracy than MFCC; 4. Performs better for noisy dataset than any other feature extraction methods; 5. It is very robust to noise as it works with localized time and frequency information; 6. In comparison with MFCC, a reduction of 20% in WER was achieved; 7. Achieve good accuracy in clean environments; 8. Show better results in noisy environments; 9. It makes the ASR more robust; 10. Same as above; 11. To boost the performance of wavelet-based algorithms; 13. Solve the problems highlighted in 12; 15. Good for speech emotion recognition; 16. Good for vowel recognition; 17. Training and testing speed is faster than that of MLP; 18. It is more robust and achieve better results than a system involving only RBF; 20. Solve the problem of 19 by finding the alignment between the input audio and its transcribed output; 21. Membership functions are effective for mapping speech signals with unclear boundaries, yielding better results on small datasets as they converge during learning; 22. In comparison with CDHMM, it was proven to be more successful in a noisy environment; by achieving 15.2% more accuracy [98]; 23. Performs better than a conventional FNN [112]; 25. It has less processing time in the training phase, and also achieved a higher accuracy in comparison to MLP [99]; 26. It requires less training data, lower computational cost [100], [102]; and 27. It has an accuracy of 77.6%, which was 4% better than the HMM.	1. Not adaptive to noise; 4. It may not perform well for speech signals obtained in clean environments; 7. Do not perform well in noisy environments; 12 They overtrain and face the local minima problem, and also ignore the time variability content of the speech signal; 14. Inability to handle dynamicity of the input speech signal because they only take input of fixed length., the algorithm can only deal with small vocabularies, which makes them not an efficient word recognizer but a good phoneme recognizer; 19. It is only good for frame-wise classification of the input audio signal; 24. Cannot take varying inputs as this is the case for speech recognition data, has high computational cost when classifying two classes; and 26. It develops a relatively higher number of binary SVMs
3	Dubey and Shah [3], 2022	1. RNN training system utilizing multiple GPUs [Deep Speech]; 2. HMM, N-Gram; 3. CNN; 4. RNN; and 5. ANN+HMM	1. It can learn robustness to noise or speaker variation automatically; 3. works very well for classifying the vowel sounds with stationary spectra; 4. It can cope with the time varying information like time-varying spectra of speech sounds; and 5. It gives optimized results.	3. It poorly performs for phoneme discrimination of consonants, which are characterizes by variations of their short-term spectra
4	Huang *et. al.* [4], 2014	1. HMM+Gaussian density; 2. HMM+DNN; 3. FFT, Filter bank: for feature extraction; 4. Deep learning; and 5. RNN	2. It overcomes the inefficiency in data representation via the DNN; Deep learning can also be used to learn powerful discriminative features for a traditional HMM speech recognition system; It produces significant error reduction; 4. It significantly improves acoustic modelling quality; and 5. It significantly improved the N-gram language model.	1. There is inefficiency in data representation by the GMM.

