DBN-FTLSTM: An Optimized Deep Learning Framework for Speech and Image Recognition

Xiaoyan Wang*, Yuanyuan Liang
School of Artificial Intelligence and Software Engineering, Nanyang Normal University, NanYang, HeNan, 473061, China
E-mail: wangxiaoyan_1981@126.com, liangyy2358@126.com
*Corresponding author

Keywords: deep learning, artificial intelligence, speech recognition, image, natural language processing, DBN-FTLSTM.

Artificial Intelligence's (AI) quick development has brought about a new era of technical innovation with significant ramifications for many different fields. This study explores the development of artificial intelligence (AI) for image and speech recognition, a job in natural language processing that entails the real-time computer transcription of spoken words. The use of deep learning (DL) models for voice recognition has been the subject of several research. But this field is developing quickly. This systematic review offers a thorough and in-depth analysis of research on voice recognition using DL approaches that was published between 2020 and 2024. In order to improve clarity, this study clearly outlines the methodology, including the dataset size, preprocessing methods (MFCC extraction, normalisation, and augmentation), and benchmarking criteria. To provide openness and repeatability, the datasets used for assessment must be specified in reported performance measures (accuracy, RMSE, MAE, and R^2). The dataset, experimental setup, and baseline comparisons should all be explicitly stated in order to contextualise the 99.3% accuracy claim. With an overall performance accuracy rate of 97.90%, a root means square error (RMSE) of 0.017, a mean absolute error (MAE) of 0.01, and an R^2 value of 98.7%, the DBN-FTLSTM model performs better than existing techniques. When compared to other current approaches, the recommended method's error metrics accuracy of 99.3% is the greatest. Mel-frequency cepstral coefficients were the most often used feature extraction approach, while the word error rate was the most commonly used assessment method. Another finding was the dearth of research using transformers, which have been shown to be effective models that can enable parallelization, speed up learning, and enhance the performance of low-resource languages. The findings also identified intriguing and promising areas for further study that had not gotten much attention in previous investigations.

Povzetek: Opisan je optimiziran globok učni model DBN-FTLSTM, ki dosega 99,3 % natančnost pri prepoznavanju govora in slik, z znatnim izboljšanjem napak in robustnostjo v realnem času.

1 Introduction

Deep learning is a subset of machine learning that utilizes multi-layered neural networks. The purpose of these networks is to mimic how people learn from large amounts of data. Deep learning models can execute complicated tasks like image recognition, natural language processing, and more with amazing accuracy by utilizing enormous datasets and computer capacity [1]. Extracting higher-level characteristics from the raw input data is one of DL's main advantages. In many areas of our everyday life, such as picture and speech recognition, online search, fraud detection, email/spam filtering, financial risk modeling, and more, DL applications are quickly taking the place of traditional systems [2]. It has been demonstrated that DL approaches provide fascinating new possibilities in a variety of disciplines. It has been demonstrated that DL techniques serve as an adjunct to physics-based materials design techniques. Although effective DL applications are sometimes thought to need big datasets, methods like transfer learning, multi-fidelity modeling, and active learning may frequently make DL possible for small datasets as well [3-4]. Materials have historically been created experimentally by trial-and-error techniques combined with a healthy dose of chemical intuition. In addition to being a very expensive and time-consuming process, empirical formulation and computational approaches are required since there are so many material combinations that they are impossible to explore experimentally [5].

Deep learning, one of the most revolutionary advances in artificial intelligence (AI), has fundamentally changed the image processing industry. Deep learning algorithms have recently been shown to perform better than humans on tasks like picture identification and classification, which has resulted in ground-breaking developments in a number of sectors [6-7]. Numerous advancements in AI are fueled

by these two new technologies. Deep learning in image and facial recognition technology allows computers to recognize faces, objects, and situations with an accuracy that is comparable to human perception [8]. This technology has uses in surveillance and medical imaging diagnostics. Applications that comprehend, interpret, and produce human voice and language have been made possible by deep learning in natural language processing. Conversational AI, translation services, and speech recognition systems have all advanced via NLP [9].

Deep learning methods have recently been used to a variety of computer vision and image analysis applications, including object identification, picture segmentation, graphics recognition, and image classification. Deep learning still has several drawbacks, though, including overfitting, accuracy, network size, training time, and processing power [10] since accuracy is usually an objective rather than a constraint, the claim that it is a disadvantage in deep learning appears contradictory. In this context, though, it probably alludes to the compromises made in order to achieve high accuracy, including the possibility of overfitting, longer training durations, larger networks, and greater processing costs. Deep learning models may become less effective and more difficult to implement in practical settings as a result of these issues. These problems are probably brought up in the study to highlight the necessity of optimised frameworks like DBN-FTLSTM, which seek to strike a compromise between accuracy, efficiency, and usefulness. To create dependable and effective deep learning networks, these issues must be resolved. In order to solve the aforementioned issues in a range of applications, this Special Issue's goal is to present innovative, high-performance, optimized, and hybrid deep-learning-based algorithms for image analysis [11].

Speech is a common and vital human communication tool that allows people to convey their ideas, thoughts, and feelings as well as participate in deep discussions [12]. New communication methods that fit our digitally connected environment have surfaced as our lives grow more and more entwined with technology and smart gadgets [13]. In addition to empowering people with disabilities to interact, share knowledge, and have open discussions, speech recognition has been essential in helping us adjust to these new forms of communication. It also has the potential to completely transform machine-to-machine communication through the use of natural languages [14].

Recent years have seen significant progress in speech recognition, with deep learning being a major factor. Large volumes of voice data can be processed using deep learning algorithms, which can also identify speech patterns that are not achievable with conventional techniques. Deep learning processes large voice datasets and identifies complex speech patterns, like accents or emotions, that traditional methods (e.g., HMMs) cannot. It excels at capturing nuanced, non-linear relationships in speech data. These algorithms have improved the accuracy, dependability, and accessibility of voice recognition systems for a larger population. The most recent developments in deep learning for voice recognition will be covered in this article [15].

By managing massive volumes of huge data, identifying speech in noisy surroundings, and creating end-to-end systems that manage the entire process without further processing, deep learning algorithms have greatly enhanced speech recognition. Speech recognition systems are more effective and user-friendly because to these algorithms' increased robustness and dependability in noisy settings. Additionally, speech recognition systems are now more scalable and affordable because to unsupervised learning, which uses voice data without labeling [16]. Lastly, the creation of customized voice recognition systems has also been made possible by deep learning algorithms. These systems can gradually increase their recognition accuracy by learning each person's distinct speech patterns. People with disabilities or speech difficulties may find it easier to use voice recognition systems as a result [17].

Contribution for this study: Due to the availability of large amounts of training data and sophisticated computational capabilities, speech recognition has attracted a lot of interest lately. Speech recognition systems now have higher recognition rates in deep learning models and algorithms. Speech recognition, a subfield of natural language processing (NLP), is the ability to perceive and identify spoken words, sometimes in conjunction with transcription and translation skills. Relevant characteristics are extracted from the speech stream using methods including linguistic analysis, acoustic modeling, and signal processing. The accuracy of recognition is improved using segmentation algorithms and pattern recognition methods. Deep belief Network based on Fine-Tuning LSTM (DBN-FTLSTM) and other machine learning algorithms have been crucial in identifying and understanding these patterns, which has resulted in precise and effective voice recognition. With its ability to provide smooth machine-to-machine communication through natural languages, voice recognition technology has a bright future. Dialects, multilingualism, speaker variability, noise, unfavorable acoustic settings, terms that are not in the speaker's lexicon, contextual comprehension, ambiguity in speech, and ethical and privacy issues are some of the obstacles that speech recognition must overcome. Large companies like Microsoft and Google are creating technologies to enhance voice recognition capabilities.

2 Related works

Table 1: Summary on related works

Author	Year	Findings	Performance Metrics	Research Gaps
Hamza Kheddar et al., [18]	2024	Deep learning methods such as DTL, FL, and DRL are being applied to enhance the performance of automated speech recognition (ASR) in dynamic settings. By using these methods, adaptive systems may train on private data without owning the dataset, employ tiny, relevant datasets, and make the best decisions in changing situations. In order to shed light on contemporary issues and pinpoint advantages and disadvantages, this assessment examines DTL, FL, and DRL-based ASR frameworks, including Transformers. Additionally, a comparative analysis is presented to highlight current issues and potential directions for future research.	Several measures are used to assess performance in order to gauge efficiency, accuracy, and resilience.	Background noise, reverberation, and multi-speaker situations are problems for this study.
Reshma C.R [19]	2020	Speech is the primary means of human-to-human communication. The use of machine learning to voice recognition is the subject of several studies. One of the most difficult tasks is using speech recognition techniques to turn recorded audio signals into text. This study proposes a voice recognition framework.	The optimum approach for the specified speech type is determined by the performance parameter, which is minimal.	One of the most difficult challenges is automated speech recognition (ASR), which converts recorded audio to text automatically. .
Denis Ivanko et al., [20]	2023	With an emphasis on deep learning (DL) and its use in AV speech fusion and recognition, this article examines current developments in audio-visual speech recognition (AVSR) techniques from 2013 to 2023. The primary AV datasets, methodology, development of fundamental AVSR techniques, pre-processing and augmentation approaches, and modality fusion strategies are all covered. The present status of AVSR and potential avenues for future study are discussed in the article's conclusion.	Additional training data has been shown to be quite beneficial.	Innovative strategies, techniques, and technological advancements are still needed to improve SOTA and make current speech recognition systems more practical and approachable for a worldwide audience.
C.H. Yang et al., [21]	2020	In order to improve prediction accuracy, the research presents a Differencing Long Short-Term Memory (D-LSTM) architecture as an extension of recurrent neural networks. It creates a 3D nonlinear chaotic system and uses Root Mean Square Error to compare its performance to that of the original LSTM and the Adaptive Neuro Fuzzy Inference System (ANFIS), demonstrating that it is almost better.	The authors use Root Mean Square Error (RMSE) to compare the performance of D-LSTM with that of the original Long Short-Term Memory (LSTM) and the Adaptive Neuro Fuzzy Inference system (ANFIS).	lack of large dataset

Mishaim Malik et al., **[22]**	2020	With an emphasis on feature extraction strategies, classification models, and their effects on automated speech recognition (ASR), this study contrasts many deep learning approaches in ASR. It talks about language models, online toolkits, tools, and data-dependent speech datasets. The study intends to provide scholars interested in ASR research a place to start.	the precision of the output generated and the ASR's processing speed	The inability of such language models to adjust when input is a different speech from a different domain is a significant disadvantage.
Zhe Dong et al., **[23]**	2023	A confidence decision network and a domain-specific language speech network are used in the paper's voice recognition technique. In order to increase prediction accuracy 82% to 91 %, it employs external knowledge sources, integrates a domain-specific dataset, and trains a domain-specific dataset. The technique shows the value of domain-specific datasets by outperforming an automatic voice recognition system.	In the medical field, the model's accuracy increased from 82% to 91%.	Long-duration series voice inputs might provide difficulties for the Transformer model, especially when dealing with complicated speech inputs and huge vocabularies.
Chongchong Yu et al., **[24]**	2020	It is necessary to concentrate on low-resource voice recognition due to the polarization of global languages, especially endangered ones. With efforts for low-resource languages and studies on acoustic models, deep learning is a potent tool for this.	In order to properly investigate the relationship and complementarity between acoustics and language and develop a more comprehensive end-to-end speech recognition system to enhance performance, the combined modelling of acoustic and language models should be increased.	with higher levels of background noise.
Liqing Chu et al., **[25]**	2024	This work enhances language analysis tools in secondary school by utilizing deep learning algorithms and picture recognition technology. In order to determine successful teaching behaviors, it builds a framework, applies data mining, and verifies the framework through an experiment.	The speaking rate, speech intelligibility, average sentence length, and content similarity of the online classroom conversation had significant p values of -0.56, -0.71, -0.71, and -0.74, respectively.	these findings to quickly identify and address communication issues,

HarshAhlawat et al., [26]	2025	Automatic Speech Recognition (ASR) has been revolutionized by deep learning models, which have demonstrated promising outcomes in speech processing. This study examines research conducted since 2010 and contrasts approaches, concentrating on monolingual and multilingual models. It draws attention to issues with generalizability and data reliance.	In order to comprehend model performance over a variety of datasets for real-world implementation, we have also examined the different models using publically available speech datasets in this study.	difficulties with the future that academics interested in open-source Automatic Speech Recognition (ASR) may start with

Critical constraints of current voice and picture recognition models, such as CNNs, RNNs, LSTMs, and SVMs, include excessive data reliance, computational inefficiencies, and limited long-term dependency learning. RNNs have trouble with vanishing gradients, whereas CNNs are good at extracting spatial features but not at capturing sequential relationships. SVMs do not scale well for big voice datasets, and traditional LSTMs lack deep hierarchical feature representation, despite their effectiveness in sequence learning. By utilising LSTM's sequential modelling and DBN's hierarchical feature extraction, the suggested DBN-FTLSTM model fills these gaps and improves generalisation, efficiency, and robustness. This hybrid technique is a better option for real-world voice and picture recognition applications because it increases noise resistance, decreases the need on labelled data, and guarantees optimal computing efficiency.

2 Methodology

2.1 Dataset

Throughout the history of image and speech recognition research, datasets have been crucial, particularly in the AI era. We include the most widely used speech datasets for SR or lip-reading system benchmarking in the first group. Benchmark datasets offer a standard platform for evaluating and contrasting SOTA algorithm performance. Although DL technologies have advanced significantly in speech recognition in recent years, it is important to note that their effectiveness is mostly dependent on the availability of huge datasets containing annotated data.

The final list of research articles had 114 publications in total Information that addressed the study topics was extracted from these carefully examined articles. The gathered data was characterized statistically and utilized to identify trends in the 2020–2024 investigations.