Table 11. Findings of research questions 3 *(continued)*

S/N	Authors and year of publication	AI algorithms/technologies	Benefits	Disadvantages
5	Saksamudre *et al.* [32], 2015	1. ANN; MFCC [96]; 2. Kernel based feature extraction method; 3. LPC; 4. Wavelet; 5. RASTA-PLP (relative spectral transform); 6. PCA; 7. LDA; 8. ICA; 9. CMS; 10 Filter Bank analysis	1. It minimizes the modelling unit, generally in the phoneme modelling so as to advance the RR of the entire system by improving the RR of phonemes; It is good for finding features; 2. It is used to remove noisy and redundant features. It helps to improve the classification error; 3. It allows the use of fixed resolution along a mel frequency scale for spectral analysis; 4. "It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allows better time resolution at high frequencies than Fourier transform." [32]; 5. It is suitable for extracting features from noisy data; 6. It is suitable for gaussian data; 7. It is good for PCA classification; 8. "Blind than PCA for classification." [32]; 9. It is similar to MFCC but works on the mean statistical parameter; and 10. Provision of any frequency resolution (wide or narrow) for spectral analysis.	1. They are not robust in the presence of additive noise; 2. It has a slow similarity classification speed; 3. On the linear scale, frequencies are weighted equally while the frequency sensitivity of the human ear is close to the logarithmic; 4. It requires longer compression time; 5. The dependency of the data on the previous context is increased; 6. Maximizing information is not equivalent to the direction of maximizing variance; 7. Inability to preserve any complex structure of the data, which may be needed for classification if the distribution is significantly non-Gaussian; 8. Unordered of extracted components; and 10. It takes more processing time and calculation than discrete Fourier analysis.
6	Desai *et al.* [30], 2013	1. LPC; 2. LPCC; 3. MFCC; and 4. HMM	1. To determine the basic parameters of speech, it provides a computational model of speech and precise estimation of speech parameters; and 2. It allows easy incorporation of knowledge sources into organized architecture	2. They are highly sensitive to quantization noise; and 4. It does give much insight on the recognition process.
7	Karpagavalli and Chandra [17], 2016	1. HMM-GMM; 2. HMM-ANN; and 3. HMM-DNN	1. It offers better generalization properties and lower memory requirements, and variable length data sequences can be handled by the HMM. HMM-based models are simple and computational feasible to use; 2. Short-time units such as individual phones and isolated words can be classified effectively by the ANN. The NN can be used for pre-processing, and dimensionality reduction in the HMM-ANN based; and 3. The DNN is used to characterize the properties (high-order correlation) of the data. The DNN provides discriminative power for classification of patterns.	1. The Gaussian mixture in the HMM-based model is statistically inefficient to model data that lie on or near a non-linear manifold in the data space; and 2. For continuous recognition tasks, the neural networks are rarely successful.
8	O'Shaughnessy [25], 2008	1. Filter banks were combined with dynamic programming; 2. LPC; 3. DTW; 4. HMM; 5. Mel-frequency cepstrum; 6. Language models; 7. Neural networks; 8. Wavelet transform; 9. Kernel-based classifiers; and 10. Dynamic Bayesian networks	1. It is mostly used for words spoken in isolation (i.e., with pause after each word), so as to simplify the task; 2. It is beneficial for automatic, simple speech compression; 3. It reduces search while it allows temporal flexibility; 4. It treats both temporal and spectral variation statistically and more flexible than DTW; 5. It has an improved auditory-based speech compression; 6. Improves the accuracy of ASR; 7. It is an excellent static nonlinear classifier; 8. The variability in the time-frequency that is tiling more closely matches human perception; 9. Better discriminative training; and 10. More general statistical networks.	4. It has the frame-independence assumption, from the use of first-order Markov models; and 7. They are relatively inflexible to handle timing variability.

Table 11. Findings of research questions 3 *(continued)*

S/N	Authors and year of publication	AI algorithms/technologies	Benefits	Disadvantages
9	Hagen and A. Morris [5], 2005	1. HMM+GMM; 2. HMM+MLP; and 3. HMM+ANN	1. It is good for modelling ASR system that takes model adaptability into consideration; 2. Unlike GMM, MLP is capable of handling nonlinear processing; and 3. Since ANNs model posterior probabilities p(w/x) whereas GMMs model likelihoods p(x/w), ANNs are better suited for multi-expert combination.	1. Not suitable for an expert system that estimates the posterior probability; and 3. They are not suitable for modelling noise or speaker adaptability in an ASR system.
10	Palaz *et al.* [35], 2019	1. CNN; 2. CNN+BLSTM+DNN (CLDNN); 3. MFCCs/PLP+ANN+HMM; 4. CNN+MLP+HMM; 5. Filter bank or critical band energies; 6. Short-term magnitude spectrum features; 7. Improved MFFCs: MFCC+LDA+MLLT+FMLLR ; and 8. DNN	1. It achieves performance comparable or better than feature extraction methods; 2. It is used to classify phones. It yields performance comparable to the case where the input to CLDNN is log filter bank energies; 3. It yields better performance than GMM+HMM; 4. It gives better performance than hybrid MFCCs/PLP+ANN+HMM; 5. It is used as input to the CNN system-based classifiers; 6. It is used as input to the DNN classifiers; and 7. It gives better performance than the conventional MFCCs.	8. It yields an inferior system when compared to standard acoustic modelling.
11	Karita *et al.* [29], 2019	1. Transformer; and 2. RNN	1. It outperforms the DNN + HMM based system in large dataset, noisy dataset, low resource dataset, far-field dataset; and also outperforms the RNN-based end-to-end system; and 2. It performs better than HMM based systems.	1. It is slow in decoding process; hence a faster decoding algorithm must be developed for transformer for its comparison with the DNN+HMM
12	Cutajar *et al.* [33], 2013	1. MFCCs; and 2. SVM	1. They have the ability to model the time distribution of speech signals; and 2. It can achieve, either comparable or even better results than the HMMs	1. It lacks robustness to noise, struggles to decode multiple phonemes in continuous speech, and relies solely on the power spectrum, ignoring phase information that is valuable for human speech perception.
13	Pandey [111], 2022	Soft computing approach (GA)-using GA to prepare ANN	The speech acknowledgment speed of ANN is much faster than HMM's.	The HMM has a little higher acknowledgment rate than ANN.