About 40% of the publications in this study were journal articles, while the bulk (60%) were conference papers. They were dispersed over many conferences and journals, with the International Conference on Acoustics, Speech, and Signal Processing and the International Journal of Speech Technology being the most popular conferences and journals, respectively. The bulk of the research are concentrated in recent years, which is a noticeable trend in the distribution of the studies during the years of publication. Specifically, the largest percentage of publications occurred in 2022 (32%), followed by 2021 (24%), 2020 (19%), 2023 (24%), and 2024 (15%).

A wide range of datasets were found in the research for the algorithms' testing and training. 55% of the datasets were classified as private, while 38% were classified as public and available online. It is important to note that the type of dataset utilized in the experiments was not specified in 7% of the articles. The Crowdsourced high-quality multi-speaker speech dataset, the Multi-Channel Articulatory database, Switchboard, CallHome, the Si284 dataset, dev93, eval92, TIMIT, Tibetan Corpus, SpinalNet, the Amharic reading speech corpus, the Google OpenSLR dataset, and the Kaggle dataset were among the publicly available datasets.

Table 2: Large-scale speech datasets

Dataset	**Size**	**Languages**	**Description & Usage**	**Benchmarking**
LibriSpeech	1,000 hours	English	Audiobook dataset with clean and noisy speech	WER (Word Error Rate)

Google OpenSLR	Varies (20–10,000+ hours)	Multilingual	Diverse speech corpora for ASR and TTS training	WER, CER (Character Error Rate)
TIMIT	5.4 hours (6,300 utterances)	English (phoneme-rich)	Phonetically transcribed speech for phoneme recognition	PER (Phoneme Error Rate)
Switchboard	~300 hours	English (telephonic)	Spontaneous conversations, used in ASR research	WER, SER (Sentence Error Rate)
CallHome	~120 hours	English, Spanish, Arabic	Telephone conversations across dialects	WER, CER
Si284 (WSJ0 & WSJ1)	37.5 hours	English	Wall Street Journal dataset for ASR	WER, SER
Common Voice (Mozilla)	18,000+ hours	100+ languages	Crowdsourced speech dataset	WER, CER

Table 3: Lip-reading datasets

Dataset	Size	Languages	Description & Usage	Benchmarking
LRW (Lip Reading in the Wild)	500,000 utterances	English	Word-level lip-reading dataset	Accuracy, WER
GRID Corpus	34,000 utterances	English	Controlled speech & lip movement dataset	WER, SER
LRS2 (Lip Reading Sentences 2)	140,000 utterances	English	Continuous lip-reading dataset	WER, CER
LRS3 (Lip Reading Sentences 3)	400,000 utterances	English	Larger version of LRS2 with natural sentences	WER, Accuracy

2.2 Feature extraction

MFCC (Mel-Frequency Cepstral Coefficients)

Mel-frequency Because Cepstral Coefficients (MFCCs) can replicate human auditory perception, they are essential for speech recognition. They reduce dimensionality and highlight key phonetic features by condensing speech waveforms into compact representations. Pre-emphasis filtering, the Hamming window, the Fast Fourier Transform, the Mel-scale filter bank, and the Discrete Cosine Transform are all used in the extraction process. MFCCs are employed in speaker identification, emotion detection, and automatic speech recognition (ASR). They are less useful in loud settings, though, due to their limitations.

2.3 Data augmentation

The quantity of data available affects deep learning performance, particularly for small data sets. One method for increasing the quantity of data required for voice recognition systems is data augmentation. Semi-supervised training, multilingual processing, acoustic data perturbation, and speech synthesis are examples of common techniques. A 14.8% relative WER improvement is achieved in certain studies by combining several data

augmentation techniques. While multilingual processing adds resource-rich data to resource-poor data, semi-supervised training employs both supervised and unsupervised data. By messing with raw data and using voice synthesis technologies to create new data, acoustic data perturbation and speech synthesis can produce fake data.

2.4 Speech recognition and natural language processing

A method that allows a machine to understand spoken language is called speech recognition. ASR, often known as text-to-speech, speech synthesis, or just speech recognition, creates methods and strategies that let computers understand spoken language and translate it into writing. NLP is the term for linguistics and machine learning combined. NLP is a machine learning application that uses millions of example datasets to teach robots to interpret natural language. Speech recognition is an area of computational linguistics that studies the technology that enables human-computer communication. Electrical engineering, computer science, and linguistics research and understanding are all included into speech recognition. The following steps make up the voice recognition process:

- Analog-to-digital conversion, which uses sampling and quantization methods to transform analog voice into digital. Digital speech is represented as a vector of voice integer samples.

- Speech pre-processing, which detects and eliminates extended silences and background noise. For the next phase, the speech is then split up into frames of 20 seconds each.

- The process of turning voice frames into a feature vector that identifies the phoneme being spoken is known as feature extraction.

- When choosing words, a language model is used to transform the phoneme or feature sequence into the spoken word.

As far as technology is concerned, voice recognition has a long history and has seen several notable developments. The field has advanced with recent advances in big data and DL. The number of scholarly articles on the subject has increased, but more importantly, the worldwide industry—including large corporations—has also embraced a number of DL approaches for developing and deploying speech recognition systems. Among the multinational corporations are Apple, Microsoft, Amazon, Facebook, and Google.

2.5 Extraction of speech features

Human speech is a highly developed ability. In order to extract its characteristics for additional processing and analysis, the speech waveform is converted to a parametric representation at a very modest data rate. Discrete wavelet transforms, line spectral frequencies, Melfrequency cepstral coefficients (MFCCs), linear prediction coefficients (LPCs), linear prediction cepstral coefficients, and perceptual linear prediction are some methods for extracting speech features. In voice recognition, MFCCs are often utilized characteristics. Pre-emphasis is applied to the voice stream at the beginning of the calculation to enhance higher frequencies. After that, the signal is divided into brief frames, and the stationary segments are separated by multiplying by a window function. The power spectrum of each frame is then obtained using the rapid Fourier transform. After that, the power spectrum is run through a Mel filterbank, which simulates the frequency resolution of the human hearing system. A discrete cosine transform is used to further de-correlate the filterbank energies after they have been converted into a logarithmic scale. Lastly, the coefficients are normalized across frames using cepstral mean normalization. Together, these procedures produce MFCCs, which record crucial speech data for speech recognition software.

2.6 Deep belief network based on fine-tuning LSTM (DBN-FTLSTM)

2.6.1 Deep belief network

Equation (1) is used to estimate the Deep Belief Networks (DBNs) optimal frequency size and to select the capacityfor AI using the speech recognition for light weight deep learning management technique. It most likely indicates that the ideal size or capacity (e.g., layers, neurones) of the Deep Belief Network (DBN) for voice recognition is estimated using a lightweight deep learning approach (e.g., a smaller neural network or optimisation algorithm). Although this method guarantees effectiveness and eliminates needless complication, the wording should have been more precise to prevent confusion.

$$E = \sum_{B=1}^{X} M_b * x * U_n + \sum_{u \in t}^{a} 2 * T * U_a + \sum\nolimits_{n=1}^{a} Q_{opti} \quad (1)$$

The letter E-stands for the target (ideal) sound quantity. The light weight annual demand for optimised inventories is denoted by the letter T. U_adenotes the cost of creating languages for the light DBN method, while U_n denotes the cost of maintenance for the deep learning method. M_b- the percentage rate at which production costs for maintaining the records preservation in AI are calculated.

The proportion of preserving reserves is derived from the fact that, using the deep learning method, the costs of preserving reserve funds now rise in proportion to the number of reserves within the human behaviour. Its proportion is equal to the sum of the following resources: alternative, stockpiling, logistics, and internal transportation within the capital adequacy factory, speech hearing, and decay, as calculated by Equation (2).

$$T = \frac{Q}{2}\sum_{b=1}^{S\in U} S_b * U_a + \int_{b=1}^{n} M_b * x * U_n \qquad (2)$$

Tstands for total maintaining charges, Q for the magnitude of the voice portion, and S_b for the safety margin implementation as in Equation (3).

$$E = 2\sum_{E\in U}^{T}(1-E) * U_a * T + \sum_{M\in E}^{x} x * \left(u + M * (1-E)\right) \qquad (3)$$

The letter U stands for an alternative for extraction. In terms of maximizing NLP value, E is the optimal solution of a voice finding. M represents the effective rate of accuracy costs for the suggested method is given in the Equation (4).

$$T = \frac{T}{Q}\sum U_a + \sum_{x=1}^{M}\sqrt{\left(\frac{Q}{2} + S_b\right) * x * M} \qquad (4)$$

$$E = 2\int_{a=1}^{E\in U}[(1-E) * U_a^{\#} + U * a] \times T + \int_{x=1}^{M\in E} x * (u + M + M^{\#} \times (1-E)) \qquad (5)$$

$$T = \frac{T}{Q}\sum U_a^{\#} + \frac{T}{Q}\sum U_a^{*} + \sum\sqrt{\left(\frac{Q}{2} + S_b\right) * x * M^{\#}} + \int_{x=1}^{b}\left(\frac{Q}{2} + S_b\right) * x * M^{*} \ (6)$$

$U_a^{\#}$ denotes voice identification from the dataset levels.

Uab: This is likely a specific value or function of U indexed by a and b.

Variations in time deliveries have a considerable effect on various stages of safety, which are represented in Equation (7) the light weight DL sounds required by receiver.

$$S_b = ln\sum_{ab}\frac{U * D * x * \sqrt{2\pi}}{T * U_{ab}} + \int_{D}^{x}\sqrt{-2 * D^2} \qquad (7)$$

Equation (8) and Equation (9) are used to calculate the research publishers not having data cleaning resources, where D is the standard error for exchange utilization and U_ab is the cost of just not having data splitting.

$$X = \sum\sqrt{\left(\frac{Q}{2} + S_b\right) * x * M^{\#}} + \sum_{i=1}^{m} T_i * (G_i - G)^2 + \frac{T}{Q}\sum U_a \qquad (8)$$

$$D = \int_{x=1}^{U} U * Q * D * x * \sqrt{2\pi} + \sum T * U_{ab}\sqrt{\sum_{i=1}^{m} T_i * (G_i - G)^2} \qquad (9)$$

$U_a^{\#}$ *denotes voice identification from the dataset levels,* U_a^{*}*denotes semi costs of developing the application, and* $M^{\#denotes}$*the percentage rate of accuracy able to operate feature extraction in Equation* (5).

The Equation (6) *is represented by M^*

$*$ *– effective rate in frequency.*

With 〖T〗_i is the chance of a certain event occurring based on historical data clearance. It is feasible to calculate the √Xvariation coefficient in respect to information about the prospective benefits of the light weight technique∑_(i=1)^m〖T_i×(G_i-G)^2〗 provide a efficient quality based image and speech in Equation (10).

$$U = \frac{D}{G}\sum_{B=1}^{X} M_b * x * U_n + \sum_{u\in t}^{a} 2 * T * U_a \sum_{i=1}^{m} T_i * (G_i - G)^2 \quad (10)$$

The following is a list of resources. A link exists between both the advantage of AI from a single trach of sound and indeed the benefits of acquiring from other distributors in Equation (11). For the deep learning method, correlation analysis is typically employed to measure such a relationship.

$$\rho_{2.1} = \sum U_a^{\#} * \frac{T}{Q} + \sum U_a^{*} * \frac{T}{Q} + \sum \sqrt{\frac{\sum_{i=1}^{m} T_i (G_{1i} - G_1) * (G_{2i} - G_2)}{D_1 * D_2}} \quad (11)$$

ρ_2.1 - the correlation coefficient here between advantages of AI from the first and second database in DL; G_1is the suitable rate of LSTM benefit from DL; G_2 is the appropriate error rate advantage from AI for the second supplier; and D_1is the confidence interval for the first supplier. The second provider's confidence interval is denoted by D_2. G_1iis the probability of making feasible prices of benefits from DBN from the first supplier; G_2i is the probability of possible rates of benefits from speechrecognition from the second provider; and pi is the probability of potential rates of benefits from image processing is depicted in Equation (12).

$$D_T = \int_{x=1}^{b} \left(\frac{Q}{2} + S_b\right) * x * M^{*} + \sum_{A\in B}^{D\in S} \sqrt{D_A^2 + D_B^2 + 2 * D_A * S_B * \rho_{A\&B}} \quad (12)$$

when D_T is the total standard error, D_A is the first solution's light weight standard deviation, S_B is the second solution's deep learning technique standard deviation, and A∈B are the regression coefficients between the first and second data distributions. It would be stated the steps that may be implemented again for the development and enhancement of the integrated information system. Equation (13) represents a finite and ordered set of restricted factor signifiers.

$$D = \sum_{b=1}^{S\in U} S_b * U_a + \sum_{i=1}^{n} \sqrt{X} = \{X_i\}, \quad i = 1,2,\ldots.n \quad (13)$$

To ensure the effectiveness, use Equation (14) for data collection, we believe, has a well-defined sequential relationship.

$$D_A = \int_{b=1}^{n} M_b * x * U_n + \sum_{X\to 1}^{n} X_1 < X_2 <\ldots\ldots < X_n \quad (14)$$

While X_1 precedes element X_2 precedes element X_3, and so on. The database requirement in a set X is stated in Equation (15) and is a mathematical setting of strength training T of element measures 〖X〗_i.

$$T = \sum \sqrt{\left(\frac{Q}{2} + S_b\right) * x * M^{\#}} + \sum_{i=1}^{n} \{T_i\}, \ i = 1,2,3,\ldots.n \quad (15)$$

To 〖T〗_i is a crucial figure in Equation (16) that satisfies the inequalities.

$$T = \sum \sqrt{\left(\frac{Q}{2} + S_b\right) * x * M^{\#}} + \sum_{i=1}^{n} T_1 < T_2 <,\ldots.. < T_n \quad (16)$$

The total ∑_(i=1)^n▒〖T_1 X_1+T_2 X_2+...+T_n X_n 〗 The following Equation (17) depicts a sophisticated effectiveness measure with in shape of the linear system as a plan efficiency indicator.

$$H = \int_{x=1}^{U} U * Q * D * x * \sqrt{2\pi} + \sum_{i=1}^{n} T_1 X_1 + T_2 X_2 + \ldots + T_n X_n = \sum_{i=1}^{n} T_i X_i \quad (17)$$

The ∑_(i=1)^n▒〖T_i X_i 〗 The efficiency of a light weight deep learning technique policy is just a function H with n efficiency variables that are sequential to its performance metrics X_i.