Table 12. Findings of research questions 4

S/N	Authors and year of publication	Traditional probabilistic approach to ASR	Artificial intelligence approach	Performance improvement of ai over traditional probabilistic approach
1	Nassif *et al.* [1], 2019	1. HMM+GMM; and 2. HMM+GMM	1. Neural network for pre-processing e.g. feature transformation, dimensionality reduction for the HMM based recognition; Deep recognition base on deep learning; and 2. MFCC+DNN	1. The WER reduced by 30% in comparison to the state-of-the-art models based on Gaussian mixtures; and 2. Advancement in speech spectrogram features
2	Malik *et al.* [2], 2021	1. HMM; and 2. HMM	1. ANN+SVM; and 2. SVMs	1. The technique (ANN+SVM) can be employed independently or as a hybrid model with HMM to obtain optimal results of ASR; and 2. Modification strategies like one-against-all and one-against-one enable SVM to handle multi-class classification effectively, achieving results that are equal to or better than HMM.
3	Huang *et. al.* [4], 2014	HMM+GMM	DNN HMM	It produces significant error reduction.

Table 12. Findings of research questions 4 *(continued)*

S/N	Authors and year of publication	Traditional probabilistic approach to ASR	Artificial intelligence approach	Performance improvement of ai over traditional probabilistic approach
4	Hagen and Morris [5], 2005	HMM+GMM	1. Multiband hybrid HMM+ANN; 2. All combination multiband hybrid HMM+ANN; 3. All combination multi stream hybrid HMM+ANN; 4. Multi stream tandem hybrid HMM+ANN; 5. Narrow band tandem hybrid HMM+ANN; and 6. SVM	1. It is a multi-expert system that involves the feature combination, posterior probabilities combination of ANN and hypothesis combination, which are used for performance improvement in HMM+ANN model by making the system more robust to unpredictable signal distortion; 2. Same as above; 3. Same as above; 4. Same as above; 5. Same as above; and 6. It was developed for use with high dimensional data.
5	Palaz *et al.* [35], 2019	HMM+GMM	1. MFCCs/PLP+ANN+HMM; and 2. CNN+MLP+HMM	1. There is performance improvement in WER and PER over the probabilistic approach; and 2. Robust to noise, thereby improving on one of the variabilities issues in ASR, has performance improvement over both the probabilistic approach and the hybrid MFCCs/PLP+ANN+HMM.

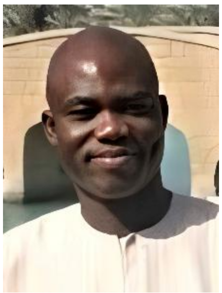
Table 13. Findings of research questions 5

S/N	Authors and year of publication	Technologies at the beginning	Present day technologies for ASR	Comparison of the two sets of technologies in terms of performing efficiently as a human listener terms of performance improvement
1	Malik *et al.* [2], 2021	HMM	ANN	HMM alone cannot achieve the ultimate goal of ASR. Hybrid models and ANN can help achieve much better results [103], [105].
2	Palaz *et al.* [35], 2019	GMM+HMM	1. MFCC/PLP+ANN+HMM, MFCC/PLP+MLP+HMM, MFCC/PLP+TDNN+HMM, MFCC/PLP+RNN+HMM, MFCC/PLP+CNN+HMM; and 2. CNN+MLP+HMM	1. Has better WER, and PER performance than GMM+HMM based system; 2. – CNNs with deep architectures improve ASR performance by leveraging the large number of hidden layers. Combining CNN for feature extraction, MLP for classification, and HMM for decoding scales effectively to continuous ASR, consistently outperforming conventional cepstral feature-based systems across all tested corpora. It yields consistently a better system with fewer parameters compared to the conventional MFCC/PLP+ANN+HMM approach, where MFCCs/PLP are used for feature extraction, ANN for classification, and HMM for decoding; and – In the CNN+MLP+HMM framework, the first convolutional layer automatically learns 'in-parts,' and the subsequent filtering produces intermediate feature representations that are more discriminative than traditional cepstral features used in MFCC/PLP+ANN+HMM systems.