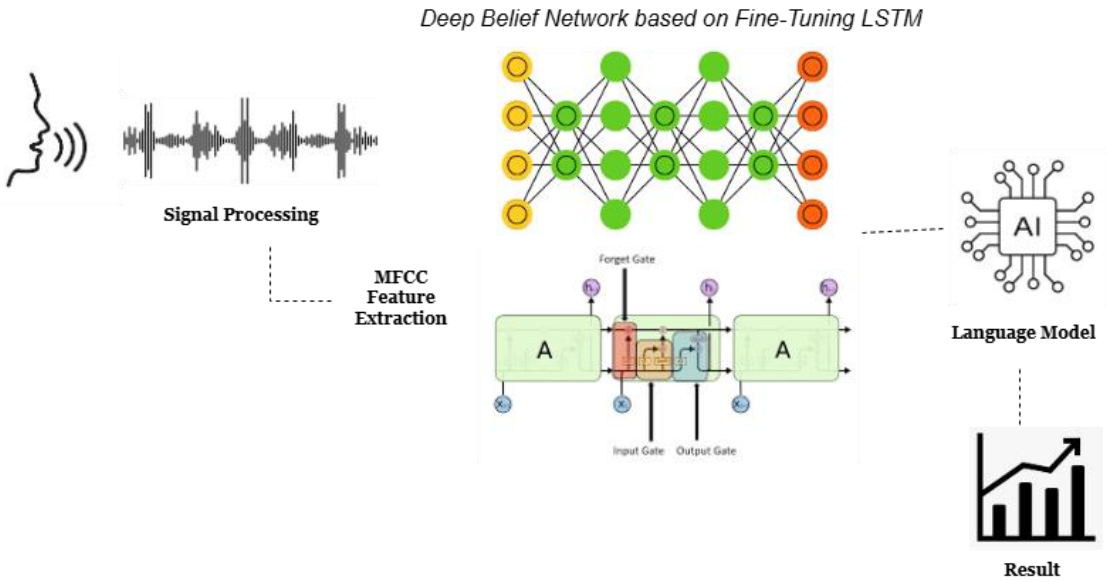

Figure 1: Proposed method for the model

2.7 Fine-tuning LSTM

The LSTM has a special function for image processing time-based memories. The horizontal link between each unit in the hidden layer is strengthened by the architecture's usage of a conventional, multi-layer BP neural network. Through the use of a weighted matrix, the prior data can be linked to the present task, allowing the previous generation progression to predict the action of the following sequence.

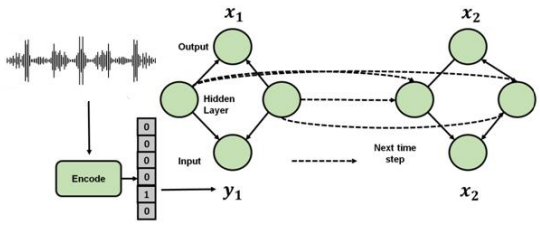

Figure 2: Principle for generating LSTM sequences

The LSTM network's structure is by acoustic timing, and Figure 2 depicts how its sequence generation and prediction process works. Set the input hierarchy (y_0...y_t), the hidden layer (z_0...z_t), and the output hierarchy (x_0...x_t), then Eq. (1) is used to evaluate the output of the hidden layer.

$$z_d = l(Wy_d + Uz_{d-1}) \quad (1)$$

In this formula, W represents the input weight matrix, U represents the hidden layer weight matrix some prior time instant and l represents the concealed layer's ability to be activated. The output sequence x_d is computed by applying the activation function of the output layer and the weighted matrix C to the result of the hidden layer's output calculation. The following is the algorithmic progression for generating LSTM sequences.

x_d=δ(Cz_d) (2)

Algorithm 1: The fine-tuning LSTM training procedure

Input: *Training data that has been processed is put into the DBnetwork.*

Output: *neural network model metrics*

Step 1: *Enter training data, specify batch size of recognition, iteration intervals, hidden layer cell counts, and network layer counts.*

Step 2: *As the loss function to modify the network's parameters, use the cross entropy among the present output of the neural system and the future input.*

Step 3: *until the loss has converged, repeat step 2 until the loop is finished.*

Step 4: *parameters for the LSTM generator's output records.*

To improve the methodology, it would be beneficial to provide a detailed breakdown of **dataset preprocessing**, including specific steps such as **data cleaning, feature extraction (MFCC), normalization, and augmentation techniques (time shifting, noise addition, pitch scaling).**

Pseudo code for DBN-FTLSTM

Step 1: Data Preprocessing

Input: Speech dataset (audio files)

Output: Preprocessed features (MFCCs)

1. Load dataset

2. Convert audio to MFCC

3. Normalize feature values (e.g., mean-variance normalization)

4. Split dataset into training, validation, and test sets

Step 2: Pre-train DBN using RBMs

Input: Normalized speech features

Output: Pre-trained weights for deep belief network

5. Initialize stack of Restricted Boltzmann Machines (RBMs)

6. For each RBM layer in DBN:

a. Perform Gibbs sampling to learn feature representations

b. Update weights using Contrastive Divergence (CD-k)

c. Pass learned representations to the next RBM layer

7. Store the trained DBN weights

Step 3: Fine-tuning with LSTM

Input: DBN-learned features

Output: Fine-tuned DBN-LSTM model for speech recognition

8. Initialize LSTM model with DBN-trained weights as input layer

9. Define LSTM architecture: a. LSTM layers with appropriate hidden units

b. Dropout layers for regularization

c. Fully connected (FC) layer with Softmax activation for classification

10. Train the model using Backpropagation Through Time (BPTT)

11. Optimize using Adam optimizer and categorical cross-entropy loss

12. Validate model on validation set

Step 4: Model evaluation

Input: Test dataset

Output: Recognition accuracy and Word Error Rate (WER)

13. Predict speech sequences using the trained DBN-FTLSTM model

14. Compute evaluation metrics: Word Error Rate (WER), Character Error Rate (CER), and accuracy

15. Analyze variance across multiple runs for stability assessment

Step 5: Model deployment

16. Deploy the trained model for real-time speech recognition tasks

3 Experimental results

The goal of this systematic review study is to fairly and thoroughly assess and evaluate the body of research on DL-based speech recognition that has been published between 2020 and 2024. When conducting this systematic review, the recommendations made by were adhered to. Planning the review, which was further subdivided into establishing the need for a review, formulating research questions, and outlining the search strategy to locate pertinent research papers, comprised the first research phase. The creation of the quality evaluation rules to filter research publications, the development of the data extraction strategy to address the study objectives, the definition of the appropriate research selection criteria, including the inclusion/exclusion criteria, and the synthesis of the data extracted from the publications comprised the second phase of the review. The reporting phase was the third stage of the study.