BIOGRAPHIES OF AUTHORS

Gabriel Oluwatobi Sobola completed his first degree from the Department of Electrical and Electronics Engineering, Federal University of Agriculture, Abeokuta, Nigeria with a First-Class Honours in 2014. He thereafter finished his Master of Science in 2018 from the University of Ibadan, Nigeria with an interest in wireless communication and signal processing. He is presently a researcher and a lecturer at Covenant University, Ota, Nigeria. He can be contacted at email: gabriel.sobola@covenantuniversity.edu.ng.

Emmanuel Adetiba is IEEE Member. He holds a Ph.D. in Information and Communication Engineering from Covenant University, Nigeria. He served as ICT Center Director (2017-2019), full professor and department head (2021-2023), and now leads CApIC-ACE as deputy director and FEDGEN project Co-PI (World Bank/AFD-funded). Founder of ASPMIR Group, he has over 100 Scopus/ISI publications on machine intelligence, biomedical signal processing, and cloud computing, with grants from Google, NSF, and others; he is a COREN-registered engineer, IITP member, and research associate at Durban University of Technology. He can be contacted at email: emmanuel.adetiba@covenantuniversity.edu.ng.



Olabode Idowu-Bismark is a senior lecturer at Covenant University, Ota, Nigeria. He earned a B.Eng. in Electrical and Electronics Engineering from the University of Benin, an M.Sc. in Telecommunications Engineering from the University of Birmingham, UK, and a Ph.D. in Information and Communication Engineering from Covenant University. With prior experience as an engineer, senior engineer, and technical manager in various companies, he is a registered COREN engineer, member of the Nigerian Society of Engineers, and MIEEE. His research focuses on mobile communication, mmWave, and MIMO, with numerous publications in peer-reviewed journals and conferences. He can be contacted at email: idowubismarkolabode@gmail.com or olabode.idowu-bismark@covenantuniversity.edu.ng.

Abdultaofeek Abayomi earned a B.Sc. (Hons) in Computer Science from the University of Ilorin, Nigeria, an M.Tech. from Federal University of Technology Akure, a Ph.D. in Information Technology from Durban University of Technology, South Africa, and completed postdoctoral research at Mangosuthu University of Technology. With over a decade in financial services and three years in IT in Nigeria, he has lectured in Computer Science, IT, and Information Systems at Federal University Oye, Nigeria, and Durban University of Technology. His research spans artificial intelligence, machine learning, wearable sensors, big data analytics, computer vision, image processing, bioinformatics, affective computing, HCI, and data mining; he is a member of IITPSA and SAICSIT. He can be contacted at email: taofeek.abayomi@summituniversity.edu.ng.

Raymond Jules Kala is a distinguished Ph.D. holder in Computer Science from the University of KwaZulu-Natal and currently serves as an assistant professor at the International University of Grand Bassam. He specializes in image processing, pattern recognition, information systems, decision support systems, and artificial intelligence. He can be contacted at email: raymondkala1@gmail.com.

Surendra Colin Thakur is an associate professor in the Department of Computer Science at the University of South Africa, Pretoria. He also serves as Director of the NEMISA KZN e-Skills Co-Laboratory and the KZN Digital Co-Laboratory at DUT, focusing on e-skills, e-government, e-democracy, and e-participation. He founded InvoTech, an innovation incubator at DUT, where one of his patents is being registered. He is an international expert in e-voting, with additional expertise in social media and big data. His research interests include data science, e-voting and public participation, and social media. He has served on executive bodies such as exco, senate, and faculty boards, and held leadership roles in the Computer Society of South Africa, including National Treasurer and KZN Chair and Vice-Chair. He can be contacted at email: thakur@dut.ac.za.

Sibusiso Moyo received the Ph.D. degree in mathematics with a focus on symmetries of differential equations and their application from the University of Natal (currently University of KwaZulu-Natal), Durban, and the master's degree in tertiary education management from the University of Melbourne. She is currently the Deputy Vice-Chancellor of research, innovation and postgraduate studies with Stellenbosch University, South Africa. She has published widely in the Mathematical Sciences. She can be contacted at email: smoyo@sun.ac.za.