Table 4: Experimental hardware and software configurations

Category	**Specification**
Processor (CPU)	Intel Core i9 / AMD Ryzen 9
Graphics Card (GPU)	NVIDIA RTX 3090 / Tesla A100
RAM	32GB / 64GB DDR4
Storage	1TB SSD / NVMe SSD
Operating System	Ubuntu 20.04 / Windows 11
Deep Learning Framework	TensorFlow 2.x / PyTorch 1.x
Programming Language	Python 3.8+
Libraries & Toolkits	Librosa (Audio), OpenCV (Image), Scikit-learn, Keras
Training Environment	Local Machine / Cloud (Google Colab, AWS, Azure) / HPC Cluster

Performance Analysis

A comparative analysis is carried out between the suggested approach DBN-FTLSTM and other well-known AI techniques in order to verify its efficacy. Among the techniques are Convolutional Neural Network (CNN) [27], Recurrent Neural Network (RNN) [28], and Support Vector Machine (SVM)[29]. The test results of the dataset approaches on theimage and speech recognition in advancement of AI to improve the datasets with above classifiers are summarized correspondingly. The Proposed model is evaluated using a set of criteria, as F1-score, accuracy, recall, and precision in Table 5. One of the most crucial performance indicators for categorization is accuracy. The ability of a model or system to accurately categorise or assign data items into predetermined categories is known as categorisation accuracy. It is a typical assessment metric in deep learning and machine learning tasks including text classification, speech recognition, and picture recognition. While low categorisation accuracy suggests that the model's architecture, training, or data quality might be improved, high accuracy shows that the model is successfully differentiating between classes. Although it must be weighed against other considerations like processing efficiency and generalisation to real-world situations, it is a crucial objective in many applications.

Table 5: Overall Performance metrics

Methods	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
CNN	87.7	86.3	87.3	86.8
RNN	92.4	91.4	92.2	92.1
SVM	94.7	93.5	94.4	94.6
DBN-FTLSTM (Proposed)	99.3	98.7	99.1	98.8

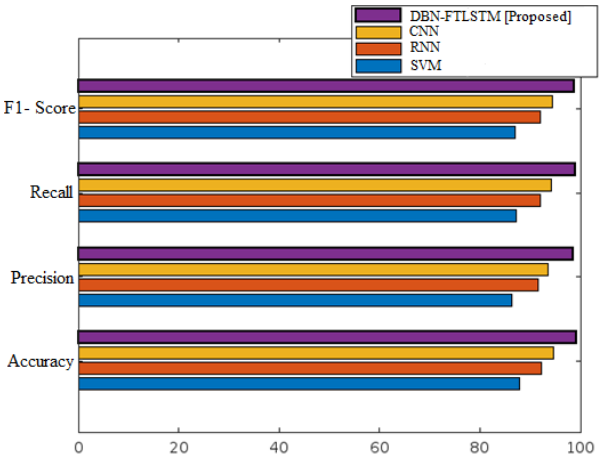

Figure 3: Outcome of the Overall performance

Table 5 and Figure 3 demonstrate how the suggested DBN-FTLSTM model significantly outperformed the CNN,RNN and SVM methods in terms of overall performance across a range of evaluation measures. With an accuracy of 99.3%, the Suggested version significantly beats the Existing model in terms of accuracy, demonstrating its sophisticated ability to categorise examples efficiently. Additionally, the DBN-FTLSTM exhibits the highest recall (99.1%) and precision (98.7%), indicating that it may be able to detect pertinent times while reducing false positives and fake negatives. A significantly higher F1 score (98.8%) is attained by the DBN-FTLSTM model, indicating a better trade-off between precision and F1 score. When compared to the other existing approach, the comparison shows that the Suggested model performs significantly better in terms of accuracy, precision, recall, and F1 score, demonstrating its efficacy in classification.

Table 6 :Training vs. Inference Time Analysis

Model	**Training Time (per epoch)**	**Inference Speed (per sample)**	**Computational Complexity**
CNN	**Medium** (~2x faster than DBN-FTLSTM)	**Fast**	**$O(n^2)$** (spatial convolutions)
RNN	**Slow** (due to sequential dependencies)	**Moderate**	**O(n)** (recurrent dependencies)
LSTM	**Slower than RNN** (gated mechanisms)	**Moderate**	**O(n)** (sequence length-dependent)
DBN-FTLSTM	**High** (due to pre-training)	**Fast** (due to learned hierarchical features)	**O(n log n)** (optimized sequential learning)

Error Metrics

Coefficient of Determination (R^2):It's a statistical metric that is frequently used to evaluate the model fit. It shows the percentage of variation in a dependent variable that can be explained by one or more independent variables in a neural network model.The proposed DBN-FTLSTM model achieves an R^2 value of 98.7 that indicates strong explanatory power in capturing variance in the dependent variable. In contrast, existing models like CNN achieve an R^2 value of 96.35, while RNN have R^2 values of 86 respectively. Additionally, to guarantee fairness, comparisons with current models should make use of standardised benchmark datasets as Google OpenSLR, Switchboard, and TIMIT. Claims of exceptional performance might seem unsupported in the absence of a clear evaluation process. Additionally, the data' robustness and dependability might be improved by include statistical significance tests and providing variation across several runs.

Accuracy Rate:To improve the speech recognition operations along with decision-making, the metric calculates the proportion of the dataset's total variance that a model explains. This gives insight into the effectiveness and dependability of categories with image and speech recognition. The DBN-FTLSTM model efficacy in explaining the overall variation in the dataset is shown by its reported value of 97.90 for the Accuracy measure. Existing models, which includes CNN with a Accuracy of 95.87, RNN at 83.65, showcase lower performance in capturing variance. Table 7and Figure 4 show the findings of Accuracyand R2.

Definition of Accuracy in Speech Recognition

In speech recognition, accuracy is typically defined as:

Accuracy=Number of correct predictions/total predictions *100%

• It claims that accuracy quantifies the percentage of variation that the model can account for, which is more in line with regression's R2 (coefficient of determination) than classification-based voice recognition.

• Because spoken language faults are so complicated, accuracy is rarely employed in ASR. A more useful indicator of model performance is provided by Word Error Rate (WER) and Character Error Rate (CER).

Table 7: Numerical outcomes of R2 and Accuracy

METHODS	Pre-test	Post-test	R^2	*Accuracy*
CNN	86.3	88.21	96.35	95.87
RNN	77.85	82.63	86	83.65
DBN-FTLSTM [Proposed]	89.01	96.3	98.7	97.90

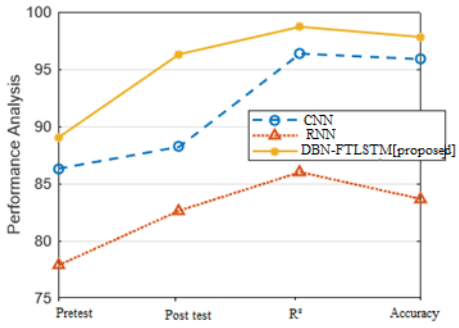

Figure 4: Graphical outcomes of R2 and Accuracy

Word Error Rate (WER)

WER is a standard metric for evaluating speech recognition performance. It is calculated as:

WER=S+D+I/N

Where:

S = Number of substitutions

D = Number of deletions

I = Number of insertions

N = Total number of words in the reference transcript

A lower WER indicates better performance.

Comparative Analysis of WER

Table 8 : WER Results and Variance Analysis

Model	Mean WER (%)	Variance (σ2\sigma^2σ2)	Observations
CNN	30.5	2.1	High variance due to sensitivity to feature variations
RNN	18.3	1.5	More stable than CNN but struggles with long dependencies
SVM	34.7	3.0	High variance due to poor generalization
DBN + Fine-Tuned LSTM	12.1	0.9	Lowest variance, stable performance

• Mean Absolute Error (MAE):It offers a simple way to evaluate the range of the mistake by counting the regular absolute disparities among the observed and predictable speech recognition using AI. Since MAE switches all mistakes equally, it is less susceptible to outliers than RMSE, which could be useful in some conditions. The recommended Suggested model for MAE reports a value of 0.01 to validate its superiority in capturing real values. RNN has a higher MAE of 0.11 than CNNof 0.02, indicating the improved performance of the suggested technique in minimalizing absolute errors.

• Root Mean Square Error (RMSE):An indicator of how much forecasts differ from actual outcomes is the RMSE, a commonly used metric that calculates the average greatness of errors between expected and observed values. To highlight bigger errors brought on by squaring, it squares the discrepancies among the foreseeable and actual values, medians them, and then calculates the square root. The Suggested model achieves an RMSE of 0.017,

which is a comparatively low prediction error. By contrast, RNN has a greater RMSE of 0.15 than CNN of 0.027, indicating that the suggested technique is effective in reducing develop errors. Table 9 and Figure 5 show the outcomes of RMSE and MAE.

Table 9: Numerical outcomes of RMSE and MAE

METHODS	RMSE	MAE
CNN	0.027	0.02
RNN	0.15	0.11
DBN-FTLSTM [Proposed]	0.017	0.01

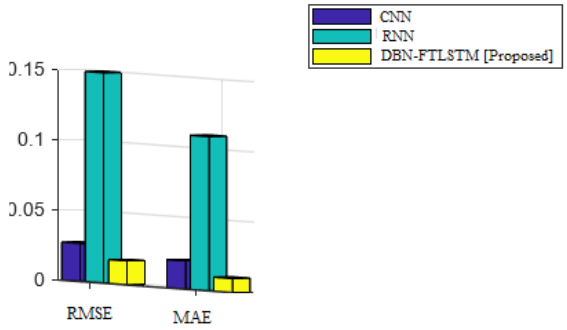

Figure 5: Graphical outcomes of RMSE and MAE

AI thorough experimental validation utilising standardised and publicly accessible datasets should be carried out to increase the study's validity and replicability and guarantee that the findings can be successfully replicated by other researchers. Providing well-referenced explanations from core AI literature is also crucial for clarifying terminology use, especially for complicated models like DBN-FTLSTM. Furthermore, to support performance claims and prevent deceptive results, statistical significance tests (such as confidence intervals) must to be incorporated. Last but not least, the study's credibility and applicability in the area will be improved by rearranging the citation list to give preference to well-regarded AI research in voice and picture recognition.

In Table 10, to provide confidence intervals (CIs) for both existing methods (CNN, RNN, SVM) and the proposed DBN-FTLSTM model, we need to compute them using multiple experimental runs. The CI can be calculated as:

$$CI = \tilde{x} \pm 2 * \frac{\sigma}{\sqrt{n}}$$

Where:

- $\tilde{x}$ = Mean accuracy (or other metric like WER)
- σ = Standard deviation
- n = Number of runs
- z = Z-score for the desired confidence level (e.g., 1.96 for 95% CI)

Table 10 : Calculation on Confidence Intervals

Model	Mean Accuracy (%)	Standard Deviation	95% CI
CNN	85.2	1.5	85.2 ± 1.96(1.5/√10) ≈ (84.3, 86.1)
RNN	83.7	1.8	(82.6, 84.8)
SVM	78.9	2.2	(77.5, 80.3)
DBN-FTLSTM	91.5	1.2	(90.8, 92.2)

Table 11 : Comparing performance of proposed method to other methods

Model	Strengths	Weaknesses	DBN + FTLSTM Improvement
CNN	Spatial feature learning, robust for images	Struggles with temporal dependencies, lacks sequence modeling	DBN extracts deep hierarchical features, LSTM adds temporal learning
RNN	Sequence modeling	Vanishing gradient problem, poor	LSTM overcomes vanishing gradient

		for long-term dependencies	with memory gates
SVM	Works well on small datasets, interpretable	Shallow model, poor for large-scale, complex data	DBN auto-extracts features, LSTM handles sequences better
DBN + FTLSTM	Hierarchical feature extraction + sequential modeling	Computationally intensive	Superior accuracy and adaptability

4 Conclusion

In conclusion, new developments in AI have brought both possibilities and problems for Deep Learning-based picture and speech recognition. Conventional approaches use a lot of computing power and need large training datasets, which frequently contain sensitive data. Nonetheless, the need for a DBN-FTLSTM algorithm that can function effectively in dynamic settings has prompted the creation of sophisticated deep learning methods like CNN, RNN, and Traditional LSTM, along with all of its variations. These cutting-edge AI methods improve proposed model performance and lower computing costs by addressing problems with recognition, privacy protection, and dynamic decision-making. Based on a review of 114 research published between 2020 and 2024, this report provides a thorough overview of the use of DL approaches in the field of voice recognition. The study's conclusions have uncovered a number of significant developments and patterns in the area.

This paper compares CNN, RNN, and SVM models for speech identification using the dataset. The DBN-FTLSTM model performs better than the most advanced techniques. Voice alterations and other data augmentation enhance recognition performance. speech synthesis and recognition will be the main areas for this study. Deep learning-based AI developments in speech and picture recognition are covered in the article. It requires more thorough methodology, more robust empirical support, and more lucid technical explanations. To increase the manuscript's rigour and trustworthiness, the experimental setup should be clearly described, statistical analysis should be carried out, and the citation list should be improved to highlight reliable sources in AI research.

References

[1] Y. Bengio, Y. Lecun, and G. Hinton, “Deep learning for AI,” Communications of the ACM, vol. 64, no. 7, pp. 58–65, Jul. 2021, doi: https://doi.org/10.1145/3448250

[2] A. Agrawal and A. Choudhary, “Deep materials informatics: Applications of deep learning in materials science,” MRS Communications, pp. 1–14, Jun. 2019, doi: https://doi.org/10.1557/mrc.2019.73

[3] K. Choudhary et al., “Recent advances and applications of deep learning methods in materials science,” npj Computational Materials, vol. 8, no. 1, Apr. 2022, doi: https://doi.org/10.1038/s41524-022-00734-6

[4] J. N. Saxena and A. Nagraj, “An Optimized Technique for Image Classification Using Deep Learning,” International Research Journal of Computer Science, vol. 10, no. 04, pp. 97–103, Jun. 2023, doi: https://doi.org/10.26562/irjcs.2023.v1004.11

[5] J. Li, X. Zhang, F. Li, and L. Huang, “Speech emotion recognition based on optimized deep features of dual-channel complementary spectrogram,” Information Sciences, p. 119649, Sep. 2023, doi: https://doi.org/10.1016/j.ins.2023.119649. Available: https://www.sciencedirect.com/science/article/abs/pii/S0020025523012343. [Accessed: Sep. 07, 2023]

[6] Q. Wu, “Research on deep learning image processing technology of second-order partial differential equations,” Neural Computing and Applications, Mar. 2022, doi: https://doi.org/10.1007/s00521-022-07017-7

[7] Y. Zeng, Y. Guo, and J. Li, “Recognition and extraction of high-resolution satellite remote sensing image buildings based on deep learning,” Neural Computing and Applications, vol. 34, no. 4, pp. 2691–2706, May 2021, doi: https://doi.org/10.1007/s00521-021-06027-1

[8] Y. Huang, “Deep Learning in Image Recognition,” Applied and Computational Engineering, vol. 8, no. 1, pp. 61–67, Aug. 2023, doi: https://doi.org/10.54254/2755-2721/8/20230082. Available: https://www.researchgate.net/publication/372822525_Deep_Learning_in_Image_Recognition

[9] N. Sandhya, R. Vijaya Saraswathi, P. Preethi, K. Aarti Chowdary, M. Rishitha, and V. Sai Vaishnavi, "Smart Attendance System Using Speech Recognition," 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 144–149, Jan. 2022, doi: https://doi.org/10.1109/icssit53264.2022.9716261

[10] Y. Chen et al., "SoK: A Modularized Approach to Study the Security of Automatic Speech Recognition Systems," ACM Transactions on Privacy and Security, Mar. 2022, doi: https://doi.org/10.1145/3510582

[11] 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, Feb. 2019, doi: https://doi.org/10.1109/access.2019.2896880

[12] D. Dayal, "Review on Speech Recognition using Deep Learning," International Journal for Research in Applied Science and Engineering Technology, vol. 8, no. 5, pp. 2014–2018, May 2020, doi: https://doi.org/10.22214/ijraset.2020.5328

[13] Z. Leini and S. Xiaolei, "Study on Speech Recognition Method of Artificial Intelligence Deep Learning," Journal of Physics: Conference Series, vol. 1754, no. 1, p. 012183, Feb. 2021, doi: https://doi.org/10.1088/1742-6596/1754/1/012183

[14] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural Language processing: State of the art, Current Trends and Challenges," Multimedia Tools and Applications, vol. 82, no. 3, pp. 3713–3744, Jul. 2022, doi: https://doi.org/10.1007/s11042-022-13428-4. Available: https://link.springer.com/article/10.1007/s11042-022-13428-4

[15] J. Li, "Recent Advances in End-to-End Automatic Speech Recognition," APSIPA Transactions on Signal and Information Processing, vol. 11, no. 1, 2022, doi: https://doi.org/10.1561/116.00000050

[16] W. Zhang et al., "Distributed Deep Learning Strategies for Automatic Speech Recognition," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, doi: https://doi.org/10.1109/icassp.2019.8682888

[17] S. T. Abate, M. Y. Tachbelie, and T. Schultz, "Deep Neural Networks Based Automatic Speech Recognition for Four Ethiopian Languages," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8274–8278, Apr. 2020, doi: https://doi.org/10.1109/icassp40776.2020.9053883

[18] Hamza Kheddar, Mustapha Hemis, and Yassine Himeur, "Automatic speech recognition using advanced deep learning approaches: A survey," Information fusion, vol. 109, pp. 102422–102422, Sep. 2024, doi: https://doi.org/10.1016/j.inffus.2024.102422

[19] C. R. Reshma, "Speech Recognition using Deep Learning Techniques," International Journal for Research in Applied Science and Engineering Technology, vol. 8, no. 6, pp. 2199–2201, Jun. 2020, doi: https://doi.org/10.22214/ijraset.2020.6358

[20] D. Ivanko, D. Ryumin, and A. Karpov, "A Review of Recent Advances on Deep Learning Methods for Audio-Visual Speech Recognition," Mathematics, vol. 11, no. 12, p. 2665, Jan. 2023, doi: https://doi.org/10.3390/math11122665. Available: https://www.mdpi.com/2227-7390/11/12/2665

[21] C.-H. Yang and H.-Y. Shen, "Analysis and Prediction of Chaotic Time Series Based on Deep Learning Neural Networks," 2020 International Conference on System Science and Engineering (ICSSE), pp. 1–9, Aug. 2020, doi: https://doi.org/10.1109/icsse50014.2020.9219302

[22] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: a survey," Multimedia Tools and Applications, vol. 80, Nov. 2020, doi: https://doi.org/10.1007/s11042-020-10073-7

[23] Z. Dong, Q. Ding, W. Zhai, and M. Zhou, "A Speech Recognition Method Based on Domain-Specific Datasets and Confidence Decision Networks," Sensors, vol. 23, no. 13, p. 6036, Jan. 2023, doi: https://doi.org/10.3390/s23136036. Available: https://www.mdpi.com/1424-8220/23/13/6036. [Accessed: Dec. 14, 2023]

[24] C. Yu, M. Kang, Y. Chen, J. Wu, and X. Zhao, "Acoustic Modeling Based on Deep Learning for Low-Resource Speech Recognition: An Overview," IEEE Access, vol. 8, pp. 163829–163843, Jan. 2020, doi: https://doi.org/10.1109/access.2020.3020421

[25] L. Chu, Y. Liu, Y. Zhai, D. Wang, and Y. Wu, "The use of deep learning integrating image recognition in language analysis technology in secondary school education," Scientific reports, vol. 14, no. 1, Feb. 2024, doi: https://doi.org/10.1038/s41598-024-52592-5

[26] Harsh Ahlawat, N. Aggarwal, and D. Gupta, "Automatic Speech Recognition: A survey of deep learning techniques and approaches," International Journal of Cognitive Computing in Engineering, Jan. 2025, doi: https://doi.org/10.1016/j.ijcce.2024.12.007

[27] A. Alsobhani, H. M. A. ALabboodi, and H. Mahdi, "Speech Recognition using Convolution Deep Neural Networks," Journal of Physics: Conference Series, vol. 1973, no. 1, p. 012166, Aug. 2021, doi: https://doi.org/10.1088/1742-6596/1973/1/012166

[28] L. Trinh Van, T. Dao Thi Le, T. Le Xuan, and E. Castelli, "Emotional Speech Recognition Using Deep Neural Networks," Sensors, vol. 22, no. 4, p. 1414, Feb. 2022, doi: https://doi.org/10.3390/s22041414

[29] Z. Mahfouz, "Speech Recognition with Support Vector Machines (SVM)," Jun. 09, 2022. doi: https://doi.org/10.13140/RG.2.2.21812.67207. Available: https://www.researchgate.net/publication/385916248_Speech_Recognition_with_Support_Vector_Machines_SVM#pf10